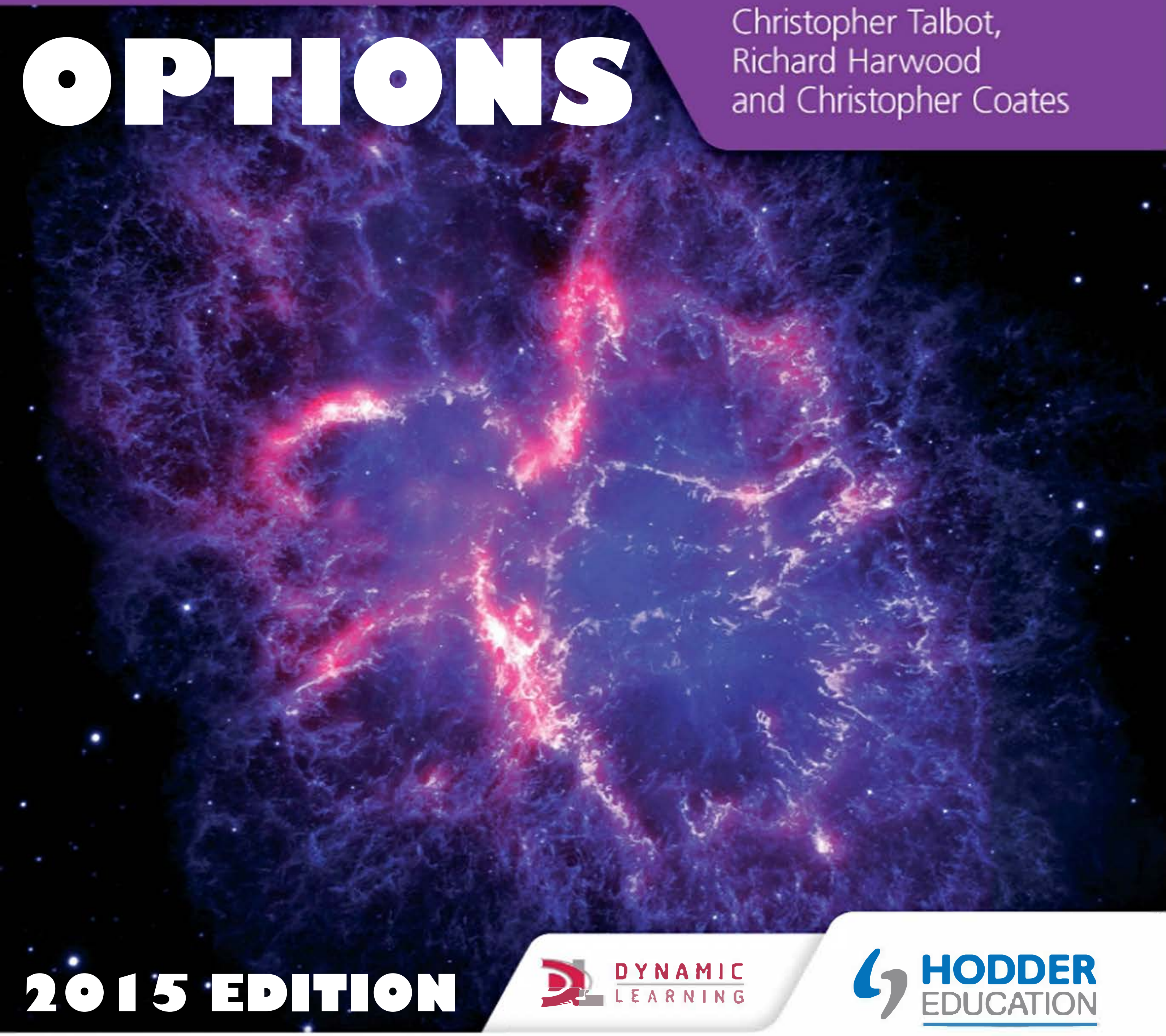
FOR THE IB DIPLOMA

SECOND EDITION

Chemistry

OPTIONS

Christopher Talbot,
Richard Harwood
and Christopher Coates



2015 EDITION

DYNAMIC LEARNING

HODDER EDUCATION

22 Option A Materials

ESSENTIAL IDEAS

- Materials science involves understanding the properties of a material, and then applying those properties to desired structures.
- Metals can be extracted from their ores and alloyed for desired characteristics. ICP-MS/OES spectroscopy ionizes metals and uses mass and emission spectra for analysis.
- Catalysts work by providing an alternative reaction pathway for the reaction. Catalysts always increase the rate of the reaction and are left unchanged at the end of the reaction.
- Liquid crystals are fluids that have physical properties which are dependent on molecular orientation relative to some fixed axis in the material.
- Polymers are made up of repeating monomer units which can be manipulated in various ways to give structures with desired properties.
- Chemical techniques position atoms in molecules using chemical reactions whilst physical techniques allow atoms/molecules to be manipulated and positioned to specific requirements.
- Although materials science generates many useful new products, there are challenges associated with recycling of, and high levels of toxicity of, some of, these materials.

Additional higher level (AHL)

- Superconductivity is zero electrical resistance and expulsion of magnetic fields. X-ray crystallography can be used to analyse structures.
- Condensation polymers are formed by the loss of small molecules as functional groups from monomers join.
- Toxicity and carcinogenic properties of heavy metals are the result of their ability to form coordinated compounds, have various oxidation states and act as catalysts in the human body.

22.1 Materials science introduction *– materials science involves understanding the properties of a material, and then applying those properties to desired structures*

Introduction

History has characterized civilizations by the materials they used: Stone Age, Bronze Age, Iron Age and now the Plastics or Polymer Age and perhaps later the Nano-materials Age. The use of materials, especially alloys, was developed based on observations and measurement of their chemical and physical properties before a physical and chemical explanation had been hypothesized. Knowledge of chemical bonding and chemical structures is used to prepare new useful materials or to modify the properties of currently used materials. *Materials science is the scientific study of the structure and properties of materials (substances).*

Nature of Science

Classification and uses of materials

From the start of human existence, materials have been fundamental to the development of civilization. Anthropologists define the historical eras based on the materials used by the different civilizations, for example the Stone, Copper, Bronze and Iron Ages. The different rates of progression towards more sophisticated materials between cultural groups is connected with different levels of innovation and the local availability of those materials, and led to different standards of living.

Early lack of exchange of technological information resulted in significant differences in advancement between cultures at any one time. For example, in 1500 BCE people in Asia Minor (Turkey) were already experimenting with iron, whereas in Mesopotamia (Iraq) they were still in the Bronze Age. The Europeans, Palestinians and Egyptians were in the Copper and early Bronze Age; the Chinese had melted iron and were advanced in the development of Bronze while in North Africa there was still evidence of the late Stone Age.

Metallurgy, defined as the science and art of processing and adapting metals, has been around for approximately 6000 years from when early man recovered and used metals through observation and deduction. Metals, glass and many other materials, such as porcelain and rubber, were used for different purposes before the development of a scientific understanding of their properties (based on chemical bonding theory and experimental evidence such as X-ray crystallography).

Understanding of how materials behave like they do and why they differ in properties was possible only with the understanding allowed by quantum mechanics, which first explained the electronic properties of atoms and then solids starting in the 1930s. The combination of physics, chemistry, and the focus on the relationship between the properties of a material and its microstructure is a branch of science known as materials science. The development of this science allowed the design of materials and provided a knowledge base for the engineering applications (materials engineering).

Classification of materials based on bonding and structure

There are a number of ways of classifying materials. One approach is to classify them into four groups on the basis of their bonding and structure.

Crystalline materials have their particles (atoms, ions or molecules) arranged into a lattice (Chapter 4) a regular repeating arrangement of particles. The type of lattice can be determined using X-ray crystallography (see section 22.8). The majority of solids, including metals and their alloys as well as ceramics, are crystalline. A few plastics have some degree of crystallinity, for example high-density polythene (up to 70 per cent).

Amorphous (disordered) materials have their particles randomly arranged; there is little order or symmetry. Glass is a familiar example of an amorphous material. Some plastics are also amorphous or have amorphous regions, for example, low-density polyethene. However, high-density polyethene has a high degree of crystallinity making it stronger and less easily deformed by heat (compared with low-density polyethene).

Semi-crystalline materials have crystalline and amorphous regions. This is typical of many polymers because the semi-crystalline structure gives a good balance of stiffness and toughness. Metals and ceramics are polycrystalline, meaning they are made up of a large number of small crystals. Liquid crystals (see section 22.4) are another well-studied group of materials.

A composite material is one that is composed of two or more different materials bonded together, with one serving as the supporting matrix surrounding particles or fibres of the other. Well-known composites include concrete, foams, fibre-reinforced plastics and laminates. Cellular materials consist of stacks of hollow cells. Wood is a good example of a cellular material and is also a composite material.

One approach to material classification is on the basis of their bonding and structure as well as their properties.

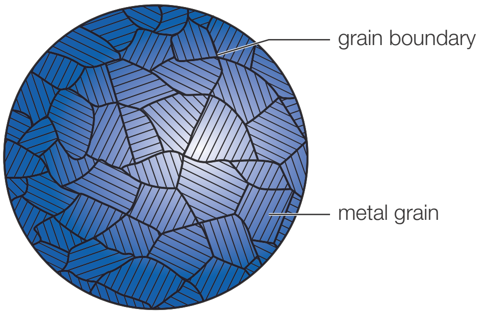
Metals

Metals are generally stiff, hard, strong and shiny (Figure 22.1) (when polished). They change shape without breaking (ductile and malleable) and are strong in tension (stretching) and compression (squeezing). They are excellent thermal and electrical conductors. The more active metals undergo oxidation with oxygen in the air and corrosion by water and acids.

Figure 22.1 Liquid mercury

These properties arise from metallic bonding (Chapter 4) and the crystalline structure of metals. Metallic bonds are non-directional, in contrast to covalent bonds that are directional. Metals in bulk are strong in tension and compression because metallic bonding is strong. Metals are giant structures in which the delocalized valence electrons are free to move and conduct heat and form an electric current (when a voltage is applied).

Metals can deform and change shape without breaking because the metallic bond allows layers of cations to slip over one another when the metal is under stress (due to the application



■ **Figure 22.2** Arrangement of metal grains

of a force). Pure metals are made stronger by alloying (Chapter 4) and by heat treatment, but as they become stronger their ductility is reduced.

Metal objects are formed by casting. Hot metal is allowed to cool down in a cast. As it cools, small nuclei of the solid appear in the liquid and as cooling continues small crystals are formed. The crystals grow and meet to form a solid mass of small crystals, a polycrystalline solid. These crystals are called grains (Figure 22.2).

1 Find out about the discovery of dislocations in metals by the transmission electron microscope. Find out about their importance in explaining the ductility of metals.

Ceramics

■ **Figure 22.3** A porcelain cup and saucer

Ceramics (Figure 22.3) are a group of materials that are very hard and brittle. They are strong in compression but weak in tension. This means it is very hard to change their shape, but they are easily snapped. They are electrical and thermal insulators and have very high melting points. They are chemically unreactive and do not react with oxygen, water or acids.

Ceramics are crystalline compounds of metallic and non-metallic elements (usually silicon and oxygen). In most ceramics, the atoms which form the framework (lattice) are linked by covalent bonds. The structure of ceramics often includes metal cations that are linked to the framework by ionic bonding.

The structure of a ceramic is therefore more rigid and less flexible than that of a metal. The structure makes ceramics harder than metals and also more brittle. Ceramics have lower densities and higher melting points than many metals. Ceramics do not contain delocalized electrons and hence are poor electrical and thermal conductors.

Water may be absorbed in the pores of ceramics, and if the water freezes and expands, damage will occur. The low thermal conductivity of ceramics can result in large thermal differences being set up, causing stress.

Glasses

■ **Figure 22.4** A glass floor

Glasses are a subset of ceramics, but have lower melting points than other ceramics. They are generally transparent and allow light to pass through. The structure of most glasses is a covalently bonded framework of silicate tetrahedra with metal cations linked by ionic bonding.

Glass is transparent (Figure 22.4) because it is non-crystalline (amorphous), so light rays can pass through without meeting any reflecting surfaces. Glass is a strong material and can withstand mechanical loading because the covalent bonding is strong. However, glass is brittle because of the rigid nature of the covalent bonds and amorphous structure and hence it shatters easily. Glasses are impermeable to water.

Plastics

2 Find out why glasses are described as super-cooled liquids. Find out what structural changes occur at the glass transition temperature.

Plastics are usually strong with a low density and usually soft, flexible and not very elastic (unless they are elastomers). Many of them soften easily when heated and melt or burn. They are thermal and electrical insulators because they lack delocalized electrons.

Plastics are long chain polymers and the bonding is covalent. Many plastics are resistant to chemical attack but may be damaged by exposure to ultraviolet radiation (sunlight). Many plastics burn easily, often releasing toxic fumes (see section 22.7). They are usually impermeable to water.

In thermosoftening plastics, the intermolecular forces between adjacent chains are weak London (dispersion) forces, though collectively they can be significant. In thermosetting plastics, covalent bonds form strong cross-links between chains (see section 22.5).

Composite materials

Composite materials are heterogeneous mixtures. In a composite material, the properties of the components combine to give a material which is more useful for a particular purpose than the individual components.

Composites consist of a matrix phase with fibres or rods or particles of another material (reinforcing phase). Examples of composites are reinforced concrete, carbon fibres and glass fibre-reinforced polyester. A very early example of a composite made of natural materials was straw and clay used to make huts.

Bullet-proof vests have the ability to absorb and disperse (spread out) the kinetic energy of a bullet. This requires a material which consists of a polymer resin reinforced with fibres of a high molar mass polyethene. The fibres are stretched only slightly when the stress is very high. The structure consists of layers of aligned fibres. The direction of the fibres is rotated through 90 degrees in alternate layers. With a number of alternate layers a rigid armour that can be used for vehicles and riot shields is obtained. For a ballistic vest worn by riot police, greater flexibility is required and this is achieved by sandwiching alternate layers of fibre-reinforced resin between films of low-density polyethene.

Classification of materials based on properties and uses

The choice of a material for a particular purpose depends on the conditions under which the product is used, for example whether it needs to be soft and whether it needs to be corrosion resistant. Another issue, apart from cost, is the method of manufacture, for example whether a material can be easily moulded into a complex shape.

Important physical properties of materials are density, tensile strength, compressive strength, toughness, hardness, and electrical and thermal conductivity. Chemical properties include its reaction (if any) towards water, air (oxygen) and acids (dilute and concentrated).

For many uses (applications), electrical and thermal properties are important in determining the choice of material.

3 Find out what materials are used as dental restorative materials. Identify the important properties of these substances that make them suitable for this use.

Designer and smart materials

Understanding bonding, atomic structure and microscopic structure allows material scientists to design, synthesize and manufacture new materials with specific chemical and physical properties. For example, Gore-Tex is a waterproof breathable fabric that allows perspiration (sweat) to evaporate while protecting the wearer from rain.

Gore-Tex is the patented name for a porous form of the polymer PTFE made by stretching the polymer fibres in a controlled way to create fine pores. Gore-Tex itself is actually one layer in the fabric design for a particular clothing application such as a raincoat or wetsuit. Gore-Tex fabric is made up of a layer of a plastic based on expanded PTFE and this is laminated on to a layer of another fabric. The layer contains 14 million pores per square millimetre. It is the tiny pores in the PTFE layer that let the water vapour molecules through – that is, it is breathable (Figure 22.5). However, the layer is waterproof because liquid water droplets cannot pass through in the opposite direction; in fact the fabric surface repels water – it is hydrophobic ('water hating').

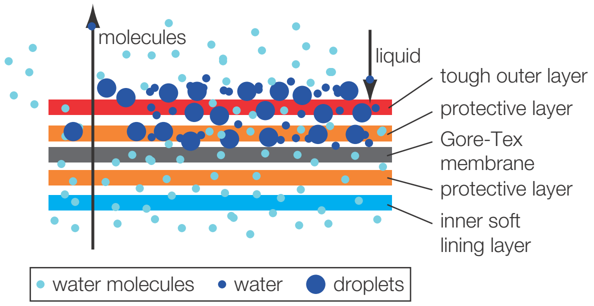


Figure 22.5 The structure of Gore-Tex

Each pore is too small for water droplets to pass through, but large enough to let water vapour molecules from sweat to go through ('breathability'). So, if you sweat in this 'breathable' material, the water vapour can escape, keeping you cooler, and you do not get the discomfort from sweat condensate. In addition, because water droplets cannot pass through the outer tough protective layer, you should keep dry in wet weather. Bonding, atomic structure and microscopic structures are strongly linked with the properties of materials.

Smart materials are materials that have a property that can be significantly changed in a controlled way by external stimuli. These include photochromic materials that change in response to the light (Chapter 9) and electrochromic materials that change their colour or opacity (how much light they let through) on the application of a voltage. This effect is used in liquid crystal displays (LCDs).

Materials in ancient civilizations

What materials were used by ancient civilizations, such as the Andeans, Romans and Chinese? Even though these ancient civilizations were located in very geographically diverse locations, how were the materials they used similar?

The Andean culture was a loose collection of different cultures that developed from the highlands of Colombia to the Atacama desert in Chile, Peru, Bolivia and Argentina. It was based mainly on the cultures of Ancient Peru, and the Inca Empire marked the end of Andean culture and conquest by Spain. A wide variety of materials were used by the Andeans, including gold, semi-precious stones, clay, adobe (mud, sand and clay brick), silver, bronze, feather, cotton and wool (from llamas and alpacas).

ToK Link

Although it is convenient to classify materials into categories, no single classification is 'perfect'. How do we evaluate the different classification systems we use in the different areas of knowledge? How does our need to categorize the world help and hinder the pursuit of knowledge?

Materials can be classified into different categories. For example, they can be classified at the atomic level: different arrangements of atoms give rise to different structures and hence different properties, for example graphite and diamond. Materials can also be classified at the microscopic level: arrangement of small grains that can be identified by microscopy. For example, transparent and frosted glasses differ in their silicate grain sizes. Some categories of materials are not based on bonding. A particular microstructure identifies composites made of different materials in intimate contact, for example fibreglass and wood, to achieve certain properties. Biomaterials can be any type of material that is biocompatible and used to replace human body parts. Nanomaterials are a whole new group of new materials derived from the science of nanotechnology.

Materials can also be classified according to their properties, which are the ways the material responds to the environment. For example, the mechanical, electrical and magnetic properties are the responses to mechanical, electrical and magnetic stress, respectively. Other important properties are thermal (transmission of heat and specific heat capacity), optical (absorption, transmission and scattering of light), and the chemical reactivity, for example corrosion resistance. Many materials, especially metals, can be processed by heat treatment or mechanical treatment which affects their microstructure and hence their properties. The classification in the text has introduced metals, ceramics, glasses and plastics, but there are many other types of materials, such as semiconductors, superconductors and even conducting polymers.

Structure of ceramics

Ceramics consist of a regular arrangement of atoms, and these repeating structural units are atoms, ions or covalently bonded structures arranged in a regular three-dimensional structure. The bonding may be covalent (e.g. silicon dioxide), ionic (e.g. magnesium oxide) or ionic with some covalent character (e.g. aluminium oxide).

Silicon(IV) oxide, SiO_2, is the basis for a wide range of ceramics. A silicon atom can form four single covalent bonds (σ bonds) to form the silicate unit, SiO_4^{4-}, with a tetrahedral distribution of bonds.

Each of the four oxygen atoms in the SiO_4^{4-} ion has an unshared electron which it can use to bond with other SiO_4^{4-} units. Anions such as $Si_2O_7^{2-}$, $Si_3O_9^{6-}$ (Figure 22.6) and $Si_6O_{18}^{12-}$ occur in many minerals.

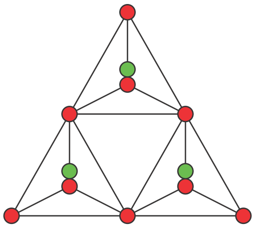
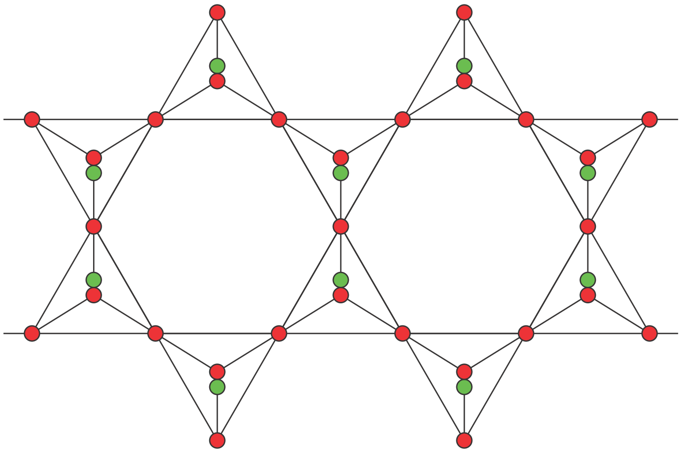


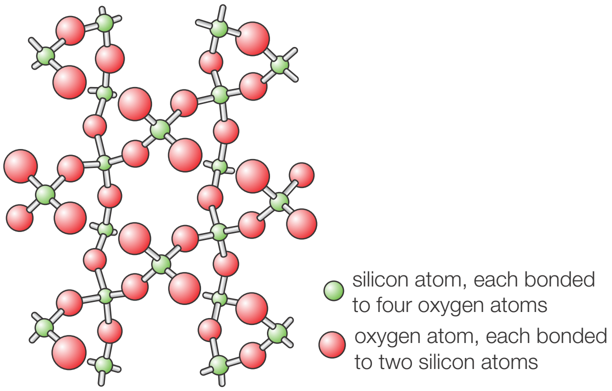
Figure 22.6 The $Si_3O_9^{6-}$ ion

4 Draw a dot-and-cross diagram for the ion $Si_2O_7^{6-}$ made by linking two silicate tetrahedra together. Show clearly the charges on the individual oxygens.

There are silicates with long chains of linked tetrahedra, and some silicate minerals have single silicate strands of formula $(SiO_3)_n^{2n-}$ which are bonded to metal cations that balance the negative charges. Asbestos has the double-stranded structure shown in Figure 22.7. The double stands are bonded to other double strands by ionic bonding to the Na^+, Fe^{2+} and Fe^{3+} ions packed around them. These can be broken relatively easily and hence asbestos has a fibrous feel (texture). Aluminium atoms replace silicon atoms in many silicate minerals. For every aluminum atom that replaces a silicon atom, a singly charged metal ion, for example sodium, is needed to balance the charge. Silicate tetrahedra bond to form silicates with sheet structures (Figure 22.7). Clay minerals are also silicates with the sheet structure, but in clays 25 per cent of the silicon atoms have been



■ **Figure 22.7** Double strands of silicate tetrahedra


silicon atom, each bonded to four oxygen atoms
oxygen atom, each bonded to two silicon atoms

■ **Figure 22.8** The three-dimensional network of silicate tetrahedra in quartz

replaced by aluminium. Replacing silicon (oxidation state +4) with aluminium (oxidation state +3) means that additional positive charges (e.g. Na^+) are needed to balance the charges on the oxygen atoms. Layers of sodium and aluminium cations between the layers hold them together by ionic bonding. These layer structures have an inner surface which can absorb large amounts of water.

Quartz (Figure 22.8) has a three-dimensional network of SiO_4 units, in which all four oxygen atoms of the silicate tetrahedra are shared with other silicon atoms. With each silicon atom covalently bonded to four oxygen atoms and each oxygen atom bonded to two silicon atoms, the formula is $(SiO_2)_n$ and the empirical formula is SiO_2.

Cement and concrete

Portland cement is made by strongly heating limestone (calcium carbonate) with silica (silicon dioxide). Clay may also be added during the production of cement. Concrete (Figure 22.9) is artificial stone made from a mixture of cement, water and fine and coarse aggregate (usually sand and coarse rock).

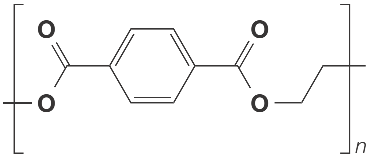
The manufacture of cement requires two raw materials, one rich in calcium, for example limestone or chalk, and one rich in silica, for example clay. Sand is sometimes added and the raw materials are finely ground and mixed before being heated. A number of reactions take place, including the thermal decomposition of calcium carbonate and at the end of the reaction between calcium oxide and silica to form a calcium silicate, Ca_3SiO_5. Tiny quantities of resinous materials that have pockets which trap air are added to cement. The air pockets help hardened concrete to withstand repeated freezing and thawing without cracking.

■ **Figure 22.9** Palace Assembly building, Chandigarh, India is made from concrete

5 Find out about 'concrete cancer'.

6 Classify the following materials into suitable categories. Find out one use for each material:

Gold; polyurethane; solder; nitrocellulose nitrate; silicon; brass; gutta percha; polystyrene; titanium; plywood; silicon nitride (SiN); porcelain; carbon-fibre reinforced epoxy resin; terracotta; talcum powder; magnadur; borosilicate glass; nylon; muntz metal; Pyrex; rayon; silicon rubber; bronze; silk; nitinol; thinsulate; Kevlar; mica; rayon; Teflon (PTFE); polyacetylene; asbestos; Bakelite; carbon fibre; Cellophane; Dacron; ebonite; cement.



■ **Figure 22.10** Structure of PET

Relating physical characteristics to structure and bonding

The physical characteristics of materials can be related to its bonding and structure. For example, consider the use of polyethene benzene-1,4-dicarboxylate or poly(ethylene terephthalate) (PET) (Figure 22.10) commonly used as a container for carbonate drinks.

It is a long chain covalently bonded polymer and hence a poor electrical and thermal conductor due to the absence of delocalized electrons.

It has a relatively high range of temperatures at which it softens because of the relatively strong and rigid London (dispersion) forces and dipole–dipole forces operating between adjacent PET chains. These intermolecular forces and partially crystalline structure also make it strong and tough. The presence of hydrophobic benzene rings makes it water resistant and impermeable.

PET needs to be very impermeable to gas since it is often used to contain carbonated drinks (Figure 22.11). This property is related to its structure since crystalline polymers are less permeable to gases than amorphous polymers because crystalline polymers have fewer spaces in the structure than amorphous polymers.

Figure 22.11 PET bottle with carbonated drink

Testing materials

Stress

The ways in which a material behaves when forces are applied are described as mechanical properties. When an external force (push, pull or twist) is applied to a material, internal forces act in a direction opposite to that of the force applied. A material which is stretched by external forces is under **tension**. A material which is being squeezed by external forces is in **compression**.

Materials used in construction, for example, concrete, must be able to sustain a large force while undergoing a small change of shape (minor deformation). Engineers and material scientists are interested in the force required to produce a definite amount of deformation in a material. The **stress** is the force acting per unit cross-sectional area of the material. The deformation produced is the **strain**.

The **tensile stress** is the stress which stretches a material in a particular direction. Tensile stress = force/original cross-sectional area. Tensile stress has the units of pressure: newtons per square metre, $N\,m^{-2}$, or Pascal, Pa. The tensile strain is the elongation (increase in length) per unit length. Since tensile strain is a ratio of two lengths, it has no dimensions.

7 a Investigate the stretching of a rubber band. Does it obey Hooke's law?

b Find out how to make cement and design an experiment to investigate the effect of composition on breaking strength after drying.

Strength

The strength of a material is its ability to resist an applied force without breaking. The **tensile strength** of a material is the maximum tensile stress it can withstand without breaking. It has the units of stress, Pa.

When a force is applied to a material the material may be deformed (change shape). This may be elastic deformation (the material will return to its original length), plastic deformation (the material will stay permanently stretched) or fracture (snap). Metals can undergo plastic deformation, but many non-metallic materials are brittle and will fracture and break.

Bond triangle diagrams

Bonds between metals and non-metals vary from ionic to covalent (via polar covalent) as a consequence of the electronegativity difference between the two atoms. The greater the electronegativity difference, the greater the ionic character. Compounds with a high degree of ionic character are crystalline solids, are non-conductors of electricity (unless molten or in aqueous solution), and have high melting and boiling points. Simple molecular or molecular covalent compounds will have low melting and boiling points and usually are non-conductors of electricity (unless they undergo hydrolysis with water). A substance with polar covalent bonds exhibits properties intermediate between highly ionic and highly covalent character. A useful way to visualize these ideas is with a simple bond triangle diagram (Ketelaar's triangle) in which the vertices (corners) are labelled metallic, covalent and ionic (Figure 22.12).

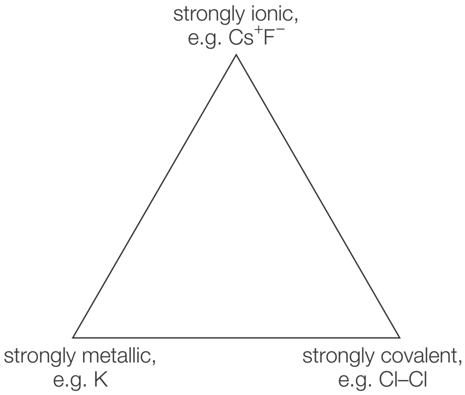

strongly ionic, e.g. Cs^+F^-
strongly metallic, e.g. K
strongly covalent, e.g. Cl–Cl

Figure 22.12 Simple bond triangle diagram

A more sophisticated version is known as the van Arkel diagram and is a plot of the difference in electronegativity ($\Delta\chi$) of two elements on the *y*-axis and average electronegativity ($\Sigma\chi = (\chi_a + \chi_b)/2$) on the *x*-axis.

There is no need to consider the stoichiometry of the binary compound. So the oxides sulfur(IV) oxide (SO_2) and sulfur(VI) oxide (SO_3) will both appear at the same coordinate on the van Arkel diagram

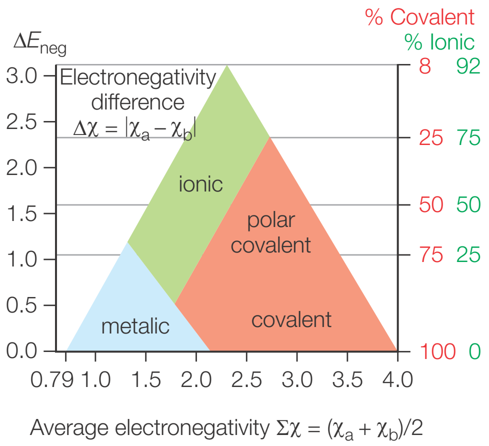


Figure 22.13 Van Arkel diagram

(Figure 22.13). It is assumed that the electronegativity of an element does not vary between compounds.

As the *x*-axis gives information about the electronegativity average, it is a measure of the degree of localization of the bonding electrons. It provides information on the degree of covalency. At the left-hand side of the *x*-axis is the most electropositive (or least electronegative) element (caesium). It is highly metallic with delocalized valence electrons. At the right-hand side of the *x*-axis is the most electronegative (or least electropositive) element (fluorine). In fluorine, the bonding valence electrons in this purely covalent molecule are localized within the sigma bond.

As the *y*-axis gives information on electronegativity differences, it indicates to what degree the bonding electrons are unevenly (asymmetrically) distributed between the two bonding atoms. It provides information on the degree of ionic character. It is for this reason that at the bottom of the triangle, where $y = 0$, the elements are found. At the top of the van Arkel diagram there is the greatest degree of asymmetry of the electron distribution in the bond. This is where caesium fluoride (CsF [Cs^+F^-]) is to be found at the extreme of ionic bonding as the elements in this binary compound are the most electropositive and the most electronegative.

Advantages and disadvantages of the van Arkel diagram

The van Arkel diagram gives a quantitative appreciation of intermediate bonding. It allows chemists to view the three main models for chemical bonding in a single triangular-shaped chart. It allows prediction in bonding for unfamiliar compounds. These predictions are easier when the resulting point plotted is close to one of the corners of the triangle.

All chemical models tend to break down at some point and whilst useful, they do have their limitations. The search by chemists for interesting new materials such as high-temperature superconductors continues, but it is not possible to use the van Arkel diagram to suggest new binary compounds that have these interesting properties. Although the van Arkel diagram provides information on bonding it does not predict the crystal structures of solids which depend on stoichiometry and electron configuration.

This is clearly shown by elements with allotropes, such as carbon. Allotropes of the same element all have the same electronegativity values but very different structures in the case of carbon. The same is also true when there are several forms of a binary compound because one of the elements is capable of existing in several oxidation states. A good example of multiple oxidation states are the chlorides of lead. Lead(IV) chloride ($PbCl_4$) is covalent whereas lead(II) chloride ($PbCl_2$) is ionic.

8 Calculate where the following compounds will appear on a van Arkel diagram. Then use the van Arkel diagram to determine the type of bonding predicted to be present: gallium nitride, GaN, hydrogen fluoride, HF, caesium hydride, CsH, aluminium chloride, $AlCl_3$, sulfur(II) chloride, SCl_2, and potassium chloride, KCl.

Worked example

Calculate where phosphorus(V) chloride, PCl_5, will appear on a van Arkel diagram. Deduce the nature of the bonding from the van Arkel diagram.

The average electronegativity = (2.2 + 3.2)/2 = 2.7

The difference in electronegativity = 3.2 – 2.2 = 1.0

On the van Arkel diagram, the coordinates for PCl_5 are (2.7, 1.0).

22.2 Metals and inductively coupled plasma (ICP) spectroscopy – *metals can be extracted from their ores and alloyed for desired characteristics. ICP-MS/OES spectroscopy ionizes metals and uses mass and emission spectra for analysis*

Reduction of metals

Some unreactive metals, such as gold, platinum and silver, are found pure (native) and can be mined directly as the element (Figure 22.14). However, the more reactive metals exist in the earth

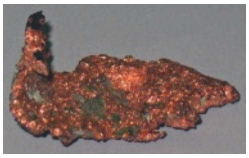


Figure 22.14 A sample of copper in its native state

in their oxidized states in compounds; for example, iron is often found as the ore hematite, impure iron(III) oxide, Fe_2O_3 [$2Fe^{3+}$ $3O^{2-}$] and aluminium as the ore bauxite (hydrated aluminium oxide, Al_2O_3 [$2Al^{3+}$ $3O^{2-}$]). These metals can be extracted via chemical reduction from their ores and can then be alloyed with carbon and other metals to give them useful physical properties.

Reactive metals in ores are in an oxidized state, and they need to be reduced to the elemental form (oxidation state zero). Chemical reduction (smelting) by coke (carbon), a redox reaction (replacement) with a more active metal, and electrolysis of a molten ionic compound are methods used to obtain metals from their ores (Figure 22.15).

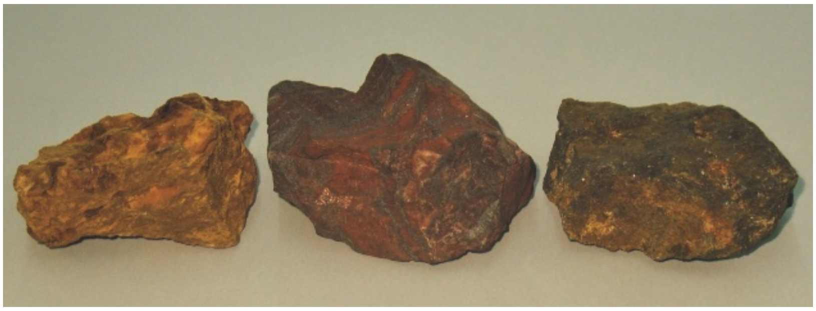


Figure 22.15 Samples of the major iron ores. **a** limonite, **b** hematite and **c** magnetite

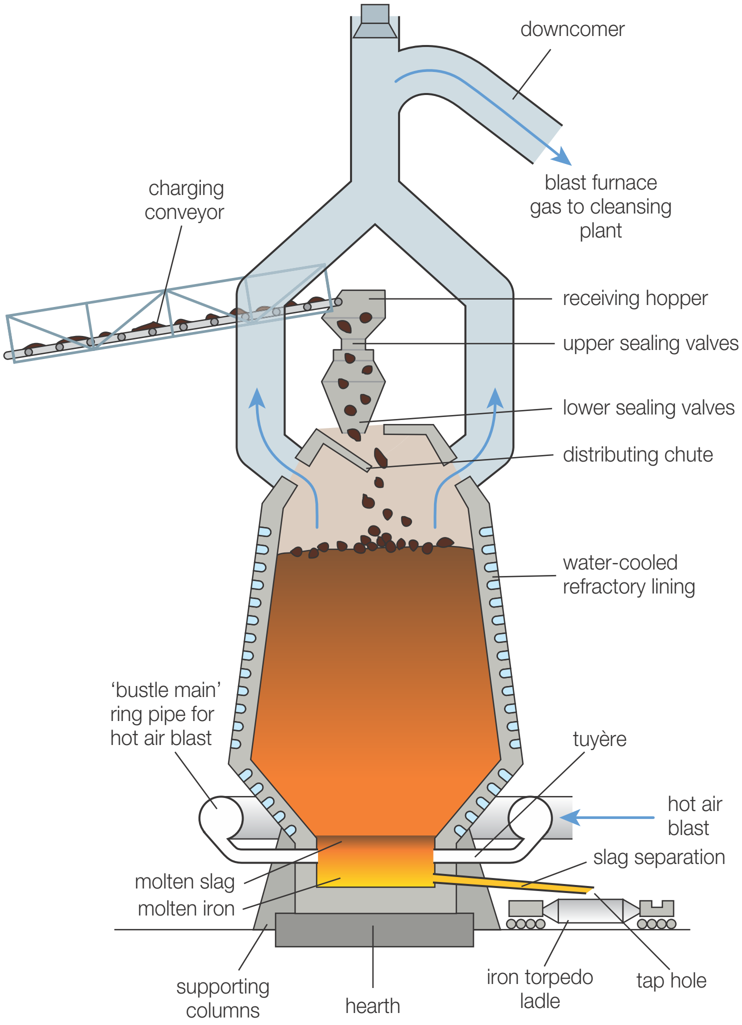


Figure 22.16 A diagram of the blast furnace for extracting iron

Chemical reduction of iron ore in the blast furnace

Reduction is carried out on a large scale industrially in the blast furnace (Figure 22.16) to obtain iron from iron ore (impure oxides of iron). Most of the iron extracted is then processed to produce steel. Iron ore is mainly the oxides hematite-iron(III) oxide, Fe_2O_3, and magnetite, Fe_3O_4, a mixed oxide which behaves like $FeO.Fe_2O_3$, and these reduced by carbon in the impure form of coke in a blast furnace. Coke is heated to form carbon dioxide which reacts with the excess coke to form carbon monoxide in the bottom of the furnace where air is blasted in:

$$C(s) + O_2(g) \rightarrow CO_2\ (g);$$

$$CO_2\ (g) + C\ (s) \rightarrow 2CO(g)$$

Carbon monoxide is a powerful reducing agent (it is easily oxidized to carbon dioxide) and reacts with the iron which is collected from the base of the furnace as a liquid:

$$Fe_2O_3(s) + 3CO(g) \rightarrow 2Fe(l) + 3CO_2(g)$$

and

$$Fe_3O_4(s) + CO(g) \rightarrow CO_2(g) + 3FeO(s)$$

then

$$FeO(s) + CO(g) \rightarrow CO_2\ (g) + Fe(l)$$

At the very high temperature in the blast furnace the coke can react directly with the iron ore and carbon itself can also act as a reducing agent.

$$Fe_2O_3(s) + 3C(s) \rightarrow 2Fe(l) + 3CO(g); Fe_3O_4(s) + 3C(s) \rightarrow 2Fe(l) + 3CO(g)$$

The carbon monoxide produced in this reaction can reduce more iron oxide from the iron ore.

The iron produced is collected in the form of pig iron, ready for further processing, such as production of steels.

Reduction by a more reactive metal

A second means of obtaining elemental metals is reduction by a more chemically active metal (Chapter 9), that is, a more powerful reducing agent. Pure nickel can be obtained from nickel(ɪɪ) sulfate by a single replacement reaction with solid magnesium:

$$Mg(s) + NiSO_4(aq) \rightarrow MgSO_4(aq) + Ni(s)$$

Other redox reactions can be used to reduce the oxidized metal. For example, passing hydrogen gas over heated silver(ɪ) oxide reduces silver(ɪ) oxide to elemental silver, while the hydrogen is oxidized to the +1 state in water:

$$Ag_2O(s) + H_2(g) \rightarrow 2Ag(s) + H_2O(l)$$

Reduction by a more active metal or by carbon cannot be used to reduce metals near the top of the activity series (see *IB Chemistry data booklet*, section 25) such as group 1, group 2 and aluminium. In this case electrolysis (Chapter 9) allows chemists to obtain the metals in a very pure state. Once obtained the elemental metals must not be exposed to air (oxygen) or they become oxidized again.

Lithium is used in high-voltage lithium and lithium ion batteries (Chapter 23) and obtained by the electrolysis of molten lithium chloride to produce lithium metal and chlorine gas:

Cathode: $Li^+(l) + e^- \rightarrow Li(l)$

Anode: $2Cl^-(l) \rightarrow Cl_2(g) + 2e^-$

Overall: $2LiCl(l) \rightarrow 2Li(l) + Cl_2(g)$

Relating the method of extraction to the position of a metal on the activity series

Metals above carbon in the activity series cannot be reduced by carbon; those below carbon in the activity series can be reduced by heating with carbon (smelting). Metals below hydrogen in the activity series can be reduced by heating hydrogen; those above hydrogen cannot be reduced by hydrogen.

9 Predict which of the following metal oxides can be reduced by heating with (i) carbon and (ii) hydrogen: lead(II) oxide, gold(I) oxide, lithium oxide, manganese(IV) oxide and zinc oxide.

Deduction of redox equations for the reduction of metals

The activity series of metals needs to be consulted to determine whether the redox reaction occurs. If the metal is below carbon or hydrogen the redox reaction occurs and a balanced equation can be written.

For example, if gold(ɪ) oxide is heated with carbon or hydrogen then reduction occurs because gold is below both these reducing agents in the activity series:

$$Au_2O(s) + H_2(g) \rightarrow 2Au(s) + H_2O(l)$$

$$Au_2O(s) + C(s) \rightarrow CO(g) + 2Au(s)$$

Gold(I) oxide can also be reduced by heating it with a metal higher up the activity series – that is, reacting it with a more powerful reducing agent, for example:

$$Au_2O(s) + Mg(s) \rightarrow MgO(s) + 2Au(s)$$

In contrast, zinc is located between carbon and hydrogen in the activity series meaning that it will undergo reduction with carbon but not with hydrogen.

Beryllium is located above carbon and hydrogen, meaning that it will not undergo reduction with either reagent.

10 Write balanced equations for the reaction of mercury(II) oxide, HgO, with carbon, hydrogen and chromium.

Quantitative electrolysis

The quantity of metal ions reduced at the cathode during electrolysis can be calculated using the current passed through the electrolytic circuit, the time it is passed for, and the Faraday constant. This is the charge in coulombs (C) carried by 1 mole of electrons and has the value 96 500 C mol^{-1}. For example, in the reduction of magnesium from its cation:

$$Mg^{2+} + 2e^- \rightarrow Mg$$

the equation shows that 2 moles of electrons are required to reduce 1 mole of magnesium ions. Providing 1 mole of electrons requires 96 500 C of charge from the electrolysis circuit.

The amount of charge (coulombs), Q, transferred can be calculated from the current I (in amperes, A) and time, t (in seconds): $Q=It$.

The SI unit of current, the ampere, is one coulomb per second; 1 A = 1 C s^{-1}. To reduce 1 mole of aluminium atoms by electrolysis would require 3 moles of electrons:

$$Al^{3+} + 3e^- \rightarrow Al$$

Solving stoichiometric problems using Faraday's constant based on mass deposits in electrolysis

Most stoichiometric problems involving electrolysis can be solved without the explicit use of Faraday's laws (Chapter 18). The following relationships are needed: quantity of charge (C) = current (A) × time (s) and amount of electrons = quantity of charge (C)/Faraday constant (96 500 C mol^{-1}).

Worked example

Calculate the mass of copper that will be deposited if a current of 0.22 A flows through a cell for 1.5 hours.

The amount of charge passing through the cell is (0.22 A) × (5400 s) = 1200 C; (1200 C) ÷ (96 500 C mol^{-1}) = 0.012 mol (of electrons).

Since the reduction of 1 mole of copper(II) ions requires the addition of 2 moles of electrons, the mass of copper deposited will be: (63.54 g mol^{-1}) × (0.5 mol Cu/mol of electrons) × (0.012 mol of electrons) = 0.39 g of copper.

11 Calculate the mass of cadmium metal that plates on to the cathode during a 3.00 hour period of electrolyzing aqueous cadmium(II) sulfate using a current of 755 mA.

12 Calculate the period of time for which a current of 2.68 A must be passed through a solution of gold(I) cyanide in order to plate out 5.00 g of gold.

Determining the Faraday constant

The Faraday constant (the amount of charge carried by 1 mole of electrons) can be determined from the quantitative electrolysis of copper(II) sulfate solution, using weighed copper electrodes.

The reaction at the cathode is $Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$ and the reaction at the anode is $Cu(s) \rightarrow Cu^{2+}(aq) + 2e^-$. Hence the concentration of copper(II) ions remains constant and there is a transfer of copper anode to the cathode.

The copper electrodes should be cleaned with fine sandpaper and dried thoroughly. A small steady current is passed through the electrolyte through a known length of time.

The electrode to be plated is weighed before (m_i = initial mass) and after (m_f = final mass) the experiment. The difference in these masses represents the mass of plated metal:

$$m = m_f - m_i$$

The electrical charge that flows through the system during electrolysis, Q, can be calculated using the following equation:

$$Q = It$$

where I is the current (in amps) and t is the total running time (in seconds).
Avogadro's constant, N_A, can then be calculated using this equation:

$$N_A = \frac{QM}{nmq_e}$$

where M is the atomic mass of the metal, n is the number of electrons in the half-reaction and q_e is the charge on one electron.

Faraday's constant, F, can be calculated using this equation:

$$F = \frac{QM}{nm}$$

13 Calculate the value of the Faraday constant from the following data:

Current passed through the cell = 0.500 A

Time current was passed through cell = 30.0 minutes

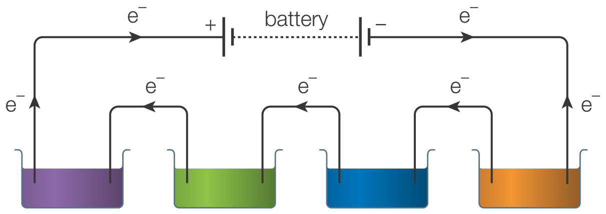
Initial mass of copper cathode = 52.243 g

Final mass of copper cathode = 54.542 g

Electrolytic cells in series

When two or more electrolytic cells are connected in series (Figure 22.17) the same electric current and operating time apply to all the electrolytic cells; that is, the current, I, and the time, t, are identical for cells connected in series and as the charge is the product of the charge and time, $Q = I \times t$, this means that cells in series all receive the same quantity of electric charge, Q value.

Figure 22.17 Four electrolytic cells in series



That is, for electrolytic cells 1, 2, 3 and 4 in series, Q_1 (for cell 1) = Q_2 (for cell 2) = Q_3 (for cell 3) = Q_4 (for cell 4).

14 Two electrolytic cells are connected in series so that the same current (quantity of charge) flows through both cells for the same length of time.

Descriptions of the two cells are:

In cell 1 the electrolyte is silver(I) nitrate ($AgNO_3(aq)$) and the electrodes are inert.

In cell 2 the electrolyte is chromium(III) nitrate ($Cr(NO_3)_3(aq)$) and the electrodes are inert.

If 0.785 g of silver metal (Ag(s)) plates on to the cathode in cell 1 during electrolysis, deduce the corresponding mass of chromium metal that plates on to the cathode of cell 2.

Explanation of the production of aluminium by the electrolysis of alumina in molten cryolite

Occurrence of aluminium

Aluminium is the most abundant metal and the third most abundant element on Earth. It is in compounds that form approximately 8% of the Earth's crust. Aluminium is relatively reactive and does not occur naturally as the metal. It is found in the form of hydrated aluminium silicates in rocks, such as clays and micas, but the percentage of aluminium is too low for commercial extraction. The main ore is bauxite, hydrated aluminium oxide, $Al_2O_3 \bullet xH_2O$ (where x ranges from 1 to 3). It is formed by the weathering of clays – large bauxite deposits are found in Jamaica and Australia.

Purification of bauxite

Crude bauxite contains iron(III) oxide, silicon dioxide (silica) and other impurities. In the Bayer process, the bauxite ore is first crushed and then mixed with concentrated aqueous sodium hydroxide solution. The mixture is pumped into a 'digester' where it is heated under high pressure. The soluble complex tetrahydroxoaluminate(III) ion is formed:

$$Al_2O_3(s) + 2OH^-(aq) + 3H_2O(l) \rightarrow 2[Al(OH)_4]^-(aq)$$

Iron(III) oxide is removed from the mixture as 'red mud' by allowing it to settle out. The solution is then filtered and transferred to a precipitation tank where it is seeded with crystals of pure aluminium hydroxide. On seeding, the aqueous sodium tetrahydroxoaluminate(III) solution decomposes, forming aluminium hydroxide. This grows into large crystals on the seed crystals:

$$[Al(OH)_4]^-(aq) \rightarrow Al(OH)_3(s) + OH^-(aq)$$

The sodium hydroxide formed in this process is recycled. The aluminium hydroxide crystals are filtered, washed and then 'roasted' in a rotary kiln at about 1000 °C. Pure aluminium oxide (alumina) is formed in a dehydration reaction:

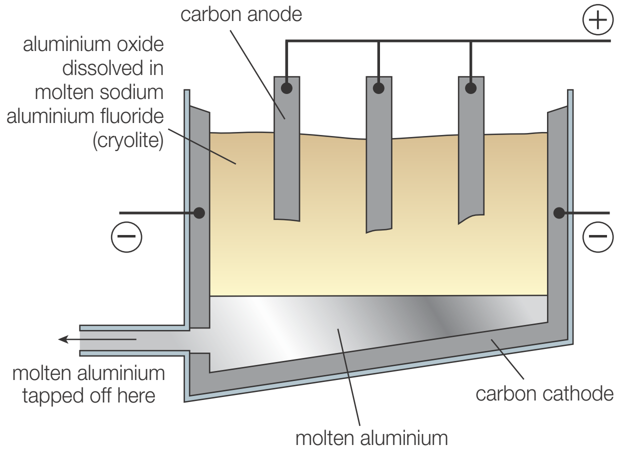
$$2Al(OH)_3(s) \rightarrow Al_2O_3(s) + 3H_2O(l)$$

It is worth noting that this purification process is based on the fact that iron(III) oxide is, like the majority of metal oxides, a basic oxide and does not react with sodium hydroxide solution. In contrast, aluminium oxide is an amphoteric oxide and so reacts with alkali to produce a salt, sodium tetrahydroxoaluminate(III) – sometimes called sodium aluminate.

Reduction of aluminium oxide by electrolysis

The melting point of aluminium oxide is 2045 °C. The use of pure molten aluminium oxide as an electrolyte is thus not practicable or economical. The temperature is significantly decreased by using a solution of aluminium oxide in molten cryolite as the electrolyte. Cryolite is an aluminium ore with the formula Na_3AlF_6.

The electrolysis is carried out in a Hall–Héroult cell (Figure 22.18). Liquid aluminium is produced at the graphite cathode (negative electrode) where it is tapped off. The aluminium is over 99% pure. The molten cryolite in the electrolyte remains unchanged, and so more aluminium oxide can be added as required. Oxygen is produced at the graphite anodes (positive electrodes) which gradually burn away and have to be replaced periodically.


carbon anode
+
aluminium oxide dissolved in molten sodium aluminium fluoride (cryolite)
−
−
molten aluminium tapped off here
carbon cathode
molten aluminium

■ **Figure 22.18** Cross-sectional diagram of an electrolysis cell for extracting aluminium

The half-equations at the electrodes are:

- cathode (negative electrode): $Al^{3+}(l) + 3e^- \rightarrow Al(s)$
- anode (positive electrode): $2O^{2-}(l) \rightarrow O_2(g) + 4e^-$
  $2C(s) + O_2(g) \rightarrow 2CO(g)$
  $C(s) + O_2(g) \rightarrow CO_2(g)$

The overall equation for the cell reaction is:

$$4Al^{3+}(l) + 6O^{2-}(l) \rightarrow 4Al(s) + 3O_2(g)$$

Alloys

An alloy is typically a homogeneous mixture of metals or a mixture of a metal and non-metal, usually carbon but sometimes phosphorus. Alloys usually have different properties from those of the component elements, and melt over a range of temperature. Alloys are generally made by mixing the metal with the other elements in molten form, and allowing the mixture to cool and the alloy to solidify.

Alloys are usually divided into ferrous and non-ferrous alloys. Ferrous alloys include the steels, which are alloys of iron containing up to 2% carbon. The majority of non-ferrous alloys are based on copper. Familiar alloys of copper are brass (copper and zinc) and bronze (copper and tin). Bronze is an alloy that humans have used considerably over long periods of history. It is still used to produce statues and sculpted artwork.

The composition of steels (after reaction with acid) can be determined by using inductively coupled plasma optical emission spectroscopy (ICP-OES). The intensities of the lines are related to the concentrations of atoms of each element.

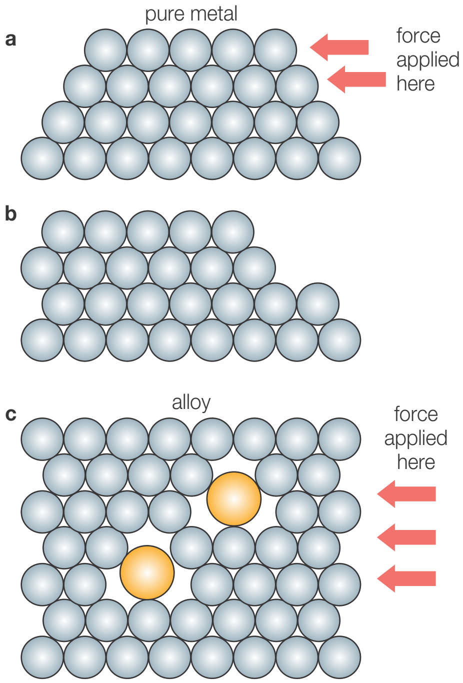


Figure 22.19 **a** The position of the atoms in a pure metal before a force is applied; **b** after the force is applied, slippage has taken place; **c** in an alloy, slippage is prevented because the atoms of different size cannot easily slide over each other

Alloying

Most pure metals are not used in engineering because they do not have the required properties. For example, the pure metal may readily undergo corrosion or be too soft. However, the properties of a metal may be improved by the formation of an alloy. Alloys are often harder than the original metals because the irregularity in the structure helps to stop rows of metal atoms from slipping over each other (Figure 22.19).

For example, brass is stronger than copper and is more easily worked because of its higher malleability – brass is also more resistant to corrosion. Duralumin (an alloy of aluminium with magnesium and copper) is much stronger than aluminium, but the presence of copper makes it prone to corrosion. Stainless steel contains iron with chromium, nickel and a small amount of carbon. It is extremely hardwearing and resistant to corrosion even when heated – its resistance to corrosion can be improved by increasing the chromium content.

The structure of metals

Metals are giant structures of atoms held together by metallic bonding (Chapter 4). The term 'giant' implies that large but variable numbers of atoms are involved – depending on the size of the piece of metal. Metals are crystalline and the atoms pack into various types of lattices. However, not all the atoms in a piece of metal are arranged in a regular way. Any piece of metal is made up of a large number of crystal **grains**, which are regions of regularity (Figure 22.20). Hence metals are described as polycrystalline. At the grain boundaries, atoms are misaligned. The grains of a piece of polished metal can be seen easily with a microscope – however, the best place to see metallic crystal grains is on a galvanised lamp post (Figure 22.21) where the large grains of zinc are clearly visible.

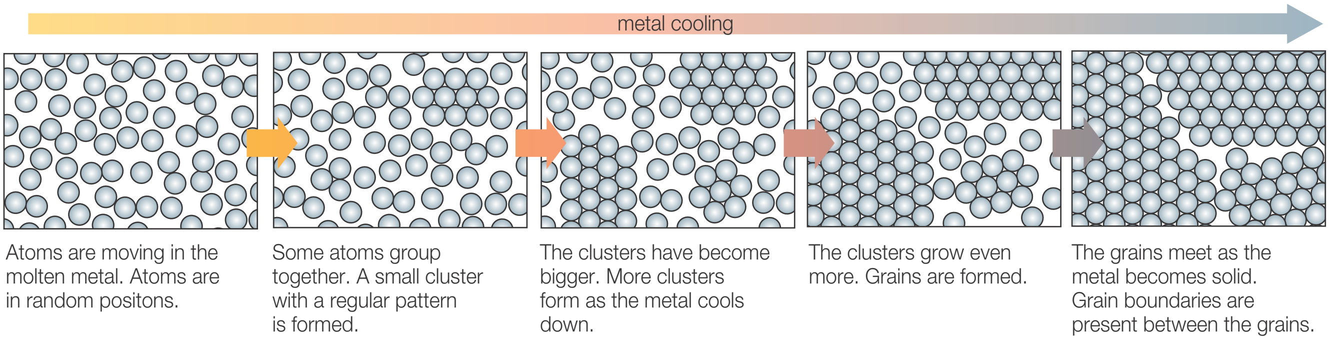

metal cooling
Atoms are moving in the molten metal. Atoms are in random positons.
Some atoms group together. A small cluster with a regular pattern is formed.
The clusters have become bigger. More clusters form as the metal cools down.
The clusters grow even more. Grains are formed.
The grains meet as the metal becomes solid. Grain boundaries are present between the grains.

Figure 22.20 The process of formation of grains as a metal cools

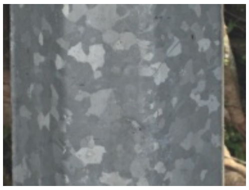


Figure 22.21 The grains of zinc on a galvanized lamp post can be seen clearly

The properties of steel depend on its grain structure. The structure of steel may change with its composition and with mechanical working (e.g. rolling, forging, heating, hammering and drawing).

A third important process for changing a metal's structure, and hence changing and controlling its properties, is known as heat treatment. The remarkable versatility of steel is in large part due to its response to heat treatment. If some steel is to be formed into an intricate shape, it can be made very soft and ductile by heat treatment; on the other hand, if it is to resist wear then it can be made very hard by heat treatment.

Many methods of heat treatment are used, but they all involve heating and keeping the metal at a selected temperature until the structure becomes uniform, and then cooling it at a controlled rate to produce or keep the required microcrystalline structure of grains.

The rate of cooling is the main difference in these treatments. Iron and carbon combine reversibly at high temperatures forming a hard brittle compound known as iron carbide (cementite), Fe_3C.

$$3Fe(s) + C(s) \rightleftharpoons Fe_3C(s)$$

The mechanical properties of steel result from laminated structures of iron and iron carbide. The forward reaction is endothermic, so that on slow cooling, the position of equilibrium (Chapter 7) shifts towards the formation of carbon, which separate as small flakes of graphite, and iron. Rapid cooling, however, prevents adjustment of equilibrium. The iron carbide is preserved, making the steel harder and much more brittle. The three main methods of heat treatment are known as **annealing**, **quenching** and **tempering**.

Annealing involves maintaining the sample of steel at a specified temperature for a specific length of time, and then gradually cooling it at a controlled rate. The steel is softened, becoming very ductile, but the hardness, toughness and tensile strength are gradually reduced. Annealing relieves internal stresses, producing a more or less uniform grain structure throughout the metal. Annealing is often used to soften steel and prepare it for machining, cold working or further heat treatment. It may also be used to develop a particular microstructure in steel.

Quenching involves plunging a piece of heated steel into a liquid bath, causing sudden cooling. The product has strong internal stresses and is very hard and brittle, being easily fractured. To be useful, the steel needs to be toughened.

Tempering is carried out by reheating quenched steel to a specified temperature for a given time to permit the structural changes to reach equilibrium, and then cooling slowly. By tempering, it is possible to remove the internal stresses and to replace brittleness by toughness, while retaining most of the hardness.

Nature of Science

History of metallurgy

The process of metallurgy is one of the oldest applied sciences. Its history can be traced back to 6000 BCE. Currently there are 86 known metals. Before the 19th century only 24 of these metals had been discovered and, of these 24 metals, 12 were discovered in the 18th century. Therefore, from the discovery of the first metals – gold and copper – until the end of the 17th century, some 7700 years, only 12 metals were known. Four of these metals, arsenic (a metalloid), antimony, zinc and bismuth, were discovered in the 13th and 14th centuries, while platinum was discovered in the 16th century. The other seven metals, known as the Metals of Antiquity, were the metals upon which civilization was based. These seven metals and their dates of first use were:

1 Gold, 6000 BCE 2 Copper, 4200 BCE 3 Silver, 4000 BCE 4 Lead, 3500 BCE

5 Tin, 1750 BCE 6 Iron, smelted, 1500 BCE 7 Mercury, (approximately) 750 BCE

These metals were known to the Mesopotamians (in Iraq), Egyptians, Greeks and the Romans. Of the seven metals, five can be found in their native states – gold, silver, copper, iron (from meteors) and mercury. However, the occurrence of these metals was not abundant and the first two metals to be used widely were gold and copper.

Sometimes the ores of copper and tin are found together, and the casting of metal from such natural alloys may have provided the accident for an important step forward in metallurgy – the use of alloys. It was discovered that these two metals, cast as one substance, are harder than either metal on its own. Initially metallurgy was more of an art than a science, relying upon 'trial-and-error' experimentation, but later it followed a more scientific route based upon the many methods used to investigate indirectly, such as X-ray diffraction and electron microscopy.

The cast alloy of copper and tin is bronze, a substance so useful to human beings that an entire period of early civilization has become known as the Bronze Age. A bronze blade will take a sharper edge than copper and will hold it longer. And bronze ornaments and vessels can be cast for a wide variety of purposes.

In the 11th century it was discovered that iron can be much improved. If it is reheated in a furnace with charcoal (containing carbon), some of the carbon is transferred to the iron, forming steel. This process hardens the metal, and the effect is considerably greater if the hot metal is rapidly reduced in temperature, usually achieved by quenching it in water. It can be worked (or 'wrought') just like softer iron, and it will keep a finer edge, capable of being honed to sharpness.

Paramagnetic and diamagnetic materials

One important physical property of a material (often a metal, alloy or complex ion) is its response to a strong magnetic field. Paramagnetic materials are attracted to a magnetic field but diamagnetic materials create a magnetic field opposed to the applied magnetic field and are therefore weakly repelled by an external magnetic field. A third type of magnetism is known as ferromagnetism. Ferromagnets will tend to stay magnetized to some extent after being subjected to an external magnetic field.

Liquid oxygen is an example of a paramagnetic substance (Chapter 14) and liquid nitrogen is an example of a diamagnetic substance.

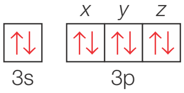


Figure 22.22 Valence electron configuration of an argon atom

In the atoms of a diamagnetic substance the electrons are spin paired; for example, argon atoms are diamagnetic with the electron configuration $1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6$ (the valence electron configuration shown in Figure 22.22) and the 18 electrons exist as spin pairs.

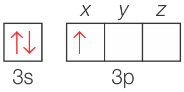


Figure 22.23 Valence electron configuration of an aluminium atom

Aluminium atoms (Figure 22.23), electron configuration $1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^1$, have one unpaired p electron that is attracted to an external electric field. Aluminium (in bulk) is paramagnetic. The electron has a magnetic dipole, and a spinning electron creates a local magnetic field (Figure 22.24).

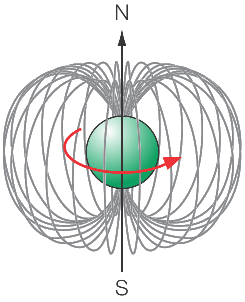


Figure 22.24 A spinning electron and its magnetic field; its magnetic dipole is pointing in the S to N direction

Because electron spin is quantized, an electron has only two possible spin states. The magnetic field produced by an electron occurs in one of two directions. In one spin state the electron produces a magnetic field with the North pole in one direction. In the other spin state the North pole is in the opposite direction.

The spins of unpaired electrons in atoms, molecules and ions can be temporarily aligned in an external magnetic field, causing the substance to be attracted to the applied magnetic field. This is what happens in paramagnetic substances.

In the absence of an external, strong magnetic field, the magnetic fields generated by the individual particles in paramagnetic substance are arranged in random directions and the magnetism produced by each atom, ion or molecule will be cancelled by the magnetic fields around it.

In a ferromagnetic material the electron alignment induced by the magnetic field can be retained, making a permanent magnet. For example, a sample of iron atoms (electron configuration [Ar] $4s^2\ 3d^6$) can be heated and cooled in a strong magnetic field (such as a solenoid) and as the metal cools the unpaired electrons in the iron atoms align themselves such that the magnetic field created by their spin is aligned with the applied field.

Banging or heating a permanent magnet (above a certain temperature) or placing it near alternating current can disrupt this alignment and weaken the strength of the magnet. Paramagnetic materials do not form permanent magnets in this way; their electrons are only temporarily aligned by the external magnetic field. Nickel and chromium also exhibit ferromagnetic behaviour.

In diamagnetic materials (Figure 22.25) all the electrons are paired. In an external magnetic field the paired electrons orientate themselves such that the magnetic fields created by their spin opposes the applied magnetic field (Lenz's law of electromagnetism, which is studied in IB Physics) and so the material will weakly repel the external magnetic field. A superconductor exhibits perfect or extreme diamagnetism (see section 22.8).

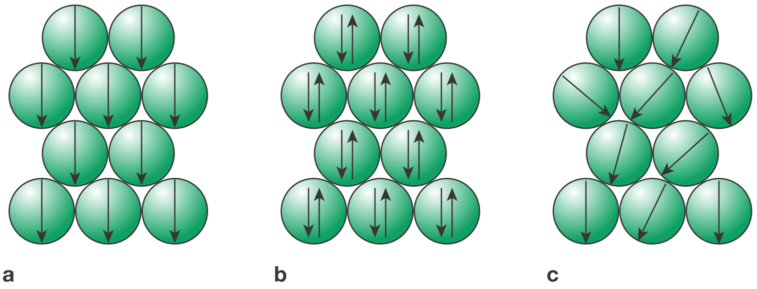
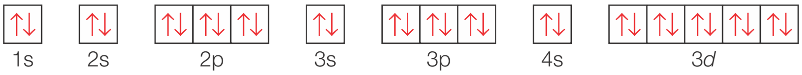


Figure 22.25 Electron spin representations of a material with **a** diamagnetic, **b** paramagnetic and **c** ferromagnetic properties

Discussion of paramagnetism and diamagnetism in relation to electron structure of metals

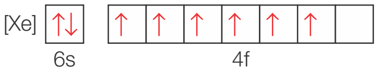
The magnetic character of a metal can be determined from its detailed electron configuration.

For example, zinc has the atomic number 30 and has the following detailed electron configuration:



Zinc is diamagnetic because its electron configuration shows no unpaired electrons.

However, samarium (atomic number 62) is paramagnetic because its electron configuration shows six unpaired electrons:



15 Classify sodium, magnesium, tin, lead and the first row transition elements from scandium to zinc as paramagnetic, diamagnetic or ferromagnetic.

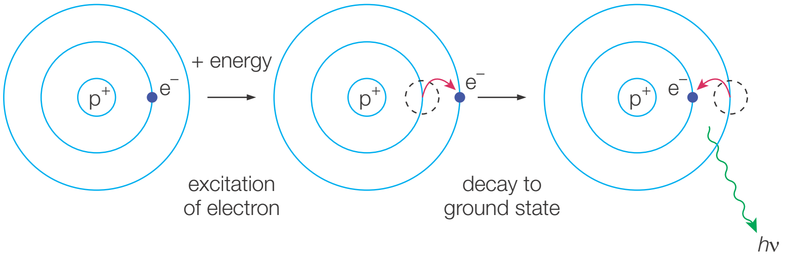
16 Find out about anti-ferromagnetic substances.

Spectroscopic methods

Trace concentrations of elements such as heavy metal atoms or their ions in polluted water are difficult to determine by chemical tests (e.g. titration) but can be detected by spectrophotometric techniques. Qualitative analysis to determine which metals or cations are present in a sample can be done by exciting electrons to higher energy levels and detecting and measuring the radiation (photons) emitted as these electrons undergo a transition to lower energy levels or the ground state.

Concentrations of trace metals and their cations can be determined by optical emission spectroscopy (OES), which is based on the direct relationship between analyte concentration and atomic/ionic emission intensity (visible or ultraviolet radiation) (Figure 22.26).

Figure 22.26 Emission of radiation from excited electrons as they return to the ground state



In mass spectrometry (MS), an ionization source converts gaseous species into cations (traditionally by electron bombardment). These cations are accelerated and enter the mass spectrometer, where they are separated according to their mass to charge ratio by a combination of electric and magnetic fields. The detector receives a signal proportional to the concentration of each analyte. The cation may be a monoatomic or polyatomic ion.

These spectroscopic techniques require gaseous atoms in an excited state or cations in the gas phase. Substances that are solids or liquids must be atomized for spectroscopic analysis and this is achieved by heating (often at low pressure) and/or electrical discharge, which bombards atoms with electrons of high kinetic energy to excite them or ionize them (by removing one or more valance electrons).

Inductively coupled plasma

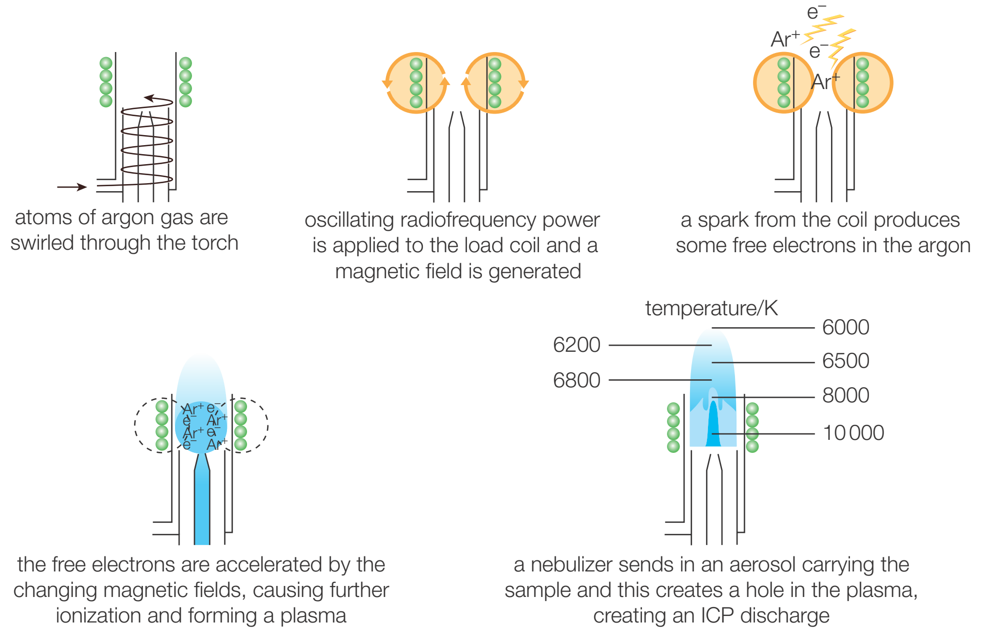
When argon is heated to temperatures above 6000 K a plasma is formed. This is a gas with a high percentage of positive ions (cations), $Ar^+(g)$, and free electrons. The plasma may be produced by a direct current discharge – in effect, passing a powerful electric spark through the argon.

A plasma may also be produced by the use of an inductively coupled plasma (ICP) torch. Plasma is hotter and much more effective at atomizing, ionizing and exciting chemical species. Since it contains charged particles (cations and electrons) it can be contained and directed using strong magnetic fields.

Discharge of a high voltage (from a coil) through flowing argon gas will supply free electrons which will 'ignite' the gas to form a plasma which is conducting due to the presence of charged particles. If the plasma is surrounded and enclosed by an electromagnetic field (oscillating at high frequency), then the ions and electrons will be accelerated and they will collide with the argon atoms (known as the support gas) and the analyte, the chemical to be analysed. The collisions cause the temperature to increase to about 10 000 K and a self-sustaining plasma is formed. It is held in place by the magnetic field in the form of a fireball.

The analyte is introduced as an aerosol (droplets of liquid in a gas) and enters the fireball at high speed. It is pushed through it, and becomes heated and forms a plume which contains free atoms or free ions. The atoms or ions of the analyte cool and at around 600–700 K they enter the ground state from their excited states. They come back to the ground state because of the excitation lifetime related to each element (these are known as Einstein coefficients). Atomic and ionic emissions take place at very high temperatures, for example 6000–7500 K. At low temperatures such as 600–700 K, recombination reactions take place, which result in molecules that do not emit radiation. This process is known as relaxation. Figure 22.27 shows a cross-section of an ICP torch and coil showing an ignition sequence and formation of a plasma.

Figure 22.27 A cross-section of an ICP torch and coil showing an ignition sequence and formation of a plasma



atoms of argon gas are swirled through the torch

oscillating radiofrequency power is applied to the load coil and a magnetic field is generated

a spark from the coil produces some free electrons in the argon

the free electrons are accelerated by the changing magnetic fields, causing further ionization and forming a plasma

a nebulizer sends in an aerosol carrying the sample and this creates a hole in the plasma, creating an ICP discharge

Characteristic emission spectral lines are produced. This technique is known as inductively coupled plasma optical emission spectroscopy (ICP-OES). If part of the plasma plume is directed into a mass spectrometer, the isotopic masses of individual elements present can be identified. This technique is known as ICP-mass spectrometry (ICP-MS).

Argon gas is supplied at 10–15 $dm^3\,min^{-1}$ through the three concentric quartz tubes of the torch. The tangential flow of gas in the outer tube contains the plasma, while the central tube carries the nebulized sample droplets suspended in argon. Tangential means the swirling argon gas is flowing at right angles to the orientation of the vertical quartz tubes.

The plasma is produced by high-voltage ignition and sustained by the magnetic field of the radio frequency generator. The sample is pumped into the nebulizer and the smallest droplets are carried forward by the gas, while other, larger drops flow to waste from the spray chamber. High-solids nebulizers, where particulate matter is introduced into the ICP, have been developed. Laser ablation, where the sample is vaporized by a laser, is also used. Electrothermal vaporization may also be used for solids and liquids. The sample undergoes a sequence of processes to generate excited atoms and cations:

- The production of an aerosol from the solution (nebulization)

  Removal of solvent: $MX(aq) \rightarrow MX(s)$

  Vaporization of the sample: $MX(s) \rightarrow MX(g)$

  Atomization: $MX(g) \rightarrow M\bullet(g) + X\bullet(g)$

  Excitation: $M(g) \rightarrow M\bullet^{*}(g)$

  Emission: $M\bullet^{*}(g) \rightarrow M\bullet(g) + h\nu$

- Ionization may also occur:

  Ionization: $M(g) \rightarrow M^{+}(g) + e^{-}$

  Excitation: $M^{+}(g) \rightarrow M^{+*}(g)$

  Emission: $M^{+*}(g) \rightarrow M^{+}(g) + h\nu$

Figure 22.28 shows how metal atoms and ionic compounds containing metals can be vaporized, ionized and excited in an ICP discharge (or flame). These stages depend on the experimental variables used in the ICP instrument, such as the viscosity of the solvent, the nature of the solvent, the rate of fuel flow and the plasma temperature (which depends on the radio frequency applied power and sample introduction flow rate).

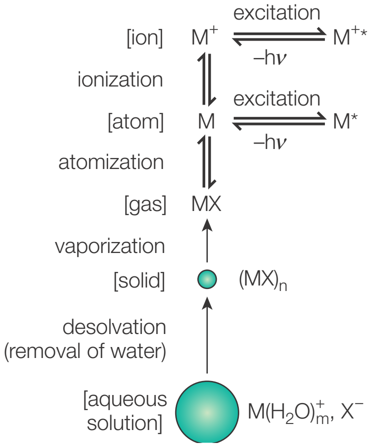
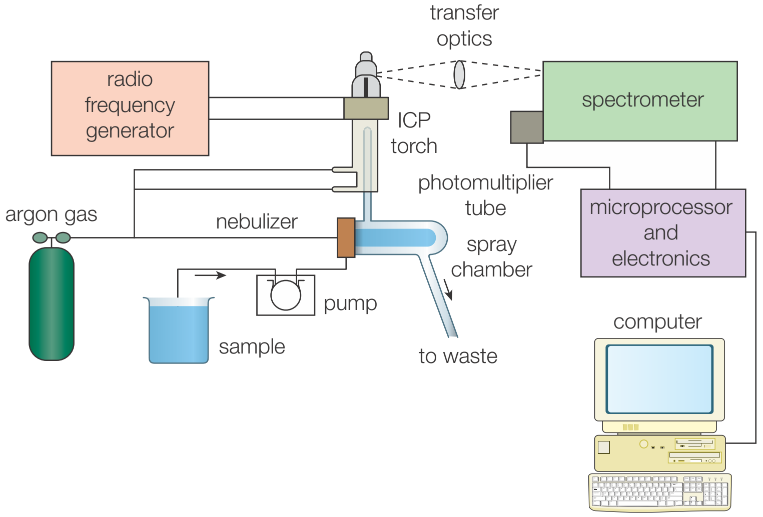
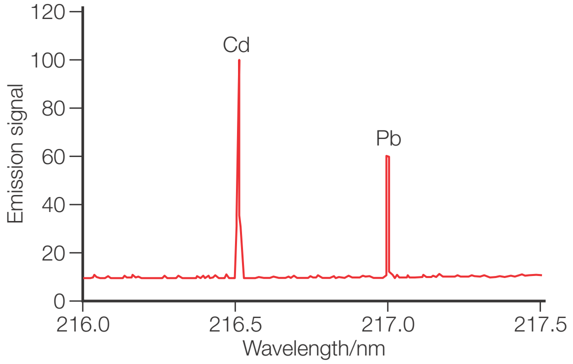


Figure 22.28 The vaporization and ionization of metals and metallic compounds in an ICP discharge (assuming the compound is in aqueous solution)

The optics of the ICP-OES spectrometer are aligned with the base of the plasma plume where the majority of the atomic and ionic relaxation is occurring. The emitted radiation from the ICP torch is focused into the spectrograph (composed of a grating that separates the different wavelengths of radiation, and a prism that separates the different wavelengths orders) and detected by a photomultiplier tube (most common in modern ICP-OES instruments is a CCD camera) (Figure 22.29).

Figure 22.29 Schematic of an ICP-OES spectrometer





■ **Figure 22.30** ICP-OES spectrum of a solution containing 60 ng mL^{-1} Pb and 100 ng mL^{-1} Cd

The output is then processed and displayed under computer control as the inductively coupled plasma-atomic emission spectrum (ICP-AES) (Figure 22.30).

ICP-OES can detect a greater number of elements than other atomic emission of atomic absorption techniques, such as atomic absorption spectroscopy (AAS). However AAS is much more sensitive than ICP-OES. While ICP-OES reaches high parts per billion (ppb) levels, AAS can reach parts per trillion (ppt) levels. The main drawback of AAS is that it is a mono-element method, while ICP is a simultaneous multi-element method. For example, at 1–10 ppb ICP-OES can measure over 30 elements, while AAS is restricted to around 10 elements.

Inductively coupled plasma spectrometry (ICP-MS)

By extracting the atoms from the cooling plasma, the high sensitivity and selectivity of the mass spectrometer may be used.

A horizontal ICP torch is placed next to a water-cooled aperture or small hole (although modern ICP-MS instruments do not use water-cooled sampler and skimmer cones any more) placed in the sampling cone. The positive ions, initially at atmospheric pressure in the plasma, are skimmed down through water-cooled metal cones through small orifices into progressively lower-pressure regions until the sample ions enter the mass spectrometer.

Explanation of the plasma state and its production in ICP-MS/OES

The plasma plays a very different role in ICP-MS than it does in ICP-OES. In both techniques, the plasma is produced by the interaction of an intense magnetic field (produced by radio waves passing through a copper coil) on a tangential flow of argon gas flowing through a concentric quartz tube (torch).

■ **Figure 22.31** ICP torch

This ionizes the gas and, when supplied with a source of electrons from a high-voltage spark, forms a very high temperature plasma discharge (~10 000 °C) at the open end of the tube (Figure 22.31). A Tesla coil provides the seed electrons to ignite the coil. From there, collisions between electrons, argon species, and sample constituents caused by the varying radio frequency field maintain the plasma.

In ICP-OES, the plasma, usually oriented axially, is used to generate photons of light by the excitation of electrons of a ground-state atom or ion to a higher energy level. When the excited electrons return to the ground state, wavelength-specific photons are emitted that are characteristic to specific elements.

In ICP-MS, the plasma torch is positioned horizontally and is used to generate cations rather than photons. It is the production and detection of large quantities of these cations that helps gives ICP-MS its ultra-trace detection capability – about three to four orders of magnitude better than ICP-OES. The sensitivity has to do more with the direct detection of analytes of ICP-MS. While ICP-OES depends on each element's emission intensity (it is an indirect method since radiation from the analyte is detected), in ICP-MS the analytes are directly detected as ions. Because the temperature in the ICP is high, most elements are near 100 % ionized, and the process is much more efficient than in ICP-OES. In addition, high background signals are observed in ICP-OES (the plasma itself is an important source of background), which hinder sensitivity. In ICP-MS, in most cases the background signal is quite low.

Explanation of the separation and quantification of metallic ions by MS and OES

Mass spectrometers use the difference in mass-to-charge ratio (m/z) of ionized atoms or molecules to separate them. Therefore, mass spectroscopy allows quantitation of atoms or molecules and provides structural information by the identification of distinctive fragmentation patterns. The general operation of a mass spectrometer is to create gas-phase ions, separate the ions in space (or time) based on their mass-to-charge ratio and measure the quantity of ions of each mass-to-charge ratio. The current detected by the mass spectrometer for a specific ion is proportional to its concentration in the sample.

Sufficient energy is often available during ICP-OES to convert the atoms to ions and subsequently promote the ions to excited states. Both the atomic and ionic excited state species may then relax to the ground state via the emission of a photon. These photons have characteristic energies that are determined by the quantized energy level structure for the atoms or ions. Thus the wavelength of the photons can be used to identify the elements from which they originated. The total number of photons is directly proportional to the concentration of the originating element in the sample.

Calibration

The most important issue in ICP-MS or ICP-OES is the accuracy of the calibration curve (line). Another factor is the quality of the samples being analysed due to chemical interference. Known standards of metal compounds or metals in a solvent (usually water) are used to construct a calibration line assuming a linear relationship between analyte concentrations and analytical signals (similar to the Beer-Lambert law).

Inductively coupled plasma emission spectrometry is used to measure the concentrations of the anti-cancer drug *cis*-platin (Chapter 13). The high temperatures mean that the samples can be injected directly, without the need to separate it from any organic material. Concentrations of platinum as low as 0.1 mg dm^{-3} in tissue or body fluid (for example, blood plasma) can be measured.

Construction of a calibration line that is accurate at very low concentrations requires a series of solutions of known concentrations to be prepared by serial dilution from a more concentrated solution.

Uses of ICP-MS and ICP-OES

Categories	Examples of samples
Agricultural and food	Animal tissues, beverages, feeds, fertilizers, garlic, nutrients, pesticides, plant materials, rice flour, soils, vegetables, wheat flour
Biological and clinical	Brain tissue, blood, bone, bovine (cow) liver, fishes, milk powder, orchard leaves, pharmaceuticals, pollen, serum, urine
Geological	Coal, minerals, fossils, fossil fuel, ore, rocks, sediments, soils, water
Environmental and water	Brines, coal fly ash, drinking water, dust, mineral water, municipal wastewater, plating bath, sewage sludge, slags, seawater, soil
Metals	Alloys, aluminium, high-purity metals, iron, precious metals, solders, steel, tin
Organics	Adhesives, amino acids, antifreeze, combustion materials, cosmetics, cotton cellulose, dried wood, dyes, elastomers, epoxy glue, lubricant, organometallic, organophosphates, oils, organic solvent, polymers and sugars
Other materials	Acids, carbon, catalytic materials, electronics, fibre, film, packaging materials, paints and coatings, phosphates, semiconductors and superconducting materials

Identify metals and abundances from simple data and calibration curves provided from ICP-MS and ICP-OES

A calibration curve (line) is used to determine the unknown concentration of an element, for example lead, in a solution. The instrument is calibrated using several solutions of known concentrations. A calibration curve is produced which is continually rescaled as more concentrated solutions are used – the more concentrated solutions absorb more radiation up to a certain absorbance.

The calibration curve shows the concentration against the amount of radiation absorbed.

The sample solution is fed into the instrument and the unknown concentration of the element, for example cadmium (in urine), is then displayed on the calibration curve (Figure 22.32).

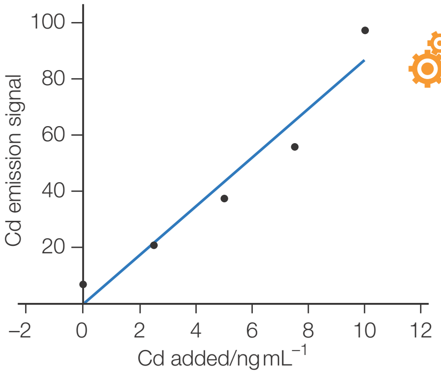


Figure 22.32 Calibration curve (line) for ICP-OES

Nature of Science

Development of ICP

In 1963 the British chemist Stanley Greenfield and his co-workers invented the inductively coupled plasma (ICP). This instrument has had a huge impact on the development of instrumental analysis. In 1969 low-power ICP was developed and in 1974 the first commercially available ICP instruments were introduced.

ICP is a type of excitation source that produces excited atoms and ions that emit electromagnetic radiation at wavelengths characteristic of a particular element. The intensity of this emission is indicative of the concentration of the element within the sample.

ICP-OES was developed by Greenfield because of the advantages of plasma emission sources over flames, ac sparks and dc arcs for the production of excited atoms and ions. A plasma source has a high degree of stability, is capable of exciting several elements and gives a high sensitivity of detection. His work was based on a sound understanding of scientific principles and recognition of the limitations of flames for controlled production of excited species.

Rare earths

The lanthanoids are also known as the 'rare earths'. Despite their name many of them are not particularly rare and have a wide variety of uses. MRI (Chapter 11) is a technique used in diagnostic medicine to produce a two-dimensional image of a 'slice' through a patient's body. It is based on the principles of nuclear magnetic resonance (NMR) and is especially useful for imaging soft and delicate tissues, such as the brain and spinal cord. Contrast agents are used in MRI to help doctors distinguish between healthy and diseased tissues. Complexes of gadolinium, Gd^{3+}, which are strongly paramagnetic are widely used. Small amounts of europium compounds are used in colour TV screens where they are responsible for their luminescence. Cerium is the most abundant rare earth and is used as a catalyst in catalytic converters and is added to diesel fuel to make it burn more efficiently. Yttrium is used to make red LEDs and high-temperature superconductors.

Use of rare earth metals

The use of rare earth metals, or exotic minerals, has grown dramatically. They are used in green technology, medicines, lasers, weapon technology and elsewhere; they are expensive to obtain but growing in demand. What happens if rare earth reserves are controlled by a few countries but required by many?

Rare earths are an important part of the global economy: over 100 000 tonnes of rare earth oxides are mined every year. China is the biggest producer of rare earth ores, with 95% of the world production. The USA is the second most important producer, with 5% of the world production, but lacks the manufacturing capacity found in China.

High-tech weapon technology depends on rare earth metals. Precision guided missiles, night vision goggles, lasers and radar all contain rare metals. Countries that control rare earth reserves will be able to influence military budgets and spending in other countries.

Green technologies, such as hybrid cars, solar panels and wind turbines also depend on rare earth metals. The control of rare earth exports from countries with deposits will affect the price and hence the adoption of green technologies around the world. This could increase the use of fossil fuels and worsen global warming.

ToK Link

What factors/outcomes should be used to determine how time, money and effort are spent on scientific research? Who decides which knowledge is to be pursued?

For most scientists, a powerful motivation for performing scientific research is an intellectual curiosity about 'how things work' and a taste for intellectual stimulation. The joy of scientific discovery is captured in the following excerpts from letters between Max Planck and Erwin Schrödinger involved in the development of quantum mechanics:

[Planck, in a letter to Schrödinger, says] '*I am reading your paper in the way a curious child eagerly listens to the solution of a riddle with which he has struggled for a long time, and I rejoice over the beauties that my eye discovers.*'

Some scientists try to achieve personal satisfaction and professional success by intellectual collaborations with colleagues and by seeking respect and rewards, status and power in the form of publications, grant money, employment, promotions and honours, including the Nobel prizes.

When a theory (or a request for research funding) is evaluated, most scientists may be influenced by the pragmatic question, 'How will the result of this evaluation affect my own personal and professional life?' Maybe a scientist has publicly taken sides on an issue and there is ego involvement with a competitive desire to win the debate; or time and money have been invested in a theory or research project, and there will be higher payoffs if there is a favourable evaluation by the scientific community.

Ideological principles are based on subjective values and on political goals to achieve an ideal or better society. These principles include socioeconomic structures, race relations, gender issues, social philosophies and customs, religions, morality, equality, freedom and justice.

A dramatic example of political influence is the control of Russian biology, from the 1930s into the 1960s, by the 'ideologically correct' theories and research programs (based on the idea of inheritance of acquired characteristics) of the biologist Lysenko, supported by the power of the Soviet government.

Opinions of authorities can also influence evaluation and direction of scientific research. Authority can be due to an acknowledgment of scientific expertise, a response to a dominant personality, and/or involvement in a power relationship. Authority that is based at least partly on power occurs in scientists' relationships with employers, tenure committees, colleagues, professional organizations, journal editors and referees, publishers, grant reviewers, and politicians who decide on government funding for science.

■ Colorimetric determination of manganese

This analysis is carried out by dissolving the steel sample, converting all of the manganese to the intensely coloured manganate(VII) ion, and then determining the percentage of light absorbed for a set of conditions, such as degree of dilution and wavelength.

The manganate(VII) ion strongly absorbs green light, leaving the colour of the solution to be determined by the red and blue light that is transmitted; it is magneta (purple). The manganate(VII) ion is so intensely coloured that the conversion of all the manganese in the sample to manganate(VII) ions indicates the amount of manganese in the steel by the amount of light it absorbs (via its absorbance or transmittance).

A calibration curve is plotted by taking solutions containing known concentrations of manganate(VII) ions, measuring the absorbance of each solution, and then constructing a linear graph in which measured absorbance is plotted against concentrations of the solutions.

When the absorbance of the test solution (from the oxidized steel sample) is determined using the spectrophotometer, the absorbance can then be compared to the calibration line (curve) to find the concentration of the manganate(VII) ions. A dilution factor may need to be applied to obtain the original concentration of manganese(II) ions in the steel sample.

A solution of concentrated phosphoric(V) acid is used to dissolve the steel sample. In order to oxidize the manganese(II) ions to manganate(VII) ions, the very strong oxidizing agent potassium iodate(VII) is used:

$$5IO_4^-(aq) + 2Mn^{2+}(aq) + 3H_2O(l) \rightarrow 2MnO_4^-(aq) + 5IO_3^-(aq) + 6H^+(aq)$$

■ Gravimetric analysis of copper

Gravimetric analysis (Chapter 1) involves determining the mass of an element or compound by chemically changing that substance into another substance of known chemical composition that can be easily isolated, purified and weighed. One common approach is via the formation of an insoluble precipitate.

A copper coin (of known mass) can be dissolved in concentrated nitric(V) acid forming a solution of copper(II) ions. The solution of copper(II) ions is then treated with aqueous potassium iodide. Copper(I) iodide is precipitated as a white solid and iodine is produced:

$$2Cu^{2+}(aq) + 4I^-(aq) \rightarrow 2CuI(s) + I_2(aq)$$

The copper(I) iodide is a precipitate that can be filtered from the solution and weighed using an analytical electronic balance. A calculation can then be carried out to determine the mass and percentage of copper in the coin.

22.3 Catalysts *– catalysts work by providing an alternative reaction pathway for the reaction. Catalysts always increase the rate of the reaction and are left unchanged at the end of the reaction*

■ Catalysis

A **catalyst** is a substance which increases the rate of a chemical reaction without undergoing any permanent change in chemical composition or mass. The phenomenon of an increase in the rate of a reaction in the presence of a catalyst is known as **catalysis**.

A catalyst accelerates the rate of reaction by providing an alternative pathway or mechanism with a lower activation energy barrier. The catalyst lowers the activation energy by interacting with the reactants to form an intermediate of lower enthalpy (potential energy).

Catalysts obey the laws of thermodynamics: they only increase the rate at which a reaction reaches its equilibrium position, and cannot affect the magnitude of the equilibrium constant, K_c.

General characteristics of catalytic reactions

The catalyst remains unchanged in amount and composition at the end of the reaction.

Only a small quantity of catalyst is generally needed. However, in some homogeneous catalysed reactions, the rate of the reactions is proportional to the concentration of the catalyst. For example, the rate of inversion of sucrose catalysed by hydrogen ions varies with the concentration of the hydrogen ions present in the solution.

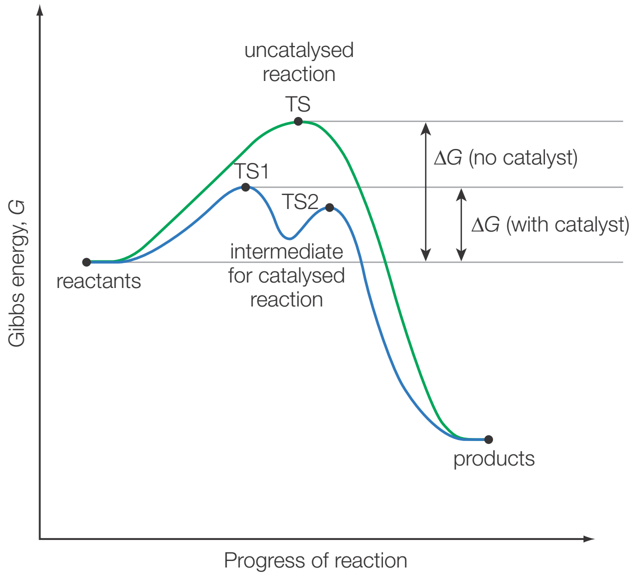
In certain heterogeneous reactions, the rate increases with an increase in the surface area of the catalytic surface. This explains why the efficiency of a solid catalyst increases when it is present in a finely divided state (powdered form).

The catalyst does not affect the position of equilibrium in a reversible reaction.

The catalyst does not initiate or start the reaction. The reaction is already occurring, though extremely slowly, in the absence of the catalyst. The reaction in the presence of the catalyst takes place via an alternative pathway (mechanism) with a decreased activation energy.

The catalyst is generally specific in its action. Manganese(IV) oxide, for example, can catalyse the decomposition of potassium chlorate(V) but not that of potassium chlorate(VII) or potassium nitrate(V). Hence, manganese(IV) oxide is specific in its action. Enzymes (Chapter 23) are highly specific in their action and very effective at lowering activation energy barriers.

The catalyst cannot alter the products of the reaction. For example, the reaction of nitrogen and hydrogen always results in the formation of ammonia (or hydrazine, N_2H_4), whether a catalyst is present or not.



■ **Figure 22.33** The effect of a catalyst on the Gibbs free energy change of a reaction. TS, transition state or activated complex

Catalysts and thermodynamics

A catalyst provides an alternative pathway (mechanism) that has a lower activation energy. A better explanation is offered by transition state theory on the basis of the decrease in the Gibbs free energy of activation. Figure 22.33 shows that the Gibbs free energy of activation is lowered in the presence of the catalyst.

It can be seen that the Gibbs free energy of activation for the reverse reaction is also lowered in the presence of the catalyst without changing the overall free energy change of the reaction. This implies that a catalyst increases the rates of both the backward and forward reactions.

Since the Gibbs free energy change, ΔG, is not changed, the equilibrium constant, K_c, remains unchanged in the presence of the catalyst ($\Delta G = -RT \ln K_c$). Hence a catalyst helps in attaining the equilibrium position more rapidly, but does not change the relative proportions of reactants and products at equilibrium.

A catalyst is poisoned by certain substances known as catalytic poisons. For example, the rate of reaction between sulfur dioxide and oxygen (in the Contact process) is slowed down significantly if traces of arsenic compounds are present.

Table 22.1 shows the activation energy for a specific reaction in the absence of a catalyst, in the presence of a catalyst and with an enzyme (biological catalyst).

■ **Table 22.1** The effect of a catalyst and enzyme on the decomposition of hydrogen peroxide

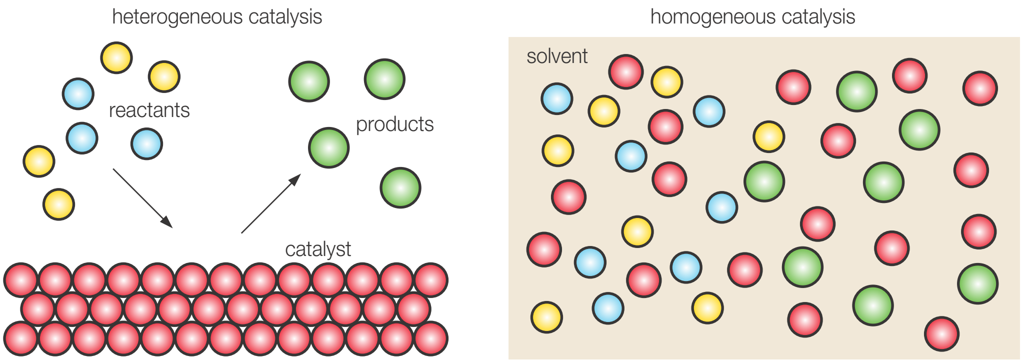
Reaction	Catalyst	Activation energy/kJ mol^{-1}
$2H_2O(l) \rightarrow 2H_2O(l) + O_2(g)$	None	+75
	Platinum	+49
	Catalase (enzyme)	+23

Enzymes

Enzymes are proteins that catalyse a wide range of chemical reactions in the cells of organisms (Chapter 23). The enzyme interacts with the reactant, the substrate, at a specific location in the enzyme known as the active site. The three-dimensional shape of the active site fits the shape of the substrate. The substrate is bound to the active site by intermolecular forces and possibly ionic interactions that depend on the three-dimensional shapes of the enzymes and substrates. The chemical reaction then takes place at the active site. The overall result may be bond breaking or bond forming. The products are released from the active site and the enzyme is free to react with another molecule of substrate.

Types of catalysis

Catalysts may be classified into three broad categories: homogeneous catalysts, heterogeneous catalysts and enzyme catalysts (Chapter 23). Homogeneous catalysts function in the same physical state (phase) as the reactants (Figure 22.34b). For example, the esterification reaction to synthesize an ester uses concentrated sulfuric acid to provide protons (H^+) to act as a catalyst (Chapter 20). The ester, alcohol and sulfuric acid are all in aqueous solution. The depletion of ozone by chlorofluorocarbons in the presence of ultraviolet radiation is another example of homogeneous catalysis.



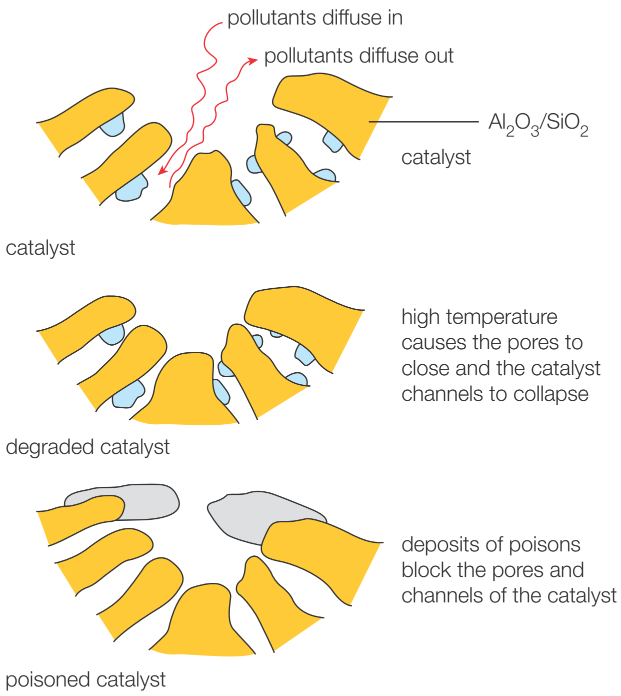
■ **Figure 22.34** The difference between **a** heterogeneous and **b** homogeneous catalysis

In contrast, a heterogeneous catalyst is in a different physical state (phase) from the reactants (Figure 22.43a). The catalyst is usually solid and the reactants are often gases. This is why heterogeneous catalysis is often called surface catalysis. For example, the Haber process involves the reaction between adsorbed reactant molecules and molecular fragments on the surface of a solid iron catalyst. The Contact process involves the reaction between sulfur dioxide and oxygen gases to form sulfur trioxide gas in the presence of a solid vanadium(v) oxide catalyst (Chapter 16).

Many catalysts, though not all, are transition metals or compounds of transition metals – this is because these elements have a variety of stable oxidation states and can form complex ions (Chapter 13) via the formation of coordinate bonds (Chapter 4).

The surface of a heterogeneous catalyst contains many active sites (Figure 22.35). These are areas where one or more of the reactants can be temporarily fixed to the catalyst surface. There is some sort of interaction between the surface of the catalyst and the reactant molecules which makes them more reactive. This might involve an actual chemical reaction with the surface, or some weakening of the bonds in the attached molecules (Chapter 16). The catalyst also holds the reactant molecules in a fixed orientation. Poisoning results in the blocking of the pores and channels where catalysis occurs.

Figure 22.35 Structure and operation of a heterogeneous catalyst


pollutants diffuse in
pollutants diffuse out
Al_2O_3/SiO_2
catalyst
catalyst
high temperature causes the pores to close and the catalyst channels to collapse
degraded catalyst
deposits of poisons block the pores and channels of the catalyst
poisoned catalyst

Homogeneous catalysis

All the reagents and the catalyst are in the same physical phase (state); that is, all are gases or all are liquids (or in solution).

For example, in the historical lead chamber process the reactants sulfur dioxide and air were mixed with nitrogen monoxide (a catalyst) in the presence of water. The following elementary steps are believed to occur:

Elementary step 1 $NO(g) + \frac{1}{2}O_2(g) \Rightarrow NO_2(g)$

Elementary step 2 $SO_2(g) + NO_2(g) + H_2O(l) \Rightarrow H_2SO_4(l) + NO(g)$

Overall
(sum of elementary steps 1 and 2) $SO_2(g) + \frac{1}{2}O_2(g) + H_2O(l) \rightarrow H_2SO_4(l)$

The catalyst, NO, is converted to an intermediate, NO_2, which is then converted back to NO. This is known as regenerating the catalyst.

Other examples of homogeneous catalysis include:

- preparation of ethoxyethane (ether) from ethanol using concentrated sulfuric acid as the catalyst.

  $$2C_2H_5OH(l) \rightarrow C_2H_5\text{-}O\text{-}C_2H_5(l) + H_2O(l)$$

- hydrolysis of ethyl ethanoate in the presence of concentrated sulfuric acid.

  $$CH_3COOC_2H_5(aq) + H_2O(l) \rightarrow CH_3COOH(aq) + C_2H_5OH(aq)$$

- hydrolysis of sucrose in the presence of dilute sulfuric acid to form glucose and fructose.

  $$C_{12}H_{22}O_{11}(aq) + H_2O(l) \rightarrow 2C_6H_{12}O_6(aq)$$

Acid–base catalysis

Acid–base catalysis includes reactions in solution which are catalysed by acids or bases or both. A reaction which is catalysed by hydrogen ions (H^+ or H_3O^+) but not by other Brønsted-Lowry acids (proton donors) is said to be specifically proton-catalysed. Examples include hydrolysis of esters and inversion of sucrose and keto-enol transformation (iodination of propanone).

Homogeneous catalysis: demonstration

In the catalytic reaction of potassium sodium tartrate with hydrogen peroxide, the cobalt chloride catalyst is in the same phase as the reactants. This is a visual demonstration of homogeneous catalysis (Chapter 6).

An observer can see the progress of the reaction via the formation of the green activated complex. An observer will see that the catalyst is regenerated by the reappearance of the pink cobalt (II) chloride colour.

The tartrate ions are oxidized by the hydrogen peroxide to carbon dioxide and water.

$$5H_2O_2(aq) + KNaC_4H_4O_6(aq) \rightarrow 4CO_2(g) + NaOH(aq) + KOH(aq) + 6H_2O(l)$$

Without the catalyst the evolution of carbon dioxide is quite slow. With the cobalt(II) chloride solution the reaction proceeds with the rapid evolution of carbon dioxide. This is an excellent demonstration of the formation of an intermediate species. The colour change from pink to green to pink is easy to observe and can be timed to see the effect of temperature on the reaction rate.

Hydrated cobalt(II) ions are pink. The hydrogen peroxide initially oxidizes the cobalt(II) ions, Co^{2+}, to cobalt(III) ions, Co^{3+}, which are green. The cobalt(III) bonds to the tartrate ion (the IUPAC name is 2,3-dihydroxybutandioate ion), allowing the oxidation to take place. The cobalt(III) ions, Co^{3+}, are then reduced back to cobalt(II) ions and the pink colour returns.

The cobalt(II) chloride catalyst provides an alternative route for the reaction to occur. This alternative route has a lower activation energy and the reaction proceeds much more quickly.

When the reaction is complete (no more bubbling), the catalyst is regenerated. This is shown by the formation, once again, of the pink colour, indicating the regeneration of the (pink) cobalt(II) chloride catalyst.

Heterogeneous catalysis

The catalyst is in a different physical phase (state) from the reactants. The most common form of this occurs in contact catalysis in which the reactants are gases and the catalyst is a solid. It is the most common form of catalysis in the chemical industry and examples include:

- the Haber process (Chapter 17) where ammonia is synthesized from its elements in the presence of an iron catalyst.

  $$N_2(g) + 3H_2(g) \rightarrow 2NH_3(g)$$

- the Contact process (Chapter 17) where sulfuric acid is manufactured from sulfur. The Contact step involves the conversion of sulfur dioxide to sulfur trioxide in the presence of vanadium(V) oxide.

  $$2SO_2(g) + O_2(g) \rightarrow 2SO_3(g)$$

- the industrial manufacture of nitric(V) acid which involves the conversion of ammonia to nitrogen monoxide in the presence of platinum.

  $$4NH_3(g) + 5O_2(g) \rightarrow 4NO(g) + 6H_2O(g)$$

- the hydrogenation of vegetable oils to form margarine and the conversion of ethane to ethane, both of which used heated nickel as the catalyst.
- the polymerization of ethene using titanium(IV) chloride and trialkylaluminium (Ziegler-Natta process) as a catalyst to form high-density polyethene.

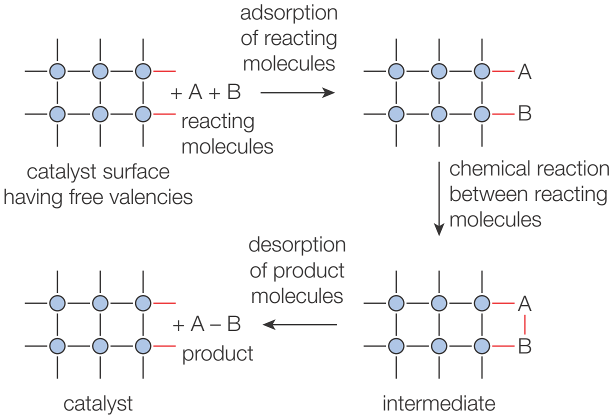
17 Design an experiment to collect and process data and determine the effect of temperature on the decomposition of hydrogen peroxide in the presence of manganese(IV) oxide acting as a catalyst.

Adsorption theory of heterogeneous catalysts

The action and behaviour of heterogeneous catalysts is explained by the adsorption theory. The catalytic activity of the heterogeneous catalyst is localized on the surface of the catalyst. The mechanism of the catalysis involves the following steps (simplified):

- Diffusion of reactants to the surface of the catalyst.
- Adsorption of reactant molecules on the surface of the catalyst – usually a transition metal or transition metal compound.
- A chemical reaction occurs on the surface of the catalyst and an intermediate is formed as shown in Figure 22.36.
- The products then undergo desorption from the catalyst's surface (they desorb) thereby making the catalyst surface available for another pair of reacting molecules.
- The products diffuse away from the catalyst's surface.

Figure 22.36 Adsorption theory


adsorption of reacting molecules
+ A + B
reacting molecules
catalyst surface having free valencies
A
B
chemical reaction between reacting molecules
desorption of product molecules
+ A – B
product
catalyst
A
B
intermediate

Heterogeneous catalysts are typically 'supported' which means that the catalyst is dispersed on a second material that enhances the effectiveness and/or minimizes their cost. Sometimes the support is merely a surface across which the catalyst is spread to increase the surface area. More often, the support and the catalyst interact, affecting the catalytic reaction.

Features of solid catalysts

Activity

The ability of catalysts to accelerate the rates of chemical reactions is called activity.

The activity of a solid catalyst depends mainly upon the strength of chemisorption. During chemisorption, chemical bonds are formed between the atoms or ions on the surface of the catalyst (adsorbent) and the reacting molecules (adsorbate). A solid catalyst must adsorb the reactants fairly strongly, but not so strongly that they are immobilized.

Physical adsorption

When the forces of attraction existing between adsorbate and absorbent are London (dispersion) forces, the adsorption is called physical adsorption or physisorption. Since the forces existing between adsorbent and absorbate are very weak, this type of adsorption can be easily reversed by heating or by decreasing the pressure.

Chemical adsorption

When the forces of attraction existing between adsorbate particles and adsorbent are chemical bonds, the absorption is called chemical adsorption or chemisorption. Since the forces of attraction existing between adsorbent and absorbate are relatively strong, this type of adsorption cannot be easily reversed.

Physisorption and chemisorption are compared in Table 22.2.

Table 22.2 Comparison of physisorption versus chemisorption

Physisorption	Chemisorption
Low enthalpy of adsorption	High enthalpy of absorption
Forces of attraction are London (dispersion) forces	Forces of attraction are chemical bonds
It usually takes place at low temperature and decreases with increasing temperature	It takes place at high temperature
It is reversible	It is irreversible
It is related to how easy it is to liquefy the gas	The extent of adsorption is generally not related to how easy it is to liquefy the gas
It is not very specific	It is highly specific
It usually forms multi-molecular layers on the adsorbent	It usually forms monomolecular layers on the adsorbent
It requires a very low activation energy	It requires a high activation energy
The molecular state of the adsorbate generally remains unaltered	The molecular state of the adsorbate undergoes a change at the surface of adsorbent

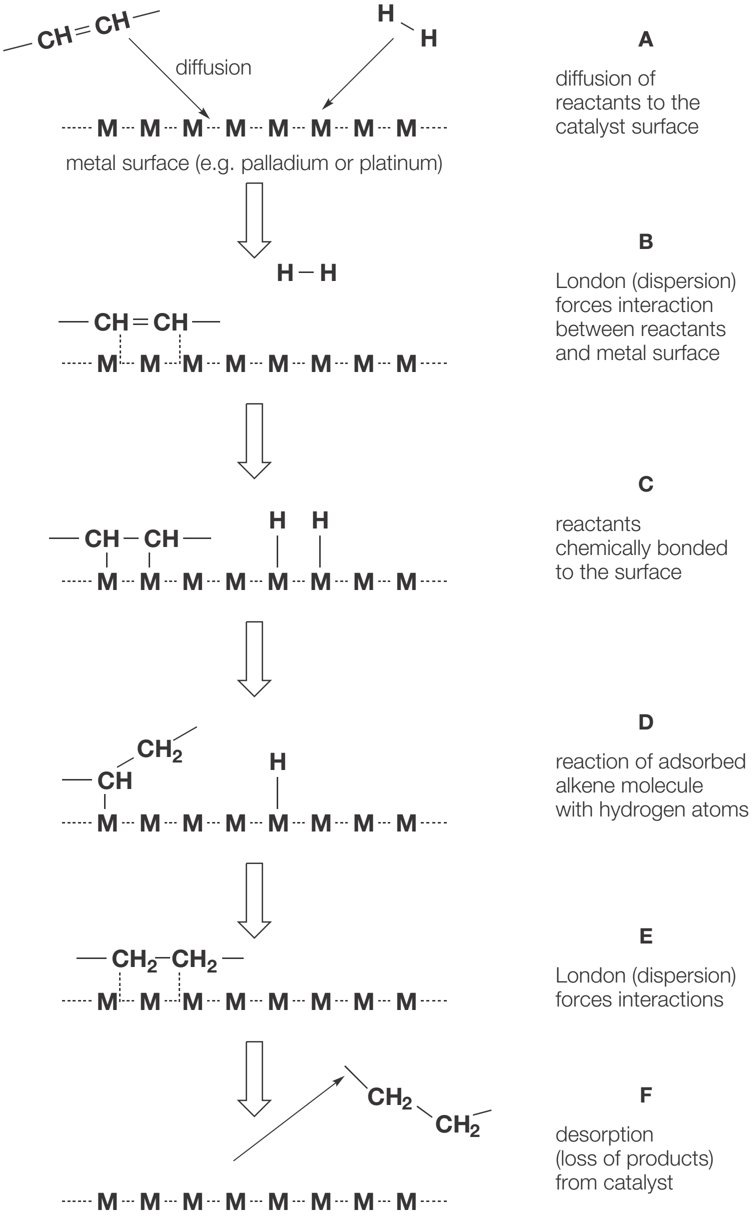

CH=CH
H
H
diffusion
M M M M M M M M
metal surface (e.g. palladium or platinum)
A
diffusion of reactants to the catalyst surface
H–H
CH=CH
B
London (dispersion) forces interaction between reactants and metal surface
CH–CH
H H
C
reactants chemically bonded to the surface
CH₂
CH
H
D
reaction of adsorbed alkene molecule with hydrogen atoms
CH₂–CH₂
E
London (dispersion) forces interactions
CH₂
CH₂
F
desorption (loss of products) from catalyst

Figure 22.37 Mechanism of heterogeneous catalysed hydrogenation

Hydrogenation

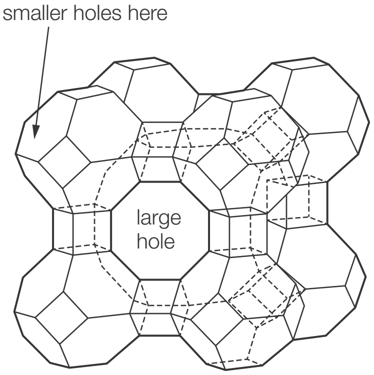
The catalytic hydrogenation of alkenes by meta surfaces, such as platinum or palladium, has been well studied since its discovery in 1900. The currently accepted mechanism is shown in Figure 22.37.

The reactant molecules diffuse to the palladium surface (A) where they first interact by London (dispersion) forces (B). This is followed by formation of chemical bonds to the surface of the palladium catalyst (C); this results in the loss of the π-bond between the carbon atoms of the alkene, and breakage of the H–H bond (so that the adsorbed hydrogen is in the form of atoms). These hydrogen atoms can then react with the adsorbed alkene (step C to D), forming the alkane (E). The alkane has no spare electrons with which to bond to the catalyst surface, so it can only interact by London (dispersion) forces. Since this is physisorption, the forces are so weak that the alkane rapidly leaves the surface.

It is interesting that the first hydrogenation step (C–D) is reversible. If D_2 is used instead of H_2, the complete range of possible deuterated alkanes and alkenes is formed, in amounts that depend on the catalyst used. This reversibility also allows the position of the double bond to alter, and the *cis* isomer to be converted into the *trans* isomer.

Zeolites

A catalytic action that depends upon the pore structure of the solid catalyst and the sizes of the reactant and product molecules is known as



■ **Figure 22.38** Outline of the cage structure of a zeolite, with formula $Na_{12}(Al_{12}Si_{12}O_{48}).27H_2O$

shape-selective catalysis. Zeolites (Figure 22.38), a class of ceramics, show this type of heterogeneous catalysis because of their well-defined porous structures.

Zeolites are aluminosilicates of the general formula $M_{x/n}[(AlO_2)_x(SiO_2)_y].mH_2O$, where n is the charge of the metal cation, M^{n+}. They are three-dimensional silicates in which some silicon atoms are replaced by aluminium ions. They are found naturally, but can also be synthesized in the laboratory. There is a wide range of structures but all have SiO_4 and AlO_4 tetrahedra linking together through their corner oxygen atoms to form giant covalent structures. The porous nature of zeolites and the presence of replaceable M^+ cations on their internal surface is a critical factor in their uses.

Zeolites become catalytic when they are heated in a vacuum to remove water molecules and create catalytic cavities and thereby make the material porous. The catalytic cavity and selectivity of zeolites depend on the size of the cavities (cages) present in them. These cavities can only adsorb and trap molecules whose size and shape allow them to enter and leave the cavities; larger molecules are unable to enter. Hence, zeolites can act as molecular 'sieves' or selective adsorbents.

When zeolites are used as catalysts, M^+ is normally a proton (H^+), and the zeolite acts as a strong Brønsted-Lowry acid. The cracking of crude oil or petroleum (Chapter 24) – the decomposition of long-chain alkane molecules into smaller molecules – is an example of a reaction that often uses zeolites as catalysts. Zeolites can also be used to convert alcohols to hydrocarbons.

Another major use of zeolites is ion exchange, where M^+ is normally the sodium cation, Na^+. Zeolites are often used to remove calcium and magnesium ions from so-called hard water. This is to allow washing powders to work well with such water. The function of the zeolites is to remove calcium and magnesium ions, which otherwise would precipitate from the water in the presence of washing powder or soap. The calcium and magnesium ions are retained in the zeolites, releasing sodium ions into the solution. The resulting water is described as soft and will readily lather with soap and washing powder.

Homogeneous versus heterogeneous catalysis

Homogeneous catalysts are potentially the most efficient type of catalyst because every catalyst molecule is potentially accessible to the reactants. In contrast, only the surface atoms, ions or molecules of a solid heterogeneous catalyst are accessible. However, in practice industrial solid catalysts are often coated, in the form of small solid particles, on to the surface of a cheap and inert support such as aluminium oxide (alumina) or silicon dioxide (silica). A powder can be used to maximize the surface area.

Hence, for the same amount of substance, a homogeneous catalyst provides a greater effective concentration of catalyst than a heterogeneous catalyst. Consequently, by employing a homogeneous catalyst, industrial chemists can use milder, and hence cheaper, reaction conditions (for example, lower temperatures and pressures). This is a generalization and there are some situations where a homogeneous catalyst, or its promoter, is in fact a compound that requires particular care in reactor design. Promoters are substances which enhance the activity of heterogeneous catalysts.

For example, the catalyst in the Monsanto process for making ethanoic acid from methanol and carbon monoxide requires iodomethane as a promoter, which in turn requires the reaction vessel to be made of a special alloy. An additional benefit of heterogeneous catalysis is greater selectivity; that is, the catalyst will only catalyse a single reaction or a small group of related reactions.

In practice, many industrial catalysts are of the heterogeneous type. This is because at the end of a reaction, the catalyst must be separable from the products. In a heterogeneous reaction, the solid catalyst can be removed from the reaction mixture by simple filtration. This means that the chemical process used must be a batch process, rather than a more efficient continuous process. This is not an issue when gaseous or liquid products are flowing over a solid catalyst surface – as in a continuous, heterogeneous process. However, if a homogeneous catalyst is used then the catalyst and products must be separated by distillation. Distillation requires heat (thermal) energy and a high distillation temperature may cause the catalyst to decompose. However, a common industrial approach is to let the reactants flow over a solid bed containing the heterogeneous catalyst.

Description of how metals work as heterogeneous catalysts

Many transition metals and their compounds act as catalysts. Their catalytic effect is often due to their ability to exist in more than one stable oxidation state. For example, the vanadium(v) oxide catalyst in the Contact process is believed to operate via the following mechanism:

18 Find out about the Monsanto process: the reaction catalysed and its industrial importance, the action of the catalyst and the conditions employed during the reaction.

$$V_2O_5(s) + SO_2(g) \rightarrow V_2O_4(s) + SO_3(g)$$

$$\frac{1}{2}V_2O_4\ (s) + O_2(g) \rightarrow V_2O_5(s)$$

The mechanism is supported by experimental data. If vanadium(v) oxide is heated with sulfur dioxide, it turns blue and forms V_2O_4. This is converted back to V_2O_5 by heating in air.

Many hydrogenation reactions are catalysed by metals that form interstitial hydrides where the hydrogen atoms are held in the spaces between the metal ions in the lattice, forming a type of alloy whose properties are only slightly different from the pure metal. Interstitial hydrides, like other alloys, are non-stoichiometric, meaning they do not have a fixed formula. The ratio of hydrogen atoms to metal ions depends on the conditions, especially the external pressure of the hydrogen.

The ability of transition metals to act as catalysts also depends on the presence of empty d orbitals. For example, nickel atoms are able to form a coordinate bond with ethene molecules. The pi bond of the ethene molecule overlaps with the empty d orbital with a nickel atom on the surface of the catalyst (Figure 22.39). The places on the nickel surface where the geometry allows molecules to bond in this way are called active sites.

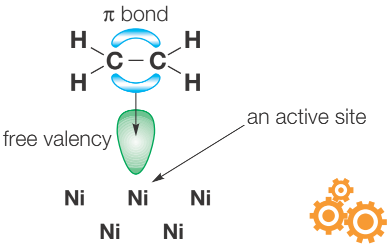


Figure 22.39 Ethene molecules can bond through their pi electrons to a nickel atom

Explanation of factors involved in choosing a catalyst for a process

The choice of catalyst for an industrial process will depend on several factors – all of which are designed to maximize yield and profit.

Selectivity

Selectivity is the ability of a catalyst to direct a reaction to yield a particular product. For example, ethyne on reaction with hydrogen in the presence of platinum as a catalyst forms ethane. However, in the presence of Lindlar's catalyst (a palladium-based heterogeneous catalyst) ethyne is converted only to ethene (Figure 22.40).

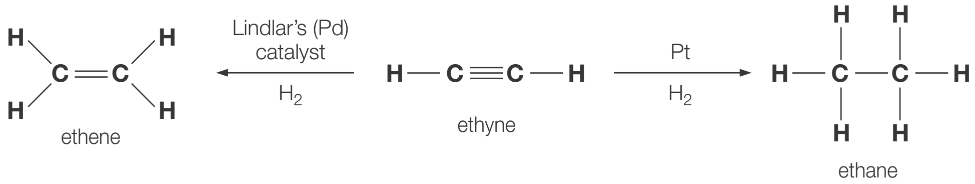


Figure 22.40 An illustration of catalyst selectivity

Efficiency

A highly efficient catalyst is preferred that causes a considerable increase in the rate. Catalyst efficiency is determined by two factors:

- turnover number (TON), which measures how many reactant molecules can be converted into product molecules by a catalyst molecule;
- turnover frequency (TON per unit time), which is a measure of the rate of turnover. Industrial chemists try to maximize catalyst efficiency without compromising costs.

Robustness

Many industrial processes require the use of high temperatures, high pressures (if gases are involved) and organic solvents or concentrated acids. The catalyst chosen must be able to function under these severe conditions without undergoing decomposition. A catalyst may become covered in soot or other surface coating, and may need to be cleaned in a regenerator. Also, at high operating temperatures a heterogeneous catalyst may melt, thus reducing its surface area and/or efficiency.

Environmental impact

Many catalysts are transition-metal based and many of these are toxic at relatively low concentrations in water. Strong acids and strong alkalis are also used as catalysts in many sectors of the chemical industry. Waste water (effluent) with a very low or very high pH will affect many freshwater organisms.

Potential for poisoning

Many heterogeneous catalysts are easily poisoned by chemicals that bind to the active sites and prevent catalysis. Catalyst poisoning can be minimized by the use of very finely divided solid particles and by purifying the feedstock. Catalyst poisoning is generally undesirable because it leads to a loss in usefulness of expensive metals or their complexes. However, partial poisoning of catalysts can also be used to improve the selectivity of reactions.

Nature of Science

The scientific study of catalysis

In 1835, Berzelius applied the term catalytic agent – now termed a catalyst – to describe substances which change the rate of chemical reaction, but without undergoing any permanent chemical change.

Catalysis is one of the most common and one of the most valuable chemical phenomena. Enzymes are biological catalysts and many industrial processes rely upon catalysis. Research into catalysts is an active area of research and two broad types of catalytic mechanisms or models have been proposed, homogeneous and heterogeneous catalysis. These theories are constantly being refined as new experimental data is obtained.

There have been tremendous economic benefits from the use of industrial catalysts, but some are toxic and polluting, such as heavy metals and concentrated sulfuric acid. Chemical catalysis benefits especially from nanoparticles, due to the extremely large surface to volume ratio. The application potential of nanoparticles in catalysis ranges from fuel cells (Chapter 24) to catalytic converters and photocatalytic devices (Chapter 24).

A great deal of knowledge has been gained in the mechanism of surface catalyst by studying examples of 'exchange reactions'. This is the term used to describe the substitution on a compound of one isotope for another of the same element, for example deuterium for hydrogen.

When a mixture of ethane, C_2H_6, and deuterium, D_2, is exposed at moderate temperatures to the surfaces of transition metals, for example nickel, copper, or palladium, which are efficient catalysts in hydrogenation reactions, deuterium is exchanged for hydrogen in successive reactions, of which the first is:

$$C_2H_6(g) + D_2(g) \rightleftharpoons C_2H_5D(g) + HD(g)$$

It is found that the experimental data is explained by supposing that the metal surface covers itself with a layer of activated molecules of deuterium. The advantage of studying such exchange reactions is that the stabilities of hydrides and deuterides are equal, and hence attention may be focused exclusively on the surface conditions.

Catalysis is said to be negative when the catalyst reduces or stops the reaction. The preferred term is **inhibition** and it can be useful in industry for controlling or stopping unwanted reactions.

In some cases, inhibitors are believed to work by interfering with chain reactions (Chapter 10). For example, the reaction between hydrogen and chlorine in the gas phase is believed to take place by the following chain reaction:

$$Cl_2 \rightarrow 2Cl\bullet \text{ (initiated by ultraviolet radiation)}$$

$$H_2 + Cl\bullet \rightarrow HCl + H\bullet$$

$$H\bullet + Cl_2 \rightarrow HCl + Cl\bullet$$

Nitrogen trichloride is an inhibitor for this reaction since it reacts with the chlorine atoms (radicals).

$$NCl_3 + Cl\bullet \rightarrow \tfrac{1}{2}N_2 + 2Cl_2$$

In heterogeneous catalysis, the reactions take place on the surface of the catalyst. Therefore, the atoms or ions buried inside the catalyst play no role in the catalysis, and an effective catalyst should have as many atoms or ions on the surface as possible.

Nanoparticles, such as gold nanoparticles, are ideal catalysts since they consist of clusters of only a few hundred atoms. Most of the gold atoms are located at the nanoparticle surface, ready to be involved in the catalytic reactions.

ToK Link

Some materials used as effective catalysts are toxic and harmful to the environment. Is environmental degradation justified in the pursuit of knowledge?

Catalytic converters

Vehicle exhaust fumes contains a number of toxic pollutant gases. Under ideal conditions the fuel, for example petrol, would undergo complete combustion to form carbon dioxide and water vapour only:

$$C_8H_{18}(l) + 12\frac{1}{2}O_2(g) \rightarrow 8CO_2(g) + 9H_2O(l)$$

However, incomplete combustion leads to the production of the toxic gas carbon monoxide and unburnt hydrocarbon molecules which are carcinogenic (cancer forming).

At the high temperatures and pressures in an engine, oxygen combines with nitrogen from the air to give toxic and acidic oxides of nitrogen:

$$N_2(g) + O_2(g) \rightarrow 2NO(g)$$

$$2NO(g) + O_2(g) \rightarrow 2NO_2(g)$$

Many cars (automobiles) are now fitted with catalytic converters which reduce both types of air pollution. A converter consists of an expansion chamber which contains a mixture of catalysts supported on a fine honeycomb aluminium oxide mesh. Rhodium metal (a transition metal) catalyses the reduction of the oxides of nitrogen to nitrogen:

$$2NO(g) \rightarrow N_2(g) + O_2(g)$$

The other catalyst, usually platinum or palladium, catalyses the oxidation of carbon monoxide and unburnt hydrocarbons into carbon dioxide:

$$2CO(g) + O_2(g) \rightarrow 2CO_2(g)$$

The oxygen for this process is supplied by the first process.

A catalytic converter must be used with unleaded petrol. Leaded petrol contains lead compounds (to promote smooth burning) which will irreversibly destroy the efficiency of the catalytic converter.

Platinum, palladium and rhodium are expensive metals and are frequently recycled when the car is scrapped. However, new research shows that they are polluting the atmosphere instead. Italian and French researchers have found dangerous heavy metals from converters far away from their sources, in remote regions of Greenland. They found that concentrations of the metals in the ice rose steadily since 1976. The ratio of platinum to rhodium resembles the ratio of these metals in car exhausts. This suggests that the pollution comes from cars.

The use of metals in catalytic converters

Palladium, platinum and rhodium are common catalysts that are used in catalytic converters. Because of the value of these metals, catalytic converter thefts are on the rise. How do crimes such as this influence the global economy?

Thieves around the world have targeted catalytic converters (Figure 22.41) located in the exhaust system under the vehicle – because of the high value of the precious metals they contain. The prices of these metals have been steadily rising and the demand for catalytic converters is rising. Metal thefts often correlate with high metal prices. The effect on the global economy during times of theft is an increase in insurance claims by car owners and an increase in business for garages which have to repair the damaged cars and install a new catalytic converter. Some states in the US have passed laws requiring that someone wanting to sell a catalytic converter provide documentation of legitimate ownership.

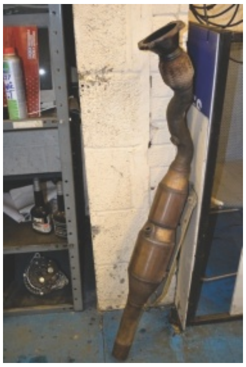


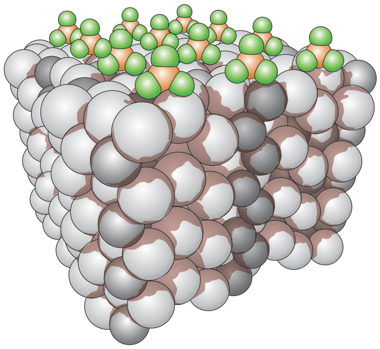
Figure 22.41 Old catalytic convertor removed from car

Nanocatalysts

A nanocatalyst is a substance or material with catalytic properties that has at least one nanoscale dimension, either externally or in terms of internal structures. Generally, catalysts that are able to function at the atomic scale are nanocatalysts.

Physical and chemical properties

The particle position can be controlled increasing the reaction stability and controlling mechanism of formation. Nanocatalysts have a controllable exposed atomic structure and a uniform dispersion. Nanocatalysts (Figure 22.42) show strong catalytic activity and great stability.



■ **Figure 22.42** Reacting molecules on the surface of a nanocatalyst

Catalytic activity

This is a very important factor in choosing a nanocatalyst. The porous nanostructure provides a high surface to volume ratio and hence greatly increases the catalytic activity.

For example, in a direct methanoic fuel cell, carbon monoxide poisoning significantly limits the catalytic activities of platinum/ruthenium and platinum/palladium alloys for methanoic acid oxidation. One solution to this poisoning is to use a nano-based catalyst on a carbon support.

Stability

Stability is one of the most important properties of nanocatalysts. The stability helps in achieving the desired size of nanoparticles with uniform dispersion on a substrate, such as carbon. Nanocatalysts, such as platinum, can be stabilized by stabilizing agents such as surfactants, ligands or polymers.

Effects of temperature and pressure on nanocatalysts

The melting point may be lower than that of the original metal species. For example, platinum has a melting point of around 2000 K but a nanocatalyst made up of platinum has a melting point of around 1000 K.

There is the possibility of using these nanocatalysts in the liquid phase. In the case of fuel cells it may penetrate through the layers to increase the surface area of reaction.

Advantages of nanocatalysts

These advantages are related to the intrinsic properties of the material: surface area, charge and size. Since nanocatalysts are composed of very small particles, this property creates a very high surface to volume ratio. This increases the activity of the catalyst since there is more surface to react with the reactants. Some nanocatalysts can develop partial and net charges that help in the process of making and breaking bonds at a more efficient scale.

Description of the benefits of nanocatalysts in industry

The benefits of nanocatalysts in industry include the increased selectivity and activity of catalysts by controlling pore size and particle characteristics and replacement of precious (expensive) metal catalysts by catalysts tailored at the nanoscale and use of base (cheaper) metals, thus improving chemical reactivity and reducing process costs. Catalytic membranes can be designed and synthesized that can remove unwanted molecules from gases or liquids by controlling the pore size and membrane characteristics.

22.4 Liquid crystals *– liquid crystals are fluids that have physical properties which are dependent on molecular orientation relative to some fixed axis in the material*

■ Introduction

■ **Figure 22.43** LRT coach with 'smart windows'

'Smart windows' based on liquid crystal films are currently being used in Singapore in the light rail transit (LRT) train (Figure 22.43). The main objective is to fog the windows when the LRT train passes residential flats to protect the privacy of the residents. The liquid crystals are dispersed as micro-droplets in a transparent plastic film between glass plates. The liquid crystals react to an application of a voltage by aligning in a parallel manner and letting light pass. The reverse is true – when no voltage is applied the liquid crystals in the film orient themselves randomly and the windows become opaque. Although this technology allows for manual control, there are no intermediate settings. In other words, the windows can only be transparent or opaque, with no gradation between. This is a function of the way the device has been constructed and is typical of polymer dispersed liquid crystal technology.

Liquid crystals are used as temperature and pressure sensors (though they are generally less sensitive to pressure) and as the display element in digital watches, calculators, TVs and laptop computers. They can be used for these applications because the weak intermolecular forces that hold the molecules together in the liquid crystalline state are easily overcome ('broken') by changes in temperature and applied fields.

Nature of Science

Discovery of liquid crystals

In 1888 the Austrian botanist Friedrich Reinitzer (1857–1927) (Figure 22.44) noticed that crystals of cholesteryl benzoate melted at 145.5 °C to form a cloudy liquid, which remained in existence up to 178.5 °C, where it changed again to form a clear liquid. The cloudy liquid was the first example of a liquid crystalline phase. Figure 22.45 shows the molecular structure of cholesteryl benzoate, an ester of cholesterol and benzene carboxylic acid. With the help of a German physicist, Otto Lehmann, the two scientists were able to determine that this cloudy liquid was a new state of matter and gave it the name 'liquid crystal'.

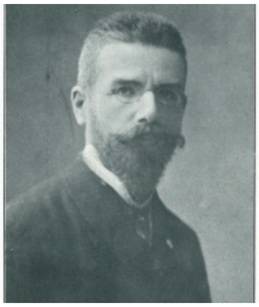


Figure 22.44 Friedrich Reinitzer

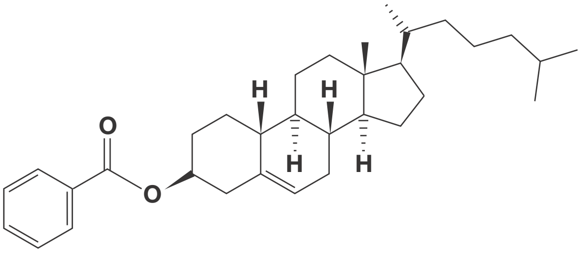


Figure 22.45 The structure of cholesteryl benzoate

Liquid crystals

The liquid crystal state

The solid and liquid states of matter were discussed in Chapter 1. When a crystalline solid melts, the ordered lattice arrangement of the particles is broken down, being replaced by the more disordered state of the liquid. Whereas in a solid the particles can only vibrate about a fixed point in a lattice, in a liquid they are able to move. However, some crystalline solids, when heated, melt to give a turbid (cloudy) phase which is fluid but retains some of the order of the solid state. On further heating this turbid phase changes to the normal clear liquid.

$$\text{solid crystal} \underset{\text{melting temperature}}{\rightleftharpoons} \text{liquid crystal} \underset{\text{clearing temperature}}{\rightleftharpoons} \text{liquid}$$

19 Find out about the enthalpy changes that accompany the formation of liquid crystals and the enthalpy changes that occur when there are changes involving different types of liquid crystals.

This turbid state of matter has properties intermediate between those of the solid and liquid states, and is called the **liquid crystal** state. It should be noted that a liquid crystal is a thermodynamically stable state of matter and there is a defined temperature transition; the 'intermediate properties of solid and liquid' do not mean it is a mixture of liquid and solid corresponding to a slow transition over a large interval of temperature.

It should also be noted that a molecule is not a liquid crystal in the same way that a molecule is not a solid or a liquid. Liquid crystal is a bulk property which cannot be attributed to a single molecule. The correct terminology is that this material forms a liquid crystal phase over a certain temperature range (or certain conditions). If a material can form a liquid crystal phase, then chemists say it is mesomorphic, a name also often given to the molecule itself. This is particularly important for lyotropic liquid crystals such as soap and cell membranes. They are not liquid crystal materials in usual conditions (home use, in the body) because they are not bulk material. Nevertheless, their molecular component can form liquid phases in certain given conditions. The applications of liquid crystals can now be seen in many areas of modern living and in the biological world.

There are many different types of liquid crystal phases, which can be distinguished by their different optical properties. When viewed under a microscope using a polarized light source, different liquid crystal phases will appear to have distinct textures. The sample is illuminated with polarized light and then viewed through an analyser, with polarizer and analyser being at 90° to one another.

The contrasting areas in the textures correspond to domains, where the liquid crystal molecules are orientated in different directions. Within a domain, however, the molecules are well ordered.

The effect of liquid crystals on polarized light is central to the application of these molecules in liquid crystal displays.

Explanation of liquid crystal behaviour on a molecular level

Instead of passing from the solid phase to the liquid phase when heated, some substances pass through an intermediate liquid crystalline phase that has some of the structure of crystalline solids and some of the freedom of the motion (flow) of liquids. Because of the partial ordering, liquid crystals may be viscous and have properties intermediate between those of solids and liquids. The region in which they show these properties is marked by sharp transition temperatures.

Liquid crystal displays

The displays on many everyday items, such as calculators and digital alarm clocks, depend on liquid crystal technology. Two classifications of liquid crystals are known, which are referred to as thermotropic and lyotropic liquid crystals – they both consist of molecular species or salts. **Thermotropic liquid crystals** exhibit transitions between the solid, liquid crystal and normal (isotropic) liquid phases with variation in temperature (as discussed for cholesteryl benzoate). **Lyotropic liquid crystals** exhibit phase transitions as a primary function of the concentration of the liquid crystal molecules in a solvent (typically water) in addition to the action of temperature. Liquid crystal display (LCD) devices are based on thermotropic liquid crystals and the combination of the order of a solid and the fluidity of a liquid confers unique and useful properties. For example, in LCDs, a small electric field can alter optical properties by changing the orientation of some of the molecules and, as a result, some areas of the display become dark while others remain light, allowing the shapes of the different numbers to be displayed. Over the past 40 years liquid crystals have gone from being an academic curiosity to the basis of big business and, in 2014, global sales of LCDs was in excess of $200 billion.

On heating, the directional consistency of orientation is lost and the normal liquid state is formed where the molecules have too much kinetic energy to be constrained in the same alignment by the intermolecular forces. The challenge to research in the early stages of this technology was to find molecules capable of this behaviour, particularly over a temperature range that included room temperature. Examples of liquid crystals can be found both in the natural world and in technological applications. Most modern electronic displays are liquid-crystal based.

20 Find out about the differences between monotropic and enantiotropic liquid crystals.

The issue in making commercially useful displays is in having low viscosity materials with the correct optical and dielectric properties and chemical stability. More than that, it was recognized that these properties would be extremely unlikely to occur together in a single compound and so displays use mixtures of compounds.

Lyotropic liquid crystalline phases are abundant in biological systems and also in the world around us, turning up as soaps and detergents. For example, many proteins and components of cell membranes are liquid crystals. Other well-known liquid crystal examples are solutions of soap and various related detergents, and the tobacco mosaic virus. DNA solutions and the concentrated protein solution extruded by spiders to form silk fibres (Figure 22.46) were found to form liquid crystal states under certain conditions. Intriguingly, in this latter case the water molecules appear to act as a plasticizer in enabling the silk fibres to move over each other as the web is woven. This phenomenon may well be related to the level of organization required for the self-organization of certain complex biological structures.

Figure 22.46 Liquid crystal properties appear to play a key role in the processing of silk fibres as a spider spins its web

Further, the fact that the fibres are spun in the liquid crystal phase means that the polymer chains of which they are constituted are aligned parallel to one another over large length scales and it is this that gives the silk its tensile strength (the same is true for the plastic Kevlar).

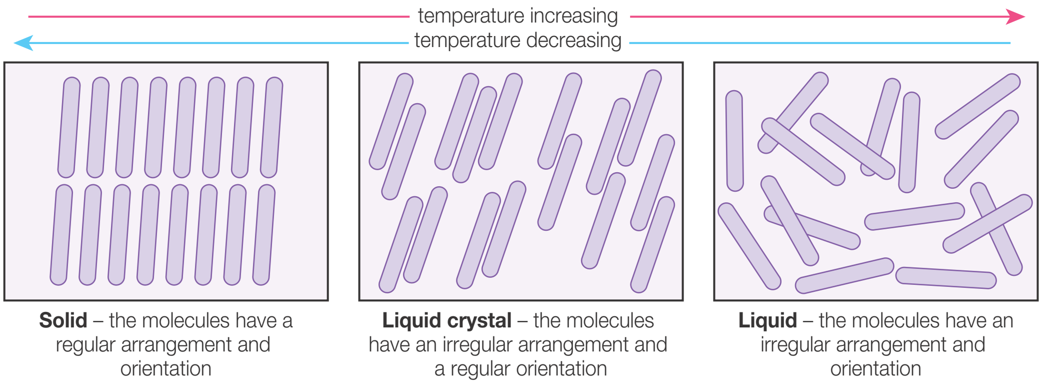
There are several specific terms applied to the liquid crystal state. **Thermotropic** is when phase changes are accomplished using temperature in the absence of solvent. In this state the molecules retain orientational order and, in some cases, partial positional order. Where there is only orientational order, the phase is termed **nematic** and this liquid crystal state produces thread-like patterns when viewed in polarized light (or better – between crossed polarisers) under a microscope. This is known as a nematic liquid crystal state (from the Greek *nemat* meaning 'thread').

A **thermotropic liquid crystal** state is where a substance displays a turbid state over a short temperature range after the solid has melted. In this fluid, but turbid, state the molecules retain a degree of organization or orientation in one dimension (Figure 22.47).

Certain molecules produce liquid crystal states in solution (usually in water). Here the liquid crystal state is a function of both concentration and temperature and is referred to as a **lyotropic liquid crystal** state.

Thermotropic and lyotropic liquid crystals

Molecules that show thermotropic liquid crystals are all typically long, thin, rigid, polar organic molecules with a rigid core and one or more flexible chains. Figure 22.47 shows the behaviour of such 'rod-like' molecules as the temperature increases around the transition between solid and liquid. As the substance changes from the solid state to the liquid crystal state, the arrangement of the molecules is more irregular, but the orientation is approximately the same. The analogy of putting a large number of pencils in a closed rectangular box has been used – imagine them being shaken.


temperature increasing
temperature decreasing
Solid – the molecules have a regular arrangement and orientation
Liquid crystal – the molecules have an irregular arrangement and a regular orientation
Liquid – the molecules have an irregular arrangement and orientation

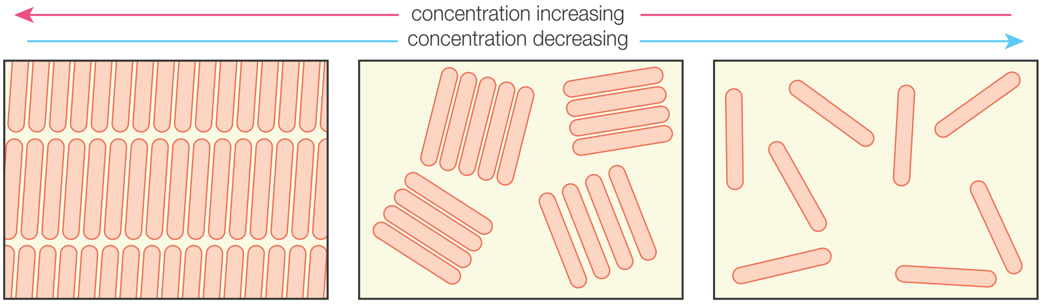
Thermotropic liquid crystals are formed in a temperature range between the solid and liquid state.

Figure 22.47 A representation of the transition of a substance to the thermotropic liquid crystal state

When the box is opened, the pencils will still be facing in *about* the same direction, but there will be no *definite* spatial organization. They are free to move, but generally line up in about the same direction. This gives a simple model of the nematic type of a liquid phase. The molecules are randomly distributed, as they are in a liquid, but the intermolecular forces are sufficiently strong to hold the molecules in one orientation. It is the nematic state of liquid crystals that is found in the vast majority of LCDs.

The formation of a liquid crystal state by the silk fibroin protein in solution demonstrates a further aspect of the liquid crystal phenomenon. The liquid crystals formed by pure substances over a certain temperature range after melting are called thermotropic liquid crystals (Figure 22.47). However, some substances can form a type of liquid crystal state in solution. This is a different set of circumstances in which the molecules are present as the solute in a solution. At low concentrations, the molecules generally have a disordered orientation and an irregular arrangement. If the concentration is increased sufficiently, the molecules will adopt an ordered structure and solid crystals will form. At intermediate concentrations, a lyotropic liquid-crystal state may be possible where the molecules have an irregular arrangement with a regular orientation (Figure 22.48). The level of organization in this state can be disrupted by changing either the temperature or the concentration of the system.

Figure 22.48 The formation of a lyotropic liquid crystal state by certain substances in solution



Solid – the molecules have a regular arrangement and orientation

Liquid crystal – the molecules have an irregular arrangement and a regular orientation

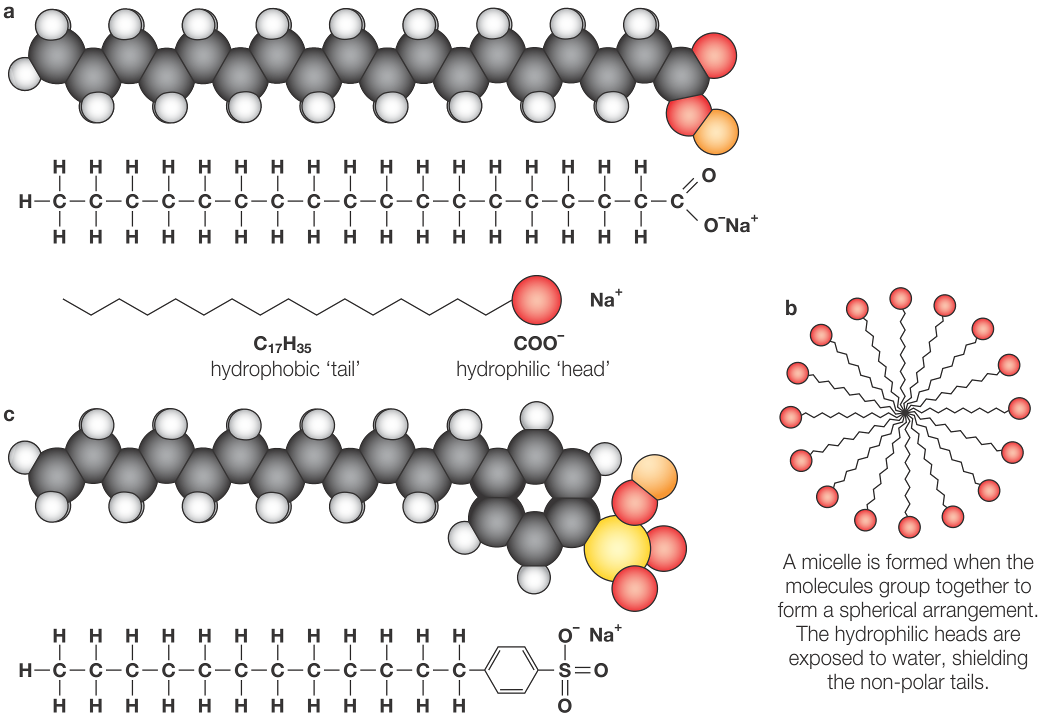
Liquid – the molecules have an irregular arrangement and orientation

The phase transitions of lyotropic liquid crystals depend on both temperature and concentration.

Lyotropic liquid crystals are found in many everyday situations. Soaps and detergents, for example, form lyotropic liquid crystals amphiphiles, meaning the molecules have a polar hydrophilic ('water loving') end and a non-polar hydrophobic ('water hating') end.

Many biological membranes also display lyotropic liquid crystalline behaviour. The molecules that can form a lyotropic liquid crystal state generally consist of two distinct parts – a polar, often ionic, 'head' and a non-polar, often hydrocarbon, 'tail' (Figures 22.49a and 22.49c). When dissolved in high enough concentrations in water, the molecules arrange themselves so that the polar heads are in contact with the polar solvent, water, in an arrangement called a micelle (Figure 22.49b).

Figure 22.49 **a** Soap molecules, such as sodium stearate (octadecanoate) shown here, have the amphiphilic structure of the type required to form a lyotropic liquid-crystal state in aqueous solution. **b** Soap and detergent molecules can form micelles in aqueous solution. **c** A molecule of a synthetic detergent shows similar structural features to a soap molecule



A micelle is formed when the molecules group together to form a spherical arrangement. The hydrophilic heads are exposed to water, shielding the non-polar tails.

Phase diagrams for liquid crystals

Figure 22.50 demonstrates the various phases a water-soluble liquid crystal can possess and their transitions between boundaries through different composition and temperatures. At certain concentrations they exhibit liquid crystal properties; at other concentrations they do not.

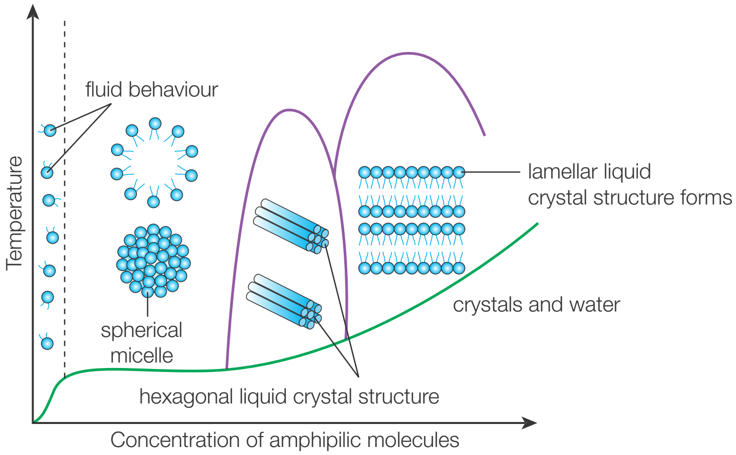

fluid behaviour
spherical micelle
lamellar liquid crystal structure forms
crystals and water
hexagonal liquid crystal structure
Temperature
Concentration of amphipilic molecules

Figure 22.50 Typical phase diagram for a typical lyotropic liquid crystal: various liquid crystal phases are observed

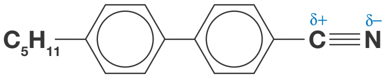
These phases are typical of amphiphiles or surfactants, such as soaps and detergents. Surfactants are compounds that lower the surface tension (or interfacial tension) between two liquids or between a liquid and a solid.

In dilute solution, the surfactants do not form any particular structure. As the concentration is increased, however, the amphiphiles condense into well-defined structures. The most readily formed structure is the micelle, where the surfactants 'hide' the hydrophobic tails inside a sphere, leaving only the water-soluble ionic heads exposed to the water molecules.

At higher concentration, surfactants can also form elongated columns that pack into hexagonal arrays. The columns have hydrophobic cores and hydrophilic surfaces. The columns are separated from one another by water molecules.

At extremely high concentration, the surfactants crystallize into a lamellar structure, with elongated sheets separated by thin water layers. The structure is very reminiscent of the lipid bilayers seen in cells.

21 Find out about the liquid crystalline nature of biological membranes and the role of liquid crystals in human diseases.



the molecule is polar because nitrogen has a higher electronegativity than carbon

Figure 22.51 The structure of 4′-pentylbiphenyl-4-carbonitrile

Liquid crystal devices

From the discussion above we have learnt about the molecular requirements for a substance to show liquid crystal properties under suitable circumstances. 4′-Pentylbiphenyl-4-carbonitrile is a commercially available nematic liquid crystal of the type used in LCDs – the compound has the structure shown in Figure 22.51.

4′-Pentylbiphenyl-4-carbonitrile is used in LCD devices (though mixtures of related compounds are often used in practice) because it shows the following appropriate properties:

- it is chemically stable
- it has a liquid crystal phase stable over a suitable range of temperatures
- its molecules are polar, allowing it able to change its orientation when an electric field is applied
- it responds to changes of voltage quickly – it has a fast switching speed courtesy of a low viscosity and high dipole moment.

The rod-like molecules of 4′-pentylbiphenyl-4-carbonitrile are suitable for LCDs because their ability to transmit light depends on their relative orientation. The molecule is polar, so its orientation can be controlled by the application of a small voltage across a small film of the material. When there is no applied voltage, light can be transmitted and the display is clear. When a small voltage is applied, the orientation of the molecules changes and light can no longer be transmitted through the film. The display then appears dark. The areas of the display that are light and dark can thus be controlled, enabling different shapes to be displayed (Figure 22.52).

Figure 22.52 The liquid crystal displays on a Mandarin–English electronic translator

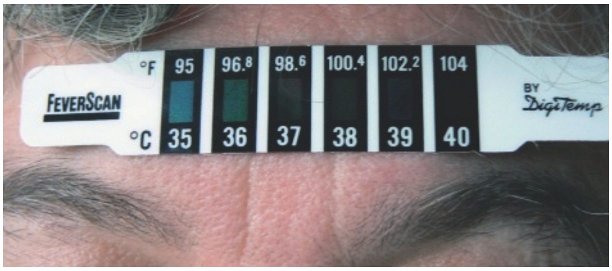


Figure 22.53 A strip thermometer for measuring skin temperature. Skin thermometers are less accurate than mercury or digital types of thermometers

The skin thermometer shown in Figure 22.53 uses materials that are chiral and show a chiral nematic phase. This adopts a helical arrangement and the wavelength of light reflected is proportional to the helical pitch, which is typically in the region 400–700 nm (visible light). This is also the basis for the colour effect on the shells of scarab beetles. The pitch is temperature dependent, decreasing as the temperature increases.

An LCD TV is a flat-panel television that uses LCD technology. Electric voltage is applied to individual pixels, which allow the liquid crystals to pass or to block light to create images. LCDs do not produce their own light, so an external light source, such as a fluorescent bulb, is needed for the image created by the LCD to be visible to the viewer. Unlike historical CRT (cathode-ray tube; Chapter 2) televisions, there are no phosphors that light up and this is why LCD panels are thin and require less power to operate. Because of the nature of LCD technology, there is no radiation emitted from the screen itself, unlike traditional televisions. Also unlike a traditional CRT television, the images on an LCD television are not scanned by an electron beam. The pixels of an LCD television, which constitute the image, are merely turned on or off in a particular sequence and at a particular refresh rate. Note that the pixels also contain colour filters.

Generation of liquid crystal displays

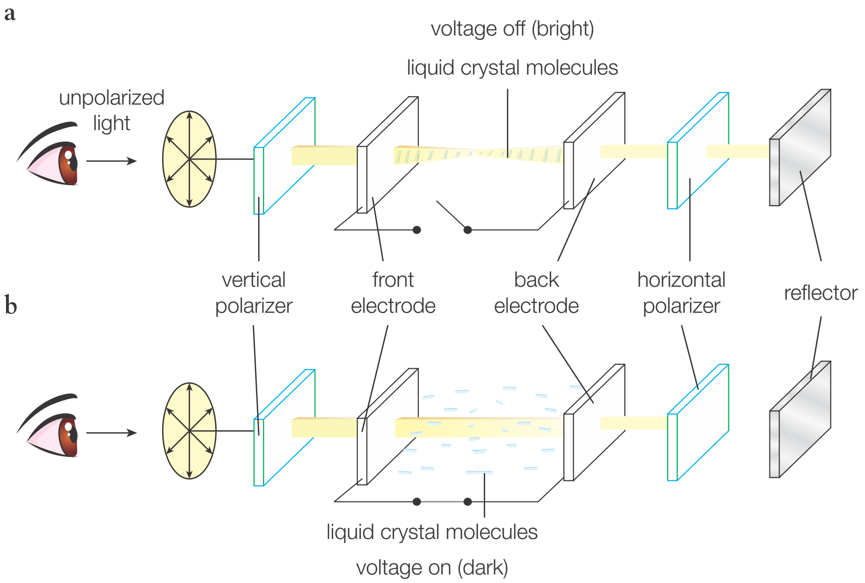
LCDs come in a variety of designs but often a thin layer (5–20 μm) of liquid-crystalline materials is placed between electrically conducting transparent glass electrodes (the conducting layer is indium tin oxide). Ordinary light passes through a vertical polarizer that permits light waves vibrating in only the vertical plane to pass.

During the fabrication process, the liquid crystal molecules are orientated so that the molecules at the front are orientated vertically and those at the back electrode horizontally. The orientation of the molecules in between the two electrodes varies systematically from vertical to horizontal, as shown in Figure 22.54a.

The plane of the polarized light is turned by 90 degrees as it passes through the liquid crystal layer and is therefore in the correct orientation to pass through the horizontal polarizer.

In an LCD watch display, a mirror reflects the light back, and the light retraces its path, allowing the device to look bright. When a voltage is applied to the plates, the liquid crystal molecules align with the voltage, as shown in Figure 22.54b.

Figure 22.54 Schematic illustration of the operation of a twisted nematic liquid crystal display (LCD)



(Strictly those molecules close to the surface do not reorientate as their surface anchoring holds them in the 'vertical' and 'horizontal' orientations, and so only those molecules in the middle of the device reorientate, but this is enough to give the necessary contrast. Having those close to the surface anchored ensures that the twisted arrangement is re-established when the voltage is removed.)

The light rays are therefore not properly orientated to pass through the horizontal polarizer, and the device appears dark. This type of device configuration is known as a twisted nematic display.

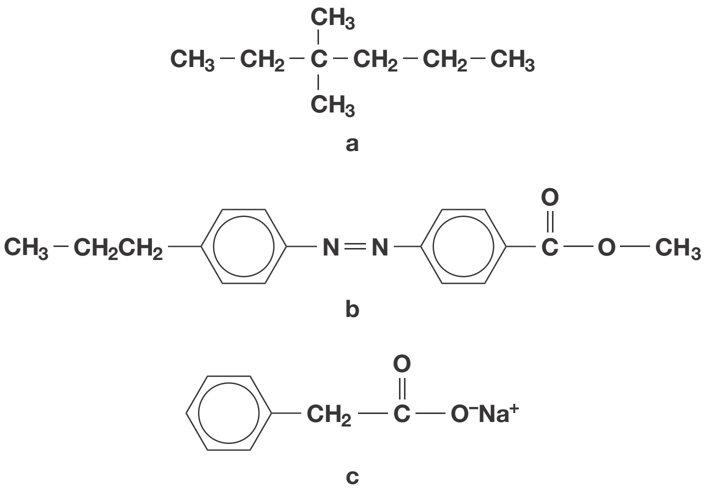
There is an energy efficiency disadvantage of LCDs, as the light source is not polarized: the necessity of the polarizer results in about half of the light not being used (it is simply absorbed by the polarizer). Also, it is necessary to have the backlight continuously on, which results in light leakage (not perfect polarizer/liquid crystal cell) and hence poor black and limited contrast. This explains why other technologies are still being researched, in particular organic light-emitting diodes (OLEDs).

Discussion of the properties needed for a substance to be used in LCDs

It essential for molecules of a potential liquid crystal to be polar so that they can be orientated in the presence of a weak electric field. The molecules also need to be long, rigid and rod-shaped which will prevent them for packing too closely together. It is also helpful if the molecule is chemically stable and will not undergo decomposition in the presence of ultraviolet radiation or at high temperatures.

Worked example

Examine Figure 22.55 and deduce which of the following substances is most likely to exhibit liquid-crystalline behaviour.



■ **Figure 22.55** Selected molecules

Molecule a is not likely to show liquid crystalline behaviour in bulk because it does not have a long axial structure.

Molecule b has the characteristic long axis and types of structural features often seen in liquid crystals, such as polarity.

Molecule c is ionic; the generally high melting points of ionic substances and the absence of a characteristic long axis make it unlikely that this substance will show liquid crystalline behaviour – though since it is amphiphilic it may show lyotropic behaviour.

Lyotropic liquid crystal polymers

Another type of liquid crystal is the lyotropic liquid crystal polymer. These polymers will enter the liquid crystal phase by exposure with solvents.

If a polymer is to be a lyotropic polymer, it must be fairly rigid and must dissolve in a solvent. These two requirements are often mutually exclusive as the rigid structure is usually not soluble. Hence, solvents such as sulfuric acid are often required.

An example of a lyotropic liquid crystal is the fibre Kevlar. Kevlar (see section 22.9) is a high-performance fibre and it is perhaps the most well-known liquid crystal polymer, with many applications from protective wear to aerospace. When dissolved in oleum (fuming sulfuric acid), at the appropriate concentration and temperature, Kevlar forms a liquid crystal phase. The liquid crystal solutions are then sheared, which aligns the polymer chains and then fibres are spun. The alignment of the polymer chains, which is possible owing to the presence of the liquid crystal state, confers the tensile strength on the material.

ToK Link

Moore's law is a famous statement made by Gordon Moore, founder of Intel, in 1965. Moore's law states that the complexity of integrated circuits (as measured by the number of components that make up 1 cm^2 of an integrated circuit) doubles every year. So clearly electronic devices can store increasing amounts of information in smaller physical spaces in digital form.

While new technologies make it possible to move more information faster than ever before, we should ask questions about the quality of the information: what is it that we are communicating? Is it relevant? And does all this information add up to knowledge and benefit for mankind?

22.5 Polymers *– polymers are made up of repeating monomer units which can be manipulated in various ways to give structures with desired properties*

Polymers are molecules of very high molar mass, each molecule being formed from a large number of small molecules known as monomers. The process by which monomers are covalently bonded to form a polymer is known as polymerization. There are two major classes of polymers: addition polymers and condensation polymers. Plastics are artificial polymers and rubber, cotton and wool are natural polymers.

Nature of Science

Development of polymer chemistry

Up until the 1920s there was little understanding of the molecular structure of these new materials. It was generally assumed that the small molecules from which they were made simply aggregated together into larger units rather than joining covalently to make larger molecules. It was the German chemist Hermann Staudinger who first recognized that polymers are made up of very large molecules. Another chemist who contributed greatly to the understanding of polymers as giant molecules was the American Wallace Carothers, the discoverer of nylon and neoprene.

The development of a comprehensive understanding of the structure and properties of polymers and the advancement in scientific techniques such as X-ray diffraction and the scanning tunnelling microscope (STM), infrared spectroscopy, NMR, neutron scattering and chemical techniques marked the start of a revolution in polymer chemistry. The 'hit or miss' approach to polymer synthesis was largely superseded and became much more rigorous and focused.

■ Addition polymers

One main use of the alkenes obtained from cracking crude oil fractions is to make addition polymers such as polyethene and polypropene. An introduction to the concept of addition polymerization is given in Chapter 10.

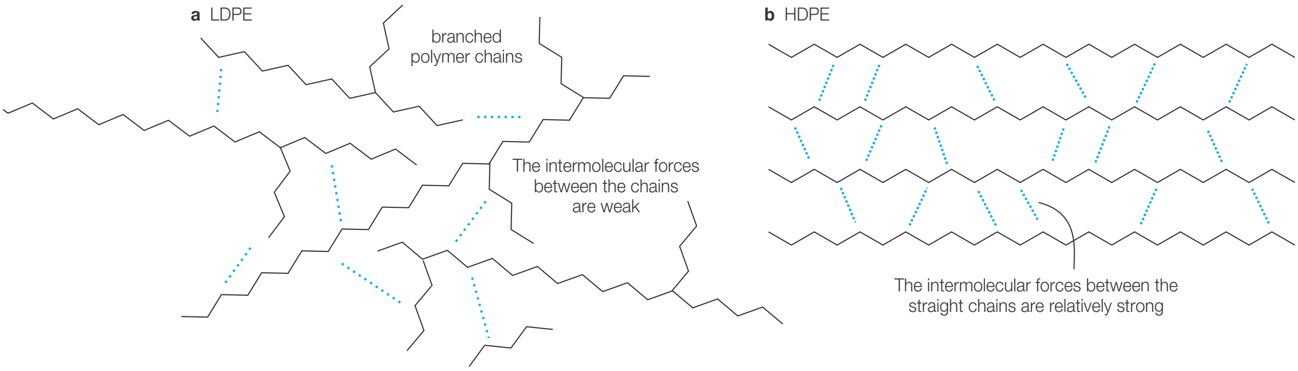
Modern society now relies heavily on these different types of addition polymers. Their properties depend not only on which functional groups are attached to the carbon–carbon double bond in the monomers, but also on the degree of branching and the way in which the side-groups are arranged in the polymers. The properties can be further modified by using more than one monomer in the same chain (copolymers), by using additives such as plasticizers and by the injection of volatile hydrocarbons during their production.

Branching

The simplest addition polymer, polyethene, has chains that can partially align in the solid state to produce regions of crystalline structure. It is this mixture of **crystalline** and **amorphous** (non-crystalline) structures that gives the plastic its mechanical properties, particularly its toughness. The ability of the chains to form this **semi-crystalline** structure depends in part on whether they are completely linear or contain branches. By controlling the reaction conditions during ethene polymerization, it is possible to form many different structures with varying degrees of branching. If ethene is polymerized at very high pressures, the reaction proceeds by a free-radical mechanism. Branched polymer chains are produced (Figure 22.56a) with many rather short (C_4) branches and a few longer ones.

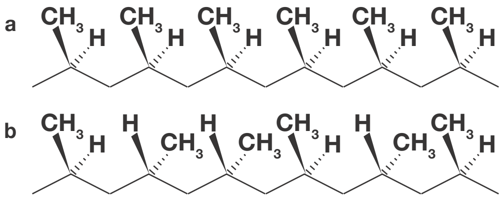
This branching makes for an 'open' structure as it limits the interaction between neighbouring chains, and the intermolecular forces between the chains are only relatively weak dispersion (London) forces. The resulting low-density polymer (0.91–0.94 $g\,cm^{-3}$) has a low melting point (around 100 °C) and is a resilient and flexible plastic. Only about 50–60 per cent of the plastic is crystalline. Low-density polyethene (LDPE) is mainly used for making films for food packaging and damp-proofing membranes, but it also finds use in items such as 'squeezy' bottles.

More linear and less branched forms of polyethene are formed by using special transition-metal based catalysts (Ziegler–Natta catalysts) in reactions at lower temperatures. Some catalysts can produce more or less perfectly linear chains with no branching (Figure 22.56b). These linear chains can pack together better than the branched chains of LDPE and give a more crystalline structure. In this high-density polyethere (HDPE) the molecules have straight chains and the plastic is more rigid because there is a higher fraction of crystalline material. This high-density form (0.95–0.97 $g\,cm^{-3}$) has a higher melting point (about 115 °C). This form of polyethene is used to make films for applications such as supermarket carrier bags, but it is also used for water pipes, containers, buckets and toys.


a LDPE
branched polymer chains
The intermolecular forces between the chains are weak
b HDPE
The intermolecular forces between the straight chains are relatively strong

■ **Figure 22.56 a** The branched nature of the chains in low-density polyethene (LDPE) makes for a flexible plastic product. In general, LDPE has about one C_4–C_6 branch per 100 carbon atoms of the chain, with the occasional longer branch. **b** The unbranched chains in high-density polyethene (HDPE) make for a more crystalline structure and a more rigid plastic

The highly branched LDPE and the non-branched HDPE are two extremes. It is possible to produce a range of types of polyethene with varying properties by modifying the extent and location of branching, and there are now a huge number of grades of polyethene available with properties varying from extremely soft and rubbery to stiff and rigid. Polyethene is the most versatile of all plastics.


a
CH3 H CH3 H CH3 H CH3 H CH3 H CH3 H
b
CH3 H H CH3 H CH3 CH3 H H CH3 CH3 H

■ **Figure 22.57 a** Planar representation of isotactic poly(propene) (with all the methyl groups on the same side of the carbon chain). **b** Part of a chain of atactic polypropene

Orientation of side-groups

The propene monomer has a methyl group in its molecule that is not present in ethene. Polypropene, therefore, has a structural feature not present in polyethene. Different orientations of the methyl side-groups can produce products with differing characteristic properties. The way in which each methyl group is stereochemically positioned, relative to the groups on each side of it, is referred to as the **tacticity** of the polymer and is vital in controlling whether and how the chains can crystallize.

A form of polypropene can be produced in which the methyl groups are randomly orientated – the **atactic** form (Figure 22.57b). In this form, the random orientation of the methyl groups prevents crystallization. This form of the polymer is soft, flexible and rubbery. It finds limited uses in sealants, adhesives and some speciality paints.

If the methyl groups can all be orientated on the same side of the polymer chain in a highly regular (**stereoregular**) manner then we obtain a polymer which is said to be **isotactic** (Figure 22.57a). As a result of its regular structure, isotactic polypropene is semi-crystalline and tough. It has a higher melting point (165 °C) than HDPE. Isotactic polypropene can be moulded into objects such as car bumpers (Figure 22.58) and plastic toys, or drawn into fibres for clothes and carpets.

It is also possible to obtain a **syndiotactic** polymer in which the methyl groups alternate stereochemically along the chain. This form of polypropene is also semi-crystalline and tough, but is more difficult to synthesize than the isotactic form and is quite new to the market.

The product of the propene polymerization reaction can be controlled by using different catalysts. This allows chemists to tailor-make polymers with precise properties. Ziegler–Natta catalysts are almost universally used and produce the isotactic form. Catalysts which make isotactic polypropene are typically heterogeneous. The monomer binds to the catalyst surface with the correct orientation to produce the more ordered polymer.

In theory, any polymer with a single substituent, such as polyvinyl chloride (PVC), can also exist in isotactic, syndiotactic and atactic forms. However, very high degrees of stereocontrol are needed for a polymer to be able to crystallize and most stereoregular polymers are very difficult or impossible to synthesize because of side reactions between monomer and catalyst which destroy the catalyst activity.

■ **Figure 22.58** Car bumpers can be made of the isotactic form of poly(propene)

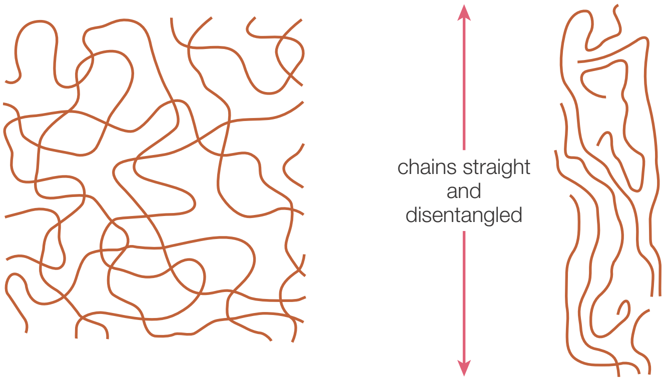
■ Elastomers

Most materials when stressed to an extension return to their original length or dimensions when the force (stress) is removed. When this critical extension (the elastic limit) is exceeded, the deformation (change in shape or length) becomes non-reversible – plastic behaviour – or causes a fracture (in glass and metals).

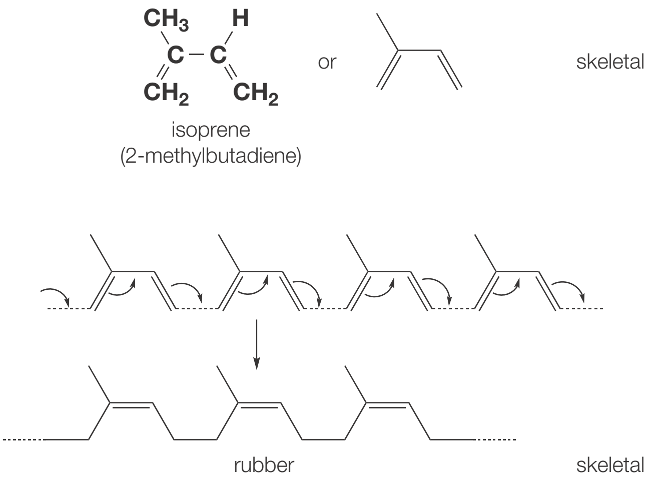
For most materials the elastic limit is very small. However, elastomers, such as natural rubber (Figure 22.59) and poly-2-methylpropene, behave differently: they can extend reversibly up to as much as ten times their original length. Elastomers are lightly cross-linked polymer networks and the structure is able to rearrange by rotation of the covalent bonds in the main polymer chain. When a polymer chain is stretched (elongated), the entropy (Chapter 15) decreases. The force which pulls the polymer chain to return to the un-deformed state comes from the return to maximum entropy on contraction.

Rubber

The rubber tree (*Hevea brasiliensis*) is indigenous to South America but was taken to South East Asia in the early 18th century by European colonists. When its bark is stripped, it oozes a sticky latex, which is an emulsion of rubber in water. The purified rubber is not useful in its natural form because it has a low melting point and low strength. Rubber is an addition polymer (Figure 22.60) of the diene 2-methylbutadiene (commonly known as isoprene) – a 'building block' of a number of biomolecules, including the carotenes.



■ **Figure 22.59** Unstretched and stretched rubber



■ **Figure 22.60** Isoprene undergoing addition polymerization to form rubber

Note that when a diene undergoes addition polymerization, a double bond is still present in the product. The double bond may be in the *cis* or *trans* configuration. In natural rubber, all the double bonds are in the *cis* configuration. The naturally occurring substance called gutta percha is an isomer of rubber, in which all double bonds are *trans*. Gutta percha is harder than rubber and thermoplastic. It was used in the past to make the cores of golf balls, drinking vessels and other moulded objects.

The presence of the double bond in rubber allows further addition reactions to take place. This is what occurs during the process known as vulcanization. Sulfur atoms can be regarded as adding across the double bonds in different chains, cross-linking the rubber molecules (see Figure 22.61). This stops the chains from moving past each other, and gives the material more rigidity.

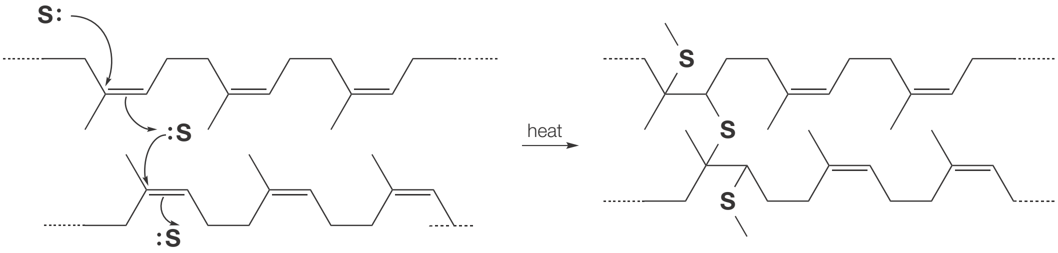


Figure 22.61 Vulcanization of rubber (simplified mechanism)

Not all double bonds have sulfur added to them since that would make the substance too hard. About 5% sulfur by mass is sufficient to give the desired physical properties. Millions of tonnes of vulcanized rubber are made each year for the manufacture of car tyres.

Synthetic rubber-like polymers were developed when rubber was in short supply during the Second World War. The most commonly used one today is a copolymer of phenylethene and butadiene, called styrene-butadiene rubber (SBR).

Deduction of structures of polymers formed from polymerizing 2-methylpropene

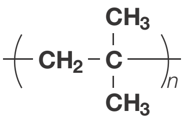
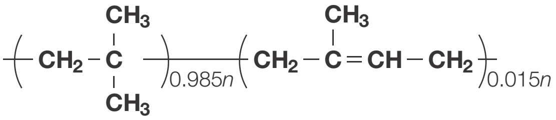


Figure 22.62 Repeating subunit of poly-2-methylpropene

2-Methyl propene can undergo polymerization to form poly-2-methylpropene (Figure 22.62), commonly known as poly-isobutylene. It is a synthetic rubber and an elastomer. It has the important and useful property of being the only rubber that is gas impermeable. It is used to make the inner liners of tyres and balls such as basketballs.

A more widely used synthetic rubber is butyl rubber, a copolymer (Figure 22.63) of isobutylene and isoprene (i.e. a polymer built of isobutylene and isoprene (2-methyl-1, 3-butadiene) units).

Figure 22.63 Repeating subunit of co-polymer of poly-2-methylpropene and 2-methyl-1,3-butadiene



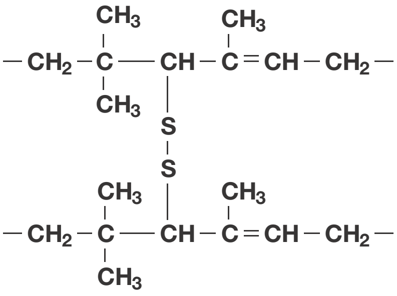


Figure 22.64 Cross-linked sulfur vulcanized butyl rubber

Butyl rubber consists primarily of isobutylene and minor amounts of isoprene. Polymerization of isoprene results in the incorporation of an alkene (i.e. carbon–carbon double bond) into the polymer chain. These double bonds serve as cross-linking sites (i.e. sites where one polymer chain can be chemically linked to another).

Vulcanization of the butyl copolymers results in the formation of a network structure in the form of a cross-linked rubber (Figure 22.64). Butyl rubber is a **thermoset** polymer and once vulcanized it cannot be reformed into a new shape. Poly-isobutylene is a **thermoplastic** polymer and can be reshaped by application of heat/pressure since none of the individual chains are chemically linked.

The use of plasticizers in PVC and hydrocarbons in the formation of expanded polystyrene

Plasticizers

Polychloroethene (more usually referred to as polyvinyl chloride or PVC) has properties that are a consequence of the side-group being a chlorine atom. The presence of the polar $C^{\delta+}-Cl^{\delta-}$ bonds in the polymer gives polyvinyl chloride very different properties from those of polyethene or polypropene. Every molecule has a permanent dipole allowing strong dipole–dipole interactions to occur between neighbouring chains. Also, chlorine atoms are relatively large and this restricts the ability of the chains to move relative to each other. The normal free-radical synthesis of PVC produces an atactic polymer which is amorphous. The consequences of these factors make the pure polymer hard, stiff and brittle. In this form it is often called unplasticized PVC (UPVC) – it has very wide applications in products such as window frames, gutters and sewage pipes.

However, when **plasticizers**, such as di-2-ethylhexylhexanedioate, are added (Figure 22.65) they act as lubricants and weaken the attraction between the chains, making the plastic more flexible. The plasticizer molecules fit between the polymer chains and separate them, allowing them to slip relative to each other more easily. By varying the amount of plasticizer added, a range of polymers can be produced with properties to suit particular purposes requiring rigidity to being fully pliable. Plasticized PVC is widely used in cable insulation (flex) and floor tiles and as a soft fabric coating.

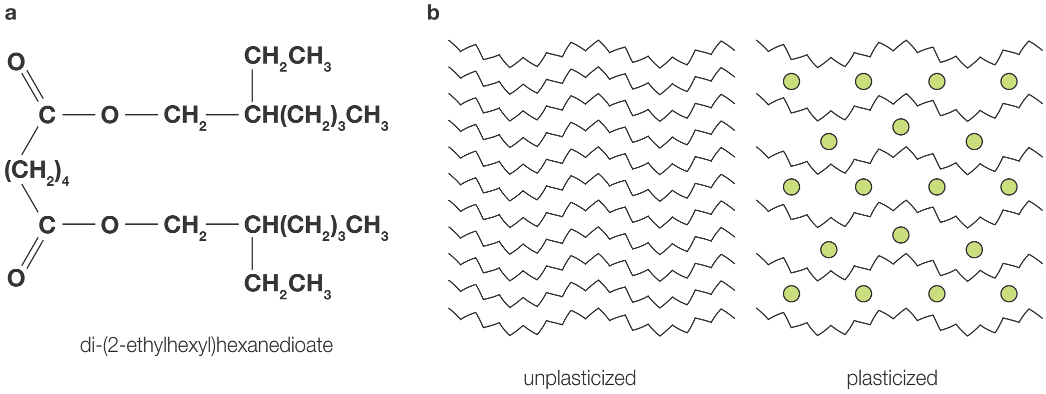


Figure 22.65 a A plasticizer used in the manufacture of PVC – di-2-ethylhexylhexanedioate. **b** The plasticizer molecules separate the polymer chains, allowing them to move freely past one another

Figure 22.66 Styrofoam cups

Volatile hydrocarbons (blowing agents)

Polyphenylethene – also known as polystyrene – is another common addition polymer. It is atactic and amorphous because it is normally made by free-radical polymerization. One version of the polymer is produced by dispersing the monomer as tiny droplets in water and polymerizing to give tiny beads of solid polymer. If this process is carried under a pressure of pentane, the hydrocarbon becomes trapped in the polymer beads. These can be loaded into a mould, which is heated. The polymer softens and the pentane vaporizes to produce an expanded foam structure with very low density, known as expanded polystyrene. This light material is a very good thermal insulator (Figure 22.66) and is used in many applications, including coffee cups. It is also used as packaging because it has good shock-absorbing properties, and in making theatre sets because it can easily be carved into shapes and is very light.

Description of ways of modifying the properties of polymers, including LDPE and HDPE

There are a number of ways of modifying the properties of polymers. The use of plasticizers and blowing agents has been described. Synthetic and natural fibres can also be blended, for example polyesters and cotton.

Another important technique used to modify the properties of polymers is copolymerization. Polypropene becomes glass-like at a temperature of about −10 °C, but for some applications a polymer is needed with the same general properties as polypropene, but with a lower glass transition temperature. The glass transition temperature is the temperature at which a glass undergoes a change from an amorphous to a crystalline phase.

One method to prepare such a material is to add some ethene to the propene during the polymerization process. This is termed copolymerization and both monomers become incorporated into the final copolymer. Figure 22.67 shows a possible structure of a small section of the copolymer.

Ion implantation can also be used to change the properties of polymers. The technique involves bombarding the polymer with ions to alter the chemical, physical and even electrical properties of the polymer. The surface is modified but the material's bulk properties are changed. The implanted ions can interact with the polar side-chains of polymers and increase the attraction between adjacent chains.

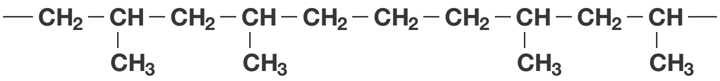


Figure 22.67 Structure of poly(propene-co-ethene) copolymer

Thermoplastic and thermosetting polymers

Most of the plastics that we use, such as polyethene ('polythene'), polychloroethene (polyvinylchloride or 'PVC') and polyphenylethene ('polystyrene'), can be softened by heating and will then flow as viscous liquids – they solidify again when cooled. Such plastics are useful because they can be remoulded – they are known as thermoplastic polymers (or thermoplastics or thermosoftening polymers).

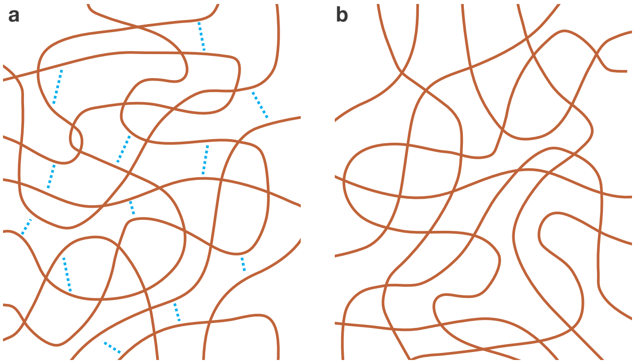


Figure 22.68 a Thermosetting and **b** thermoplastic polymers have different properties

Another, more restricted group of polymers can be heated and moulded only once, for example melamine–methanal resin (the material used in Formica). Such polymers are known as thermosetting polymers (or thermosets). The chains in these polymers are cross-linked to each other by permanent covalent bonds (Figure 22.68) during the moulding or curing process. They make the structures rigid when moulded, and no softening takes place on heating.

The term pre-polymer can be used to refer to the monomer or monomers of a thermosetting plastic. The process of polymerizing these monomers due to the formation of cross links is known as curing.

Many polymers are mixtures of **crystalline** (ordered) regions and **amorphous** (random) regions (Figure 22.69) in which the chains are further apart and have more freedom to move. A single polymer chain may have both crystalline and amorphous regions along its length. Let us examine the ability of polymers to form crystalline regions.

- Linear, unbranched, chain structures are most likely to form crystalline regions. Examples are isotactic polypropene, in which polymer chains pack closely together because the methyl groups, $-CH_3$, are regularly spaced along the chain; and Kevlar, in which the absence of both branched chains and bulky side-chains promotes the formation of crystalline regions.

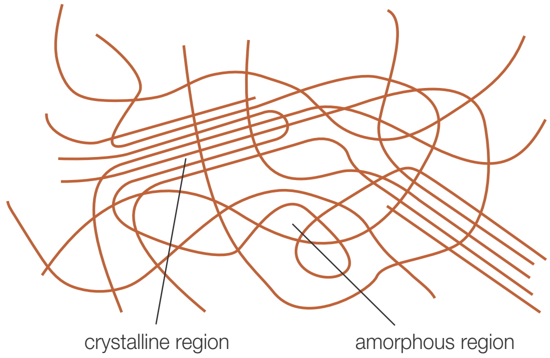


Figure 22.69 Crystalline and amorphous regions of a polymer

- Branched polymer chains are not easy to pack together in a regular manner, although if the branches are regularly spaced then some crystallinity is possible. Polychloroethene does not give rise to crystalline structures because the chlorine atoms are rather bulky and are irregularly spaced along the molecular chain.
- **Cross-linked** polymer chains cannot pack in a regular manner because of the covalent links between the chains so crystallinity is not possible.

HDPE is composed of long molecules with very little branching – fewer than one side-chain per 200 carbon atoms in the main chain. The chains can pack closely together into a largely crystalline structure, giving the polymer a higher density (Figure 22.70). Compared with LDPE, HDPE is harder and stiffer, with a higher melting temperature (about 115 °C) and greater tensile strength. It has good resistance to chemical attack, is brittle at low temperature and has low permeability to gases.

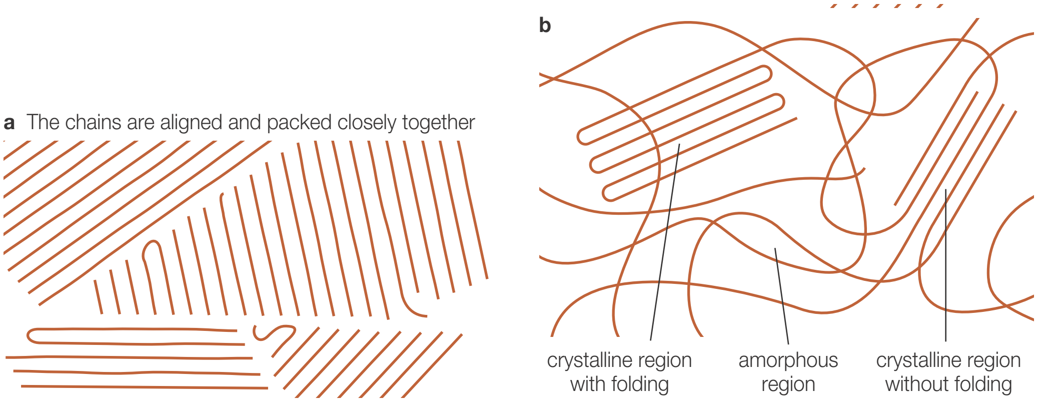

a The chains are aligned and packed closely together
b
crystalline region with folding
amorphous region
crystalline region without folding

Figure 22.70 The organization of the chains in **a** high-density polyethene and **b** low-density polyethene

The properties of polymers depend on their structural features

The physical properties of a polymer, such as its strength and flexibility, are determined by the structural features of its molecules:

Chain length: in general, the longer the chains, the stronger the polymer, but the more viscous in the molten state and thus harder to process.

Side groups: polar side-groups result in stronger attraction between polymer chains, making the polymer stronger.

Branching: straight unbranched chains can pack together more closely than highly branched chains, resulting in polymers that have a higher degree of crystallinity.

Cross-linking: if polymer chains are covalently linked together (curing), the polymer is harder and more difficult to melt. Thermosetting polymers have extensive cross-linking.

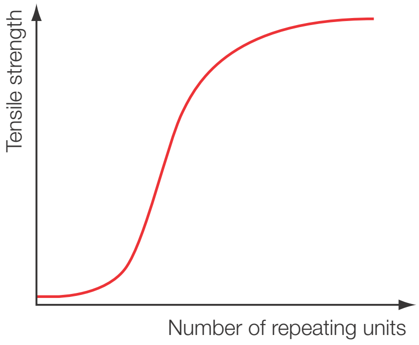


Figure 22.71 The relationship between chain length and tensile strength for a polymer

Strength of polymers

In general, the longer the chain of the polymer, the stronger the polymer will be. However, there is not a simple relationship since a critical length must be reached before strength increases. The length is different for different polymers. For polyethene this is at least 100 repeating units, but for nylon it may be only about 40 repeating units.

Figure 22.71 shows how tensile strength and chain length are related for a typical polymer whose chains are arranged randomly. There are two factors which cause the increase in tensile strength of a polymer with increasing chain length: longer chains are more tangled together and when chains are longer they have a greater surface area over which intermolecular forces can operate between adjacent chains.

22 Find out about the cold drawing of a polymer and its effect on strength.

Atom economy

Atom economy (Chapter 1) is derived from the principle of green chemistry. Atom economy is a measure of the proportion of reactants that become useful products.

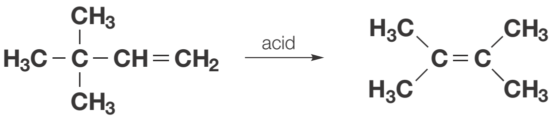
$$\%\text{ atom economy} = \frac{\text{molar mass of desired product}}{\text{molar mass of all reactants}} \times 100\%$$

In an ideal reaction, all reactant atoms end up within the useful product molecule and no waste is produced. Inefficient, wasteful reactions have a low atom economy. Efficient processes have high atom economy and are important for sustainable development. They conserve natural resources and create less waste.

Green chemistry is the sustainable design of chemical products and processes. It aims to minimize the use and generation of chemical substances that are harmful to human health or the environment. Chemists design reactions with the highest possible atom economy in order to minimize the environmental impact. Chemists want to reduce the consumption of raw material and energy.

Worked example

Deduce and comment on the atom economy in the following reaction:



C_6H_{12} C_6H_{12}

Total mass of reactants

= [(6 × 12.01) + (12 × 1.01)]

= 84.18 g

Mass of desired product:

= [(6 × 12.01) + (12 × 1.10)]

= 84.18 g

$\%\text{ atom economy} = \frac{84.18}{84.18} \times 100 = 100\%$

This rearrangement reaction has a very high atom economy as all the reactant atoms are incorporated into the desired products. The production of addition polymers also represents 100% atom economy since all of the reactant monomers end up in the polymer product.

23 Outline an investigation to determine the atom economy of nylon synthesis. Assume interfacial polymerization is used and the reacting chemicals are 1,6-diaminohexane and decanedioyl dichloride.

Solving problems and evaluating atom economy in synthesis reactions

The atom economy can be calculated from the balanced equation (without state symbols) for a synthesis reaction. The molar masses of the atoms can be used to calculate the masses of the reactants and products. The atom economy can then be calculated:

% atom economy = molar mass of desired product/molar mass of all reactants × 100%. This expression is on page 6 of the *IB Chemistry data booklet*.

Hydrazine (N_2H_4) is used for rocket fuel. Calculate the atom economy for hydrazine production.

$$2NH_3 + NaOCl \rightarrow N_2H_4 + NaCl + H_2O$$

The effects of plastics

Plastics were virtually unheard of prior to World War II. How has the introduction of plastics affected the world economically, socially and environmentally?

Plastics are all around us and are an important part of many products at home and at work (Figure 22.72). More than 300 million tons of plastics are produced worldwide every year. Plastics have a wide range of properties and have replaced a range of traditional materials (such as cotton, wool and leather) as well as metals for some uses. The majority of our plastics are synthetic polymers and thus their synthesis is ultimately dependent on crude oil (petroleum). Ethene is the starting material or

Figure 22.72 A bottle made from HDPE (high-density polyethene)

feedstock for alkene-based plastics and is a by-product of the cracking of long hydrocarbon chains. The rise in the use of plastics has fuelled, to a limited degree, the demand for crude oil and helped to increase its price on world markets, but polymers account for about 4.5% of all oil use.

Many plastics are resistant to attack to acids and alkalis and the alkene-based polymers are non-biodegradable and hence remain in the environment for a very long time. Plastic waste is about 10% of household waste, and if buried in landfills will remain largely unchanged for many years and may leach (if the lining is broken) harmful chemicals into ground water. This leads to the demand for more landfill sites. An alternative is incineration (burning at high temperature), which can be used to generate electricity, but a variety of toxic gases are generated. There have been moves in many countries to recycle plastics and to promote the use of biodegradable plastics. Although plastics are very long-lived products, their main use is as single-use items that will often enter a landfill within a year.

Plastics have changed society to make life easier and safer. The benefits of plastics include their use in cell phones, safety helmets, computers and hospital intravenous bags. However, plastics have also led to harmful imprints on the environment (Figure 22.73) and perhaps even affected human health (see section 22.7). Chemicals added to plastics are absorbed by human bodies and some of these compounds have been found to alter hormone levels or are directly toxic. Plastic debris and its chemicals are often eaten by marine animals, killing or injuring them. Floating plastic waste in the ocean (the plastic gyre) serves a transportation device for invasive species, disrupting habitats.

■ **Figure 22.73** Seaside rubbish – mostly plastic

22.6 Nanotechnology *– chemical techniques position atoms in molecules using chemical reactions whilst physical techniques allow atoms/molecules to be manipulated and positioned to specific requirements*

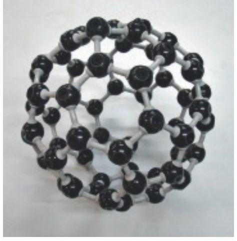
Nature of Science

Developments in nanotechnology

Richard Feynman (1918–1988), the Nobel Prize-winning physicist, gave a ground breaking talk in 1959 about the physical possibility of making, manipulating and visualizing matter on a small scale and arranging atoms 'the way we want'. He famously predicted that one day we would be able to fit an entire encyclopaedia on to the head of a pin. Feynman challenged scientists to develop a new field where devices and machines could be built from tens or hundreds of atoms. Some of the key points in the history of that field, now called nanotechnology**,** can be summarized in the following timeline. Critical to the development of nanotechnology have been improvements in apparatus and instrumentation which have allowed the visualization and individual movement of atoms and molecules. Theories to explain the formation of carbon-60 and carbon nanotubes are uncertain and still being developed and supporting evidence is being sought.

■ Some important events in nanotechnology history

1900	Max Planck proposes energy quantization
1905–30	Development of quantum mechanics by Heisenberg, Born and Schrödinger
1959	Feynman's talk – 'There is plenty of room at the bottom'
1974	Norio Taniguchi conceives the word 'nanotechnology'
1981	Invention of the scanning tunnelling microscope by Binnig and Rohrer
1985	Discovery of carbon-60 by Kroto, Curl and Smalley (Figure 22.74)
1986	Invention of the atomic force microscope by Binnig, Quate and Gerber
1989	Eigler of IBM writes the letters of his company using individual xenon atoms
1991	Discovery of carbon nanotubes by Sumio Iijima
2005	Beam of electrons used to shape metallic nanowires



■ **Figure 22.74** A geometric model of carbon-60 – an iconic image of nanotechnology

'The principles of physics, as far as I can see, do not speak against the possibility of manoeuvring things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but in practice, it has not been done because we are too big.'

— Richard Feynman, Nobel Prize winner in Physics

Nanotechnology involves research and technology development in the 1–100 nm range. It creates and uses structures that have novel properties because of their small size, and builds on the ability to control or manipulate on the atomic scale.

Nanotechnology is very much an interdisciplinary subject drawing on the approaches and techniques of chemistry, physics, biology and materials science. For the chemist, who mentally visualizes the world of atoms and molecules and models their interactions, 1 nm (10^{-9} m) is relatively large, whereas 1 μm (10^{-6} m) is considered small on an engineering scale. The general thinking about nanotechnology suggests two approaches to producing nanomaterials. The **top-down approach** starts with a sample of bulk material and breaks it into smaller pieces. The **bottom-up approach** builds the material from atomic or molecular species – by manipulating atoms or using self-assembly phenomena, for example. The second of these approaches has become more feasible with the development of ways of manipulating individual and small groups of atoms. Instruments such as the **scanning tunnelling microscope** (STM) and the **atomic force microscope** (AFM) have broken new ground in visualizing the nanoscale world and given insight into the levels of manipulation that are feasible. The understanding that materials behave very differently to their bulk properties on the nanoscale level was also important for making progress in the field. The physical rules when working at these levels are very different from those that apply to our everyday macroscopic world. Quantum effects and large surface-area-to-volume ratios can lead to the same material having a range of size-dependent properties. The colour of a material, for example, can depend on the size of the particles involved to a quite remarkable level.

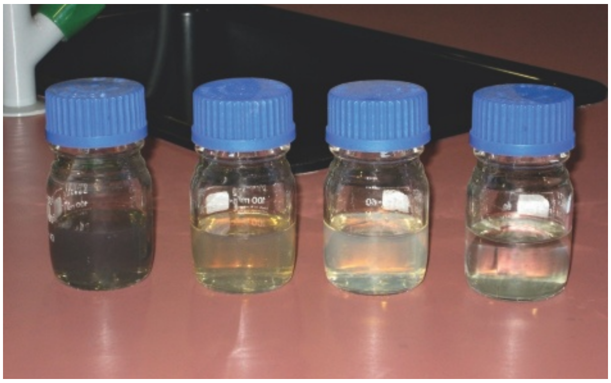


Figure 22.75 Silver nanoparticles show different colours depending on the particle size

Quantum effects

By modifying materials at the nanoscale level, physical properties such as magnetism, hardness and conductivity can be changed radically. These changes arise from confining electrons in nanometre-sized structures. On the nanoscale, electrons act like standing waves (Chapter 12). When electrons act like waves, they can pass through insulation that blocks flowing electrons (in a process known as quantum tunnelling). Some elements, such as gold and silver, can show change at the nanoscale. For example, silver nanoparticles (Figure 22.75) show changes in colour over the range of nanoparticle diameter from 20.0 nm to 90.0 nm. Physical properties – strength, crystal shape, solubility, thermal and electrical conductivity – along with magnetic and electronic properties also change as the particle size changes.

The development of nanotechnology

Making products on the nanometre scale is, and will likely become, a large and increasingly important sector of the economy for developed countries. In 2015, the global nanotechnology market was worth 27 billion dollors. Nanotechnology materials are expected to result in lighter, stronger, smarter, cheaper, cleaner and longer-lasting products. Researchers and technologists believe that nanotechnology will have several phases of development. The first consisted of using nanostructures, simple nanoparticles, designed to perform one task. In the second phase, researchers will construct nanoscale ‘building blocks’ – flat or curved structures, bundles, sheets or tubes. The third phase will feature complex nanosystems with many interacting components.

Presently there are more than 200 companies that sell a total of 700 products using nanotechnology applications. The range of nanoproducts currently available is surveyed below.

Current applications of nanotechnology

The following summary gives some idea of how nanotechnology has penetrated into a wide range of human activities.

- Sporting goods – nanoparticles made of carbon are used to stiffen certain key areas of tennis racquets and hockey sticks.

- Car paint and waxes – new car paints have improved scratch-resistant qualities compared to conventional car paint. Nano car-waxes, made with nano-sized polishing agents, provide a better shine due to their ability to fill in tiny blemishes (scratches or pits) in car-paint finishes.
- Antibacterial cleanser – a number of antibacterial cleaners use nano-emulsion technology to kill bacteria. The cleaners are non-flammable, non-corrosive and non-toxic.
- Medical bandages – special dressings for burns provide antibacterial protection using silver-based nanoparticles.
- Sunscreens and cosmetics – several companies have marketed sunscreens, deodorants and anti-aging creams all based on nanoparticles. These particles penetrate the skin rapidly.

Future applications of nanotechnology

- Environment – emerging nanotechnologies may result in the development of new approaches to detecting air pollution and cleaning polluted waste streams and groundwater. Magnetic nanoparticles have been developed that can absorb and trap organic contaminants in water. They could be used to clean up hazardous and toxic waste sites.
- Solar energy – photovoltaic cells convert solar energy to electrical energy. They are currently expensive and their efficiency is low. The use of nanoparticles may increase their efficiency.
- Vehicles – nanoscale powders and nanoparticles will be used to improve the physical properties of cars, aircraft, ships, trains and spacecraft. These vehicles will be lighter, faster and more fuel-efficient and be constructed of lighter and stronger materials. This will help in reducing the mass of the finished vehicle, energy efficiency and safety.
- Medical applications – many medical procedures could be handled by nano-machines that repair arteries and rebuild and reinforce bones. In the field of cancer nanotechnology research, scientists are testing and experimenting with new approaches to diagnose, treat and prevent cancer in the future. Nanoparticle-based approaches are being developed to target and selectively kill cancer cells.

The properties of nanomaterials

One of the first advances in nanotechnology was the invention of the scanning tunnelling microscope (STM). This instrument does not ‘see’ atoms but ‘feels’ them. An ultrafine tip scans a surface and records a signal as the tip moves up and down depending on the atoms present. The STM also provides a physical technique for manipulating individual atoms. They can be positioned accurately in just the same way as using a pair of tweezers.

In 1989, scientists at the IBM Research Centre in San Jose, California, manipulated 35 atoms of the noble gas xenon to write the letters ‘IBM’ (Figure 22.76). The letters were 500 000 times smaller than the letters used in the printing of this book. To place the atoms in the form of letters, the scientists used a special tip on the end of an STM to push them into place. The ‘bumps’ in Figure 22.76 are individual xenon atoms; each one is half a nanometre away from its neighbours.

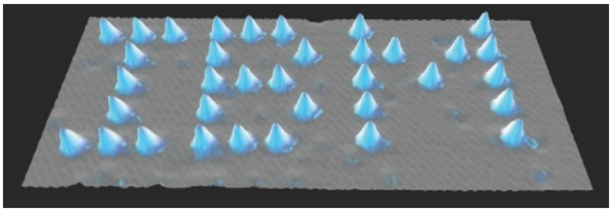


Figure 22.76 The IBM logo written in individual xenon atoms

This new-found ability to ‘see’ and manipulate individual atoms and molecules gave confidence to research in the revolutionary new field of nanotechnology research. The invention of the atomic force microscope (AFM) from its precursor the STM reinforced this growing confidence.

These inventions suggested that the ‘bottom-up’ approach to devising nanostructures was a feasible approach in this area of research. The ‘bottom-up’ approach is also encouraged by the phenomenon of self-assembly seen in several biological and chemical systems.

Distinguishing between physical and chemical techniques in manipulating atoms to form molecules

There are a wide variety of physical techniques that manipulate (move) atoms, ions and molecules at the molecular level including the AFM and the STM. Focused electron beam induced processing is a recently developed technique used to synthesize nanostructures.

The carbon atoms and other atoms that may be required for the formation of nanomaterials or nanostructures will be provided by a chemical reaction. Fullerenes and nanotubes were first synthesized by the arc discharge of graphite rods in a helium atmosphere (Chapter 4), but a new range of methods have been developed to synthesize nanotubes.

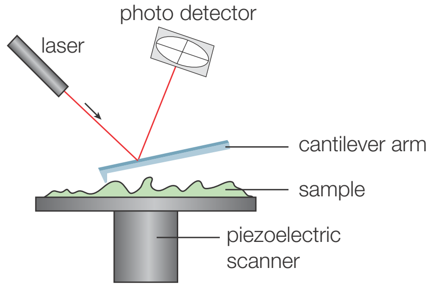


Figure 22.77 Atomic force microscopy (AFM)

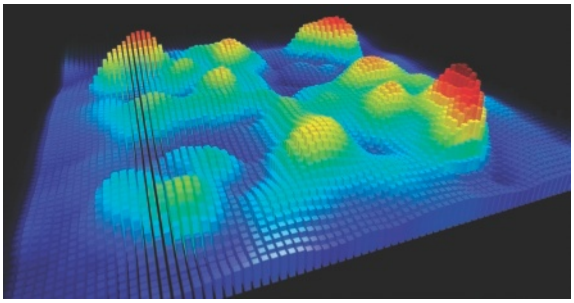


Figure 22.78 Three-dimensional representation of the experimental atomic probe force microscopy image of a single naphthalocyanine molecule (at low temperature and ultra-high vacuum)

Atomic force microscope

The key component of an AFM is a very small and light cantilevered arm, at the end of which is a sharp tip (made of silicon or silicon nitride) which acts as a probe. This is held very close to the sample (Figure 22.77). The sample is deposited on to a flat surface (often silica (silicon dioxide)), which sits on a piezoelectric block.

By changing the voltage across this block, it can be caused to move, with a distance of less than 1 nm. As the sample moves, the tip of the AFM moves up and down, tracking the features (shape) of the sample. The height of the AFM tip can be measured reflecting a laser off the back of the arm. The AFM can measure the height of the sample, but it can also respond to electrical charge or conductivity. Hence, as the tip is scanned across the sample, it maps out the height (or charge or conductivity) of the sample. One advantage of AFM over electron microscopy is that the sample can remain hydrated and the sample does not need to be dried.

An additional advantage of AFM is that it can be used in non-contact mode. The tip oscillates at a regular frequency (a harmonic) and is placed to within a few nanometres of the surface. London (dispersion) forces and other interactions interfere with the regular oscillations and provide a visualization of the surface without contact (Figure 22.78).

Scanning tunnelling microscope

The STM was invented in 1981 and gives images of individual atoms and molecules. The STM is based on a fine-tipped metal probe that scans across a small area of a solid surface. The probe's movement is controlled by three piezoelectric transducers within 1 picometre (10^{-12} m) and is at a small constant potential difference (voltage). The tip is held at a fixed height of no more than 1 nm (10^{-9} m) above the surface under analysis.

At this height, electrons 'tunnel' across the gap and the tiny current is used move the probe vertically to maintain the gap width. Hence, if the STM probe tip moves near a raised atom, the tunnelling current increases and this is used to move the probe vertically up to keep the gap width constant. The vertical resolution is of the order 1 pm, much smaller than the size of the smallest atom.

Figure 22.79 shows how the potential energy of an electron varies from the STM tip to the surface of the solid. The electron is unable to overcome the potential barrier due to the strong metallic bonding of electrons in the metal tip, even though the tip is relative to the surface. Figure 22.79 shows that the amplitude of the 'electron wave' (wave function) decreases

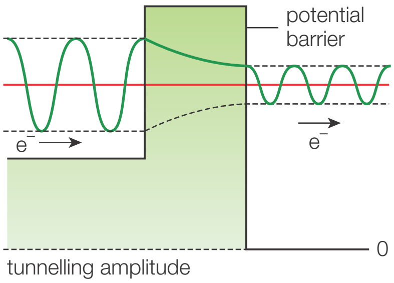


Figure 22.79 The principle of the scanning tunnelling microscope

exponentially with distance from the STM tip. In effect, electrons that escape from the tip to the surface tunnel through the barrier. This is a quantum effect and due to the uncertainty principle (Chapters 2 and 12).

As outlined in Chapter 2, the de Broglie wavelength of an electron depends on its momentum (product of mass and velocity), and is in the order of nanometres in the STM. Hence, quantum tunnelling is possible for gaps of the orders of nanometres. The tunnelling current is very sensitive to changes of the STM gap of as little as 1 pm.

ToK Link

The use of the scanning tunnelling microscope has allowed us to 'see' individual atoms, which was previously thought to be unattainable. How do these advances in technology change our view of what knowledge is attainable?

More than 30 years ago the scientific world saw the appearance of real space imaging of single atoms with never-before-seen resolution. This was the development of one of the most versatile surface probes, based on the physics of quantum mechanical tunnelling: the STM. Invented in 1981 by Gerd Binnig and Heinrich Rohrer of IBM, Zurich, it led to their award of the 1986 Nobel Prize.

Atoms, once regarded as abstract entities used by theoreticians for calculations, can be seen to exist for real with the nano 'eye' of an STM tip that also gives real-space images of molecules and adsorbed complexes on surfaces.

From a very fundamental perspective, the STM changed the course of surface science and engineering. STM also emerged as a powerful tool to study various fundamental phenomena relevant to the properties of surfaces in technological applications such as medical implants, catalysis and sensors, besides studying the importance of local bonding geometries and defects.

Atom-level probing, once considered a dream by Feynman, is a reality with the evolution of STM. An important off-shoot of the STM was the AFM for the surface mapping of insulating samples. AFM has enabled researchers in recent years to image and analyse complex surfaces on microscopic and nanoscopic scales.

The invention of AFM by Gerd Binnig, Calvin Quate and Christopher Gerber opened up new opportunities for the characterization of a variety of materials, and many industrial applications are possible. AFM observations of thin-film surfaces give scientist a 'picture' of surface topography and morphology and any visible defects. The growing importance of ultra-thin films for magnetic recording in hard disk drive systems requires an in-depth understanding of the fundamental mechanisms occurring during growth.

Focused electron beam induced processing

Focused electron beam induced processing (FEBIP) is a developing lithographic (printing) technique used in nanotechnology. It uses a beam of electrons (from an STM) to decompose transiently (for short periods of times) adsorbed molecules under low vacuum conditions (Figure 22.80). This is to enable the construction of three-dimensional structures on single-nanometre scale dimensions.

FEBIP can result in either surface etching or deposition depending upon the products formed during the electron beam's interaction with the precursor molecules. When the decomposition products react with the substrate to form volatile species, material is removed from the surface in a process referred to as electron beam induced etching.

In contrast, deposition occurs when electron-stimulated decomposition produces non-volatile fragments in a process called electron beam induced deposition. Virtually any nanostructure shape can be produced accurately using this method. It is analogous to a three-dimensional printer. It is a potential tool for the nanofabrication of magnets, wires, gears and motors, all on the nanoscale. When structures are deposited within an electron microscope they can be imaged using scanning electron microscopy or scanning transmission electron microscopy.

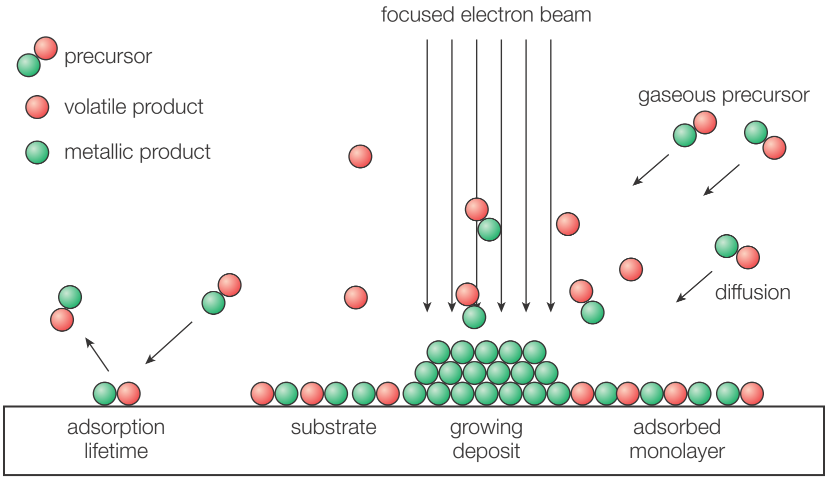


Figure 22.80 Focused electron beam induced processing

Fullerenes

Fullerenes are a structural form (allotrope) of pure carbon. Theoretically, a wide range of molecular shapes can be engineered at the molecular level using fullerenes. The structure of carbon-60 (buckminsterfullerene, C_{60}) was discussed in Chapter 4. The addition of pentagons into the hexagonal structure of graphite allows the carbon atoms to form a closed, approximately spherical cage. The discovery and synthesis of C_{60} was one of the key developments in nanochemistry. C_{70} and a range of smaller and larger fullerenes have been characterized.

Carbon-60 is a simple molecular or molecular covalent substance and exhibits properties typical of that class of substances. Unlike other forms of carbon, fullerenes may be soluble. C_{60} is pink and C_{70} is red in solution. The interactions between the solvent molecules and the surface of the cage is more energetically favourable than the interactions between adjacent C_{60} molecules.

The properties of C_{60} are a combination of the microscopic and macroscopic or bulk properties. The bonding within a single molecule of buckminsterfullerene is very strong. The carbon-60 molecule is very unlikely to be easily squashed or deformed during collision, so the molecule itself could be described as being ‘hard’. However, the fact that carbon-60 has a low sublimation point suggests that it should have the typical properties of a compound whose intermolecular forces are mainly London (dispersion) forces. So in a bulk sample, we would expect C_{60} to be soft, like a polycyclic hydrocarbon such as anthracene ($C_{14}H_{10}$), which consists of three fused benzene rings.

In C_{60} the carbon atoms are all joined by delocalized π bonds, so it would be expected that an electric current should easily pass around the molecular sphere. However, transfer of the current (pi electrons) from one molecule to the next does not occur because of the large energy gap. In bulk form, samples of the simpler molecule that contain delocalized pi electrons (such as benzene) are good insulators, although NMR shows that electrons readily move around the aromatic ring within each molecule (a ring current).

Figure 22.81 Carbon nanotubes

Nanotubes

Following the discovery of C_{60} a whole family of structurally related carbon **nanotubes** was discovered. These resemble a rolled-up sheet of graphite, with the carbon atoms arranged in repeating hexagons (Figure 22.81). The nanotubes, which have a diameter of 1 nm, can be closed at either end if pentagons are present in the structure.

A whole series of carbon-based molecules, including structures with multiple walls of concentric tubes, have been produced. Carbon nanotubes have been shown to have very useful properties. Bundles of carbon nanotubes have tensile strengths between 50 to 100 times that of iron because of the strong covalent bonding within the walls of the nanotube. Different nanotubes have different electrical properties because at the nanoscale the behaviour of electrons is very sensitive to the dimensions of the tube. Some nanotubes are conductors and some are semiconductors.

Their properties can also be altered by trapping different atoms inside the tubes. For example, silver chloride can be inserted into a tube and then decomposed by light to form an inner coating of silver. The resulting nanotube is a thin metallic electrical conductor. As these nanotubes have relatively large surface areas and can be made with specific dimensions, they have the potential to be very efficient and size-selective heterogeneous catalysts. Their mechanical (stiffness, strength, toughness), thermal and electrical properties suggest a wide variety of applications from batteries and fuel cells, to fibres and cables, to pharmaceuticals and biomedical materials.

Two types of nanotubes

There are two main types of carbon nanotubes – single-walled carbon nanotubes and multi-walled carbon nanotubes. Most single-walled nanotubes have a diameter of close to 1 nm, with a tube length that can be many thousands of times longer. In fact, single-walled nanotubes can reach a length of over 1 cm. The structure of a single-walled nanotube can be visualized by wrapping a one-atom-thick layer of graphite called graphene into a cylinder. A graphene is a two-dimensional single sheet of sp^2-bonded carbon atoms.

A single-walled nanotube is cylindrical with at least one end typically capped with a hemisphere of the carbon-60 structure. The diameter of the nanotube is only a few nanometres wide and can extend up to 50 μm in length. Nanotubes have the following physical, chemical and mechanical properties that make them such a potentially useful nanomaterial:

- Electrical conductivity – depending on their precise structure, carbon nanotubes can be either metallic conductors or semiconductors. The electrons can travel much faster in nanotubes than in metals, and they do not dissipate or scatter. Carbon nanotubes could find an important use in conducting circuits as they can sustain a current density of up to $10^9\,\text{A cm}^{-2}$ (compared to copper at $10^6\,\text{A cm}^{-2}$).
- Thermal conductivity – the thermal conductivity of nanotubes is superior to that of diamond. This is despite diamond, an electrical insulator, being an excellent thermal conductor because the increased vibrations caused by heat are rapidly transferred throughout a structure in which all the atoms are bonded to each other in an extensive lattice. In some tests, nanotubes have been shown to have a thermal conductivity at least twice that of diamond.
- Mechanical – nanotubes are the stiffest, strongest and toughest fibre currently known. With their small size, nanotubes are six times lighter than steel but up to 100 times stronger, giving them a strength : weight ratio 600 times greater than that of steel.

The carbon nanotubes described here are formed by curving individual sheets of graphite, otherwise known as **graphene**. Graphene (Chapter 4) is a single atomic plane of graphite, which is sufficiently isolated from its environment to be considered free-standing. It is the thinnest known material and the strongest ever measured. Graphene can sustain electric current densities six orders of magnitude higher than that of copper, and it also shows record thermal conductivity and stiffness. Graphene can adsorb and desorb various atoms and molecules – for example NO_2 and NH_3.

Graphene is a one-atom-thick planar sheet of sp^2-bonded carbon atoms that are arranged in a hexagonal crystal lattice.

Single-walled carbon nanotubes can be considered to be graphene cylinders – some have a hemispherical graphene cap (that includes six pentagons) at one or both ends. The manipulation of graphene and fullerenes means that a wide range of novel structures can be constructed, including 'nanobuds' (Figure 22.82).

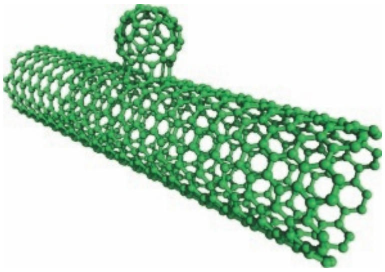


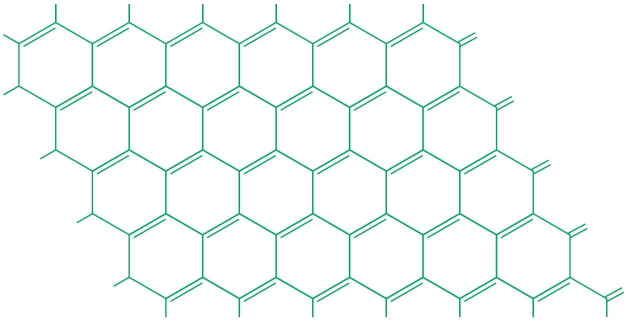
Figure 22.82 An example of a 'nanobud' structure – a fullerene 'budding' off a carbon nanotube

Structure and physical properties of carbon nanotubes

The main cylinder or tube of a carbon nanotube is composed of carbon atoms arranged into hexagons (essentially a single layer of graphite (graphene) curved into an open-ended tube). However pentagons are needed to close the structure at the ends and form closed nanotubes.

The nanotube molecule is held together by strong covalent carbon–carbon bonds that extend all along the nanotube. Single or multiple-walled tubes made from concentric nanotubes (that is, one tube inside a larger nanotube) can be formed.

Long carbon nanotubes are conducting and this is due to the presence of delocalized electrons that move along the surface of the nanotube in the presence of voltage. The bonding can be regarded as being similar to graphite with each carbon atom (sp^2 hybridization) contributing one electron to the delocalized 'cloud' of electrons.



■ **Figure 22.83** Skeletal formula of graphite or graphene

One possible skeletal formula representation of a layer of graphite or a molecule of graphene is shown in Figure 22.83 in localized Kekulé form as in aromatic compounds, such as benzene.

The carbon–carbon (C–C) bond length in graphite or graphene is 0.142 nm, which is between a single carbon–carbon bond length of 0.154 nm and a carbon–carbon double bond (C=C) of 0.134 nm. The carbon–carbon bond order in graphite or graphene is 1.33, because each carbon atom uses three valency electrons to form single bonds; the fourth valence electron is delocalized.

The bond order is 1.5 in benzene (Chapter 20), the average of a carbon–carbon single bond (bond order 1) and carbon–carbon double bond (bond order 2), but there is a C–H bond too. The C–C–C bond angle is exactly 120°, for the planar carbon hexagons (due to the presence of three negative charge centres). In graphite the planar hexagonal ring layers of carbon atoms are 0.335 nm apart. Nanotubes are essentially a single layer of graphite (or a molecule of graphene) wrapped around to form an elongated tube-like molecule; they only consist of hexagonal rings throughout the sides of the nanotubes.

24 Find out about the zigzag, chiral and armchair forms of carbon nanotubes.

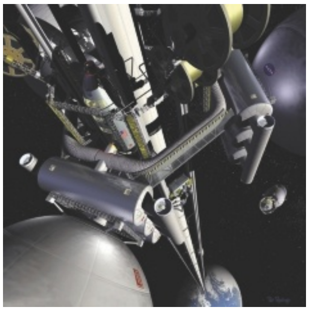
The tensile strength of carbon nanotubes is exceptionally high due to the strong covalent bonds holding the carbon atoms together. The carbon atoms are relatively small and the overlap of orbitals between adjacent atoms is effective. Short, strong bonds with partial double bond character are formed

In 2000, a multi-walled carbon nanotube was tested to have a tensile strength of 63 GPa. In comparison, high-carbon steel has a tensile strength of approximately 1.2 GPa. Under excessive tensile strain the tubes will undergo plastic deformation, which means the deformation is permanent.

Space elevator

International collaboration in space exploration is growing. Would a carbon nanotube space elevator be feasible or wanted? What are the implications?

A space elevator (Figure 22.84) is a proposed type of space transportation system. It would consist of a ribbon-like cable (tether) anchored to the surface of the Earth and extending into space. It is designed to allow vehicle transport along the cable from the Earth's surface directly into space or orbit, without the use of large rockets. The cable would be under tension due to gravity which is stronger at the lower end and centrifugal force acting at the upper end.



■ **Figure 22.84** Possible design of space elevator

The cable would need to be made of a very light but very strong material. A carbon nanotube composite or graphene ribbon might be a suitable material. A space elevator would greatly reduce the cost of carrying cargo and humans into space. This would expand the possibilities for space tourism, scientific research and even space colonization. The use of a space elevator would also reduce the pollution and space debris caused by the use of rockets.

Ribosomes and molecular motors

Self-assembly is an important example of the 'bottom-up' approach in nanotechnology. It is the spontaneous formation of precise and well-defined molecular structures from small molecules. Nature provides a number of examples of self-assembled structures, such as viruses, ribosomes (sites of protein synthesis in cells) and cell membranes, which all form via molecular recognition involving complementary shapes and favourable intermolecular forces. Ribosomes are found in all cells, including bacterial cells and consist of a number of different protein and RNA molecules that self-assemble into two subunits. When supplied with messenger RNA and amino acids ribosomes will generate proteins in a process known as translation. The structure of ribosomes has been solved using X-ray crystallography.

In recent years, chemists have developed a number of elegant syntheses that have resulted in the formation of a range of different molecular structures, many of which could form the basis of future 'molecular machines' and 'molecular motors'.

In 1999 a research group in the Netherlands made the first molecular motor (Figure 22.85). The motor is powered by light and the molecule rotates about a carbon–carbon double bond. The groups either side of the double bond are identical and ultraviolet radiation causes these to undergo *cis–trans* isomerization. Because of the large size of the groups, which are chiral, the motor can rotate in only one direction.

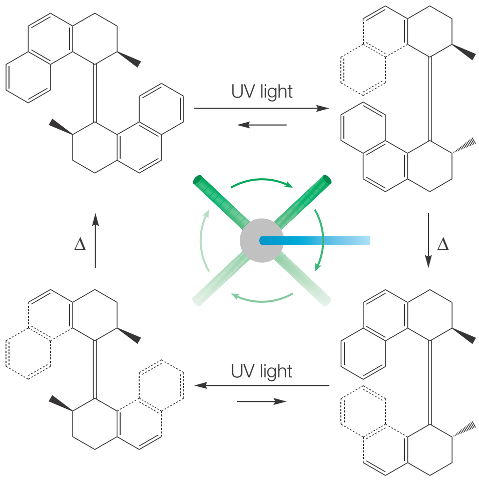


Figure 22.85 The first light-driven molecular motor

Synthesis of carbon nanotubes

Methods of synthesizing carbon nanotubes include arc discharge, chemical vapour deposition (CVD) and high pressure carbon monoxide disproportionation (HiPCO). Arc discharge was initially used to produce fullerenes (C_{60} and C_{70}) but nanotubes are also formed. It involves either electrically vaporizing the surface of a graphite electrode (using a large current) or discharging an electrical arc through metal electrodes in a hydrocarbon solvent forming a small rod-shaped deposit on the anode.

Arc discharge using graphite electrodes

Two graphite rods are placed about 1 mm apart in a bell jar of inert gas (originally helium, but argon is also used). A large direct current (DC) produces a high-temperature discharge (spark) between the two graphite electrodes, vaporizing parts of one carbon anode (positive electrode) and forming a small rod-shaped deposit on the carbon cathode (negative electrode) (Figure 22.86).

The anode may contain small amounts of a transition metal to act as a catalyst. If pure graphite electrodes are used then the formation of multi-walled nanotubes tends to be favoured. The presence of nickel, cobalt or yttrium (a lanthanoid) favours the formation of single-walled nanotubes.

However the arc discharge method needs a large-scale vacuum system and a frequent replacement of graphite electrodes, because the electrodes are consumed as a carbon source of the carbon nanotubes by the arc discharge.

The arch discharge method is not based on oxidation or any electrochemical process, as the electric arc is needed simply to vaporize graphite so that carbon nanotubes or fullerenes can be formed from carbon vapour (mainly gaseous carbon atoms).

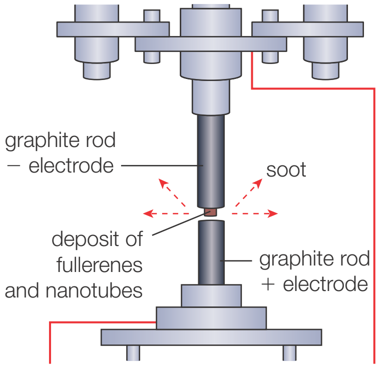


Figure 22.86 Arc discharge method for carbon nanotube production

Arc discharge using metal electrodes

Nickel or iron electrodes can be used for discharge in a hydrocarbon solvent, for example methylbenzene ($C_6H_5CH_3$) or cyclohexane (C_6H_{12}). The organic solvent is the source of carbon atoms as the hydrocarbon is decomposed by the electrical arc and soot is produced at the anode (which occurs with methylbenzene) or dispersed through the solvent.

Since the metal electrodes have high mechanical strength, the consumption of the electrodes is negligible and the exchange of the electrodes is not necessary. As a result, the arc discharge with the metal electrodes in the hydrocarbon method allows the continuous synthesis of carbon nanotubes.

The production of carbon from hydrocarbon solvents in arc discharge by oxidation at the anode

In the use of a hydrocarbon solvent, it is the solvent not the electrodes which are the source of carbon. In order to obtain carbon from a hydrocarbon the carbon in the solvent, such as cyclohexane or toluene, is oxidized, and therefore the carbon is formed at the anode. The arc discharge in this case works on the solvent rather than a graphite electrode. The solvent and metal used for the electrode seems to have some influence on production of nanoparticles. Methylbenzene with the pi electrons in the benzene ring has a greater yield than cyclohexane, which is all singly bonded carbons.

25 Find out about the laser ablation method for synthesizing carbon nanotubes.

Chemical vapour deposition

In chemical vapour deposition (CVD) carbon atoms are deposited on to an inert substrate. This is done via the decomposition of a gaseous hydrocarbon, such as methane, ethyne or carbon monoxide, or even an alcohol, in the presence of a transition metal catalyst. The C–H bonds in the gas phase molecules are dissociated by either plasma discharge or heat. In effect the hydrocarbon undergoes 'cracking' and the gaseous carbon atoms diffuse towards an inert substrate (often zeolite) which is covered with a catalytic surface. The catalyst is usually iron, nickel or cobalt and is attached to the substrate by heating or etching (due to the action of hydroxyl radicals $^\bullet$OH).

The substrate is heated in an oven to over 600 °C and the hydrocarbon or alcohol gas is gradually introduced. The gaseous molecules decompose and the carbon atoms form nanotubes on the substrate surface. The apparatus (Figure 22.87) must be free of air to prevent the formation of carbon dioxide and carbon monoxide. The carbon atoms diffuse to the substrate by diffusion and form either single-walled or multi-walled nanotubes depending on the conditions.

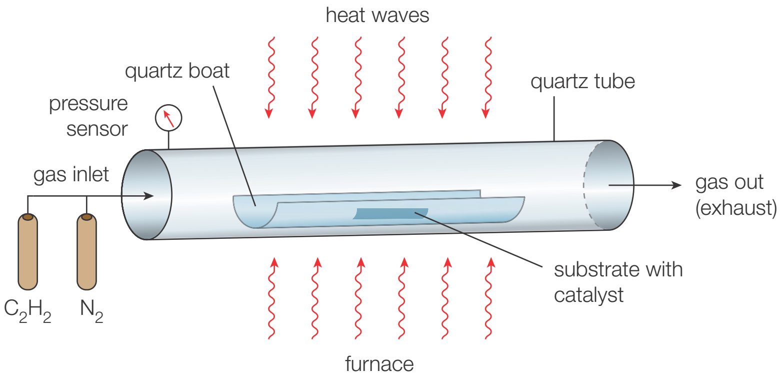


Figure 22.87 Chemical vapour deposition (CVD)

Methane or carbon monoxide is heated to over 900 °C to form single-walled carbon nanotubes, but if ethyne is heated to 700 °C multi-walled carbon nanotubes are formed. Single-walled nanotubes have a higher enthalpy of formation (per carbon atom) than multi-walled carbon nanotubes.

One method of CVD is high pressure carbon monoxide disproportionation (HiPCO). In a disproportionation reaction the same substance is both oxidized and reduced. In HiPCO hot carbon monoxide gas is continuously supplied at high pressure (50 atmosphere pressure) into the reaction mixture. The catalyst iron pentacarbonyl(0), $Fe(CO)_5$, is also fed in.

During this process the iron pentacarbonyl(0) reacts to produce iron nanoparticles:

$$Fe(CO)_5(g) \rightarrow Fe(s) + 5CO(g)$$

The iron nanoparticles provide a nucleation surface for the transformation of carbon monoxide into carbon during the growth of the nanotubes.

$$xCO(g) \rightarrow \text{carbon nanotubes (s)} + \frac{1}{2}xCO_2(g)$$

where x is typically 6000 giving a carbon nanotube containing 3000 carbon atoms.

No substrate is needed and the reaction can take place with a continuous feed making it suitable for industrial-scale production. In HiPCO, carbon monoxide is reduced to carbon, which forms nanotubes, and is also oxidized into carbon dioxide (which is removed):

$$2CO(g) \xrightarrow{Fe(CO)_5} C(s) + CO_2\ (g)$$

Deduction of equations for the production of carbon atoms from HiPCO

The Boudard reaction is responsible for the formation of carbon atoms during HiPCO:

$$2CO(g) \rightarrow C(s) + CO_2(g)$$

This is a disproportionation reaction involving a change in the oxidation number of carbon from +2 to 0 (in elemental carbon) and +4 (in carbon dioxide).

Explanation of why an inert gas, and not oxygen, is necessary for CVD preparation of carbon nanotubes

An inert gas is necessary for the CVD preparation of carbon nanotubes. The presence of oxygen would result in the combustion (oxidation) of the carbon atoms to oxides of carbon. Carbon nanotube synthesis is hence prevented. An inert gas, either a noble gas or nitrogen, is used to displace the oxygen in air and prevent oxidation.

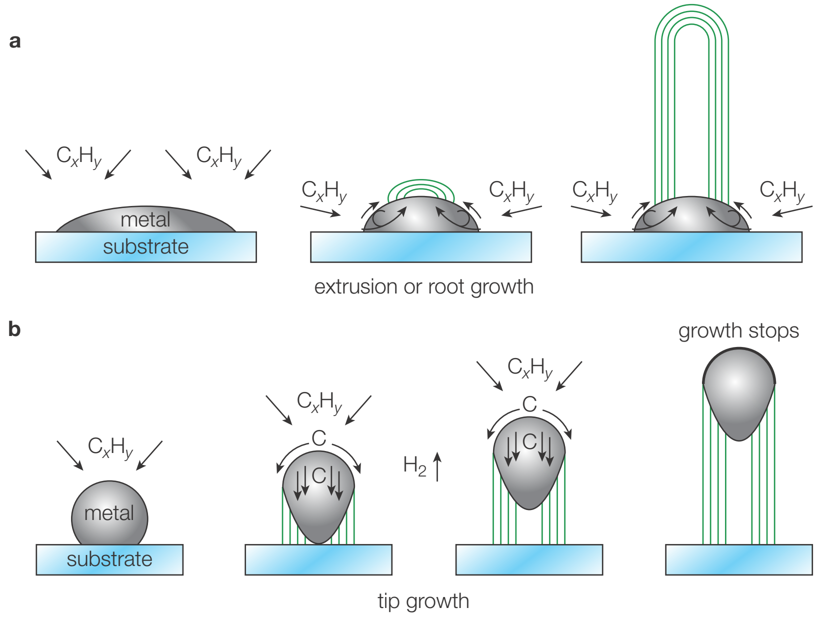
Technique	**Arc discharge**	**Laser ablation**	**Chemical vapour deposition (CVD)**	**High pressure carbon monoxide disproportionation (HiPCO)**
Date of publication	1992	1995	1993 of nanotubes	1999
Method	Electric plasma discharge vaporizes a graphite electrode, depositing it on the other electrode (a few millimetres distance) as single or multi-walled nanotubes	Laser pulse is absorbed and vaporizes graphite surface, forming gaseous carbon atoms	Uses heat of an oven to crack a gaseous hydrocarbon into carbon atoms, which are deposited on a substrate containing an etched-on transition metal catalyst	Carbon atoms produced in the disproportionation reaction from carbon monoxide in the presence of vaporized $Fe(CO)_5$ catalyst to produce nanotubes
Alternative method	In a hydrocarbon solvent using metal electrodes	Continuous wave instead of pulse laser	Plasma discharge instead	Cobalt-molybdenum catalyst instead of $Fe(CO)_5$ complex
Specific conditions	Inert gas (usually helium) low pressure atmosphere; temperature of electrodes >3000 °C; large current (100 A) passed through electrodes	Inert gas low pressure atmosphere; gaseous flow; temperature approximately 2000 °C	Catalyst etched and deposited on substrate; CH_4, C_2H_2 or CO; temperature >1000 °C to crack hydrocarbons or decompose CO	High carbon monoxide pressure; temperature >1000 °C to crack hydrocarbon molecules (lower in the presence of Co–Mo catalyst
Yield	About 50% per batch of graphite electrodes; replaced each time; about 10 g per day	About 70% per batch before replacing electrode; powdered graphite; less than 1 g per day	About 50%; large quantities produced (over 1000 kg) due to continuous flow and substrate size	Greater than 95% yield, can be run continuously with gas flow, producing about 1 kg per day
Advantages	Mainly defect–free short nanotubes	Bundle of very high-quality single-walled nanotubes that can be produced with desired diameters (via control of conditions)	Relatively easy to scale up to industrial production	Very high yields; few carbon impurities other than nanotubes
Disadvantages	Small nanotubes with random orientations (chirality) and sizes; difficult to purify	Very expensive (high cost)	Produces mostly long multi-walled nanotubes; difficult to separate single- from multi-walled nanotubes	Some defects in nanotubes

Table 22.3 Summary of methods of carbon nanotube production

Nature of Science

Growth of nanotubes

The way in which carbon nanotubes are formed is not exactly known. The growth mechanism (Figure 22.88) is still a subject of debate and active research, and more than one mechanism might be operating during the formation of carbon nanotubes.



■ **Figure 22.88** Visualization of possible nanotube growth mechanisms. A precursor is formed on the metal catalyst. The carbon diffuses on the sides leaving the top of it free, resulting in the hollow core of the nanotube. Out of this a rod-like carbon structure is formed. In base growth (extrusion) the nanotube grows upwards from the metal particle that remains attached to the catalyst. In tip growth the particle detaches and stays on the top of the growing nanotube

One of the proposed mechanisms consists of three steps. First a precursor to the formation of nanotubes and fullerenes, a C_2 molecule, is formed on the surface of the metal catalyst particle. From this unstable molecule, a rod-like carbon molecule is formed rapidly. Then there is a slow conversion of its wall to graphene. This mechanism is based on observations using a transmission electron microscope at different times during the synthesis. The actual growth of the carbon nanotube seems to be the same for all the major synthetic techniques.
There are several theories proposed as a growth mechanism for nanotubes. One suggests that metal catalyst particles are supported on graphite or another substrate. It assumes that the catalyst particles are spherical or pear-shaped, in which case the deposition will take place on only one half of the surface. The carbon diffuses down the concentration gradient (from high to low) and deposits on the opposite half, around and below the bisecting diameter. However, it does not precipitate from the apex of the hemisphere, which accounts for the hollow core that is characteristic of these filaments. For supported metals, filaments can form either by extrusion (also known as base growth) in which the nanotube grows upwards from the metal particles that remain attached to the substrate, or the particles detach and move at the head of the growing nanotube (tip growth). Depending on the size of the catalyst particles, single-walled or multi-walled nanotubes are grown. In arc discharge, if no catalyst is present in the graphite, multi-walled carbon nanotubes will be grown on the C_2 molecules that are formed in the plasma.

Safety and regulatory issues in nanotechnology

Some studies have shown that inhaling nanoparticle dust can be as harmful as inhaling asbestos. Should nanotechnology be regulated or will this hinder research?

There have been proposals for the tighter regulation of nanotechnology which has occurred together with a growing debate about the human health and safety risks of nanotechnology. There has been considerable debate about who is responsible for the regulation of nanotechnology. In most countries there will be regulatory agencies covering some nanotechnology and products and processes, to varying degrees. There need to be regulations to assess and control risks associated with the release of nanoparticles and nanotubes. There is insufficient funding of the human health and safety risks associated with nanotechnology and as a result many stakeholders, including scientists, have called for the application of the precautionary principle with longer times to market approval, more informative labelling and additional safety data development. There is a particular risk of nanoparticles or nanotubes being released during the disposal, destruction and recycling and hence end-of-life regulations are needed for nano-based products.

ToK Link

Some people are concerned about the possible implication of nanotechnology. How do we evaluate the possible consequences of future developments in this area? Is the knowledge we need publicly available or do we rely on the authority of experts?

There is no doubt that nanochemistry is providing materials scientists and chemists with incredibly useful atomic and molecular structures with a wide range of potential industrial applications. However, like all new discoveries and their applications to the world of technology, the very novel nature of these new materials raises safety issues and implications. The effect of nanoparticles on the body is difficult to predict and not easily understood and the long-term effects on our health and well-being are quite unknown. Since the properties of nanoparticles are not as well known or as easily predicted there are worries that nanoparticles may have undiscovered harmful side effects if, for example, they are breathed in or orally ingested, so there may well be health and safety issues which are not yet fully understood. There are genuine health concerns and toxicity regulations are difficult to apply as the properties of nanomaterials depend on the size of particles. In reality, there must be unknown health effects because all new materials have new health risks. There is concern that the human immune system may be defenceless against nano-sized particles. This poses responsibilities for the nanotechnology industries, which in turn raises political issues, for example for informed public education and debate and for the public to be involved in policy discussions about decisions related to nanotechnology.

22.7 Environmental impact – plastics *– although materials science generates many useful new products there are challenges associated with recycling of, and high levels of toxicity of, some of these materials*

The new discipline of materials science has developed many useful materials and products, but it raises challenges associated with the recycling and hazardous nature of some new materials.

Plastics are composed mainly of carbon and hydrogen atoms, though halogens, nitrogen and oxygen may also be present. They have strong covalent bonds along the chain so plastics do not decompose readily and are relatively stable. Some organic polymers can dissolve in organic solvents, but plastics are generally resistant to attack by acids and water.

Many are slowly degraded by oxidation, accelerated by ultraviolet radiation present in sunlight, and the majority are not biodegradable. Polyamides and polyesters will undergo slow hydrolysis by acids and alkalis. The hydrolysis is only significant at high concentrations and temperatures.

Some addition polymers, such as polyvinylchloride (PVC), contain chlorine and release hydrogen chloride (HCl) or dioxins (some of which are toxic) when combusted in air. Benzene and unsaturated hydrocarbons are formed during the thermal decomposition of PVC.

The presence of volatile plasticizers is another environmental issue associated with some plastics (mainly PVC and cellulose acetate), since they may evaporate or leach from the plastic. Phthalate esters are often used as plasticizers and although less toxic than dioxins are known to disrupt the endocrine system, leading to cellular and genetic changes in fish. Safer plasticizers with better biodegradability and fewer biochemical effects are being developed, for example triethyl citrate.

26 Find out what happened at Seveso in Italy in the context of dioxins.

Chlorine-free plastics are also being used as substitutes for PVC, including high-density polyethene, polypropylene and polyisobutylene. In the event of a fire these halogen-free plastics will not release dioxins or hydrogen chloride (which forms hydrochloric acid in the presence of water).

Deduction of the equation for any given combustion reaction

The equations below describe the complete combustion of pure polyethene and PVC.

$$(CH_2)_n + 1\frac{1}{2}nO_2 \rightarrow nCO_2 + nH_2O$$

$$(CH_2CHCl)_n + 2\frac{1}{2}nO_2 \rightarrow 2nCO_2 + nHCl + nH_2O$$

Marine plastic waste

The oceans have rotating current or gyres, each with a calm spot at the centre. There are five major gyres and they are caused by the Coriolis effect due to the rotation of the Earth. Chemical sludge and floating plastic rubbish including microscopic particles collect at the centre of these gyres (Figure 22.89).

Figure 22.89 Floating rubbish in an oceanic gyre

One suggestion for cleaning up the gyres would be to use the surface currents to let the debris drift to specially designed arms and collection platforms of boats. This way the running costs would be virtually zero, and the operation would be so efficient that it may even be profitable.

Organohalogen compounds

There are a number of organohalogen compounds found as pollutants in water and the atmosphere. Marine organisms, especially algae, act as natural sources of atmospheric organohalogens.

Polychlorinated dibenzo-*p*-dioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs) are pollutant compounds with the general formulas shown in Figure 22.90. Many are believed to be toxic; one of the most well-known is 2,3,7,8-TCCD or 'dioxin'.

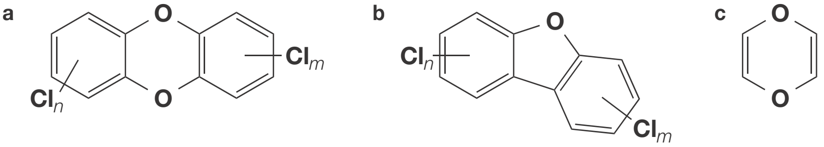
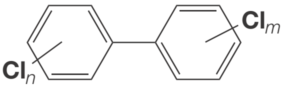


Figure 22.90 The generalized structures of **a** PCDFs, **b** PCDDs and **c** dioxin (where *m* and *n* refer to different numbers of chlorine atoms)

PCDDs and PCDFs enter the atmosphere from many sources, including car engines, waste incinerators, and steel and other metal production. The most important source is probably the burning of household waste in older municipal waste incinerators. PCDDs and PCDFs are formed when PVC plastic is burnt in the presence of metals, which act as catalysts. Hydrogen chloride, carbon monoxide and carbon may also be formed during the combustion of PVC (though modern incinerators have scrubbers).

Dioxin and other PCDDs can enter humans via eating food, specifically fish, meat and dairy products (milk), since dioxins are fat soluble and are easily passed through the food chain. Breast feeding results in the passing of dioxins from the mother to the child. Dioxins are also present in cigarette smoke.

Dioxins build up in fatty tissues over time in a process known as bioaccumulation. Small repeated exposures may eventually result in high concentration levels. Dioxins are slowly eliminated by the body over a number of years. Effects of dioxins at high-dose levels include acne, abnormal teeth development in children, thyroid disorders, damage to the immune system, diabetes and endometriosis (where uterus lining cells grow outside the uterus). Exposure to dioxins can also change the ratio of male to female births, resulting in the birth of more females.



■ **Figure 22.91** Generalized structure of PCBs

Polychlorinated biphenyls (PCBs) are synthetic organic molecules containing two linked benzene rings where more than one hydrogen atom has been substituted by chlorine atoms (Figure 22.91). They do not have a 1,4-dioxin ring in their structure, but have the same toxic effect as dioxins and so can be described as dioxin-like.

Products that have contained PCBs include transformers and capacitors, oil used in motors and hydraulic systems, thermal insulation material including fibreglass, plastics (as plasticizers), oil-based paint and floor finish. However, the European Union (EU) and many countries, including the US, no longer allow their commercial production.

The effect of plastic waste and organic pollutants on wildlife

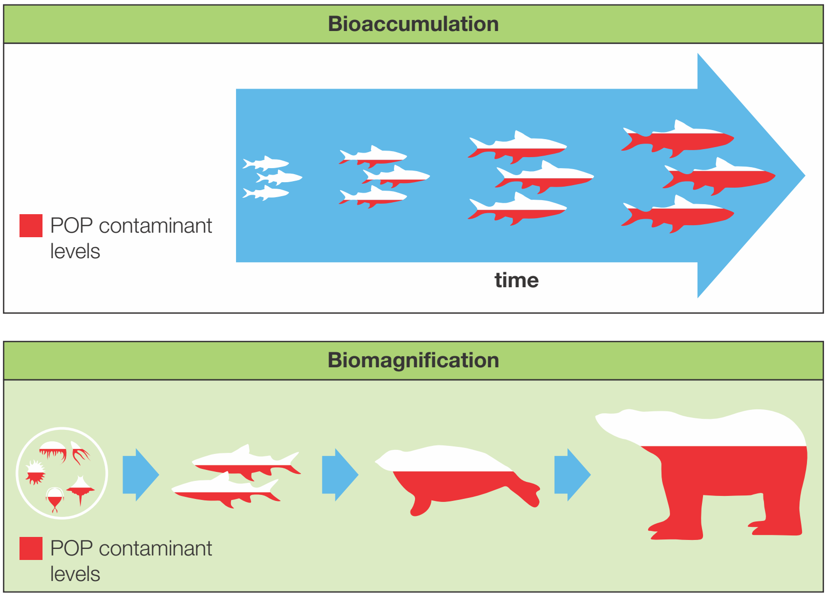
Plastic waste is a growing concern and the demand for plastics will continue to grow. Plastics are highly useful materials and their applications are expected to increase as more new products and plastics are developed to meet demands.

The increased use and production of plastic in developing countries is a particular concern, as the sophistication of their waste management infrastructure may not be developed enough to fully cope with their increasing levels of plastic waste.

Large plastic bottles and bags break down due to the action of the Sun and abrasion by the waves. These smaller pieces can be mistaken for prey by marine animals. Over a million sea birds, marine mammals and turtles are killed each year from ingesting plastic or entangled with plastic. The smaller pieces of plastic eventually form microscopic particles of plastic that float near the surface of the oceans in gyres. Other less known effects of marine plastic waste are the alteration of habitats and the transport of alien species.

Persistent organic pollutants (POPs) such as dioxins and PCBs can enter the food chain, having long-term effects on the health of animals up the food chain. Plastic waste has the ability to attract contaminants, such as POPs. This is particularly so in the marine environment since many of these contaminants are hydrophobic, which means they do not mix with water. However, in some conditions plastic could potentially act as a sink for contaminants, making them less available to wildlife, particularly if they are buried on the seafloor.

As POPs are passed along the food chain their concentrations increase and can reach very high levels in top predators (Chapter 23). This process, known as biomagnification, has been largely responsible for the extinction or significant population reduction of many birds of prey and large marine animals across the globe, including in regions far distant from the places where the POPs were released to the environment. Figure 22.92 shows the processes of bioaccumulation (increase in POPs concentration with time) and biomagnification (increase in POPs concentration when nutrients enter a new trophic (feeding) level).



■ **Figure 22.92** Bioaccumulation and biomagnification

27 Find out about the uses and possible harmful effects of bisphenol.

International symbol for recycling

The international symbol for recycle, reuse, reduce is a Mobius strip designed in the late 1960s, but global recognition of this symbol ranks well below other symbols. What factors influence the recognition of symbols?

Figure 22.93 International symbol for recycling

The universal recycling symbol for recyclable material consists of a Mobius strip comprising three changing arrows (representing recycle, reduce and reuse) used to form a triangle (Figure 22.93). The symbol was designed by American Gary Anderson, a college student in the late 1960s. It was the winning entry for an art contest sponsored by a Chicago-based recycled paperboard company to raise environmental awareness among high schools and colleges across the country.

A single image or symbol delivers considerable amounts of information in a very short time because we perceive an image all at once, whereas reading or hearing often takes significantly longer to process the same information. Important factors in the recognition of symbols are colours and shapes.

Disposal of plastics

28 Find out what happens to the plastic waste in your own country and what recycling schemes are operating in the towns and cities.

Plastic waste may form up to about 10% (by volume) of household waste. If plastic waste is buried in landfill sites it will remain unchanged for many years. An alternative to dumping is incineration, with the possibility of making use of the heat generated. Plastics are products of crude oil (petroleum), and plastic waste contains about the same amount of energy as the crude oil or petroleum it is derived from. However, some plastics burn with the formation of toxic gases, for example hydrogen chloride (HCl) from PVC and hydrogen cyanide (HCN) from polypropenenitrile, and incinerators must be designed to remove these gases from the exhaust.

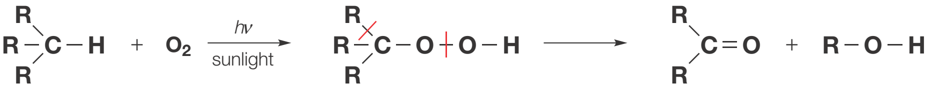
Recycling of plastics

Recycling rather than disposal into a landfill or incineration is an obvious way of reducing the environmental impact of plastics. The atom economy (Chapter 1) increases while the need for the manufacture of new plastics is decreased. However, there are significant challenges to successfully recycling plastics at low cost. Thermoset plastics, such as polyurethanes, cannot be melted down and recycled – they will decompose before melting if heated so cannot be reprocessed. The combustion of chlorine-containing polymers, such as PVC, means that dioxins may be released. Another major difficulty is that since different plastics have different properties, mixed plastic waste is of limited use. Mixtures of plastics are much weaker than individual plastics. The waste plastics need to be sorted and separated. The easiest plastic waste to recycle comes from industrial waste and is mostly relatively 'pure' plastic.

Degradation of plastics

Polyalkenes are straight-chain alkanes and therefore chemically unreactive. They biodegrade only very slowly in the environment. There are strains of bacteria that can metabolize straight-chain alkanes once they have been 'functionalized' by oxidation somewhere along the hydrocarbon chain; they find branched-chain alkanes more difficult to degrade. They can therefore attack HPE more easily than LDPE. The compact nature of the chains and the presence of weaker tertiary C–H bonds allows photo-oxidation (Figure 22.94) to take place more easily. Polychloroethene (PVC) is less inert and its carbon–chlorine bond is slowly attacked by alkali and can undergo cleavage in ultraviolet radiation (sunlight).

Figure 22.94 Photo-oxidation of tertiary polyalkanes



Resin identification codes

Plastics are recycled based on their polymer type, identified by a resin identification code (RIC). This coding system was developed by the Society of the Plastics Industry (SPI) in 1988 and used internationally. Its primary purpose was the efficient identification of plastic polymer types, but it was soon applied to the classification of plastics for the purpose of recycling. The symbols used in the code consist of the recycling symbol enclosing a number, often with an acronym representing the plastic below the triangle. The number on the code gives information about the polymer type rather than its hardness, how frequently it can be recycled, difficulty in recycling, or colour. The number does not indicate how hard the item is to recycle, nor how often the plastic has already been recycled. It is an arbitrarily assigned number that has no other meaning aside from identifying the specific plastic. Section 30 in the *IB Chemistry data booklet* provides a list of RICs.

Recycling is an energy-intensive and labour-intensive process, since some preliminary sorting is done by hand. Plastic bottles can be sorted into polyethene, PVC and PET. Plastic bottles for recycling need to be collected and separated from other material. The labels and any other debris are removed and the plastic is washed with water. The waste is automatically sorted using near-infrared scanning techniques and then manually checked again as incomplete sorting can lead to difficulties with the process. The chlorine in PVC can also be detected by means of X-rays allowing PVC bottles to be separated from polyethene and PET. Separating out PVC is important as HCl released from PVC causes major degradation of PET (but PVC bottles are not common these days). These can then be separated on the basis of their different densities. Recycled PVC is used for drains and sewer pipes and flooring. Clean PET may be recycled to make new bottles; less good material is used to make fibre filling for duvets and outdoor jackets. The separated plastics are then ground into flakes and then washed and dried and any further foreign substances such as metals are removed.

Some plastics cannot be recycled into new products. For example, the plastic cases of some cell phones contain brominated fire retardants and these plastics cannot be put through a recycling process – they are usually incinerated.

The issues associated with recycling plastics are summarized in Table 22.4.

Resin identification code (RIC)	Plastic and properties	Applications	Recycling
1 PET	Poly(ethylene terephthalate) (PET) is clear, tough and solvent resistant. Used for rigid sheets and fibres. Good barrier to gases and liquids. Resin can be spun into threads.	Carbonated soft drink bottles, food jars, carpet fibres, microwave trays, fruit juice bottles, pillow and sleeping bag filling, textile fibres.	Drinks bottles, clothes, detergent bottles, laminated sheets, carpet fibres, clear packaging films.
2 HDPE	High-density polyethene (HDPE) is hard to semi-flexible and opaque. Relatively impermeable to gas and moisture.	Freezer bags, milk bottles, bleach bottles, hard hats, buckets, 3D printing, milk crates, wire cable covering.	Pipes for farms, pallets, bins, extruded sheet, garden edging, household bags, oil containers.
3 PVC	Plasticized polyvinyl chloride (PVC) is flexible, clear and elastic; can be solvent welded. Unplasticized PVC is hard, rigid; may be clear and can be solvent welded.	Garden hoses, shoe soles, cable sheathing, car gaskets, shower curtains, gloves, pipes, blood bags and tubing, credit cards, watch straps.	Pipe and hoses fittings, garden hose, electrical conduit, shoes, road cones, drainage pipes, ducting, detergent bottles.
4 LDPE	Low-density polyethene (LDPE) is soft, flexible, surface is translucent, withstands solvent	Rubbish (garbage) bags, squeezy bottles, cling wrap, hot and cold drinks cups, flexible container lids, rubbish bins.	Concrete lining and bags.
5 PP	Polypropylene is hard, flexible and translucent (can be transparent) and has good chemical resistance	Film, carpet fibre, carts, bottles, caps, furniture, rigid packaging, yoghurt containers, takeaway containers, fishing nets, toilet seats.	Crates, boxes, plant pots, compost bins, garden edging.
6 PS	Polystyrene is clear, glassy, rigid, brittle, opaque, semi-tough and affected by fats and solvents. Expanded polystyrene is foamed, lightweight and thermally insulating.	Refrigerator bins, stationery accessories, coat hangers, medical disposables, trays, egg cartons, vending cups, plastic cutlery, yoghurt containers.	Industrial packaging, coat hangers, moulded products, office accessories, spools, rulers, video cases, printer cartridges.
7 OTHER	Other plastics includes all other resins, laminates, acrylonitrile butadiene styrene, acrylic, nylon, polyurethane and polycarbonates.	Automotive (car), aircraft, boating, furniture, electrical and medical parts.	Agricultural piping, furniture fittings, wheels and castors, fence posts, pallets and outdoor furniture, marine structures.

■ **Table 22.4** Plastic resin identification codes

Distinguishing possible resin identification codes (RICs) of plastics from an IR spectrum

Infrared spectroscopy (Chapter 11) can be used to sort plastics. Infrared radiation has a wavelength of 770 nm to 1000 µm, which corresponds to a wavenumber range of 12 900 to 10 cm^{-1}. For sorting plastics, infrared radiation with wavenumbers from 600 to 4000 cm^{-1} will be used.

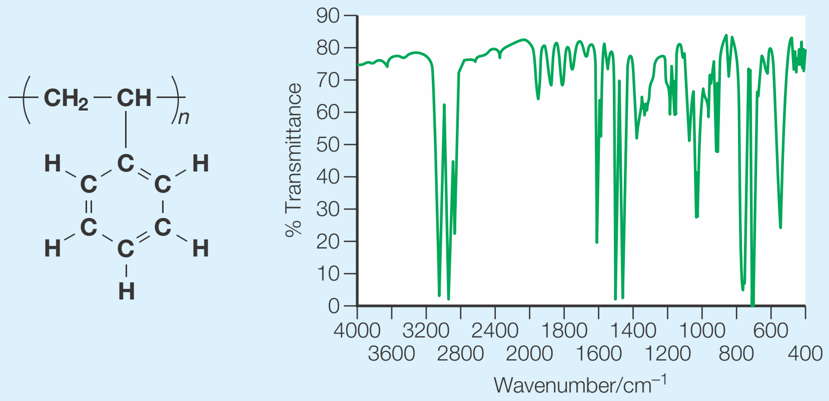
When infrared light encounters a material, some of it may be absorbed by the chemical bonds in the material. Each type of bond absorbs specific wavenumbers of infrared radiation. For examples of this, see Table 22.5.

Table 22.5 Identification of plastics from wavenumber absorptions

Functional group	Bond	Wavenumber/cm^{-1}
Alkane	C–H	2850–3000
$-CH_2-$	C–H	1465
$-CH_3$	C–H	1375 and 1400
Benzene ring	CH	3050–3150 and 700–1000
	C=C	1400–1600
Ester	C=O	1730–1750
	C–O	1000–1300
Chloride	C–Cl	600–800

It should be noted that plasticizers often have >C=O absorption at 1735 cm^{-1} or lower. In some cases, their presence does not alter the remainder of the spectrum, and the peak at 1735 cm^{-1} may simply be ignored when analyzing the spectrum. PVC is the most common example of a plastic that contains plasticizers.

29 Polystyrene is an addition polymer with the following structural formula and infrared spectrum.



Which peak(s) in the IR spectrum of polystyrene correspond to the benzene ring in polystyrene? What bond(s) cause peaks near wavenumber 3000 cm^{-1}?

Nature of Science

Risks and problems with plastics

Scientific research often proceeds with perceived benefits in mind, but the risks and implications also need to be considered. The development of biodegradable plastics illustrates these aspects of science.

One popular biodegradable plastic is polylactide, or PLA. It is made from a monomer (lactide) derived from fermentation of corn starch. PLA is a synthetic polyester that biodegrades within a year (in industrial composting conditions), decaying much faster than conventional crude oil-based plastics. However, it is non-biodegradable in the environment because its glass transition temperature (>50 °C) is too high and stops the hydrolysis. It will not, for example, degrade in home composting or if left lying around. It also prevents PET recycling if used as a bottle material.

Manufacturers use PLA because its method of synthesis is cheaper than the synthesis of other biodegradable plastics; because it is bioderived much of the economics depends on corn farming subsidy. In addition, PLA is biocompatible and can be used in biomedical applications, such as in medical plates and screws that can be degraded and absorbed by the body.

However, PLA is brittle, thermally unstable and hydrophobic. PLA degrades by hydrolysis with no need for external enzymes, but creates a large build-up of lactic acid during degradation, which can cause problems in biomedical applications due to the decrease in pH.

ToK Link

The products of science and technology can have a negative impact on the environment. Are scientists ethically responsible for the impact of their products?

Much of the investment in science in recent years has been motivated by wars, such as World War II and the Cold War. This use of human and financial resources is one factor that has helped focus the attention on ethics in science, and it is clear that since World War II, the interest in ethics in science has increased tremendously. Another factor is the discovery of abuses of power in scientific experimentation, such as the experimentation carried out by the Nazi doctors and various scandals involving disclosures of fraud, falsification of research data, and other forms of scientific misconduct.

The traditional position of the majority of scientists has been (very simply phrased) that they were seekers of objective truth who shall not be, to quote Robert Hooke (1663), 'meddling with Divinity, Metaphysics, Moralls, Politicks, Grammar, Rhetorick, or Logick'.

In modern terms, science should not as such deal with any subjective matters, notably those related to religion, politics, ethics or social responsibility. Seeking and finding scientific facts and proposing theories was supposed to be the scientists' task. The application of this scientific knowledge was not considered to be the scientists' responsibility, but the responsibility of politicians and other lawmakers and regulators.

30 Edward Teller, who was involved in the development of the atomic bomb, wrote in a letter to Leo Szilard: '...I have no hope of clearing my conscience. The things we are working on are so terrible that no amount of protesting or fiddling with politics will save our souls...'

Apply the different ethical theories from the ToK course, for example deontology and utilitarianism, and discuss whether the atomic bomb should have been developed and used.

22.8 Superconductivity and X-ray crystallography (AHL) *– superconductivity is zero electrical resistance and expulsion of magnetic fields. X-ray crystallography can be used to analyse structures*

Superconductivity

Our electronic devices depend on the use of electricity. However, when an electric current flows in a conductor, some of the electrical energy is transferred to heat. This is known as the heating effect of an electric current and is due to the resistance of the conductor.

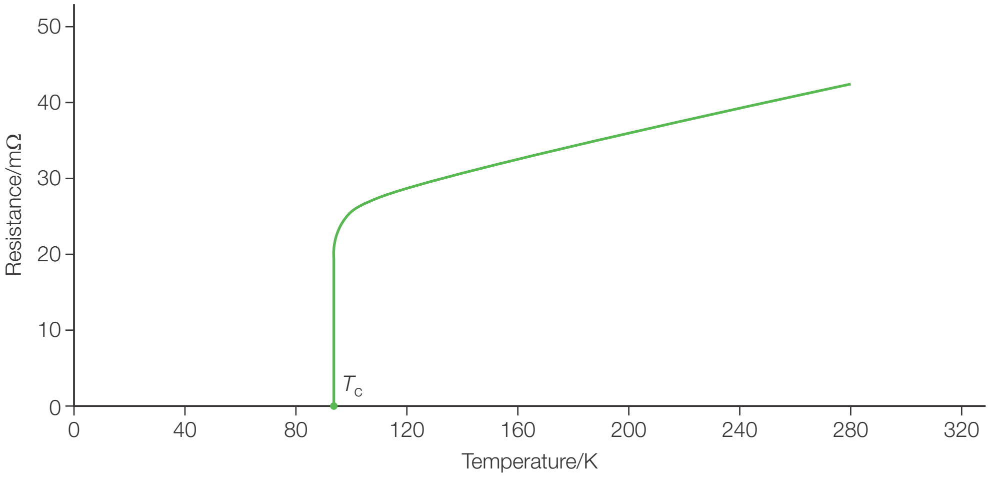
In 1911 the German physicist Heike Kamerlingh Onnes (1853–1936) discovered that mercury (when cooled in liquid helium) lost all its resistance at 4.2 K and transmitted current without any electrical energy being converted to heat. In 1913 lead was found to superconduct at 7 K, and in 1930 the metal niobium was found to superconduct at 9.3 K. A perfect superconductor exhibits perfect diamagnetic behaviour (see section 22.2).

An electric current is a flow of electrons from high potential to low potential. Electrons will flow through a conductor when a voltage (or electromotive force) from a battery or power supply is applied. When the electrons in the current collide with the metal ions (cations), they lose kinetic energy. This energy is converted to thermal energy (heat) and the temperature of the conductor increases. The conductor is said to have electrical resistance.

At higher temperatures the metal ions in the conductor vibrate even faster with greater amplitude. This increases the resistance to the electric current (electron flow) and hence resistance of a metal conductor increases with temperature. When a metal is cooled the metal ions vibrate less, offering an easier path, so the resistance decreases. Except at very low temperatures, this leads to a resistance which increases linearly with increasing temperature. The vibrations of the lattice can be viewed as waves called phonons, and the decrease in resistance at low temperatures is caused by a reduction of electron–phonon scattering.

At sufficiently low temperatures the resistance due to lattice collisions becomes smaller than the resistance due to collisions with impurities. This means that the total resistance is not linear with temperature at low temperatures, but rather tending towards a constant value. This is seen in the resistance curve against temperature for platinum in Figure 22.101.

Figure 22.95 A graph of resistance versus temperature for a superconducting material



In mercury and many other metals, for example beryllium, aluminium, zinc, gallium and some alloys, as temperature is decreased, at a certain **critical temperature (T_c)**, the resistance rapidly decreases to zero (Figure 22.95). Under these conditions an electric current could theoretically flow forever without conversion to heat. This phenomenon is known as superconductivity.

The main problem in using the superconductivity of metals is the relatively low value of the critical temperature, which is only a few degrees above absolute zero. To cool a metal to these low temperatures, liquid hydrogen or helium has historically been used. These liquids are expensive and considerable care is required with insulation to maintain the very low temperatures.

Cryocoolers are now used to reach very low temperatures without the need for liquid hydrogen or helium. Their use is increasing as the technology improves driven by the fact that demand for helium (partly driven by the growth of MRI (Chapter 21) is outstripping the global supply (helium is extracted as a by-product of oil and natural gas extraction).

High-temperature superconductors

A breakthrough in superconductivity research occurred in 1987 when so-called 'high-temperature' ceramic superconductors were discovered. Among the first to be studied was yttrium barium copper oxide, $YBa_2Cu_3O_7$-x, where x is a very small number. It is prepared by heating solid samples of $BaCO_3$, Y_2O_3 and CuO at 940 °C over a period of days followed by rapidly cooling (quenching) to room temperature. The substance forms a layered structure including planes containing copper and oxygen and is slightly oxygen deficient compared with the perfect crystal structure, $YBa_2Cu_3O_7$. This structure allows superconductivity to occur on cooling to 77 K. This is achieved by the use of liquid nitrogen which is a relatively cheap and available substance, used commonly in industry and hospitals.

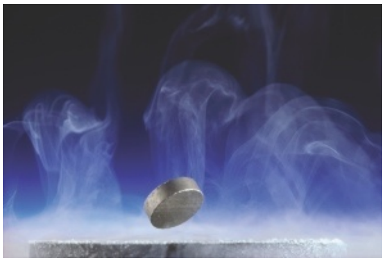
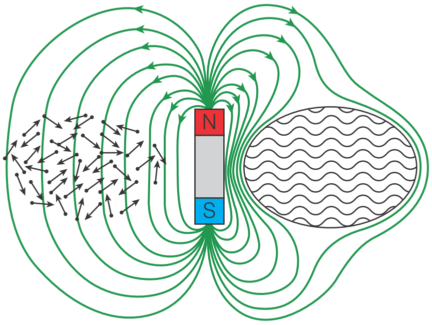


Figure 22.96 The Meissner effect: the floating magnet is only in contact with air

Meissner effect

A superconductor can be readily identified by the Meissner effect (Figure 22.96). When cooled below its critical temperature, a superconductor will create a mirror image of an external magnetic field, so that the total magnetic field inside the superconductor is zero. Electrical currents flow in the superconductor to generate a magnetic field that is equal and opposite to the applied magnetic field so that the sum of the applied and induced magnetic fields in the superconductor is zero.

When a magnet is brought near the surface of a superconductor, electrical currents flow in the superconductor close to its surface and, as in an electromagnet, this leads to a generated magnetic field. The superconductor responds to the applied magnetic field by creating a magnetic field that is the exact mirror image of the magnet's field.



■ **Figure 22.97** An ordinary conductor (left) shows random electron movement and allows magnetic field (flux) penetration, but a superconductor (right) excludes any magnetic field (flux) penetration, creating a mirror image of any nearby magnetic field

The superconductor behaves as an identical copy of the magnet with like poles facing each other. When the magnet is removed from the superconductor the magnetic field disappears (Figure 22.97).

A complete and detailed explanation of superconductivity requires quantum theory, but a simplified description is based on the ideas that electrons flow through superconductors in pairs (called Cooper pairs). The Cooper pairs are believed to move through the lattice in a concerted way with the lattice vibrations, resulting in no electron scattering and zero electrical resistance.

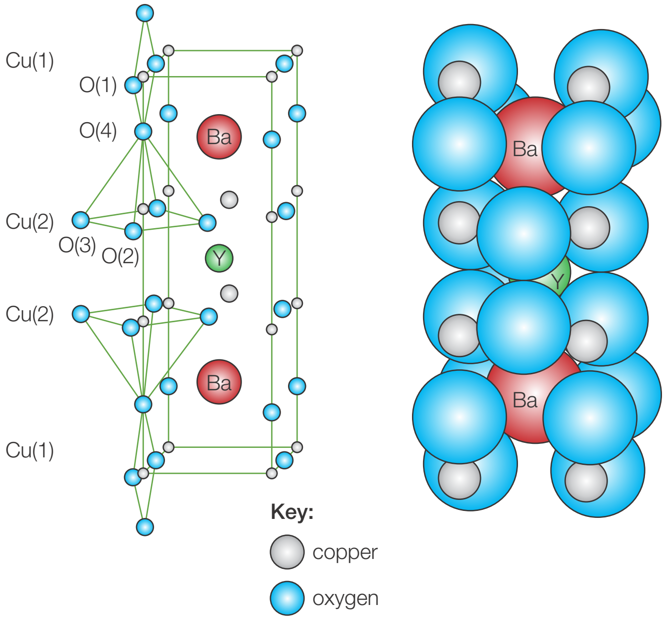
Bardeen–Cooper–Schrieffer (BCS) theory

The established theory explaining superconductivity is known as BCS theory and is named after the American physicists John Bardeen, Leon Cooper and John Schrieffer who were awarded the 1972 Nobel Prize. The BCS theory and competing theories explaining superconductivity are an active area of research. None of the theories accurately predicts which compounds might be high-temperature superconductors.

Type I superconductors are all pure metals and show zero electrical resistance below a critical temperature, zero internal magnetic field (Meissner effect) and a critical magnetic field above which superconductivity suddenly stops. The superconducting state cannot exist in the presence of a magnetic field greater than a critical value, even at temperatures near absolute zero. Type I superconductors are often termed ‘soft’ superconductors because they lose their superconductivity at relatively low temperatures and in the presence of weak magnetic fields.

The superconductivity in type I superconductors is modelled and described well by the BCS theory, which relies upon electron pairs coupled by lattice vibration interactions. At low temperatures the positive ions in the lattice are distorted slightly by a passing electron. A second electron is attracted to this slight positive deformation and a coupling of these two electrons occurs.

Superconductors made from alloys or compounds are termed type II superconductors. They are harder than type I superconductors and have a much more gradual transition to superconductivity, but more importantly they have much higher critical magnetic fields. They can be exposed to a much stronger magnetic field than a type I superconductor before they stop being superconducting and show a gradual transition with a mixture of normal and superconducting properties.



■ **Figure 22.98** Idealized structure of $YBa_2Cu_3O_7$

Structure of superconductors

The structures of superconductors have been studied using the technique of X-ray crystallography. The structure of $YBa_2Cu_3O_7$ is derived from the perovskite structure. Stacking three perovskite unit cells (of stoichiometry ABO_3) directly on top of each other produces a material with the stoichiometry $A_3B_3O_9$ (3 × ABO_3). If the A type cations are replaced by two barium ions (Ba^{2+}) and one yttrium ion (Y^{3+}) in the sequence Ba-Y-Ba-Ba-Y-Ba-Ba in the tripled perovskite and the B cations are copper(II) ions (Cu) then the compound stoichiometry is $YBa_2Cu_3O_9$. If two oxide ions are removed then a ceramic superconductor with the formula $YBa_2Cu_3O_7$ is generated.

If the oxidation state is calculated for copper in $YBa_2Cu_3O_7$ assuming normal oxidation states for the other elements (Y +3; Ba +2 and O −2), then the average oxidation state is +7/3. This calculation and chemical analysis suggests the presence of one copper(III) ion and two copper(II) ions. The Cu^{3+} ions are in the Cu(1) plan of the superconductor and the Cu^{2+} ions are in the square pyramidal Cu(2) structure (Figure 22.98).

31 Deduce the average oxidation state of copper in $YBa_2Cu_3O_6$.

In 1995 it was discovered that when carbon-60 (Chapter 4) reacts with an alkali metal, it becomes an electrically conducting material, K_3C_{60} [$3K^+$ $C_{60}{}^{3-}$], and becomes superconducting at about 40 K. Even more recently, the simple binary compound magnesium diboride, MgB_2, was found to become superconducting at 39 K. This is an important discovery since MgB_2 is a semiconductor and a relatively cheap material. Related compounds may have even higher superconducting transition temperatures.

There is considerable scope for searching for superconductivity in new materials, and imaginative chemical synthesis is an important activity in the field. Furthermore, there is no theoretical upper limit on the operating temperature for superconductivity, so there is hope that one day superconductors that work at even higher temperatures will be found, perhaps even room temperature.

32 Find out about the synthesis, properties (including transition temperature) and potential uses of K_3C_{60}.

Nature of Science

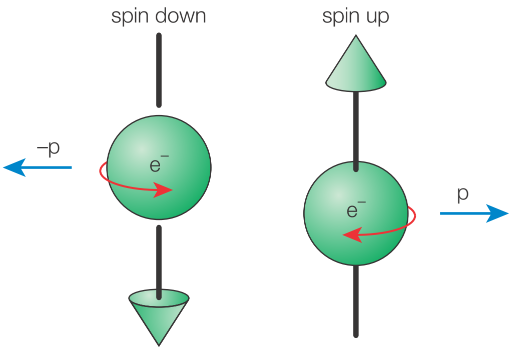
Theories of superconductivity

Superconductivity is usually described in terms of the BSC theory postulated in 1957. The key feature of this model are Cooper pairs, which consist of two electrons of opposite spin (obeying the Pauli exclusion principle) and momentum which are brought together by a cooperative effect also involving the ions (nuclei and bound electrons) in the vibrating crystal lattice.

The cooperative nature of superconductivity is a key part of its fundamental nature; that is, superconductivity is about a large number of electrons forming Cooper pairs – if you break a few, the remaining pairs are also less strongly bound.

Cooper pairs of electrons (Figure 22.99) remain bound as spin pairs below the critical temperature, forming a single quantum state, and their presence gives rise to resistance-free conductivity.

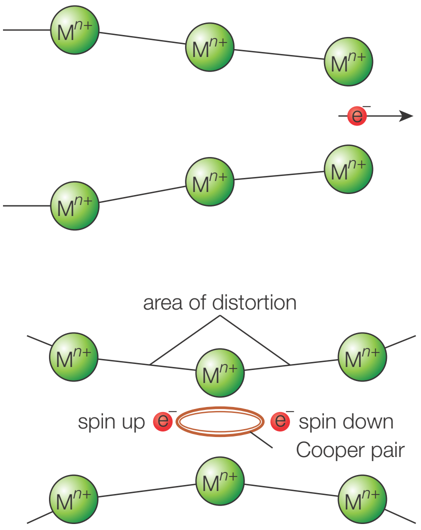
The model holds for the earliest known superconductors, such as solid mercury. Although Cooper pairs may still be significant for the high-temperature cuprate superconductors, new theories are required. Currently, no complete explanation for the conducting properties of high-temperature superconductors has been formulated.



■ **Figure 22.99** A Cooper pair: the arrows represent direction of movement and p represents momentum (*mv*); spin down and spin up can be interpreted as clockwise and anticlockwise directions of spin

Figure 22.100 shows a simple explanation of the formation of Cooper pairs in a BCS superconductor (either type I or type II) such as niobium (body-centred cubic structure). As a negatively charged electron passes between the niobium cations in the lattice, the cations are attracted inwards and this distortion creates a region of higher positive charge which attracts another electron to this position in the lattice. The two electrons then pair up to form a Cooper pair and travel through the lattice. In effect the atoms of the niobium lattice oscillate, creating temporary positive and negative regions which push and pull the Cooper pair of electrons along.

Imagine that some energy is gained by the lattice from the first electron – this then moves (propagates) through the lattice in the form of a wave called a phonon. If the phonon, which is a quantized object, is absorbed by a second electron, the second electron gains all of its energy. The electrons are paired into Cooper pairs if the overall energy state is lower as a result of this process. The Cooper pairs are in a single quantum state and are then able to move through the lattice without colliding with the ions. Since there is no loss of energy there is no electrical resistance.



■ **Figure 22.100** Formation of Cooper pairs in a BCS superconductor

However, if the substance is not at a sufficiently low temperature, lattice vibrations are strong enough that they are able to break up the Cooper pairs and so superconductivity can only occur below a critical temperature. The problem at higher temperatures is not a reduction of the strength of the pairing interaction but an increase in the thermal fluctuations that break up the Cooper pairs.

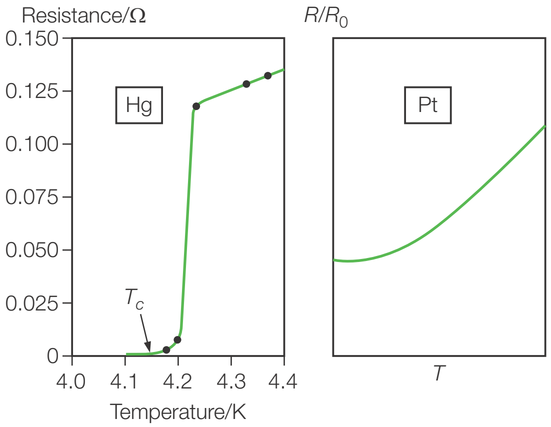
Magnetic fields play an important role in the field of superconductivity. Electricity and magnetism are closely related and electromagnetism is very relevant to how superconductors work: moving electrons induce a magnetic field (motor effect) and a changing magnetic field around a conductor induces a current (generator effect). Many of the modern instruments that are used in superconductor research involve both magnetic fields (generated or measured) and the measurement of magnetic properties of materials. All atomic magnetic effects arise from the orbital motion of the electron and form a property of the electron known as spin. A second form of spin is intrinsic to the electron and, although in reality a purely quantum effect, is often visualized as an electron spinning around its axis.

Analysis of resistance versus temperature data for type I and type II superconductors

If an external magnetic field is applied to a superconductor and steadily increased, a point is reached when the magnetic field (magnetic flux) penetrates the substance and the superconducting properties are lost. If the transition is sharp, the superconductor is classified as a type I superconductor, and the field strength at which the transition occurs is termed the critical magnetic field. A type II superconductor undergoes a gradual transition from superconductor to normal conductor.

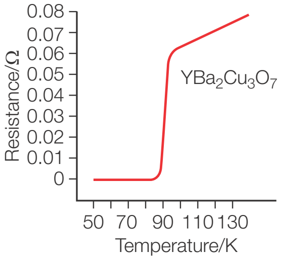
Figure 22.101 shows the plots of resistance versus temperature for mercury (a type I superconductor) and platinum (a normal conductor). Note that the resistance of mercury follows the path of a normal metal above the critical temperature, T_c, and then suddenly drops to zero at the critical temperature, which is 4.15 K for mercury. In contrast, the data for platinum shows a finite resistance R_0 even at very low temperatures.

Figure 22.101 Plots of resistance versus temperature for mercury and platinum



The resistance of type II superconductors shows similar behaviour to type I superconductors with a sudden rapid decrease in resistance at a specific temperature. Figure 22.102 shows a plot of resistance against temperature for $YBa_2Cu_3O_7$, a ceramic-based type II superconductor. The graph is similar to that of a type I superconductor except it has a much higher critical temperature.

Figure 22.102 Plot of resistance versus temperature for $YBa_2Cu_3O_7$



33 The data in Table 22.6 was obtained from a bulk sample of $YBa_2Cu_3O_7$. Calculate the resistance for each trial given that a constant current of 100 mA was flowing through the sample. Use Ohm's law, $V = I \times R$, where V is the voltage (V), I is the current (A) and R is the resistance in ohms (Ω).

Voltage/V	Absolute temperature/K	Resistance/Ω
0.0010370	118.2	
0.0010270	116.1	
0.0010600	114.8	
0.0010490	112.9	
0.0010350	110.9	
0.0010220	109.1	
0.0010090	106.9	
0.0010010	105.0	
0.0009890	103.5	
0.0009750	102.2	
0.0009670	100.0	
0.0009510	97.9	
0.0009440	95.8	
0.0009180	95.0	
0.0009110	94.3	
0.0008920	93.8	
0.0008440	93.5	
0.0007830	93.2	
0.0006390	93.0	
0.0005050	92.6	
0.0003790	92.3	
0.0002430	92.1	
0.0000930	91.7	
0.0000100	91.4	
0.0000030	91.0	

■ **Table 22.6** Data from $YBa_2Cu_3O_7$

Plot a graph using a spreadsheet of resistance against temperature and estimate the critical temperature, T_C.

Applications of superconductors

The first large-scale commercial application of superconductivity was in magnetic resonance imaging (MRI) (Chapters 21). This is a non-intrusive medical imaging technique that requires a person to be placed inside a large and uniform electromagnet with a high magnetic field. Although normal electromagnets can be used for this purpose, because of resistance they would dissipate a great deal of heat and have large power requirements.

Superconducting magnets, on the other hand, have almost no power requirements apart from operating the cooling. Once electrical current flows in the superconducting wire the power supply can be switched off because the wires can be formed into a loop and the current will persist indefinitely as long as the temperature is kept below the transition temperature of the superconductor.

Superconductors can also be used to make a device known as a superconducting quantum interference device (SQUID). This is incredibly sensitive to small magnetic fields so that it can detect the magnetic fields from the heart (10^{-10} Tesla) and the brain (10^{-13} Tesla). For comparison, the Earth's magnetic field is about 10^{-4} Tesla. As a result, SQUIDs are used in non-intrusive medical diagnostics on the brain and in other applications requiring detection of very small magnetic field variations, such as geophysical surveys. SQUIDs are also used in submarines for detecting undersea mines.

Levitating trains have been built that use powerful electromagnets made from superconductors. The superconducting electromagnets are mounted on the train. Normal electromagnets, on a guideway beneath the train, repel (or attract) the superconducting electromagnets to levitate the train while pulling it forwards.

Large and powerful superconducting magnets are also needed in particle accelerators and nuclear fusion reactors (Chapter 24). Superconductors are ideal transmitters of electrical energy, but a major challenge is that type II superconductors which operate at higher temperatures are ceramic in nature, and hence are not at all straightforward, due to their brittleness, to make into wires and electrical components. Type II superconductors can be made into tapes for carrying current.

X-ray crystallography

The structure of crystalline solids (metals, ionic solids and molecular solids) is determined by X-ray diffraction (Chapter 21). Type II superconductor cuprates have complex arrangements of ions in their crystal lattices and knowledge of their structural features can help explain their superconducting behaviour.

Lawrence Bragg proposed in 1912 that the formation of diffraction patterns could be explained by assuming that the X-rays were reflected from the various planes of particles in a crystal.

Figure 22.103 shows that the difference in the distance travelled by X-rays reflecting from two different planes is $2d \sin \theta$. For constructive interference of X-rays:

$n\lambda = 2d \sin \theta$ Bragg's equation (given in section 1 of the *IB Chemistry data booklet*)

d = interplanar distances, λ = wavelength of the X-rays, n = 1,2,3… positive integer, reflects the order of reflection (typically only first order reflections are considered) and θ is the angle of reflection (which increases with increasing order of reflection).

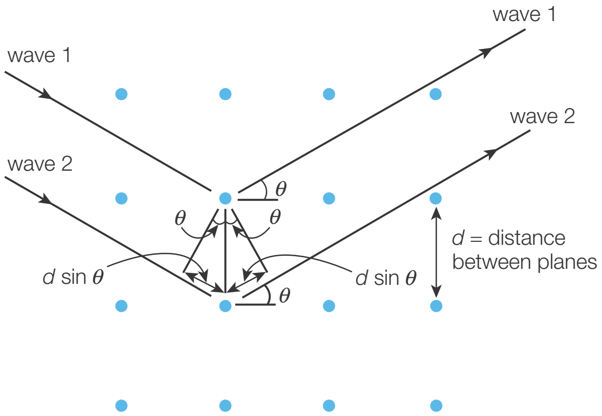


Figure 22.103 The extra distance travelled by X-ray wave 2 is 2 × $2d \sin \theta$ and for constructive interference $n\lambda = 2d \sin \theta$

34 Calculate the frequency (in Hz) of X-rays with a wavelength of 1 nm.

Worked example

X-rays from a copper X-ray tube (λ = 154 pm) were diffracted at an angle of 14.22° by a crystal of silicon. Assuming first-order diffraction (n = 1), deduce the interplanar spacing in silicon (in pm).

$d = n\lambda/2\sin\theta = (1 \times 154\,\text{pm})/[2 \times \sin(14.22°)] = 313\,\text{pm}$

Crystal lattices and unit cells

A crystal can be imagined to be generated from the repeating of some basic unit of patterns of atoms, ions or molecules. A lattice is a regular three-dimensional arrangement of identical points in space. Each point in a crystal lattice represents one particle (atom, ion or molecule) and the points are joined by straight lines to show the geometry of the lattice.

For every crystal lattice it is possible to select a group of lattice points which are repeated many times in the crystal lattice. This part of the crystal lattice is termed the unit cell. The entire lattice can be generated by the stacking of these unit cells. A unit cell may be defined as a three-dimensional group of lattice points that generates the whole lattice by stacking. Figure 22.104 shows the relationship between a crystal lattice and its unit cell. The number of nearest neighbours for a particle in a lattice is its coordination number.

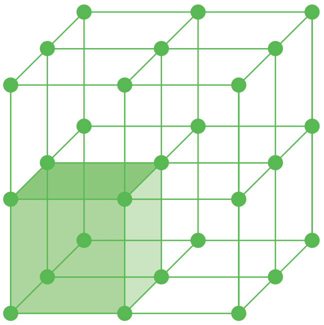


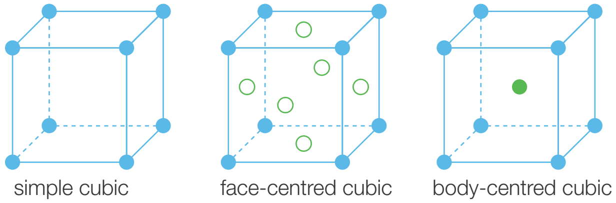
Figure 22.104 A crystal lattice with the shaded portion representing the unit cell

Types of unit cells

There are three cubic unit cells (Figure 22.105) which are present in many crystalline substances:

- simple or primitive cubic unit cell in which the atoms are present only at the corners.
- face-centered cubic in which the atoms are present at the centre of each face in addition to atoms at the corners
- body-centered cubic in which atoms are present at the centre of the cube in addition to atoms at the corners.

Figure 22.105 Cubic unit cells

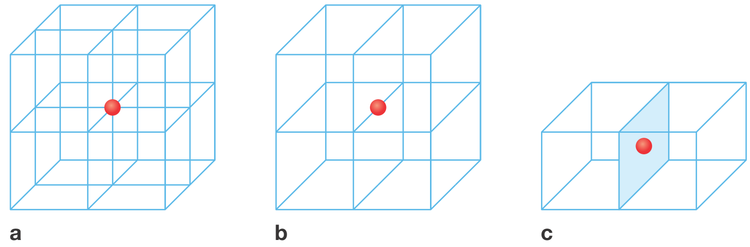


Calculation of the number of particles in a unit cell

In order to calculate the number of particles in a unit cell, the following rules (Figure 22.106) must be obeyed:

1. Each particle at the corner of a unit cell is shared by eight unit cells in the lattice and hence contributes only 1/8 to a particular unit cell (Figure 22.106a).
2. A particle at the edge centre is shared by four unit cells in the lattice and hence contributes only 1/4 to a particular unit cell (Figure 22.106b).
3. A particle at the centre of the face of a unit cell is shared by two unit cells in the lattice and contributes only 1/2 to a particular unit cell (Figure 106c).
4. A particle at the body centre of a unit cell belongs only to the particular unit cell = 1 particle.

Figure 22.106 Calculating number of particles in a unit cell



Simple cubic unit cell

In a simple cubic unit cell, there are eight atoms at the corner and each atom makes 1/8 contribution to the unit cell (Figure 22.107a). Hence, a simple cubic unit cell has

$$8 \text{ (at corners)} \times \frac{1}{8} = 1 \text{ atom}$$

Body-centered cubic unit cell

A body centered cubic unit cell has eight atoms at corners and one at the centre (Figure 22.107b). Each corner makes 1/8 contribution and the atom at the centre belongs only to the particular unit cell. Hence, a body centred cubic unit cell has

$$8 \text{ (at corners)} \times \frac{1}{8} + 1\text{(at centre of unit cell)} \times 1 = 2 \text{ atoms}$$

Face-centered cubic unit cell

A face-centered cubic unit cell has one atom at each corner (there are eight corners of a cube) and one atom at each face centre (a cube has six faces) (Figure 22.107c). An atom at the face centre is being shared by two unit cells and makes a contribution of only 1/2 to a particular unit cell. Hence, a face-centred cubic unit cell has

$$8 \text{ (at corners)} \times \frac{1}{8} + 6 \text{ (at face centres)} \times \frac{1}{2} = 4 \text{ atoms}$$

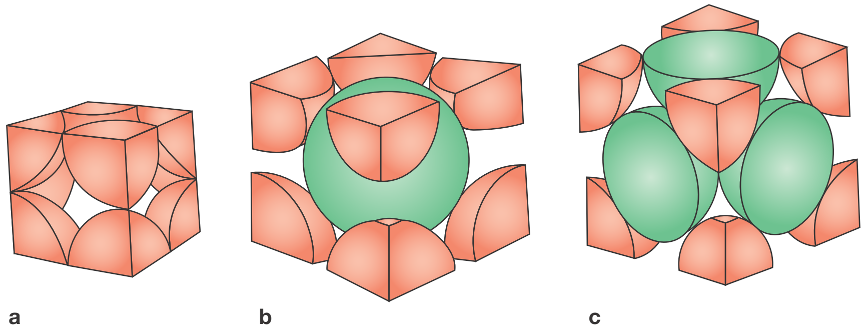


Figure 22.107 Space filling models of simple cubic, body-centred and face-centred cubic crystal structures

Table 22.7 summarizes the simple cubic, body-centered and face-centred cubic crystal structures.

Table 22.7 Summary of simple cubic, body-centered and face-centered cubic crystal structures for metals

Unit cell	Number of particles involved	Number of particles per unit cell	Coordination number	Percentage of cell volume occupied by atoms
Simple cubic	8;1 at each corner	1	6	52%
Body-centered cubic	9;1 at each corner and a central atom	2	8	68%
Face-centered cubic	14;1 at each corner and 1 on each face	4	12	74%

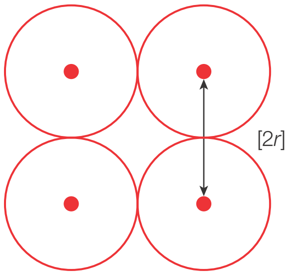
The packing efficiencies can be calculated for the various crystal structures using simple geometry and assuming the particles are perfect spheres. The remaining spaces are known as interstitial spaces. These can be occupied by small atoms in some compounds (interstitial compounds).

Worked example

A solid has a cubic structure in which atoms of X are located at the corners of the cube, Y atoms are located at the cube centres and oxygen atoms are at the edge centres. Deduce the formula of the compound.

There are eight corners in a cube and each corner has an atom of X present located there. Each atom of X at the corner makes 1/8 contribution towards a particular unit cell.

The number of atoms of X per unit cell = $8 \times \frac{1}{8} = 1$. The Y atom present at the centre belongs to a particular unit cell. The number of atoms of Y per unit cell = 1 × 1. There are 12 edges of a cube and each edge has an atom of oxygen present at the edge centre. Each atom of oxygen present at the edge centre makes 1/8 contribution towards a particular unit cell. The number of oxygen atoms per unit cell = $12 \times \frac{1}{4} = 3$. The formula of the compound is therefore XYO_3.



■ **Figure 22.108** Simple cubic unit cell

Calculations based on X-ray crystallography

The radius of an atom can be determined from the packing structure (lattice) and the distance between atoms from X-ray diffraction analysis of the solid. A simple cubic unit cell has particles touching each other as shown in Figure 22.108. The length of a side of the unit cell cube, d, is therefore equal to the diameter of the atom, hence $r = d/2$.

For a face-centered cubic structure the atoms touch along a diagonal, but not along the edge (Figure 22.109). The diagonal represents four atomic radii, so using Pythagoras's theorem:

$$(4r)^2 = d^2 + d^2 \text{ or } (4r)^2 = 2d^2$$

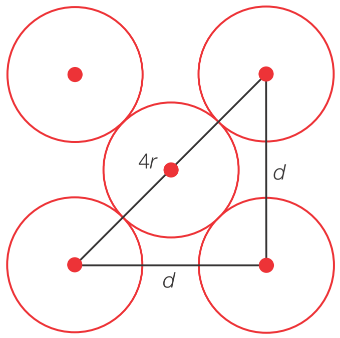
$$\text{so } r = \sqrt{2\mathrm{d}/4}$$

For a body-centered cubic structure (Figure 22.110) the atoms touch along the diagonal of the cube of the unit cell and applying Pythagoras's theorem:

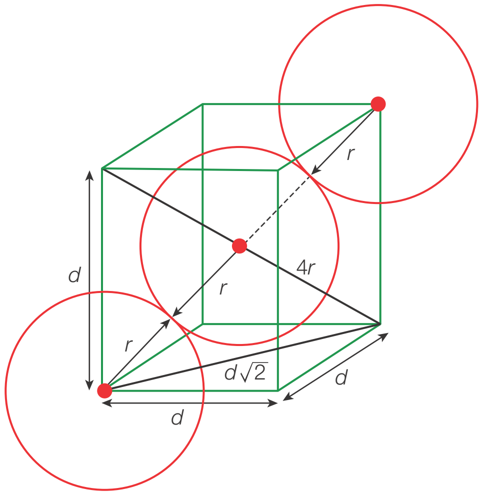
$$(\text{diagonal of cube})^2 = \text{length}^2 + \text{width}^2 + \text{height}^2$$

The unit cell cube diagonal is $4r$ (the diameter of the central atom and the atomic radii of each corner atom), so since length = width = height = d:

$$(4r)^2 = 3d^2;\ r = \sqrt{3d/4}$$



■ **Figure 22.109** Face-centered cubic unit cell



■ **Figure 22.110** Body-centered cubic unit cell

Table 22.8 summarizes the atomic radius in terms of the length of the unit cell for different metallic crystal structures. Similar approaches can be taken to estimate the size of ions assuming that the cations and anions are spherical and touch.

■ **Table 22.8** Atomic radii in terms of the length of the unit cell for three common metallic lattices (crystal structures)

Unit cell	**Atomic radius/r**
Simple cubic	$d/2$
Body-centered cubic	$\sqrt{3d/4}$
Face-centered cubic	$\sqrt{2d/4}$

Worked example

Nickel crystallizes in a face-centered cubic unit cell with a cell-edge length of 0.3524 nm. Calculate the radius (in nm) of a nickel atom.

$r = \sqrt{2d/4}$; $r = \sqrt{2} \times (0.3524\ \text{nm}/4) = 0.1246\ \text{nm}$

35 At 20 °C iron adopts the body-centered cubic structure with atoms of atomic radius 0.124 nm. Calculate the length of the cube edge of the iron unit cell.

The edge length of a unit cell can be determined by X-ray diffraction and the number of atoms per unit cell can be deduced from knowledge of the type of lattice. Knowing the edge length, the number of atoms per unit cell and molar mass of the element, the density of the unit cell can be calculated as follows:

- Edge of unit cell = a pm = $a \times 10^{-10}$ cm
- Volume of unit cell = $a^3 \times 10^{-30}$ cm^3
- Mass of unit cell = number of atoms in the unit cell × mass of each atom = $z \times m$
- Mass of each atom = molar mass/Avogadro constant = M/N_A
- Density of the unit cell = mass of unit cell/volume of the unit cell
- Density of the unit cell = $(z \times \mathrm{M})/(a^3 \times \mathrm{N_A} \times 10^{-30})$ g cm^{-3}. If a is expressed in ppm, then the density will be expressed in g cm^{-3}. The density of the unit cell is the same as the density of the substance.

Worked example

Niobium crystallizes in a body-centered cubic structure. The density of niobium is 8.55 g cm^{-3} and its molar mass is 92.91 g mol^{-1}. Calculate the atomic radius in ppm.

In a body-centred cubic structure, the number of atoms per unit cell is 2.

Density = $(2 \times 92.91)/(6.02 \times 10^{23} \times 10^{-30} \times 8.55) = 3.61 \times 10^{7} = 330.5$ ppm

But in a body-centered cubic structure the diagonal is equal to four times the radius of the atom:

$4r = \sqrt{3}a = \sqrt{3} \times 330.5$; $r = 143$ ppm

Structure of simple ionic compounds

Structure of sodium chloride

Sodium chloride consists of sodium ions, Na$^+$, and chloride ions, Cl$^-$. The chloride ions are larger than the sodium ions. The sodium ions are located at the corners and faces of a cube with the chloride ions on the edge and in the middle of the cube. The lattice, known as the rock salt structure, can be regarded as being composed of two face-centred cubic structures (one for each type of ion) that overlap. There are six chloride ions around every sodium ion (and vice versa). Hence, the coordination number of each ion is 6. The rock salt structure is described as showing 6 : 6 coordination. The ratio of 6 : 6 of sodium to chloride ions simplifies to 1:1, which agrees with the formula, NaCl.

The unit cell of sodium chloride has 4 sodium ions and 4 chloride ions as calculated below:

- Number of sodium ions = 12 (at edge centres) × $\frac{1}{4}$ + 1 (at body centre) × 1 = 4
- Number of chloride ions = 8 (at corners) × $\frac{1}{8}$ + 6 (at face centres) × $\frac{1}{2}$ = 4

Hence the number of NaCl formula units per unit cell is 4. Most of the halides of alkali metals (group 1) and oxides of alkaline earth metals (group 2) metals have the rock salt structure.

Caesium chloride structure

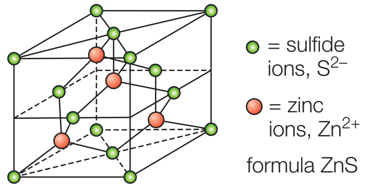
The caesium chloride unit cell consists of eight caesium ions in a simple cubic arrangement, with a single chloride ion at the centre of the cube. The caesium chloride structure can be regarded as being formed from two interlocking cubic arrangements of caesium and chloride ions.

This structure has an 8 : 8 coordination: each caesium ion is touching eight chloride ions and each chloride ion is touching eight caesium ions.

The unit cell of caesium chloride has one caesium ion and one chloride ion as calculated below:

- Number of chloride ions = 8 (at corners) × $\frac{1}{8}$ = 1
- Number of caesium ions = 1 (at the body centre) × 1 = 1

Hence, the number of CsCl formula units per unit cell is 1.



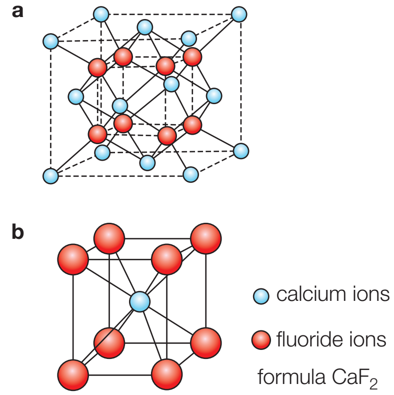
■ **Figure 22.111** The zinc blende structure

Zinc blende

Zinc sulfide can occur in two different crystalline forms: zinc blende and wurtzite. The unit cell of the zinc blende structure consists of sulfide ions in a face-centred cubic arrangement (Figure 22.111). Each cell contains four zinc ions arranged at the corners of a tetrahedron. The tetrahedron sits completely inside the cube. The coordination number of the zinc ions is the same as the sulfide ion.

Fluorite structure

This mineral is calcium fluoride, CaF_2, and has a structure based on two cubic arrangements. In Figure 22.112 calcium ions are shown in a face-centred cubic arrangement and inside this cube is a second cube of fluoride ions. Each calcium ion has eight fluoride ions for its nearest neighbours. However, each fluoride ions exists within a tetrahedral arrangement of calcium ions; so the fluoride ions have four calcium ions around them. This ratio (8 : 4) of the coordination number for calcium to fluoride ions preserves the ratio in formula CaF_2.



■ **Figure 22.112** The fluorite structure

Perovskite structure

Pervoskite is a mineral with the formula $CaTiO_3$. The calcium ions, Ca^{2+}, lie at the corners of a cube. The oxide ions, O^{2-}, lie on the faces of the cube, and the titanium(IV) ion, Ti^{4+}, lies at the centre of the cube. High-temperature type II cuprate superconductors have structures based on perovskite.

36 Draw the unit cell of the pervoskite structure based on the description given above.

ToK Link

X-ray diffraction has allowed us to probe the world beyond the biological limits of our senses. How reliable is our knowledge of the microscopic world compared to what we know at the macroscopic level?

X-ray diffraction gives researchers the detailed structures of molecules (with the exception of hydrogen atoms due to their low electron density) in the crystalline state. To 'look at' any system we need to use electromagnetic radiation that is of a wavelength comparable to the dimensions of the molecule. Hence X-rays which have wavelengths comparable to the bond lengths in molecules are ideal for studying their structures. Because X-rays cannot be focused easily (like light), the image has to be reconstructed computationally in the form of an electron density map. Electrons can also be used, as in electron microscopy, but they give a much less precise picture of molecules. Electrons can be focused using a magnetic lens so we can get a direct image of the sample. They are ideal for subcellular structures, for example the nucleus, and macromolecular assemblies, for example ribosomes.

Our empirical knowledge of the microscopic world is less reliable than at the macroscopic level. Experimental measurements have greater percentage errors as the measurements become smaller. The microscopic world of atoms, molecules and electrons is governed by quantum mechanics including the uncertainty principle. There is no definite end to the electron density of atoms and molecules – their edges are 'fuzzy' and imprecise.

22.9 Condensation polymers (AHL) *– condensation polymers are formed by the loss of small molecules as functional groups from monomers join*

Condensation polymers are formed by a reaction that covalently bonds monomers together and produces small molecules as a condensation product. Hydrogen chloride and water molecules are common condensation products. (Ammonia molecules may also be formed during the formation of an organo-silicon polymer.) A condensation reaction involves an addition reaction immediately followed by the elimination of a small molecule.

Condensation polymerization involves the reaction of two types of monomers both with reactive functional groups at the ends of the molecule. Nylon, Kevlar, polyurethanes and phenol-methanal resins are all examples of artificial condensation polymers. Proteins, starch, DNA and cellulose are all natural condensation polymers.

Condensation polymers form more slowly than addition polymers, often requiring heat, and they are generally lower in molar mass. The terminal functional groups on a chain remain active, so that groups of shorter chains combine into longer chains in the late stages of polymerization. The presence of polar functional groups on the chains often enhances chain-chain attractions, particularly if these involve hydrogen bonding.

Nature of Science

The plastic age

If an era is known by the kinds of materials its people use to build the world in which they live, then the Stone Age, the Bronze Age and the Iron Age have given way to our own Plastic Age or Age of Polymers. The first commercially successful truly synthetic polymer was made in 1907 by the Belgian-American chemist, Leo Baekeland. He mixed phenol and methanal and from the reaction mixture obtained a resinous material which he called Bakelite. During this period no technological advancement, other than the delivery of electrical power to our homes, has impacted our lives more than the widespread use of synthetic plastics. Indeed safe electrical wiring is made possible with the use of plastics.

Types of condensation polymers

Polyesters

The product formed by the condensation reaction between an alcohol and a carboxylic acid is called an ester (Chapter 10). If two bifunctional molecules, one containing alcohol groups and the other carboxylic acid groups, react, it should be possible to form a long-chain molecule, a **polyester**. Such a molecule would be a **condensation polymer** (or a **step polymer** in modern terminology).

For instance, if a dihydric alcohol or diol such as ethane-1,2-diol is reacted with a dicarboxylic acid such as benzene-1,4-dicarboxylic acid then a polyester known as polyethylene terephthalate (PET) is produced (Figure 22.113). In its fibre form this polyester is known as Terylene in the UK and as Dacron in the USA. However, it can be produced as a packaging film (Mylar, Melinex) or in a form suitable for making bottles, when it is referred to as PET (poly(ethylene terephthalate)).

The different Terylene chains are not particularly strongly attracted together as there is no capacity for hydrogen bonding between the chains. There are London (dispersion) forces

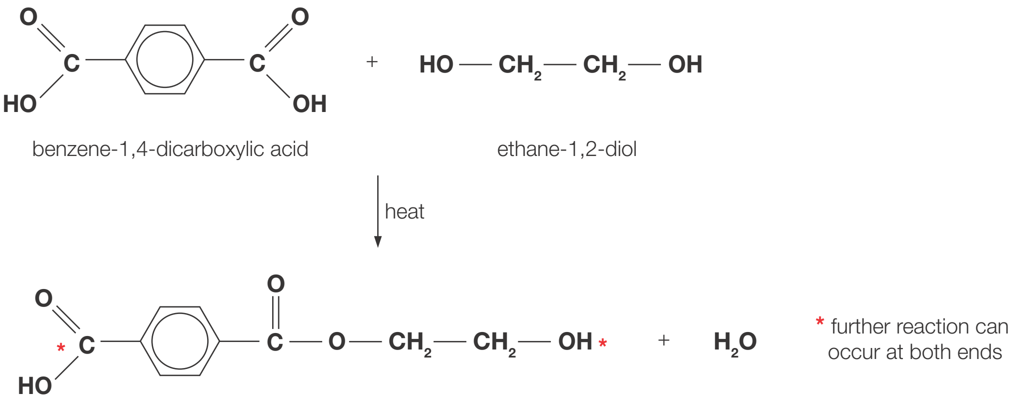

benzene-1,4-dicarboxylic acid
+
HO — CH₂ — CH₂ — OH
ethane-1,2-diol
heat
+
H₂O
* further reaction can occur at both ends

Figure 22.113 The formation of the condensation polymer PET (Terylene)

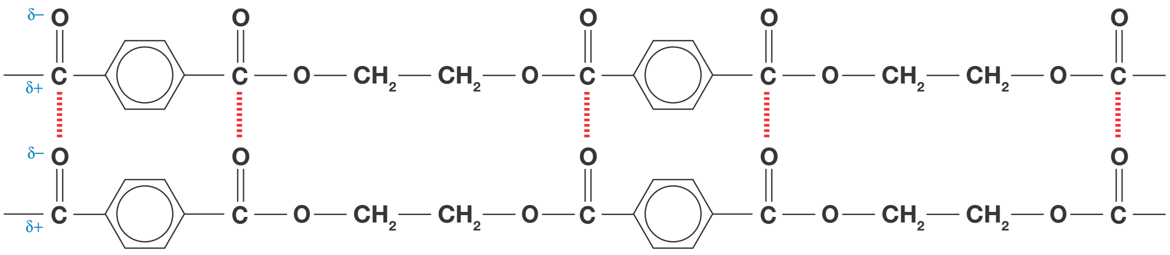


Figure 22.114 The dipole–dipole interactions between adjacent chains in polyethylene terephthalate (Terylene)

■ **Figure 22.115** Terylene (Dacron) finds considerable use in making clothing – shirts, for example. It is usually used as cotton/polyester mixtures

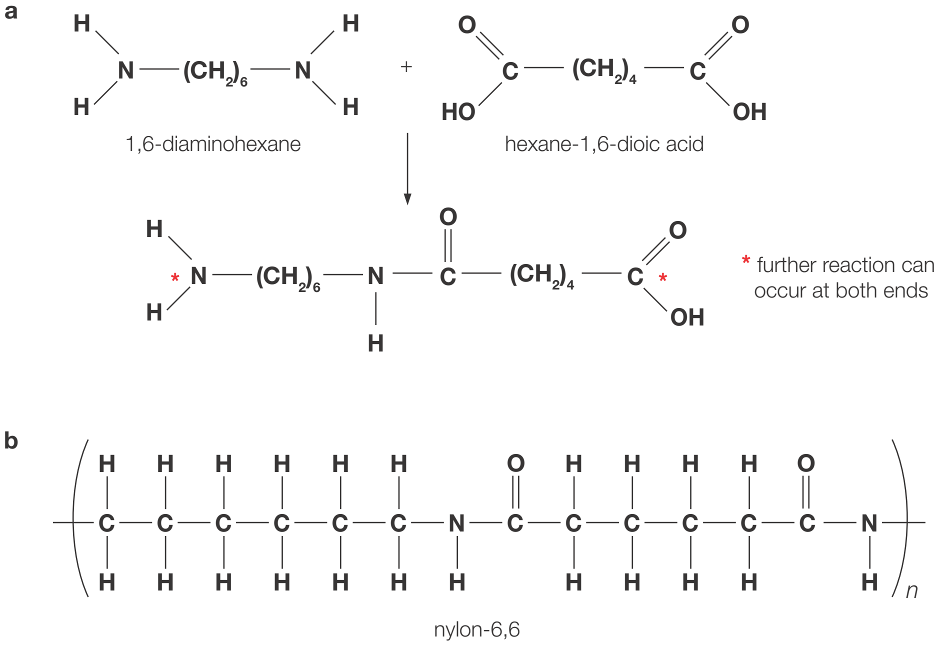
between the chains, and weak dipole–dipole interactions can take place between the polar carbonyl >C=O groups as shown in Figure 22.114. The physical properties of a polymer are dependent on three factors. The inter-chain forces mentioned here, the crystallinity of the structure and the orientation of the chains all play their part. Although the intermolecular forces in this polymer are relatively weak, the chain is stiff and rigid, with very little rotation about the bonds. Crystallization of this polyester is slow and so various forms of the polymer can be made. The molten polymer can be extruded to form fibres that are useful for materials for making clothes (Figure 22.115). Since it has the useful property of being able to form permanent creases, it has been used extensively in the production of trousers and skirts. Currently the market is for approximately 60 per cent fibre production but this material can also be used for packaging films. In addition a polyester resin (PET) can be produced which has a low permeability to carbon dioxide and so is used extensively in bottling carbonated drinks.

The production of nylon

Nylon was originally made by the reaction of a diamine with a dicarboxylic acid. The two monomers used initially were 1,6-diaminohexane and hexane-1,6-dioic acid. The polymer chain is made up from the two monomers reacting alternately and results in the chain type:

—A—B—A—B—A—B—A—B—A—B—A—B—A—B—A—

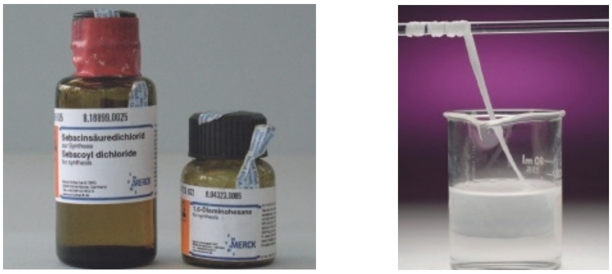
Each time a reaction takes place between the two monomers a molecule of water is lost. This is a further example of condensation polymerization (Figure 22.116a). The link formed repetitively between the monomers is an amide link and nylon is known as a **polyamide**. This link is that found in proteins, where it is often referred to as a **peptide bond**.



■ **Figure 22.116 a** The polymerization of 1,6-diaminohexane and hexane-1,6-dioic acid to form nylon. **b** Because the two monomers each contain a six-carbon chain, this form of nylon is known as nylon-6,6

There are various different forms of nylon, for example nylon-6,6 (Figure 22.116b) and nylon-6,10. The type of nylon depends on the number of carbon atoms in the monomers used. If the diamine used contains six carbon atoms and the dicarboxylic acid contains a chain of ten carbon atoms, then the resulting nylon is referred to as nylon-6,10. The formation of these different forms of nylon requires high temperatures and the presence of catalysts. The reactions are usually carried out under vacuum conditions to remove the water released.

When nylon is prepared by the reaction between a diamine and a dicarboxylic acid the reaction is actually quite slow. For a demonstration in the laboratory this can be speeded up by reacting the diamine with an acyl chloride (or acid chloride, a derivative of the acid where the –OH in the carboxylic acid group is replaced by a chlorine atom, Cl). In this case, the condensation reaction is much faster and hydrogen chloride is eliminated between the monomers instead of water (Figure 22.117a). The diamine is dissolved in water and a solution of the acyl chloride in hexane is layered carefully on top of the aqueous solution. The reaction takes place at the interface between the two immiscible solutions and the raw nylon can be spooled away as a 'nylon rope' (Figure 22.117b).

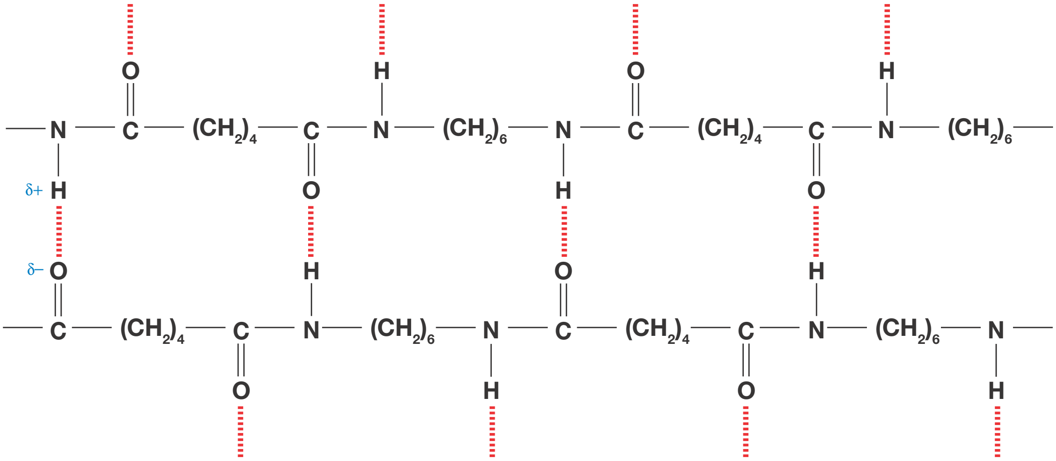


■ **Figure 22.117 a** The monomers for the 'nylon rope' experiment (decanedioyl dichloride [or sebacoyl chloride] and 1,6-diaminohexane); **b** the 'nylon rope' can be continuously drawn off from the interface between the solutions of the two monomers. This experiment produces nylon-6,10

■ **Figure 22.118** A nylon climbing rope

When nylon is made in industry it forms a solid, which is melted and then forced through fine jets and extruded. The long filaments cool and the solid nylon fibres produced are stretched to align the polymer molecules and then dried before spinning to stop high-temperature hydrolysis. The resulting yarn can be woven into fabric to make shirts, ties, sheets and parachutes or turned into ropes (Figure 22.118) or racket strings for tennis and badminton rackets.

The molecular chains in nylon fibres interact by hydrogen bonding between the hydrogen atoms of the N–H groups of the amide link of one polymer chain with the > C=O groups on adjacent polymer chains (Figure 22.119). Thus nylon fibres contain strong hydrogen bonding between the chains but the chains do show high flexibility and, because nylon crystallizes quickly, the fibres are always semi-crystalline. These factors give nylon its distinctive properties when compared to other polymers.



■ **Figure 22.119** Hydrogen bonding between adjacent nylon chains

A recent development in polymer chemistry has been to create a polyamide in which the straight-chain hydrocarbon unit within the polymer has been replaced by an aromatic benzene ring. This type of polymer is known as an aramid. The first aramid was made from 3-aminobenzoic acid. However, it was found not to be particularly strong even though it had exceptional fire resistance and could be made into fibres. The starting monomer was modified to create straighter chains in the polymer and a polyaramid was produced with exceptional properties; it is now known as Kevlar (Figure 22.120).

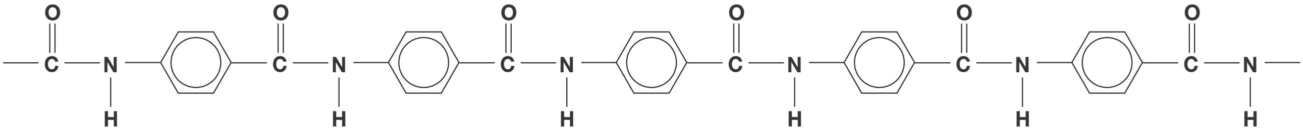


Figure 22.120 A section of the polymer chain in Kevlar

Figure 22.121 This motorcycle helmet is reinforced with Kevlar

In nylon the single covalent bonds within the polymer chain are free to rotate and this tends to make the polymer quite flexible. However, in the case of Kevlar the replacement of the straight hydrocarbon chain parts of the polyamide by rigid benzene rings makes the chains inflexible. This polymer is far more rigid than nylon. Kevlar is exceptionally strong, being five times the strength of steel on a weight for weight basis. In addition it is very fire resistant. These properties have led to a variety of uses in the aircraft and aerospace industry, for the manufacture of cables and ropes, and for protective clothing such as bullet-proof vests and motorcycle helmets (Figure 22.121).

The economic importance of condensation

Polyamides and polyesters are just two examples of condensation polymers. The world production of nylon alone is currently estimated to exceed 5 million tonnes. We are very conscious of polyesters and nylon being used to make clothes and carpets. However, there are many other uses for condensation polymers. Nylon has high strength, resists abrasion and is easy to dye. Nylon fibres are used to make climbing ropes for mountaineering, and one of the main uses of nylon is in engineering (gears, etc.). The structure of nylon can be altered to give specific properties for a defined purpose by using fillers, pigments, glass fibre and toughening agents. Nylon film finds many everyday uses, including 'boil-in-the-bag' convenience meal packaging. The chemical and physical properties can also be altered by changing the number of carbon atoms in the two condensed carbon chains. Thus nylon-6,10 has the repeating unit $-NH(CH_2)_6NHCO(CH_2)_8CO-$. Kevlar is a polyamide that finds many uses for its strength and rigidity, including reinforcing the chassis of Formula 1 racing cars.

Discovery and development of polymers

37 Find out the about the development of nylon, Kevlar, Ziegler – Natta catalysts, PEEK and polyacetylene (a conducting polymer). Discuss the roles of science, politics and economics in the development of these polymers.

Economics and politics has clearly played an important role in the discovery and development of new polymers. For example Imperial Chemical Industries (ICI) was formed in 1926 in the UK by the merging of smaller chemical companies. The aim of this merger was to form a strong competitor to the large German chemical company IG Farben. In 1993 two ICI chemists. Gibson and Fawcett, carried out a reaction between ethene and benzaldehyde using a pressure of about 200 atmospheres. They were expecting to produce a ketone and left the mixture over the weekend. On the Monday they opened the reaction vessel and found a white waxy solid. Upon analysis it was found that it had the empirical formula CH_2. Later experiments under the charge of Michael Perrin showed the importance of oxygen that had leaked into the original reaction mixture. Perrin also demonstrated that the polymer (polyethene) was formed even if the benzaldehyde was left out of the reaction mixture. Its unique properties were essential during the development of radar during World War II. It was an excellent insulator with no tendency to absorb electrical signals and, unlike rubber, was not affected by weather or water.

Distinguishing between addition and condensation polymers

Addition polymerization generally occurs with unsaturated monomers containing a carbon–carbon double bond. This usually involves a single monomer and so the polymer produced is a **homopolymer** – as with polypropene and polychloroethene (PVC) for example. Condensation polymerization occurs when the monomers contain two reactive functional groups – for example in the formation of polyamides such as nylon and polyesters, for example polyethylene terephthalate (PET). Such polymerization often involves two monomers and involves the elimination of water each time a link is made (hence the name for this type of polymerization).

Forming condensation polymers

Phenol–methanal plastics

These are prepared by adding acid or alkali to a mixture of phenol and methanal. Under these conditions the methanal is first substituted into the phenol molecule in the 2- or 4-position of the benzene ring (Figure 22.122a). Then the product undergoes a condensation reaction with another molecule of phenol with the elimination of water – further polymerization takes place to build up a long chain (Figure 22.122b), followed by covalent cross-linking to form a three-dimensional structure (Figure 22.122c).

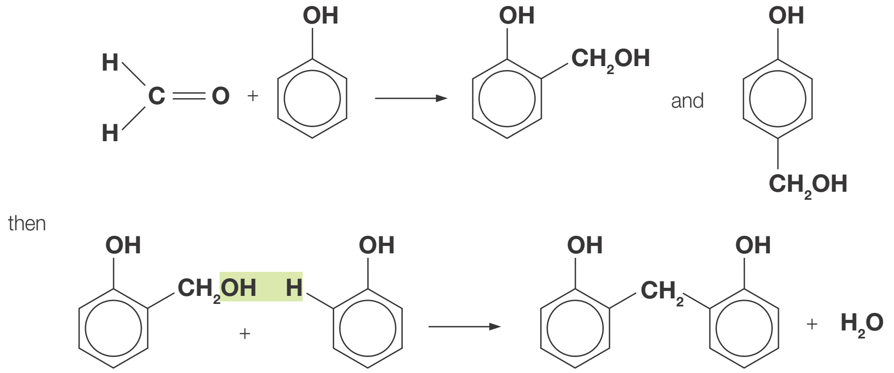


Figure 22.122a The initial reaction in the formation of a phenol–methanal plastic

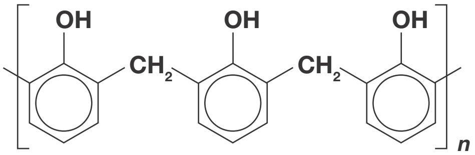


Figure 22.122b The formation of the long chain of the phenol–methanal plastic

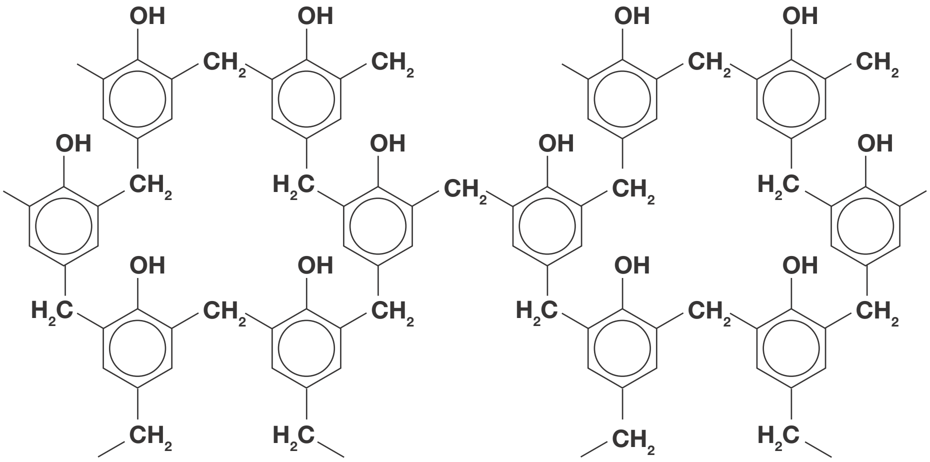


Figure 22.122c Cross-linking produces a rigid thermosetting plastic

Phenol–methanal plastic (Bakelite) is a rigid plastic used for making electrical plugs and similar fixtures. It is chemically and thermally stable, with a high melting point. Famously it was used for the classic old telephones that are now antique pieces and for the casings for radios and clocks (Figure 22.123).

Figure 22.123 Antique radio casings were made out of Bakelite (phenol–methanal plastic)

Polyurethanes

Polyurethanes are formed from the reaction of polyhydric alcohols, such as diols or triols (Chapter 20), with compounds containing more than one isocyanate functional group (–NCO). Generally the reaction is of the type shown in Figure 22.124.

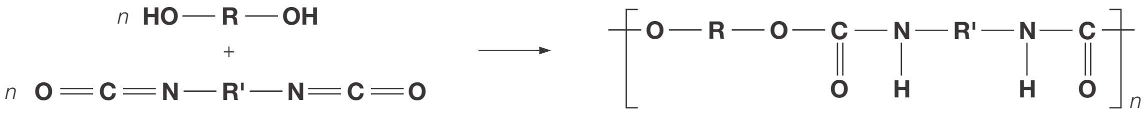


Figure 22.124 The formation of a polyurethane chain

Polyurethane chemistry is also used in the manufacture of textiles – for example, spandex or elastane (often known as lycra) is an elastomeric material based on polyurethane chemistry. Polyurethanes are often studied with condensation polymers but no small molecules are formed during the reaction. There are no urethane monomers that add together to form polyurethanes.

The properties of condensation polymers

When polymers are able to form giant three-dimensional structures, as for example in phenol–methanal plastics, the resulting polymer is extremely strong and rigid. These polymers are also insoluble and generally resistant to chemical attack. The following are some examples of addition and condensation polymers that brought novel properties to the use of such materials.

Kevlar

One example of a polymer with novel qualities is Kevlar (Figure 22.125), the material from which lightweight bulletproof vests, and composites for motor-cycle helmets and body armour are made (Chapter 20). Kevlar is a polyamide made by condensing 1,4-diaminobenzene with benzene-1,4-dicarbonyl chloride (or the corresponding carboxylic acid). Kevlar forms a strong three-dimensional structure due to hydrogen bonding between the long, rigid chains.

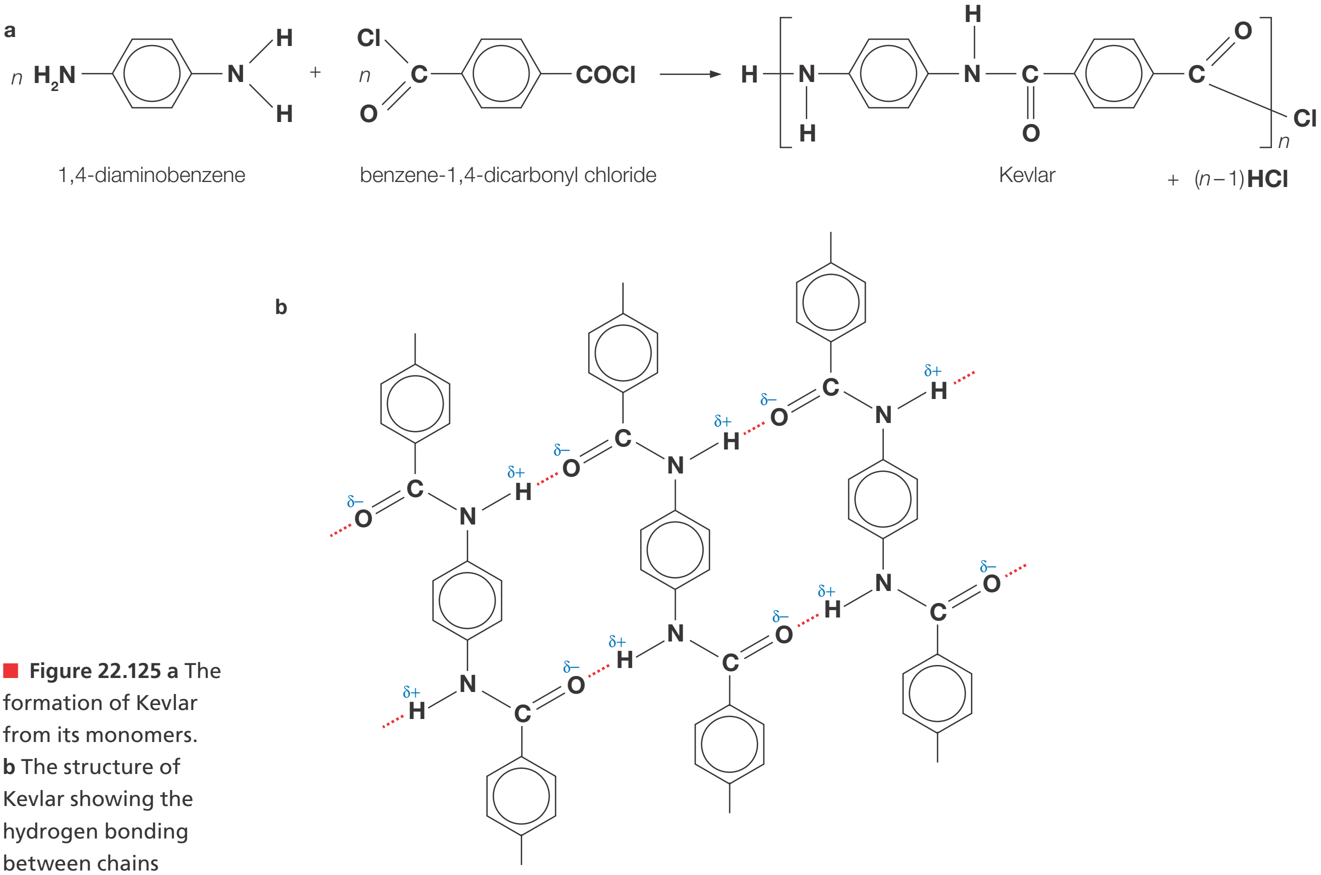
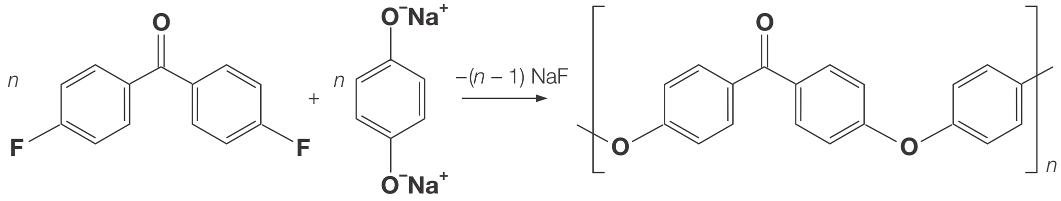


Figure 22.125 a The formation of Kevlar from its monomers. **b** The structure of Kevlar showing the hydrogen bonding between chains

PEEK

PEEK, poly(ether-ether-ketone), is a polymer that can withstand very high temperatures. It is a further example of a copolymer and its structure is shown in Figure 22.126.



■ **Figure 22.126** The synthesis of PEEK from 4,4′-difluorobenzophenone and the disodium salt of hydroquinone

The heat-resistant properties of PEEK mean that it is used for such diverse purposes as making plastic kettles, the nose cones of missiles and parts of car engines.

■ Atom economy

Addition polymerization has 100% atom economy because all the monomer ends up in the desired condensation polymer product. This is not the case for condensation polymerization as the inorganic condensation product (H_2O, HCl or NH_3) is lost from the polymer.

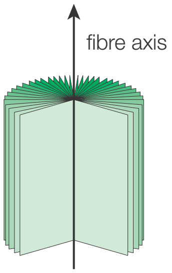
For example, consider esterification: $C_2H_5OH + CH_3COOH \rightarrow C_2H_5OOCCH_3 + H_2O$. The atom economy (to the nearest integer) = 88 × 100/106 = 83%.

Explanation of Kevlar's strength and its solubility in concentrated sulfuric acid

The individual polymer chains in Kevlar are very strong and extremely stiff. The reason for this will become clear when we look closely at Figure 22.125, which shows how the atoms are actually arranged in a single molecule of Kevlar.

The molecule is essentially flat and this is mainly due to pi electron delocalization, which is not confined just to the benzene rings but extends over the whole length of the molecule. Pi electron delocalization also strengthens the covalent bonds and prevents any rotation about individual bonds within the chain. All these factors make the individual Kevlar molecules strong, stiff and rod-like. The linear polymer chains line up parallel to one another, forming a hydrogen-bonded sheet of molecules. This packing arrangement is illustrated in Figure 22.125b.

The amide groups (–CONH–) are polar and it is through them that hydrogen bonds are set up between the molecules. Although they are considerably weaker than covalent bonds, the hydrogen bonds keep the polymer chains in alignment. This arrangement imparts even greater strength to Kevlar.



■ **Figure 22.127** An illustration of the crystalline structure of Kevlar

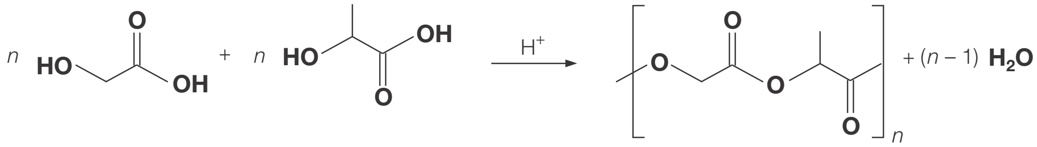
In the Kevlar fibre itself the flat sheets of molecules stack together around the fibre axis. This arrangement is shown in Figure 22.127. It is the very high degree of molecular alignment within the fibre that is the largest contributing factor in Kevlar's exceptional tensile strength. It adopts this crystalline structure because of the way the polymer is processed to produce the fibres.

Kevlar is insoluble in water but soluble in concentrated sulfuric acid. The solubility occurs because the intermolecular forces (the hydrogen bonds) are disrupted by the concentrated sulfuric acid. The nitrogen atoms in the amide bonds are protonated. The positively charged groups repel adjacent chains and are solvated by sulfate ions.

■ Biodegradable condensation polymers

Sutures, or 'stitches', are used to close wounds to help them to heal. They were traditionally made from cat gut – natural fibres from the intestines of sheep or goats. However, cat gut is relatively inert and needs to be removed after normal healing of the tissues – this is a painful and delicate procedure. A replacement for cat gut is a condensation copolymer of glycolic acid (hydroxyethanoic acid) with lactic acid (2-hydroxypropanoic acid) (Figure 22.128). Its fibres are strong and its hydrolysis products are totally absorbed by the body after the wound has healed. The sutures dissolve slowly because the pH near the wound is only slightly acidic.

Figure 22.128 Poly(glycolic-co-/lactic) acid – a copolymer used for self-dissolving sutures



Polyester: acid-catalysed hydrolysis degrades the polymer backbone by breaking ester C—O bonds

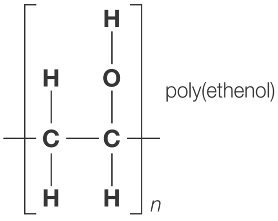


Figure 22.129 The repeating unit of poly(ethenol)

Other polymers that dissolve in particular conditions have also been developed. These include poly(ethenol), which is used to make disposable hospital laundry bags – its repeating unit is shown in Figure 22.129. This polymer dissolves in hot water and so minimizes the amount of handling of potentially hazardous contents that hospital staff are involved in – the whole bag and contents can be placed in the wash. Poly(ethenol) is more commonly referred to as poly(vinyl alcohol) (PVOH).

Modifying condensation polymers

The properties of addition polymers can be modified by:

- control of the various types of high-density poly(ethene) and low-density poly(ethene)
- the forms of poly(propene) that differ in their stereoregularity
- the use of plasticizers to produce different forms of poly(vinyl chloride).

In condensation polymerization, it is possible to produce different forms of a polymer by varying the monomer(s) used. For instance, the choice of monomer(s) can influence the properties of manufactured nylon, with nylon-6,6, nylon-6,10 and nylon-6 all being produced.

In a similar way to addition polymers, condensation polymers can be modified by a variety of means during their manufacture. For example, air can be blown into polyurethane to make polyurethane foam for use in cushions and thermal insulation. The fibres of polyesters can be blended with other manufactured polymers, or natural fibres such as cotton, for making clothes that are more comfortable, are durable and retain the colour from dyeing better.

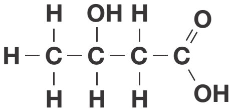
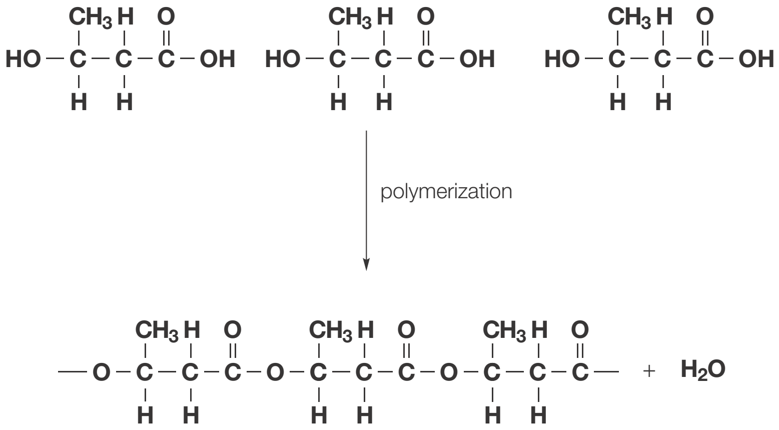


Figure 22.130 Structure of 3-hydroxybutanoic acid

Completion and descriptions of equations to show how condensation polymers are formed.

PHB (poly(3-hydroxybutanoate) is a biopolymer. Figure 22.130 shows the structure of the monomer, 3-hydroxybutanoic acid. It contains both a hydroxyl group and a carboxyl group and it is through these that the monomers are able to link together, as shown in Figure 22.131.

Figure 22.131 Polymerization of 3-hydroxybutanoic acid



At each reaction point a water molecule is eliminated and an ester linkage (–COO–) is formed; therefore PHB is both a condensation polymer and a polyester.

However, in the laboratory cyclization occurs and this type of reaction does not produce molecules with a high molar mass. The methyl groups have stereochemistry and simple routes result in the atactic polymer, which is not a useful material. Bacteria can synthesize isotactic polymer chains with a molar mass in excess of 250 000.

Deduction of the structures of polyamides and polyesters from their respective monomers

The structure of a polyamide can be deduced from their monomers by aligning the two polymers so the amine group of one diamine monomer is aligned next to the carboxylic acid (or acid chloride) group of the dicarboxylic acid (or diacid chloride) (Figure 22.132). Water (or hydrogen chloride) should then be removed and an amide linkage is formed.

The structure of a polyester can be deduced from their monomers by aligning the two polymers so the alcohol group of one diol monomer is aligned next to the carboxylic acid (or acid chloride) group of the dicarboxylic acid (or diacid chloride) (Figure 22.133). Water (or hydrogen chloride) should then be removed and an ester linkage is formed.

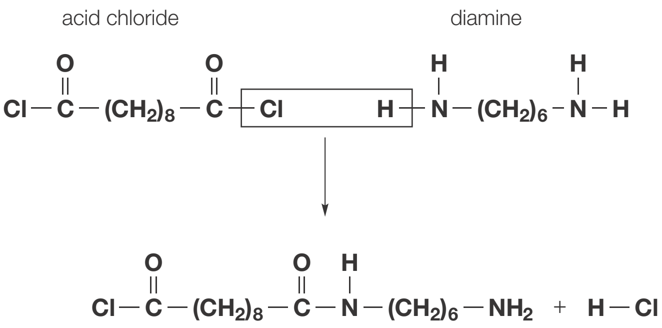


Figure 22.132 Formation of a polyamide

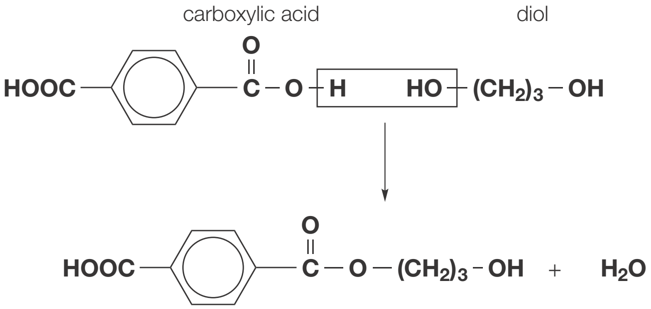


Figure 22.133 Formation of a polyester

38 Nylon-6,8 can be formed by the reaction of a diamine with a diol. Draw the structures of the two monomers and the structure of the repeating unit in nylon-6,8.

Instead of a dicarboxylic acid (diacid chloride) and a diol, the reaction may proceed with only one monomer that contains two functional groups, for example 5-hydroxy pentanoic acid which contains an alcohol and a carboxylic acid group, so it can polymerize with itself (Figure 22.134).

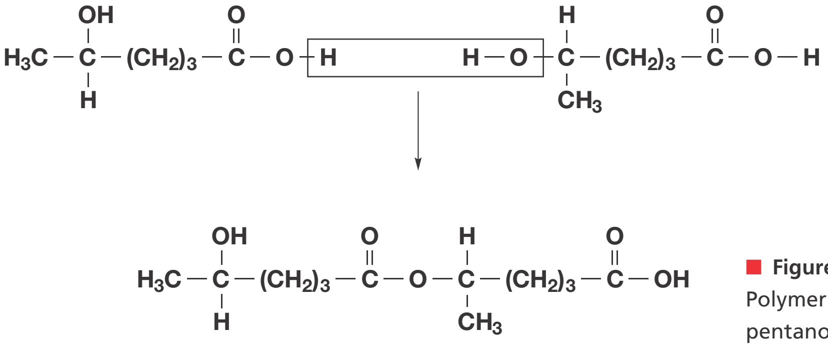


Figure 22.134 Polymerization of 5-hydroxy pentanoic acid

Summaries

Tables 22.9 and 22.10 summarize the main condensation and addition polymers.

Table 22.9 Summary of condensation polymers

Name	General repeating structure	Example(s)
Polyesters	–O–R–O–C(O)–R′–C(O)–	PET, Dacron, Terylene, packaging film (Mylar)
Polycarbonates	–O–R–O–C(O)–	
Polyanhydrides	–O–C(O)–R–C(O)–	
Polyamides	–C(O)–R–C(O)–NH–R′–NH–	Nylon
Phenol-methanal polymers	–N(H)–C(O)–N(H)–CH_2–O–	Bakelite
Polyurethanes	–C(O)–O–R–O–C(O)–N(H)–R′–N(H)–	Spandex or elastane (Lycra), polyurethane foam
Polyethers	–R–O–R–	PEEK

Table 22.10 Summary of addition and condensation polymer physical properties

Structural property	Physical property	Example
Chain length	The longer the chain, the stronger the polymer	Longer polymer chains have higher melting point, increased strength and increased impact resistance due to increased London (dispersion) forces
Branching and packing structure	Straight unbranched chains can pack more closely. A higher degree of branching keeps strands apart and weakens intermolecular forces	HDPE with no branching is more rigid than the more branched LDPE. Use of plasticizers in PVC to soften the polymer
Side-group on monomers	Hydrogen bonding can increase strength, e.g. Kevlar. Atactic and isotactic placement can influence strength, e.g. polystyrene	Polystyrene
Cross-linking	Extensive covalently bonded cross-linkage increases polymer strength	Vulcanized rubber, Bakelite

22.10 Environmental impact – heavy metals (AHL) – *toxicity and carcinogenic properties of heavy metals are the result of their ability to form coordinated compounds, have various oxidation states and act as catalysts in the human body*

Heavy metals and their uses

Some metals are naturally found in the body and are essential to health. Iron, for example, is present as an essential component of hemoglobin, while zinc is a co-factor in many enzyme-controlled reactions. Because these metals normally occur at low concentrations in the body they are known as trace metals. Indeed, at high levels these metals may be toxic or produce deficiencies in other trace metals.

Heavy metal is a rather vague term that refers to toxic metals (and their ions) with a high relative atomic mass such as lead, mercury and cadmium which have harmful effects on human health.

Such metals have many uses: lead, nickel and cadmium are used in batteries (lead acid battery and Ni-cads; see Chapter 24); arsenic (a metalloid), bismuth and antimony (a metalloid) are often found in semiconductors; and mercury has many uses such as in thermometers, barometers and dental amalgams and for the collection of gases under anhydrous conditions. Heavy metals are commonly used as catalysts; for example, palladium and platinum are used as catalysts for hydrogenation. Lead has historical uses such as lead for water pipes, lead paint (containing lead(II) chromate(VI) or lead(II) carbonate) and petrol additives (tetraethyl lead(IV)).

A large proportion of heavy metals are absorbed or precipitated on to particulate matter in what is sometimes termed a self-purification process. However, in lakes heavy metals can reach dangerously high levels. Heavy metals accumulate in biological systems over time. They are stored in living organisms and passed along the food chain by ingestion, digestion and assimilation (see biomagnification in Chapter 23). The toxicity and carcinogenic properties of heavy metals are the result of their ability to form coordinated compounds, exist in various oxidation states and act as catalysts in the human body.

Toxic metals can react with enzyme-binding sites and inhibit or overstimulate these enzymes. Many enzymes are metallo-enzymes and contain metal ions in their active sites which can be displaced by heavy metals or their ions. For example, cadmium belongs to the same group as zinc, and due to a similar atomic radius competes with zinc during absorption in to the body. Lead can compete with calcium ions in the same way. When more zinc and cadmium are taken in the diet via contaminated water or food, the heavy metals are not eliminated and tend to accumulate in the liver and kidney.

High doses of transition metals can be toxic and disturb the normal oxidation–reduction balance in cells through various mechanisms. They can disrupt the endocrine system because they compete for the active sites of enzymes and cellular receptors (often proteins on cell membranes). They have multiple stable oxidation states so they can take part in redox reactions (electron transfer), and they can initiate free radical reactions in electron transfer. Their ability to form complex ions enables them to bind with enzymes: iron(II) ions, for example, form a complex with hemoglobin in the blood, which is essential for oxygen transport. Finally, transition metals are very good heterogeneous catalysts.

The Haber–Weiss reaction

The Haber–Weiss reaction generates hydroxyl radicals, $^{\bullet}OH$, from hydrogen peroxide and superoxide ions ($^{\bullet}O_2^-$):

$$^{\bullet}O_2^- + H_2O_2 \rightarrow O_2 + OH^- + {}^{\bullet}OH$$

This reaction can naturally occur in cells and is therefore a possible source for oxidative stress: damage to biological tissue by radicals. The reaction is very slow, but is catalysed by iron(III) ions (released from toxin-injured cells). The first step of the catalytic cycle involves the reduction of iron(II) ions to iron(III) ions:

$$Fe^{3+} + {}^{\bullet}O_2^- \rightarrow Fe^{2+} + O_2$$

The second step is known as the Fenton reaction:

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH$$

The efficiency of the Fenton reaction depends mainly on H_2O_2 concentration, Fe^{2+}/H_2O_2 ratio, pH and reaction time.

The net or overall reaction is:

$$^{\bullet}O_2^- + H_2O_2 \rightarrow {}^{\bullet}OH + OH^- + O_2$$

This reaction is named after Fritz Haber and his student Joseph Joshua Weiss (1905–1972). Fritz Haber is best known for fixing nitrogen (synthesizing ammonia from its elements), and he received the Nobel Prize in chemistry in 1918 for this work. His last paper in 1934 proposed that the reactive hydroxyl radical could be generated from superoxide ions and hydrogen peroxide molecules. This greatly increased chemists' understanding of the role of radicals in biochemistry. However, the Fenton reaction is still under study, with some data suggesting the ferryl ion (FeO^{2+}) is involved.

The highly reactive $^{\bullet}OH$ radical is one of the most damaging free radicals in the body. It reacts with almost any molecule it encounters including macromolecules such as DNA, phospholipids in membranes, and enzymes.

Because it is so reactive it can be used to break down dyes and pollutants such as pesticides, aromatic amines, dyes, methanal and phenols and the Fenton reaction is carried out in waste-water treatment plants. For example, benzene derivatives, which are not very reactive, can be oxidized to less toxic and more water-soluble phenols.

$$2{}^{\bullet}OH + C_6H_6 \rightarrow C_6H_5OH + H_2O$$

The $^{\bullet}OH$ radical created by the Fenton reaction is the first step in many industrial processes. It can be used to eliminate some greenhouse gases such as methane from plant emissions, and to reduce odour (bad smells) from waste-water treatment sites. The highly reactive radical can break carbon–carbon double bonds, open up aromatic rings, abstract hydrogen atoms and even initiate polymerization by reacting with pi bonds.

Superoxide ion

It has been estimated that as much as six per cent of the oxygen taken up by cells is transformed into superoxide ions, $^{\bullet}O_2^-$. There is considerable evidence that when exposed to superoxide ions, heart mitochondria, lung tissue, synovial joint fluid and skin all show damage and loss of biological function. As it is charged, $^{\bullet}O_2^-$ is a powerful nucleophile (electron pair donor) as well

39 Find out about the use of the Fenton reagent in organic chemistry.

as a radical. In addition it is a reducing agent (being oxidized back to molecular oxygen, O_2) and an oxidizing agent (when it is converted into H_2O_2) and disproportionations are possible:

$$2{}^{\bullet}O_2^- + 2H^+ \rightarrow H_2O_2 + O_2$$

Hydrogen peroxide is unstable because of the relatively weak O–O bond (due to repulsions between the lone pairs on the oxygens), and the molecule can dissociate:

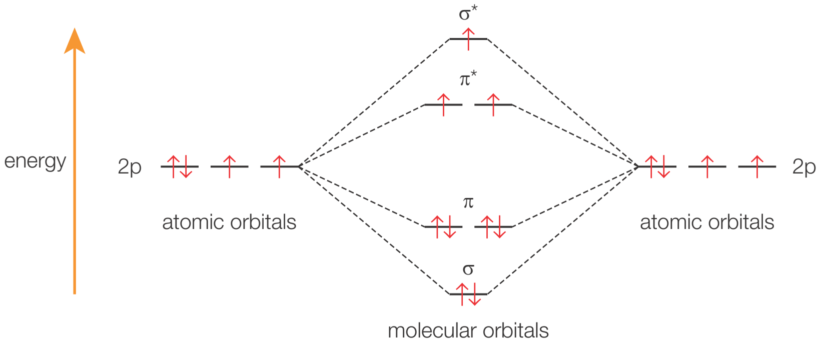
$$H_2O_2 \rightarrow 2OH^{\bullet}$$

or it can react with more ${}^{\bullet}O_2^-$:

$$H_2O_2 + {}^{\bullet}O_2^- \rightarrow O_2 + OH^- + {}^{\bullet}OH$$

so a whole range of harmful ions and radicals can be produced. The structure of the superoxide ion can be described by molecular orbital (MO) theory (Chapter 14). It can be regarded as being an oxygen molecule that has gained an extra electron, which enters the anti-bonding sigma molecular orbital (Figure 22.135).

Figure 22.135 MO diagram of the superoxide ion



Compare and contrast the Fenton and Haber–Weiss reaction mechanism

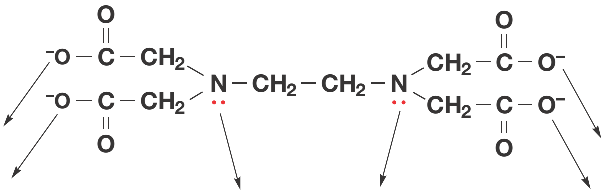
The Haber–Weiss reaction occurs at a slower rate than the Fenton reaction (under the same conditions). The Fenton reaction involves a homogeneous catalyst, the iron(III) ions. The Haber–Weiss reaction does not involve catalysis. Both reactions are redox reactions occurring in aqueous solution and involve radicals.

Chelating effects

Apart from the Fenton reaction, other methods of removing heavy metals include precipitation, adsorption and chelation. Chelation takes advantage of a metal's ability to form complex ions. The term 'chelate' refers to polydentate ligands (Chapter 13). Chelating agents are used to remove heavy metals such as lead, arsenic and mercury from the body. Once chelated the complex ion is too large to enter cells but being a water-soluble ion can be excreted from the body in the urine.

To act as ligands, chelating agents must have lone pairs of electrons that can form coordinate bonds to a central atom. The term 'polydendate' refers to their ability to form more than one such coordinate bond. Figure 22.136 shows that the chelating agent EDTA can form up to six coordinate covalent bonds with a central atom.

Figure 22.136 EDTA is a hexadentate ligand



Polydentate ligands such as EDTA are more effective than monodentate ligands and will replace them in reactions. Competition between ligands was discussed in Chapter 13; one factor influencing the reaction, is the increase in entropy involved. EDTA will replace the six molecules in this reaction, forming a large complex and releasing six smaller molecules, thus increasing the overall entropy:

$$EDTA^{4-}(aq) + [Ni(H_2O)_6]^{3+}(aq) \rightarrow [Ni(EDTA)]^{2-}(aq) + 6H_2O(l)$$

The existence of a greater number of smaller molecules rather than one larger one yields more ways of distributing the effective energy and hence represents an increase in entropy. This is one reason why chelation is effective at removing metals, as the polydentate ligands replace larger numbers of existing ligands, usually water.

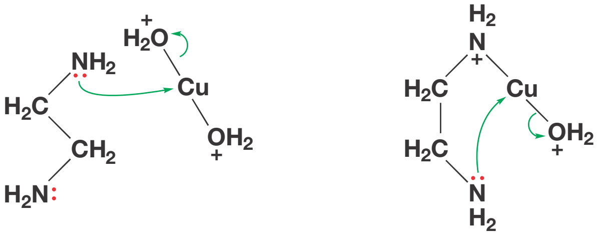
Explanation of how chelating substances can be used to remove heavy metals

Chelating agents form particular stable complex ions, partly because they form strong coordinate bonds to the heavy metal atom or ion. In addition there is an entropy additional effect that adds to the stability of the chelated complex.

For example, a chelate is formed in which three oxalate ions are coordinated to an iron(III) ion. Four chemical species becomes seven after the ligand replacement so the formation of this chelate is accompanied by an increase in entropy:

$$[Fe(H_2O)_6]^{3+}(aq) + 3C_2O_4{}^{2-}(aq) \rightarrow [Fe(C_2O_4)_3]^{3-}(aq) + 6H_2O(l)$$

Another way of visualizing the increase in entropy is to consider the effect after one end of the chelating agent has become bonded to the metal ion. Once this end is bonded, it becomes more likely that the other will be in the right position to form a coordinate bond (Figure 22.137).

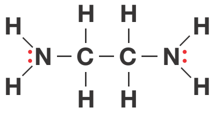


■ **Figure 22.137** Chelating effect with the 1,2-diaminoethane molecule and copper(II) hydroxide

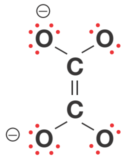
Deducing the number of coordinate bonds a ligand can form with a central metal ion

The number of coordinate bonds a ligand can form with a central metal ion can be established by deducing the number of lone pairs of electrons that are present in the ion or molecule.

For example, ethylene diamine (1,2-diaminoethane; Figure 22.138) has two lone pairs located on the two terminal nitrogen atoms. Hence this bidentate ligand can form two coordinate bonds. The oxalate ion (ethanedioate ion; Figure 22.139) has lone pairs on all four oxygens and hence is a tetradentate ligand and can form four coordinate bonds.



■ **Figure 22.138** Structure of 1,2-diaminoethane molecule



■ **Figure 22.139** Structure of ethanedioate ion

■ Chelation therapy

Every year many young children swallow iron tablets, mistaking them for sweets. Every year patients undergoing repeated blood transfusion for sickle cell anemia or thalassemia experience iron overload which causes fits, coma and death.

Disferrioxamine is a hexadentate ligand which forms a very stable complex with iron(II) ions. If excess iron pills have been swallowed then treatment will involve washing out the stomach with a solution of disferrioxamine. The protonated amine group, $-NH_3{}^+$, at one of the molecule and the >C=O and N–OH groups make the complex soluble in water so it can be excreted in

40 Find out about the roles of the peptides glutathione and metallothionein as natural chelating agents.

the urine. Unfortunately, because of the amide groups, –CO–NH–, it undergoes acid hydrolysis and cannot be taken orally and has to be injected. Disferrioxamine is also effective in the treatment of workers in the nuclear industry who have ingested small amounts of plutonium.

Adsorption of heavy metal ions

Another method of removing heavy metals or their ions in water is by physical adsorption on to a solid surface, such as activated carbon, charcoal or clays. Activated charcoal is charcoal that has been treated with oxygen to open up millions of tiny pores between the carbon atoms. This is an expensive substance and cheaper natural materials have been used such as coconut shells, crab shells, orange peels and sugar cane husks.

Biomass such as brewer's yeast, algae, bacteria and fungi has also been found to be effective at bioadsorption of heavy metals and their ions. Ion exchangers (often based on zeolites) are columns which exchange heavy metals for calcium or sodium ions. The technique of reverse osmosis using membranes that allow water molecules but not ions can be used to remove all cations from water. The treated water then undergoes further purification, such as chlorination or ultraviolet radiation treatment.

The limited solubility of many transition metal hydroxides (Chapter 13) allows us to remove the transition metals from waste water by chemical precipitation.

When the concentration of a solute exceeds its solubility at the given temperature, the excess solute will precipitate out of the solution. However, by careful choice of added substances, the solubility of the dangerous transition metal compounds can be lowered still further, to the extent that almost all of the heavy metal ions can be removed from the solution. In order to explain this effect we need to introduce a new form of the equilibrium constant (K_c) called the solubility product constant, K_{sp}.

Solubility product constant, K_{sp}

Metal ions from group 1, including Li^+, Na^+ and K^+, form highly soluble compounds, whereas the heavy metal ions form compounds of low solubility. Their salts and hydroxides precipitate easily and this means heavy metal ions can be removed during waste-water treatment. Many heavy metal hydroxides are only slightly soluble so hydroxide ions are often added to precipitate the metals as the concentration of hydroxide ions can be monitored by measuring the pH. Quicklime (solid calcium hydroxide, $Ca(OH)_2$) or lime (solid calcium oxide, CaO) can be used since it is relatively cheap.

Consider a sparingly soluble salt, MX. When excess MX is placed in water, it will dissolve to a limited extent, a dynamic equilibrium (Chapter 7) forming between the ions in solution and the undissolved solid:

$$MX(s) \rightleftharpoons M^+(aq) + X^-(aq)$$

The equilibrium expression will look like this:

$$K_c = \frac{[M^+(aq)] \times [X^-(aq)]}{[MX(s)]}$$

This is a heterogeneous equilibrium. The concentration of solid MX is a constant (Chapter 7) at a given temperature. Rearranging gives:

$$K_c\,[MX(s)] = [M^+(aq)]\,[X^-(aq)]$$

This new constant K_c [MX(s)] is called the **solubility product constant**, K_{sp}.

$$K_{sp} = [M^+(aq)]\,[X^-(aq)]$$

For the most complicated situations involving non-binary solutes, for example M_yX_z in the presence of water, these equations generalize as follows:

$$M_yX_z(s) \rightleftharpoons yM^{z+}(aq) + zX^{y-}(aq)$$

$$K_{sp} = [M^{z+}(aq)]^y\,[X^{y-}(aq)]^z$$

Thus, consider aluminium hydroxide, $Al(OH)_3$, in the presence of water. The equation for the ionic equilibrium set up is:

$$Al(OH)_3(s) \rightleftharpoons Al^{3+}(aq) + 3OH^-(aq)$$

$$K_{sp} = [Al^{3+}(aq)]\,[OH^-(aq)]^3 = 1.0 \times 10^{-32}$$

(The value is determined from experiment.)

The solubility product constant, like other equilibrium constants, is constant *at a given temperature*. Changing the temperature changes K_{sp} and therefore changes the amount of substance that will dissolve. Some solubility product values are given in the *IB Chemistry data booklet,* section 32.

Calculations involving K_{sp} as an application of removing metals in solution

Calculating a solubility product from solubility data

If the solubility of a salt (i.e. the amount that must dissolve in order for the solution to become **saturated**) is known, the solubility product can be calculated. Solubilities are usually quoted in grams per cubic decimetre, $g\,dm^{-3}$.

For example, the solubility of lead(II) bromide at 298 K is $6.15\,g\,dm^{-3}$. The molar mass of $PbBr_2$ is $288\,g\,mol^{-1}$. The solubility product equation is:

$$K_{sp} = [Pb^{2+}(aq)]\,[Br^-(aq)]^2$$

The molarity of the saturated solution of lead bromide,

$$PbBr_2 = \frac{6.15\,g\,dm^{-3}}{288\,g\,mol^{-1}} = 0.0214\,mol\,dm^{-3}$$

Upon dissociation the concentrations of each ion are as follows:

$$[Pb^{2+}(aq)] = 0.0214\,mol\,dm^{-3}$$

$$[Br-(aq)] = 0.0214 \times 2 = 0.0428\,mol\,dm^{-3}$$

Substituting these concentrations into the solubility product equation gives:

$$K_{sp} = [Pb^{2+}(aq)]\,[Br^-(aq)]^2 = 0.0214 \times (0.0428)^2 = 3.9 \times 10^{-5}$$

Calculating solubility from the solubility product

The solubility product can be used to calculate the solubility of a salt in $g\,dm^{-3}$. The solubility product constant, K_{sp} for aluminium hydroxide, $Al(OH)_3$, is 1.0×10^{-32}. The equation for the ionic equilibrium in water is:

$$Al(OH)_3(s) \rightleftharpoons Al^{3+}(aq) + 3OH^-(aq)$$

And the solubility product equation is:

$$K_{sp} = [Al^{3+}(aq)]\,[OH^-(aq)]^3$$

The dissociation of 1 mole of aluminium hydroxide (formula unit) produces 1 mole of aluminium ions and 3 moles of hydroxide ions. Setting the molarity of aluminium ions as x, the molarity of hydroxide ions becomes $3x$. We now have:

$$K_{sp} = x \times (3x)^3 = 27x^4$$

We can now substitute the value for the solubility product:

$$27x^4 = 1.0 \times 10^{-32}$$

$$x^4 = 3.7 \times 10^{-34}$$

$$x = (3.7 \times 10^{-34})^{\frac{1}{4}} = 4.4 \times 10^{-9}$$

The solubility in moles of aluminium hydroxide is therefore 4.4×10^{-9} mol dm^{-3}. The molar mass of $Al(OH)_3$ is 78 g mol^{-1}. The solubility in grams per cubic decimetre, g dm^{-3}, can therefore be expressed as:

$$4.4 \times 10^{-9}\,\text{mol dm}^{-3} \times 78\,\text{g mol}^{-1} = 3.4 \times 10^{-7}\,\text{g dm}^{-3}$$

Predicting precipitation

Phosphate(v) ions can be removed from waste water by precipitating them with aluminium or iron(III) ions:

$$Fe^{3+}(aq) + PO_4^{3-}(aq) \rightarrow FePO_4(s)$$

$$Al^{3+}(aq) + PO_4^{3-}(aq) \rightarrow AlPO_4(s)$$

When sufficient metal ions are added the product of the concentration of the added metal ions and that of the phosphate ions already present in the waste water exceeds the solubility product, so the salts precipitate out of the solution.

Solubility products can allow us to predict whether a precipitate will form. For example, aluminium ions are added to waste water in the form of hydrated aluminium sulfate, $Al_2(SO_4)_3.14H_2O$, commonly called ‘alum’. This can be delivered as a solid, or as a liquid ‘slurry’. Multiplying the concentration of aluminium ions and phosphate ions gives us the ionic product of these ions in solution. If the ionic product is greater than the solubility product, a precipitate will form.

The following worked example shows how to calculate an ionic product to predict precipitation.

Worked example

1 kg of solid alum is added to a tonne of waste water containing phosphate ions at a concentration of 10^{-4} mol dm^{-3}. Will a precipitate result?

Assume the density of the waste water to be 1 g cm^{-3} and that there is no change in volume upon adding the alum. The solubility product of aluminium phosphate is 9.8×10^{-21} at 298 K. The molar mass of $Al_2(SO_4)_3.14H_2O$ = 594.4 g mol^{-1}.

The equilibrium in question is:

$$AlPO_4(s) \rightleftharpoons Al^{3+}(aq) + PO_4^{3-}(aq)$$

$$K_{sp} = [Al^{3+}(aq)]\,[PO_4^{3-}(aq)]$$

Calculating the concentration of aluminium ions, $[Al^{3+}(aq)]$:

$$1\,\text{kg} = 1000\,\text{g}\ Al_2(SO_4)_3.14H_2O$$

$$= \frac{\text{mass (g)}}{\text{molar mass (g mol}^{-1}\text{)}} = \frac{1000\ \text{g}}{594.4\ \text{g mol}^{-1}}$$

$$= 1.7\ \text{mol}$$

amount of $Al_2(SO_4)_3.14H_2O$

1 mole of alum contains 2 moles of aluminium ions. Therefore:

amount of aluminium, Al^{3+} ions = $1.7 \times 2 = 3.4$ mol

One tonne of waste water = 1000 kg = 1000 dm^3

$$= \frac{3.4\ \text{mol}}{1000\ \text{dm}^3}$$

$$= 3.4 \times 10^{-3}\ \text{mol dm}^{-3}$$

Molarity of Al^{3+} in 1 tonne of waste water

Molarity of PO_4^{3-} in waste water = 10^{-4} mol dm^{-3}

Ionic product of Al^{3+} and PO_4^{3-} ions

$= 3.4 \times 10^{-3} \times 10^{-4}$

$= 3.4 \times 10^{-7}$

This value exceeds the solubility product of aluminium phosphate, $AlPO_4$, so a precipitate will form.

Problems with chemical precipitation methods

An example of an organic ligand present in industrial waste water is EDTA. EDTA has widespread uses including food preservation, cleaning and medical uses. It is commonly found in factory waste water.

EDTA is a chelating ligand (Chapter 13), which means that it forms multiple dative bonds to the metal ion. It is able to wrap around a central metal ion, using its four carboxylate groups and two amino acid groups to form six dative bonds to the central metal ion (it is a hexadentate ligand).

At low pH the ligand exists in the deprotonated form, often called $EDTA^{4-}$. In solution, EDTA–metal complexes exist in equilibrium with the free metal ions and the ligand:

$$Fe^{3+}(aq) + EDTA^{4\ -}(aq) \rightleftharpoons [Fe(EDTA)]^{-}(aq)$$

Iron(III) ions, Fe^{3+}, therefore become 'locked up' in EDTA complexes, meaning that more iron(III) ions must be added in order to precipitate out the phosphate ions. Later the EDTA complex decomposes, releasing the heavy metal ions into the environment.

ToK Link

What responsibility do scientists have for the impact of their endeavours on the planet?

This is an ethical issue and the view adopted by an individual and will be influenced by ethics, culture, and political, philosophical and religious beliefs.

However, reproduced below is the declaration on science and the use of scientific knowledge agreed by UNESCO:

'We all live on the same planet and are part of the biosphere. We have come to recognize that we are in a situation of increasing interdependence, and that our future is intrinsically linked to the preservation of the global life-support systems and to the survival of all forms of life. The nations and the scientists of the world are called upon to acknowledge the urgency of using knowledge from all fields of science in a responsible manner to address human needs and aspirations without misusing this knowledge. We seek active collaboration across all the fields of scientific endeavour, that is the natural sciences such as the physical, earth and biological sciences, the biomedical and engineering sciences, and the social and human sciences. While the *Framework for Action* emphasizes the promise and the dynamism of the natural sciences but also their potential adverse effects, and the need to understand their impact on and relations with society, the commitment to science, as well as the challenges and the responsibilities set out in this Declaration, pertain to all fields of the sciences. All cultures can contribute scientific knowledge of universal value. The sciences should be at the service of humanity as a whole, and should contribute to providing everyone with a deeper understanding of nature and society, a better quality of life and a sustainable and healthy environment for present and future generations'.

www.unesco.org/science/wcs/eng/declaration_e.htm

ToK Link

The production of many electronic goods is concentrated in areas of the world where the working conditions may not be ideal. Should there be internationally set labour standards for all workers? What implications would this have on the cost of consumer goods?

'International standards' implies a common denominator – for example, a minimum wage for all workers. Resistance to this idea, for example in Singapore, is related to the implications for costs of production, competitiveness and, ultimately, the price of consumer goods. It is essentially a question of equity: are governments and societies prepared to make consumers (and employers) pay to improve the conditions of workers and the working environment? The International Labour Organization would answer in the affirmative, as would the European Union, but even these organizations need to be aware of the possible repercussions for businesses and consumers. Reason dictates that minimum wages cannot be so high that costs are prohibitive to effective business and competition, and the general working environment has to be conducive to work and not play. Indeed, respecting humanity and perceived human rights could come at the price of fewer jobs; and it should not be forgotten that workers are also consumers.

It is possible that improved working conditions might incentivise workers, raise productivity and output, and allow employers to pass on the benefits to consumers in the form of lower prices. This begs the question of how people react to higher levels of appreciation, particularly if they have been used to poor working conditions. Any international code of labour standards will not be productive if responsibility is not exercised on the part of workers, and would ideally ensure a balance between the well-being of workers and the cost of producing goods and services. The code has to be written and exercised within the parameters of ethical acceptability and economic reality, and this requires a shared vision of the needs and rights of the worker, consumer and employer. This vision will be qualitative rather than quantitative, and therein lies the problem.

However, there are a number of meeting points; for example, it is generally agreed that slave labour is unethical, although the consumer would reap the benefit of lower prices, or so it could be argued. Can it be said that many workers, including those who produce electronic goods, are akin to slaves? They might not be owned by a master, but they are treated as slaves. Human dignity is the issue here, and the tension between dignity and efficiency becomes all too apparent. To enslave human dignity – whether people are aware of it or not – is to reduce the human person to the status of a machine, and for this reason alone it can be argued that internationally set labour standards are justified, regardless of the cost to the buying capacity of consumers. This argument has particular weight when the abuse of human dignity is inflicted on a developing country by a corporation from a developed country. Why should consumers in a developed economy benefit from less than satisfactory working conditions in a lower-income economy?

Examination questions – a selection

Q1 A gold single crystal has a cubic shape and the dimension of the cube is 1.000 cm. When irradiated with X-rays of wavlength 154.05 pm at the angle of 10.89° it gives a well-defined first-order diffraction pattern. The molar mass of gold is 196.97 g mol^{-1}.

a Sketch the unit cell of gold and state the coordination number of the atoms and the number of atoms to which the unit cell is equivalent. [3]

b Deduce the number of gold atoms present in the cube:

i Calculate the lattice constant of gold. Use Bragg's law, $n\lambda = 2a\sin\theta$, where a is the lattice constant (length of the unit cell). [1]

ii Calculate the volume of the unit cell of gold. [1]

iii Calculate the number of unit cells present within 1.000 cm^3 of gold. [1]

iv Calculate the number of gold atoms present within 1.000 cm^3 of gold. [1]

c Deduce the mass of the unit cell of gold:

i Calculate the mass of one atom of gold. [1]

ii Calculate the mass of the unit cell of gold. [1]

d Calculate the density of gold. [1]

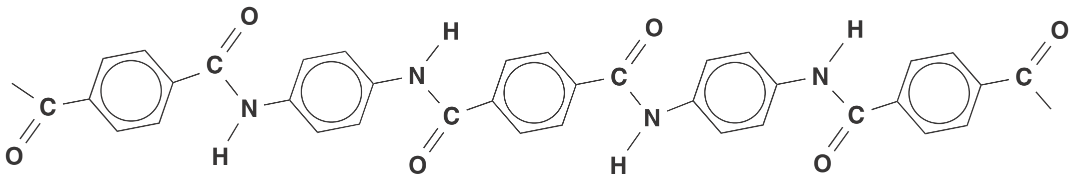
e Gold is used in nanocatalysts. Gold and some gold compounds are superconducting. Gold can be prepared by the electrolysis of solutions containing gold(I) or gold(III) ions using platinum electrodes.

i According to Bardeen–Cooper–Schrieffer (BCS) theory, Cooper pairs account for type 1 superconductivity. Describe how Cooper pairs are formed and the role of the positive ion lattice in this. [2]

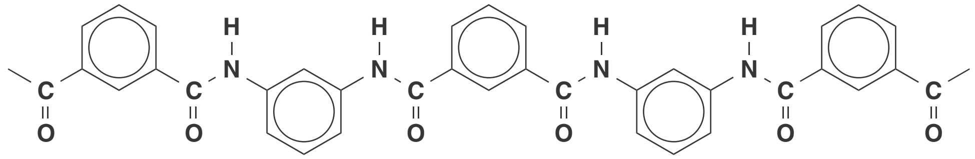
ii Outline two advantages of nanocatalysts over conventional catalysts. [2]

f Two electrolytic cells are connected in series. In cell 1 aqueous copper(II) sulfate is electrolysed and during electrolysis 0.965 g of copper metal is plated on to the cathode. Deduce the mass of gold metal plated on to the cathode of cell 2 simultaneously if the electrolyte used is aqueous gold(III) sulfate. [3]

Q2 The polymer Kevlar is used to make bullet-proof vests and cords to reinforce the walls of car tyres. Part of the structure of a Kevlar molecule and that of a related condensation polymer molecule called Nomex are shown:



Nomex



Kevlar

a Kevlar and Nomex are aromatic polyamides. Explain why these molecules can be described as aromatic. [1]

b i Draw the structural formulae of the two monomers from which Kevlar is made. [2]

ii Draw the structural formulae of the two monomers from which Nomex is made. [2]

c One reason why Kevlar is so strong is because of the close packing of the polymer molecules. State the type of bonds that operate between adjacent polymer chains. [1]

d Nomex has a lower tensile strength than Kevlar. Suggest a reason why. [2]

e Kevlar is insoluble in most solvents but dissolves in concentrated sulfuric acid. Explain how this happens. [2]

f Explain why addition polymers have higher atom economies than condensation polymers such as Kevlar. [2]

g State one characteristic of Kevlar that allows it to exist in a liquid crystal state. [1]

h Kevlar is used both as a raw fibre and in composites. Outline the meaning of the term composite. [2]

Q3 Different metal compounds are widely used in the production of stained glass windows.

a Both chromium(III) oxide and cadmium sulfide are incorporated into glass. Cr_2O_3 is green and CdS is red. Use section 8 of the *IB Chemistry data booklet* to calculate values to complete the table below. [2]

Compound	Chromium(III) oxide	Cadmium sulfide
Electronegativity difference		
Average electronegativity		

b Predict the bond type and percentage covalent character of each oxide, using section 9 of the *IB Chemistry data booklet*. [2]

Compound	Chromium(III) oxide	Cadmium sulfide
Bond type		
%covalent character		

c Waste water from leather factories may contain toxic chromium(III) ions. Its concentration may be estimated using inductively coupled plasma optical emission spectroscopy (ICP-OES). 50.0 cm^3 of waste water from a leather factory was made up to 1.00 dm^3 with distilled water and gave an emission intensity of 498 on an inductively coupled plasma spectrometer.

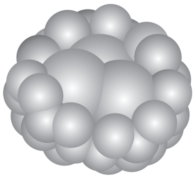
A standard aqueous solution of a chromium(III) salt with a concentration of 5.00 g dm^{-3} was used to prepare solutions of known concentration. Known volumes of the chromium(III) salt solution were diluted to 100 cm^3.

Volume of standard chromium(III) salt solution/cm^3	2.0	4.0	6.0	8.0	10.0
Emission intensity/ arbitrary units	143	285	427	569	709

i Calculate the concentration of chromium(III) ions in each of the calibration solutions. [2]
ii Plot a calibration curve with concentration along the *x* axis. [1]
iii Determine the concentration (mg dm^{-3}) of the chromium(III) ions in the sample solution. [1]
iv Determine the concentration (mg dm^{-3}) of the chromium(III) ions in the waste water. [1]

d The solubility product of K_{sp}, of silver(I) chromate(VI), Ag_2CrO_4 is 4 × 10^{-12} at 298 K. Determine its solubility (in mol dm^{-3}) at this temperature. [3]

Q4 The fullerenes are a newly discovered allotropic form of pure carbon. They are prepared in the laboratory from graphite and extracted by thermal sublimation.



Endohedral fullerenes are fullerenes that have additional atoms inside. Shown above is the interior of $Sc_3@C_{82}$. This consists of a C_{82} fullerene molecule enclosing three scandium atoms. Endohedral metallofullerenes are characterized by the fact that electrons will transfer from the metal atom to the fullerene cage. Fullerenes can be prepared by heating and vaporizing graphite in helium. They can be detected by mass spectrometry. The fullerenes can be separated from graphite in soot by heating the mixture until the fullerenes undergo sublimation. Nanotubes and graphene are two other well-studied carbon-based nanotechnology materials.

a i Write out the detailed electron configuration for scandium atoms, Sc, and scandium(III) ions, Sc^{3+}. [2]
ii State and explain whether scandium and scandium(III) compounds are diamagnetic or paramagnetic. [2]
iii Deduce the ionic formula for $Sc_3@C_{82}$. [1]

b i Suggest why fullerenes are prepared in an atmosphere of helium. [1]
ii Describe the chemical vapour deposition (CVD) method for the production of carbon nanotubes. [2]
iii Explain why C_{82} is predicted to have a higher sublimation point higher than C_{60}.[2]

c i State the type of hydridization of the carbon atoms present in carbon-60, carbon nanotubes and graphene. [1]
ii Explain why carbon-60 is a very poor electrical conductor, but graphene is a good electrical conductor. [2]

d Outline two possible hazards associated with nanotechnology. [2]

e Scandium(III) oxide is used in the manufacture of electronic ceramics and glass. It can be reduced by carbon and hydrogen. Scandium is often alloyed with aluminium.
i State the two main elements present in glass and quartz. [1]
ii Write equations showing the reduction of scandium(III) oxide by hydrogen and carbon. [2]
iii State two reasons why metals are often used as alloys rather than as pure metals. [2]

23 Option B Biochemistry

ESSENTIAL IDEAS

- Metabolic reactions involve a complex interplay between many different components in highly controlled environments.
- Proteins are the most diverse of the biopolymers and are responsible for metabolism and structural integrity of living organisms.
- Lipids are a broad group of biomolecules that are largely non-polar and therefore insoluble in water.
- Carbohydrates are oxygen-rich biomolecules which play a central role in metabolic reactions of energy transfer.
- Vitamins are organic micronutrients with diverse functions that must be obtained from the diet.
- Our increasing knowledge of biochemistry has led to several environmental problems, while also helping to solve others.

Additional higher level (AHL)

- Analyses of protein activity and concentration are key areas of biochemical research.
- DNA is the genetic material that expresses itself by controlling the synthesis of proteins by the cell.
- Biological pigments include a variety of chemical structures with diverse functions which absorb specific wavelengths of light.
- Most biochemical processes are stereospecific and involve only molecules with certain configurations of chiral carbon atoms.

23.1 Introduction to biochemistry *– metabolic reactions involve a complex interplay between many different components in highly controlled environments*

Biochemistry is the subject that seeks to find explanations at a molecular level for the key biological phenomena that sustain life. Living things extract and transform energy from their environment, build complex structures, and have the ability to self-regulate and self-replicate. Historically there were ideas that the chemistry of living matter was somehow composed of distinctive compounds separate from inorganic chemistry – summarized in the notions of 'vitalism'. We have seen how these ideas were confounded by the chemical synthesis of urea without the presence of any biological tissue (see Chapter 10). Following this synthesis Wöhler wrote to his mentor, Berzelius, saying: 'I must tell you that I can make urea without the use of kidneys, either man or dog. Ammonium cyanate is urea.' All living organisms are made of molecules that individually are no different from any other molecules found in non-living matter. There is still a subtle but general belief in the distinction between 'animal' and 'mineral' but contrary to this minerals and animals do not belong to separate kingdoms. Many minerals are produced in and by life, sometimes in crystalline form. One of the most common minerals, calcium carbonate, is formed by living marine animals such as coccolithophores as shells (Figure 23.1). When the organism dies, the plates separate and sink to the ocean floor. Individual plates have been found in vast numbers and can make up the major component of a particular rock, such as the chalk of England. *Emiliana huxleyi* is found in marine environments worldwide.

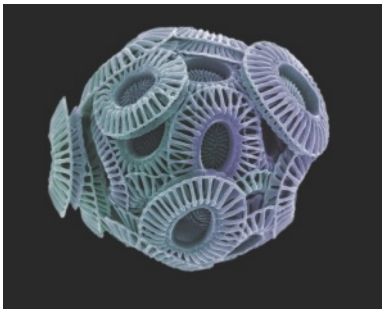
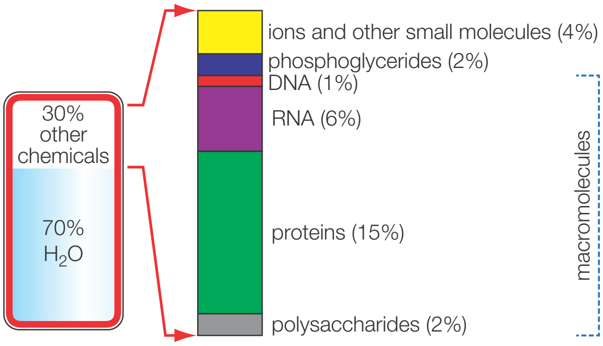


Figure 23.1 Coloured scanning electron micrograph (SEM) of *Emiliana huxleyi* – a small algal organism (coccolithophore) is surrounded by a skeleton of calcium carbonate plates

Another compound, calcium phosphate, is precipitated by the cells of human bones. All the different forms of living organism have members that produce minerals, with over 50 different minerals now known to be produced by living cells. While we can marvel at the distinct capabilities of different biological molecules, we should not forget how life is interwoven with the inanimate world.

There are many questions that remain unanswered in the study of the origin and nature of life but there is general agreement as to the medium from which life emerged. Life began in a water environment and a key event was the development of a membrane that enclosed a defined volume of that medium – a contained volume known as a **cell**. The chemistry of the living cell is predominantly the chemistry that can take place in an aqueous environment under relatively mild conditions of temperature and pH.

Figure 23.2 The chemical components of a bacterial cell



Humans, like all living matter, are mostly water. The chemistry of life at its most basic level is the chemistry of a spectrum of complex organic compounds made from a surprisingly limited number of elements (carbon, hydrogen, nitrogen, oxygen, sulfur, calcium, potassium and phosphorus are the only elements with a percentage by mass greater than 0.25 per cent in the human body), which together with a range of inorganic elements (zinc, magnesium, iron, iodine and copper, for instance) participate in carefully controlled reactions in just one solvent – water (Figure 23.2). The study of life from this perspective has revolutionized our view of ourselves and led to some of the most astounding scientific discoveries and advances of modern times.

The nature of metabolism

Cells are the basic unit of structure and function in all living organisms (Figure 23.3). Complex organisms contain vast numbers of cells, which together carry out all the life processes. Fibroblasts, for instance, are cells forming connective tissue, and are responsible for secreting connective tissue proteins such as collagen.

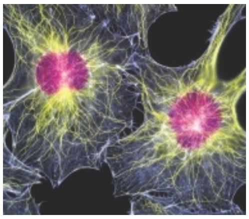


Figure 23.3 Fluorescent light micrograph of two fibroblast cells, showing their nuclei (purple) and cytoskeleton. The cytoskeleton is made up of microtubules of the protein tubulin (yellow) and filaments of the protein actin (blue)

At any one time within the microscopic volume of a living cell, thousands of chemical reactions occur involving a range of biochemically important molecules that have a series of diverse roles. These roles include:

- Controlling self-replication and carrying genetic information – DNA (deoxyribonucleic acid) molecules determine who we are; the sequence of nitrogenous bases in DNA molecules constitutes the genetic code while RNA (ribonucleic acid) molecules are involved in the transmission of the genetic message into proteins.
- The conversion of light energy to chemical energy in the process of photosynthesis – for example, one pigment, chlorophyll, facilitates this because it has an extended system of alternating single and double bonds.
- Energy production and storage – fats and sugars are involved in reactions that convert chemical energy to other forms of energy. The energy living things absorb from their environment is used to synthesize their own complex structures and to carry out functions such as movement and reproduction. The structure of glucose allows it to be soluble in water so that it can provide a readily available source of energy.
- Containing catalysts required for biochemical reactions – enzymes are proteins that accomplish this role with specificity for a particular reaction, or type of reaction, that is greater than inorganic catalysts (Figure 23.4). Their complex three-dimensional structure can accommodate the bonding of other molecules for the purpose of both catalysis and control.

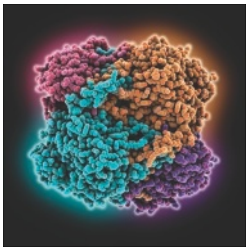


Figure 23.4 Human catalase, molecular model. This enzyme, found in the liver, kidney and blood, catalyses the breakdown of hydrogen peroxide to water and oxygen

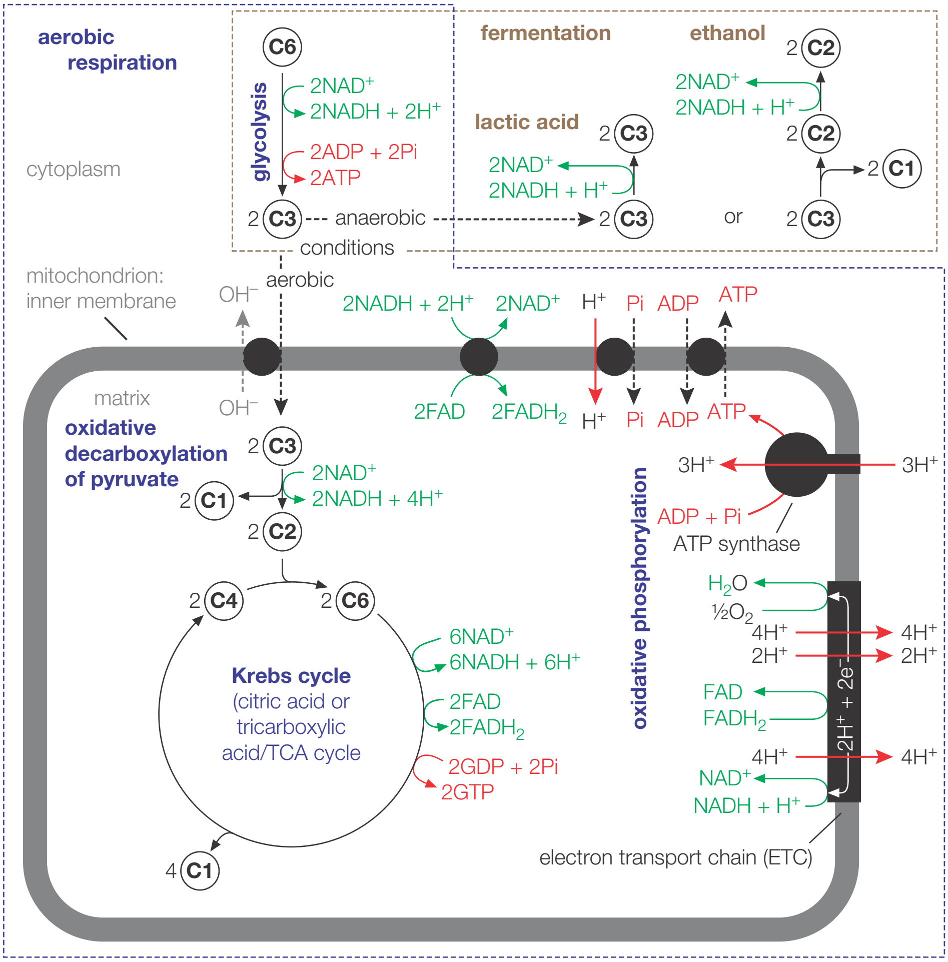
- The provision of structural support in cells (the cytoskeleton in Figure 23.3) and tissues – the cytoskeleton supports the cell's structure, allows the cell to move and assists in the transport of organelles and vesicles within the cell. In plants, cellulose is a polymeric carbohydrate made up of long chains of sugar molecules that pack together closely to provide structure and support.

These are just some of the roles carried out by biological molecules and these and other functions of biological molecules will be discussed in subsequent sections.

The sum of all these reactions taking place in an organism is known as **metabolism**. This complex chemistry is dependent on a high level of order where every compound has a distinct function. Some features of metabolism are:

- Reactions are controlled in sequences and cycles known as **metabolic pathways**. The product of each step is the reactant for the next. Compounds taking part in metabolism are known as **metabolites**. Figure 23.5 shows a small section of the many metabolic pathways in the human body, focused on those involved in **aerobic respiration**. Note the overall pattern of linked reactions and pathways, and the compartmentalization of certain pathways in the cytoplasm or mitochondria.
- Every reaction is controlled by a specific catalyst called an enzyme.
- Similar pathways and enzymes exist in a wide range of different organisms.
- Reactions can be coupled so that energy from one reaction is used to drive another.
- Many reactions are reversible, which allows for feedback control of a pathway or cycle. The products of a metabolic pathway may exert control on the overall pathway. For example, the adenosine triphosphate (ATP) produced in respiration can act as an inhibitor for one of the enzymes involved in respiration, so that ATP controls its own production. This is an example of end-product inhibition.

Figure 23.5 A small section of the many metabolic pathways in the human body, focused on those involved in aerobic respiration



Given the complexity of metabolism, it is useful to classify pathways according to their broad purpose. Later in this chapter, specific examples of metabolic processes will be discussed in more detail.

Anabolic pathways

Anabolism is the process of synthesizing molecules needed by cells – it requires energy. The reactants are small molecules called precursors, and the products are larger, more complex molecules of higher energy. Anabolic pathways therefore require energy. Examples include the synthesis of proteins from amino acids, nucleic acids from nucleotides, and carbohydrates from the process of photosynthesis.

Catabolic pathways

Catabolism is the breakdown of larger molecules into smaller ones with the release of energy. Catabolic reactions release energy and produce energy-poor end-products, such as carbon dioxide and water. Examples include the breakdown of glucose in respiration or the oxidation of fatty acids. Catabolism results in the production of ATP, and ATP is used as an energy source in anabolism.

Metabolic reactions occur in a highly controlled aqueous environment in the nuclei, mitochondria and cytoplasm of cells. Slight changes in pH and temperature, for example, can have large effects on the structures of biomolecules and the reactions that occur during metabolism.

Photosynthesis

An example of an anabolic process is **photosynthesis**. This is an endothermic reaction using light energy from the Sun to produce energy-rich molecules (glucose) from carbon dioxide and water. The overall reaction is:

$$6CO_2(g) + 6H_2O(l) \rightarrow C_6H_{12}O_6(aq) + 6O_2(g)$$

Green plants (and some other organisms) are able to capture solar energy and use it to synthesize energy-rich biomolecules. The key to photosynthesis is the absorption of sunlight by photosynthetic pigment molecules. The primary pigment is chlorophyll. The light energy trapped drives a series of redox reactions, in which remarkably water is split into hydrogen and oxygen. The oxygen is released as a waste product and the hydrogen ultimately reduces carbon dioxide to simple sugar molecules. This is a highly complex process – summarized in the equation above – which involves many intermediates, electron carriers and enzymes. In essence, photosynthesis transforms the energy-poor molecules carbon dioxide and water into energy-rich sugars with the release of oxygen (Figure 23.6).

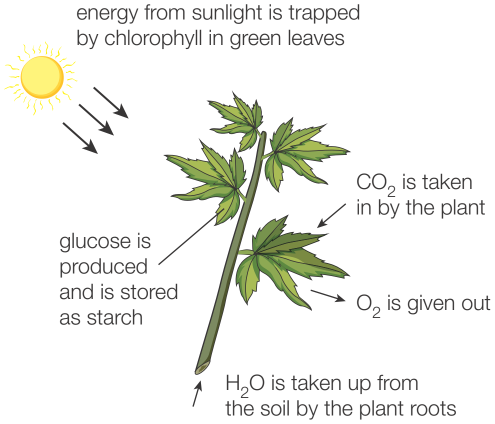

energy from sunlight is trapped by chlorophyll in green leaves
CO_2 is taken in by the plant
glucose is produced and is stored as starch
O_2 is given out
H_2O is taken up from the soil by the plant roots

Figure 23.6 An overview of photosynthesis, showing the importance of sunlight as the source of energy

Respiration

An example of catabolism is the breakdown of glucose, in a series of complex steps, into carbon dioxide and water with the release of energy (in the form of adenosine triphosphate, ATP). This process is called **respiration** and the overall equation can be represented as:

$$C_6H_{12}O_6(aq) + 6O_2(g) \rightarrow 6CO_2(g) + 6H_2O(l)$$

This reaction provides the energy for other processes in cells.

To recap, photosynthesis is carried out in a range of organisms from plants to bacteria and is the ultimate source of chemical energy in all the food we eat. Respiration is then carried out by all organisms to produce energy from food. The first stage of respiration, known as glycolysis, is common to all cells and does not use oxygen. Only a small proportion of the energy in glucose is released, as most is trapped in the products of this stage. Under anaerobic conditions,

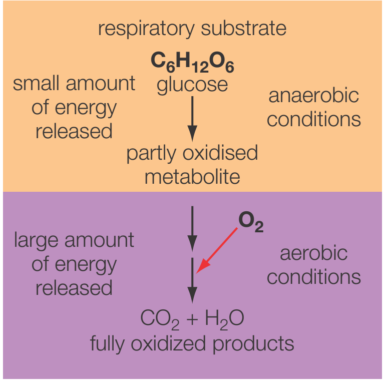


Figure 23.7 The anaerobic and aerobic phases of respiration

in the absence of oxygen, this is the only energy released, and it is enough to keep some cells alive, temporarily in the case of muscle cells and permanently in the case of some bacteria. Products of anaerobic respiration such as lactate and ethanol are therefore energy-rich molecules (Figure 23.7).

In the presence of oxygen (aerobic conditions) the oxidation of glucose is complete, and much more energy is released. This is why most cells are dependent on a continuous supply of oxygen. The end-products of aerobic respiration are the energy-poor molecules carbon dioxide and water.

Aerobic respiration involves a series of coupled redox reactions, where reactants known as cytochromes are successively reduced and then re-oxidized. Ultimately, oxygen acts as the terminal electron acceptor when it is reduced to water. The overall equation for the aerobic respiration of glucose was shown earlier. Note that the reactants and products are the same as for the combustion of glucose.

Photosynthesis and respiration are opposites – for every six molecules of carbon dioxide removed from the atmosphere in photosynthesis, six molecules are added by respiration. The same is true for oxygen in the opposite direction. Respiration is often likened to burning a fuel in oxygen, though in reality it is a much more complex and highly controlled process. Similar to burning, it does involve reactions of oxidation where the amount of energy released depends on the extent of oxidation achieved. Indeed the comparison shows how biological systems achieve the equivalent of a reaction from the everyday world under mild and controlled conditions.

These two processes of photosynthesis and respiration help to maintain the balance between carbon dioxide and oxygen in the atmosphere. Of course, there are things that disrupt this balance – for instance, burning fossil fuels, which releases carbon dioxide that was removed from the atmosphere by photosynthesis millions of years ago.

The discussion of photosynthesis and respiration places emphasis on the chemical processes known as **reduction** and **oxidation**. In the next section we consider two other important reaction types: those involving condensation and hydrolysis.

1 Define reduction and oxidation in terms of the following possible changes:
a Electron transfer
b Gain or loss of hydrogen
c Gain or loss of oxygen
d A change of oxidation state.

2 In human cells under certain conditions a series of metabolic processes can result in the following overall reaction:

$CH_3C(O)COOH$	+	$CH_3CH(OH)CH_3$	→	$CH_3CH(OH)COOH$	+	$CH_3C(O)CH_3$
2-oxopropanoic acid (or pyruvic acid)		propan-2-ol		2-hydroxypropanoic acid (or lactic acid)		propanone

a State which of the two reactants has been oxidized and which reduced.
b Deduce redox half-equations for the two processes. Use protons or water molecules to balance the equations where necessary. The two half-equations must produce the overall equation above when added together.

Atmospheric variation

Data from the Mauna Loa research station in Hawaii has provided a record of atmospheric carbon dioxide concentration since 1958. Taking into account seasonal variation, the overall trend has been an increase in carbon dioxide levels from about 315 ppm to about 390 ppm in the past 50 years. Carbon dioxide is a greenhouse gas that absorbs and re-radiates infrared radiation, and so this dramatic increase in its concentration is widely accepted as a major factor influencing climate change. Increased combustion of fossil fuels is seen as a major contributor to the progressive elevation of carbon dioxide levels, but deforestation in various regions of the world has also contributed and also resulting in prolonged periods of haze and smog in certain

■ **Figure 23.8** Living bacterial structures called stromatolites at Shark's Bay, Australia represent remaining colonies of photosynthesizing cyanobacteria

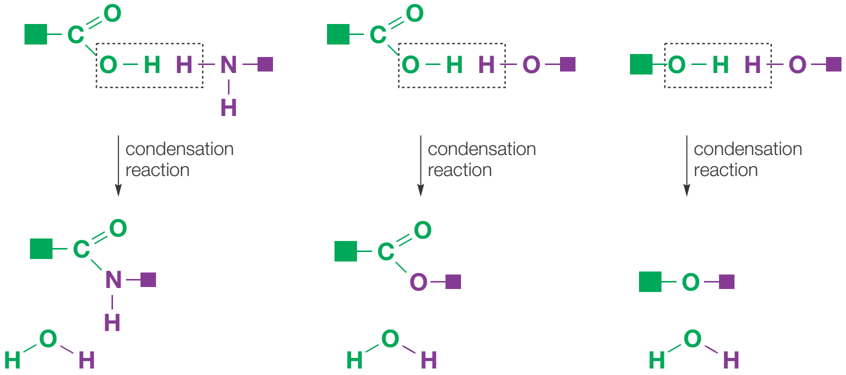
parts of the world. Part of the drive for burning of forest areas has been the need to supply beef and other animal products to the increasing human population, and this emphasizes the need for international cooperation and action to control this type of development.

Our concern over atmospheric pollution and climate change and the extent to which we can exert control and reverse or ameliorate current trends is absolutely necessary. But it is also worth pausing to consider earlier pollution events in the life of the planet. The pollution crisis caused by the rise of photosynthetic organisms (cyanobacteria) in pre-history was far greater in its effect than current developments. The degree of change on the global environment brought about by the development of the cyanobacteria (Figure 23.8) was very dramatic, introducing the highly dangerous gas oxygen into the Earth's primitive atmosphere. The remaining colonies of these organisms represent the origins of photosynthesis in the development of the planet and a reminder that pollution is a natural phenomenon.

■ Condensation and hydrolysis reactions

These are two extremely important reactions in the build-up and breakdown of molecules in biological systems.

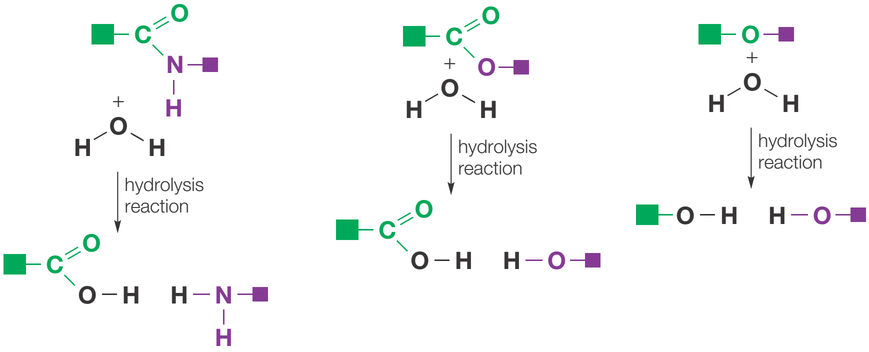
Condensation is the joining together of two molecules with the formation of a covalent bond and the elimination of a small molecule, usually water, as the bond is made (Figure 23.9).



■ **Figure 23.9** The condensation reactions involved in the formation of proteins, lipids and polysaccharides

Amino acids (to form proteins), carboxylic acids with alcohols (to form the esters present in lipids) and sugar molecules (to form disaccharides and polysaccharides) all undergo condensation reactions.

A **hydrolysis** reaction is essentially the reverse of condensation – it involves breaking a covalent bond by the addition of water (Figure 23.10).



■ **Figure 23.10** Hydrolysis reactions involve the breaking of a covalent bond by the addition of the fragments of water (H– and –OH) across the bond

The breakdown of these molecules reverses a condensation reaction, adding a molecule of water for each covalent bond broken. The water is split with –H and –OH attaching separately to the product molecules. These are hydrolysis reactions and occur during chemical digestion. Biologically they are catalysed by enzymes, but can also be favoured by heat and acidic or alkaline conditions when extracted molecules are subjected to analysis.

As mentioned, examples of condensation and hydrolysis reactions include the synthesis and breakdown of proteins from amino acids and polysaccharides from sugars. Lipids, although they do not form polymers, also involve condensation and hydrolysis reactions between their sub-components. These are all discussed in more detail later in this chapter..

Nature of Science

A case of rapid development

The study and development of biochemistry is a prime example of how scientific knowledge is ever changing and rapidly growing. It is an extremely complex field and has developed into several different specialisms such as molecular biology. The double-helix structure of DNA was discovered only 60 years ago and the use of new techniques such as X-ray crystallography has allowed the elucidation of this structure as well as the structures of many other complex biological molecules. We can now routinely sequence both protein and DNA molecules and aspects of biochemistry have become the focus of 'big science' with the mapping of the human genome and that of other organisms.

Some 60 years ago there were still those who questioned the existence and role of messenger RNA in the cellular synthesis of proteins and yet now we are able to 'engineer' proteins in foreign organisms demonstrating that key processes have been conserved over a wide range of species.

Data is extremely important in science, and the collection of large amounts of data using a variety of instruments and techniques has meant that our knowledge and understanding of the reactions that occur in metabolism has increased greatly over the past 50 years. As more and more data has been obtained on biochemical pathways, similar reaction patterns are seen in metabolic processes in species that may not be closely related. For this reason biochemistry is now a major part of the study of evolutionary biology. This is an example of how an interdisciplinary approach can contribute to deeper knowledge and understanding.

Pathway engineering

One area that illustrates the expansion of ideas possible in a growing field of research is the development of 'pathway engineering'. Pathway engineering (Figure 23.11) is defined as the improvement of metabolic pathways by the manipulation of the enzyme and regulatory functions of the cell using recombinant DNA technology. It specifically seeks to mathematically model the metabolic networks, calculating a yield of useful products and pinpointing parts of the network that constrain the production of these products. Genetic engineering techniques can then be used to modify the pathways within a suitable microorganism in order to relieve these constraints. This modified metabolic pathway can be modelled to calculate the new product yield.

Pathway engineering is viewed as a highly positive development as it has the potential to produce from simple, readily available, cheap starting materials a large number of chemicals that are currently derived from non-renewable resources or limited natural resources. Common target pathways have been glycolysis, the Krebs cycle and amino acid biosynthesis. The target molecules are usually at major branch points within these metabolic pathways.

Pathway engineering has been carried out in a range of bacteria and yeasts. In yeasts, pathway engineering has been used to develop yeast strains that can ferment xylose (a breakdown product of cellulose) anaerobically. These new strains can ferment hydrolysed cellulose as a feedstock, which is available in large amounts. The new strain of yeast has a new gene for xylose metabolism introduced by genetic engineering techniques.

A simple example of pathway engineering is the introduction of a new metabolite by inserting a new gene into a microorganism. For example, 1,3-propanediol, a useful intermediate in the synthesis of polyurethanes and polyesters, is now being produced from glucose by *Escherichia coli* engineered with genes from the yeast *Saccharomyces cerevisiae*. Recent

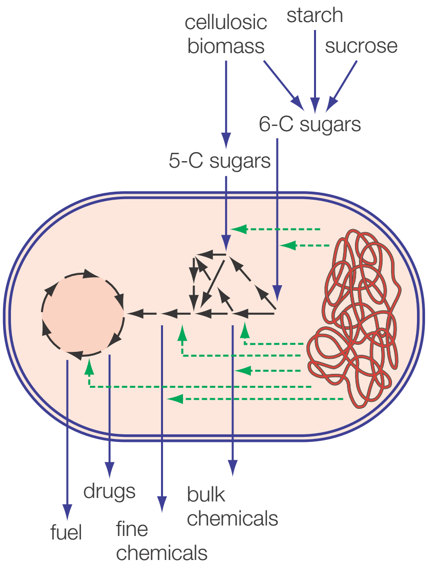


Figure 23.11 A summary of the approach to pathway engineering indicating the use of readily available feedstock resources and genetic engineering techniques to produce desired products

advances in pathway and protein engineering have made it possible to develop micro-organisms that produce hydrocarbons with properties similar to those of petroleum-derived fuels and thus compatible with our existing transportation infrastructure. Linear hydrocarbons (alkanes, alkenes and esters) typical of diesel and jet fuel can be produced by the fatty acid biosynthetic pathway.

Dietary health

The supply and availability of food have been important factors shaping the emergence and development of human civilizations for thousands of years. For the past few decades food has been cheaper in real terms and more readily available than probably at any time in history. However, it cannot be said we have a functioning global food system when one in seven people today still do not have access to sufficient food and hence are malnourished, and an equal number are over-fed and many obese. There are a number of known threats to the food system and factors that will increase the risks of a rise in hunger. Population and consumption growth will lead to the demand for food increasing for most of the current century, while increasing competition for land, water and other resources threaten the supply of food.

Many differences in human health across the world are the result of differences in the supply of nutritious food. Metabolism in the human body is dependent on a regular dietary supply of a range of diverse nutrients. For instance, proteins are the main source of amino acids for the body and so they must be present in sufficient quantity in a healthy diet. Protein deficiency results in various diseases that are widespread in different parts of the world. One of these diseases, kwashiorkor, is characterized by a swollen stomach, skin discoloration and retarded growth.

Protein deficiency in the diet is significant because, while certain 2-amino acids can be synthesized by the body, other proteogenic amino acids must be supplied in the diet. These amino acids are referred to as **essential amino acids**, and include histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), threonine (Thr), tryptophan (Trp) and valine (Val). Certain other amino acids become dietary essentials at different stages of growth. Thus, arginine, cysteine and tyrosine must be present in the balanced diet of infants and growing children. Other amino acids, hydroxyproline and hydroxylysine, synthesized post-translationally in the polypeptide chain are important for the structure of connective tissue protein and require adequate levels of vitamin C in the diet.

The above illustrates the importance of diet and it is distressing to see the effects of deprivation in different parts of the world, particularly where children are involved. However, even in relatively wealthy countries there can be regional differences in life expectancy and general health that are thought to be dependent on local provision and dietary preference. Such situations place pressure on local health and education services when attempting to rectify the regional imbalance.

23.2 Proteins and enzymes *– proteins are the most diverse of the biopolymers responsible for metabolism and structural integrity of living organisms*

Proteins make up approximately 18 per cent of the mass of an average person (Figure 23.12).

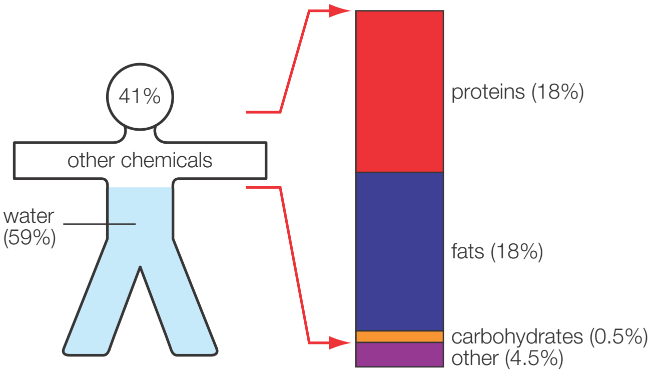


Figure 23.12 A summary of the chemical components of the average person

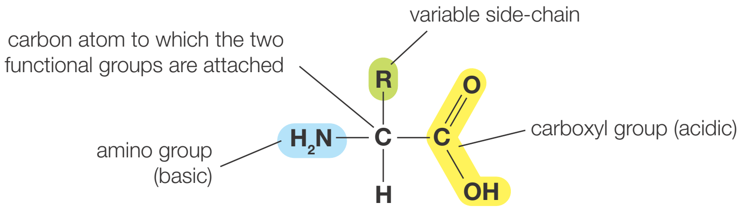
Many proteins, such as antibodies, enzymes and hemoglobin, are water-soluble molecules and can be transported easily within and between cells and tissues. Others, such as collagen and keratin, are insoluble and form very tough, stable structures. Some proteins of the human body and their functions are listed in Table 23.1.

■ **Table 23.1** Some proteins of the human body and their functions

Protein	Function	Where found in the body
Collagen	Structural protein, gives strength and elasticity	Skin, tendon
Keratin	Structural	Hairs, nails
Myosin	Muscle contraction	Muscle tissue
Actin	Muscle contraction	Muscle tissue
Chymotrypsin	Digestive enzyme, breaks down food	Small intestine
Pepsin	Digestive enzyme, breaks down food	Stomach
Insulin	Hormone, enables use of glucose for energy	Blood, pancreas
Immunoglobulins	Antibodies	Blood, lymph
Hemoglobin	Oxygen transport between lungs and rest of body	Blood

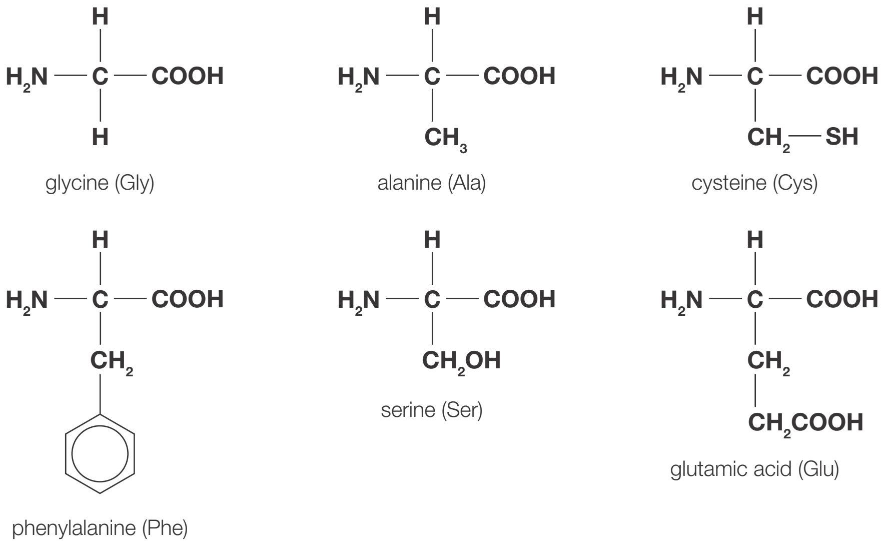
Proteins are linear polymers composed of subunits called **amino acids**. The 20 biologically important 2-amino acids (often referred to as α-amino acids, as the amino group is attached to the first carbon after that of the acid grouping in the chain) have the general formula shown in Figure 23.13. A central carbon atom is attached in turn to a hydrogen atom (–H), an amino group (–NH_2), a carboxylic acid functional group (–COOH) and a variable side-chain referred to as R. In the simplest amino acid, glycine, R represents a hydrogen atom. Amino acids are amphoteric; that is, they are able to act as acids and bases in aqueous solutions.

■ **Figure 23.13** Generalized structure of a 2-amino acid



The variable side-chain (R) has a different structure in the 20 amino acids (Figure 23.14) and determines the individual chemical and physical properties and shapes of the folded proteins.

■ **Figure 23.14** A selection of naturally occurring amino acids



A complete list of all the amino acids used in proteins is given in Section 33 of the *IB Chemistry data booklet*. Amino acids can be classified according to the chemical nature of their R group, usually on the basis of their different polarities, as shown in the examples in Table 23.2.

Table 23.2 Some examples of the different types of R groups in amino acids

Type of amino acid	R group contains	Example	Structure
Non-polar/hydrophobic	Hydrocarbon chain	Alanine, Ala	$H_2N-CH-COOH$ $\quad\quad\ \ \vert$ $\quad\quad CH_3$
Polar, uncharged	Hydroxyl (–OH) sulfhydryl (–SH) or amide (–$CONH_2$) groups	Serine, Ser	$H_2N-CH-COOH$ $\quad\quad\ \ \vert$ $\quad\quad CH_2-OH$
Basic (positively charged at pH 6.0–8.0)	Amino group (–NH_2)	Lysine, Lys	$H_2N-CH-COOH$ $\quad\quad\ \ \vert$ $\quad\quad CH_2-CH_2-CH_2-CH_2-NH_2$
Acidic (negatively charged at pH 6.0–8.0)	Carboxylic acid group (–COOH)	Aspartic acid, Asp	$H_2N-CH-COOH$ $\quad\quad\ \ \vert$ $\quad\quad CH_2-COOH$

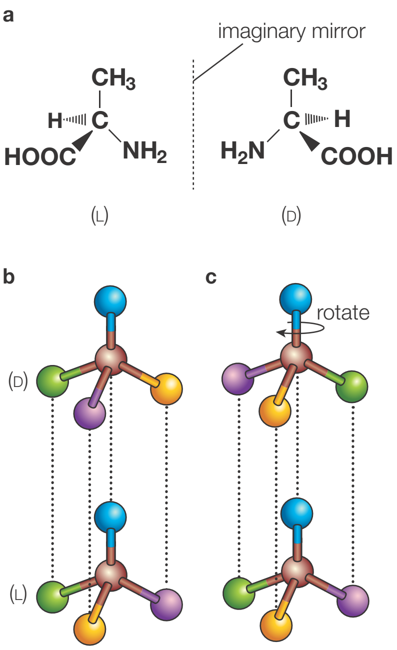


Figure 23.15 a The enantiomers of alanine are mirror images of each other. **b** The two forms are not superimposable on each other – if two groups are in position, the other two are not. **c** If the top model is rotated it cannot be made to fit with the bottom one

Enantiomers of amino acids

The central carbon atom is a chiral centre in all the 2-amino acids, except for glycine. The two possible enantiomers (Figure 23.15) are labelled D- and L- (see Chapter 20). All the naturally occurring amino acids found in proteins have the L-configuration, although the actual direction in which they rotate the plane of plane-polarized light depends on the nature of the variable side-chain, R. This finding serves to emphasize how tremendously important molecular shape is in biological systems.

The 2-carbon atom in alanine and the other amino acids is a chiral centre because it has four different atoms or groups attached to it. As a result, two different enantiomers, or optical isomers, can exist containing the same groups. These two isomers are mirror image forms of each other which cannot be superimposed on each other (Figure 23.15b and c). The crucial point here is that biological organisms use only one of these isomeric forms to build proteins.

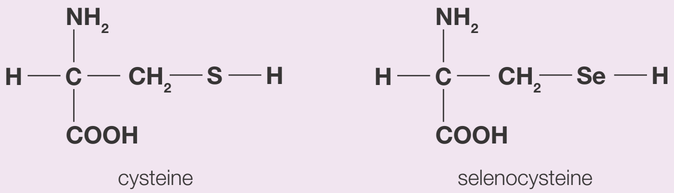
3 a Glycine is the simplest amino acid. What is the IUPAC systematic name for glycine?
b Draw the structural formula of glycine. Glycine is the only α-amino acid that does not contain a chiral centre. Explain why.
c Which type of stereoisomerism is not shown by glycine but is shown by other α-amino acids? State two physical properties of glycine.
d Write equations showing the reactions of glycine with dilute hydrochloric acid and methanol.

Additional Perspectives

Non-standard amino acids

There are 20 'standard' amino acids (i.e. those specified by the genetic code and incorporated into proteins under genetic control). Some bacteria use the non-standard amino acid selenocysteine (Figure 23.16), which is similar to cysteine except that it has selenium in the place of sulfur. In humans some antioxidant enzymes are selenoproteins. Selenium is therefore a biologically essential trace element.

Figure 23.16 Structures of the amino acids cysteine and selenocysteine



In addition β-alanine, 3-aminopropanoic acid, is a normal metabolite in the human body but is a non-chiral molecule.

Formation of zwitterions

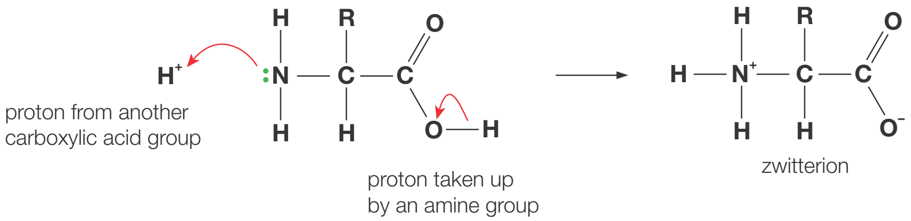
Crystalline 2-amino acids have relatively high melting or decomposition points. They are more soluble in water than in non-polar solvents, such as hexane. In these ways amino acids show the physical properties characteristic of ionic compounds. These observations and other experimental data suggest that amino acids exist as dipolar ions, known as **zwitterions** (Figure 23.17). A zwitterion (from the German word *zwitter* – 'hybrid') is a chemical compound that is electrically neutral but carries formal positive and negative charges on different atoms.

The ionic nature of amino acids explains the fact that they are fairly solubility in water – the degree of solubility depends on the nature of the R group. This relatively high solubility in water can be explained by the fact that in aqueous solution amino acids can be readily solvated (hydrated) by water molecules. The solubility in water is affected by pH, with solubility generally being lowest around the isoelectric point but higher at other pH values because the amino acid then has an overall charge.

Their low solubility in non-polar organic solvents can also be explained in terms of the zwitterion – the energy needed to overcome the stronger electrostatic forces between the ions cannot be paid back by the formation of relatively weak London forces between the amino acid and the organic solvent molecules.

The electrostatic forces of attraction between oppositely charged functional groups account for the high melting or decomposition points. The zwitterion formation can be regarded as an internal acid–base reaction: the zwitterions are sometimes referred to as **internal salts**, as the charges result from this internal acid–base reaction, with the transfer of a proton (H^+) from the acidic –COOH group to the basic –NH_2 group in the same amino acid.

Figure 23.17 Zwitterion formation by an amino acid


proton from another carboxylic acid group
proton taken up by an amine group
zwitterion

As amino acids contain both an acidic group and a basic group, they are **amphoteric** or **amphiprotic** (see Chapter 8). In aqueous solution, the zwitterion form of an amino acid can behave as either an acid (proton donor) or a base (proton acceptor).

4 a Name and give the formulas of the two functional groups present in all 2-amino acids.
b Which one of the 20 amino acids found in proteins has an unusual structure involving one of these functional groups? What is distinctive about its structure?
c Which 2-amino acids have non-polar R groups?
d Which 2-amino acids have polar R groups?
e Which two 2-amino acids contain acidic groups in their R groups?

5 Lysine contains an amino group in its R group.
a What feature of an amino group means that it can accept a proton (H^+ ion) under certain conditions?
b What property does the R group of lysine show as a result of the presence of the amino group?

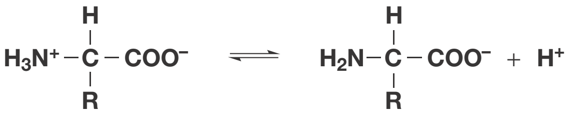
6 2-Aminocarboxylic acids (2-amino acids) have the general structure shown in Figure 23.13, where R represents a variety of different side-chains.
a Re-draw this structure, showing the main form in which this general amino acid would exist when in aqueous solution at pH 7.
b What is the overall charge on an amino acid in aqueous solution at pH 7 if R is non-polar?

7 Alanine is a white crystalline solid that readily dissolves in water. What special features about the structure of the molecule account for:
a Its crystalline nature?
b Its solubility in water?

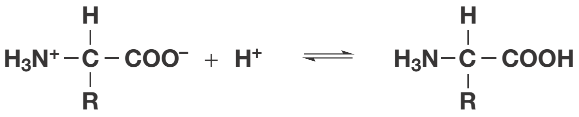
The isoelectric point of amino acids

In aqueous solution at around pH 7 both the acid and amino groups of an amino acid molecule can be ionized. In aqueous solution, this zwitterion form can behave as either an acid (proton donor) or a base (proton acceptor). As a consequence, these molecules show the properties of both an acid and a base – they are amphoteric.

- In alkaline solution (high pH/low H^+ ion concentration), the zwitterion can donate a proton (an H^+ ion), and the anionic form is favoured:



- In acid solution (low pH/high H^+ ion concentration), the zwitterion can accept a proton, and the cationic form is favoured:



The precise charged form of a particular amino acid depends on the pH of the solution. At low pH (high H^+ concentration) only the amino groups are charged ($–NH_3^+$). Conversely, at high pH (low H^+ concentration) only the acid groups carry a charge (–COOH) (Figure 23.18).

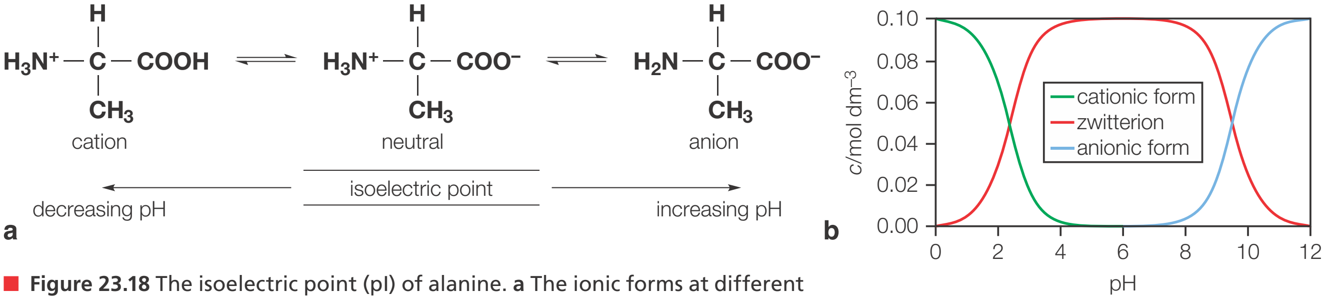


Figure 23.18 The isoelectric point (pI) of alanine. **a** The ionic forms at different pH. **b** The acid–base equilibria in a 0.1 mol dm^{-3} solution (pI = 6.0)

So amino acids tend to be positively charged at low pH and negatively charged at high pH. The intermediate pH at which the amino acid is electrically neutral is known as its **isoelectric point (pI)**. With no net charge at this pH, amino acids will not move in an electric field. Also at this point, the molecules will have minimum mutual repulsion and so be the least soluble. Different amino acids will have different isoelectric points, particularly those with additional amino groups or carboxylic acid groups in their side-chains. Section 33 in the IB *Chemistry data booklet* gives the pH of the isoelectric point of each amino acid alongside its structure, and some examples are shown in Table 23.3.

Table 23.3 The isoelectric points of several amino acids

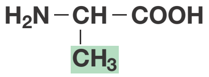
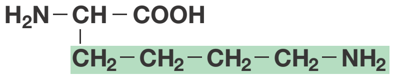
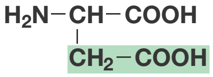
Name of amino acid	Short form	Structure	Isoelectric point
Glycine	Gly	H_2N-CH_2-COOH	6.0
Alanine	Ala		6.0
Lysine	Lys		9.7
Aspartic acid	Asp		2.8

Table 23.3 shows that amino acids such as alanine and glycine, which have uncharged R groups, have the same isoelectric point of pH 6.0. But where the R group contains an acidic or a basic group, the presence of these groups will also influence the charge as pH changes. This is why aspartic acid and lysine, for example, have very different isoelectric points. The isoelectric effect is the basis for a separation technique known as **electrophoresis** in which the different amino acids can be separated on the basis of their movement in an electric field**.** The word electrophoresis is derived from the term *electro-* and the Ancient Greek word *phoresis*, 'the act of bearing'.

The ability of amino acids or proteins to exist in various forms and neutralize both strong acids and strong bases is important in maintaining the acid–base balance in living organisms (see Section 23.7).

8 The isoelectric point (pI value) of the amino acid serine (2-amino-hydroxypropanoic acid) is 5.7. Draw the main structural formula of serine, in the solid state and in an aqueous solution at pH values of 1, 14 and 5.7.

Electrophoresis

As we have seen in our discussion, in a solution of an amino acid in water, the average charge on the amino acid molecules in the solution depends on the solution pH. This is because of the amphoteric properties of amino acids. The pH at which the net or overall charge is zero is called the isoelectric point of that amino acid.

The isoelectric point of glycine is 6.0. In aqueous solution at pH values greater than 6.0, the average charge on glycine molecules becomes negative (the anion or acidic form predominates); at pH values less than 6.0, the molecules become positively charged (the cation or basic form predominates) (Figure 23.19).

Figure 23.19 Behaviour of glycine molecules at different pH values

$$\underset{\text{at pH > 6.0}}{NH_2CH_2COO^-} \xleftarrow{-H^+} \underset{\text{at pH = 6.0}}{{}^+NH_3CH_2COO^-} \xrightarrow{+H^+} \underset{\text{at pH < 6.0}}{{}^+NH_3CH_2COOH}$$

Depending on the nature of the variable side-chains, different amino acids have isoelectric points at different values. Basic amino acids, such as lysine (Lys), have a tendency to form cations if dissolved in water and have isoelectric points greater than 7. Acidic amino acids, such as aspartic acid (asp), have a tendency to form anions if dissolved in water and have isoelectric points less than 7 (see Table 23.3).

If glycine, lysine and aspartic acid were dissolved together in the same buffer solution maintained at pH 6.0, the molecules of glycine would be neutral, the molecules of lysine would be positively charged and the molecules of aspartic acid would become negatively charged (Figure 23.20).

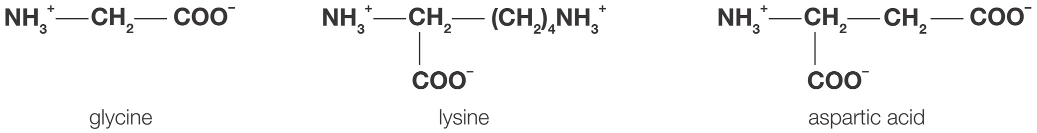


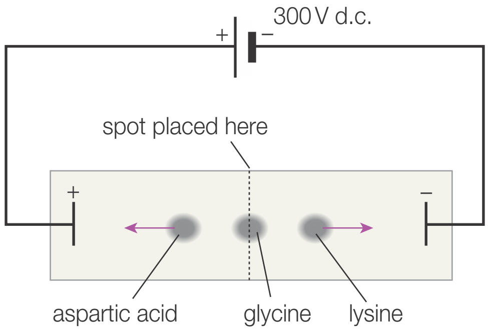
Figure 23.20 The ions formed from glycine, lysine and aspartic acid at pH 6.0

This mixture of amino acids can be separated by the process of electrophoresis, which involves separating the molecules (dissolved in a buffer and strong electrolyte) by means of an electric field (generated by a voltage). Traditionally, electrophoresis was performed on paper, but modern approaches use a variety of natural and synthetic gels.

The samples are spotted in the centre of the paper or gel, a high voltage is applied, and the spots migrate according to their charges. After electrophoresis, the separated components can be detected by a variety of staining techniques, depending upon their chemical identity.

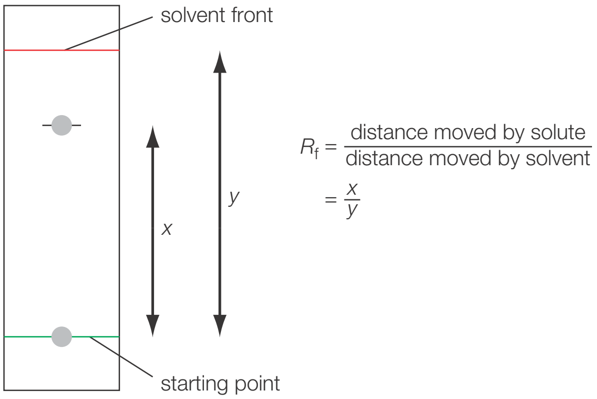
If the mixture of glycine, lysine and aspartic acid were subjected to electrophoresis, the lysine molecules would migrate to the cathode, the aspartic acid molecules would migrate to the anode and the glycine molecules, having no overall charge, would not move at all (Figure 23.21).

Figure 23.21 Electrophoresis of a mixture of glycine, lysine and aspartic acid



Chromatography

The amino acids present in a protein can be separated by chromatography. To do this, the protein is completely hydrolysed by hydrochloric acid. Samples of the hydrolysed protein are then placed on to chromatography paper. The cellulose molecules of the paper fibres are hydrated and this water remains bonded to the cellulose, forming a stationary phase for the partition chromatography. The solution of amino acids and excess hydrochloric acid moves up the vertically held paper by capillary action. The amino acids become distributed between the two phases according to their differing solubilities in the stationary and mobile phases. At the end of the separation process the different amino acids have moved different distances from the origin. Amino acids are colourless, so once the process is complete the paper is sprayed with ninhydrin. A chemical reaction occurs and coloured spots are formed. Ninhydrin is known as a locating agent. Commercial samples of appropriate amino acids are used as standards and run alongside the samples.

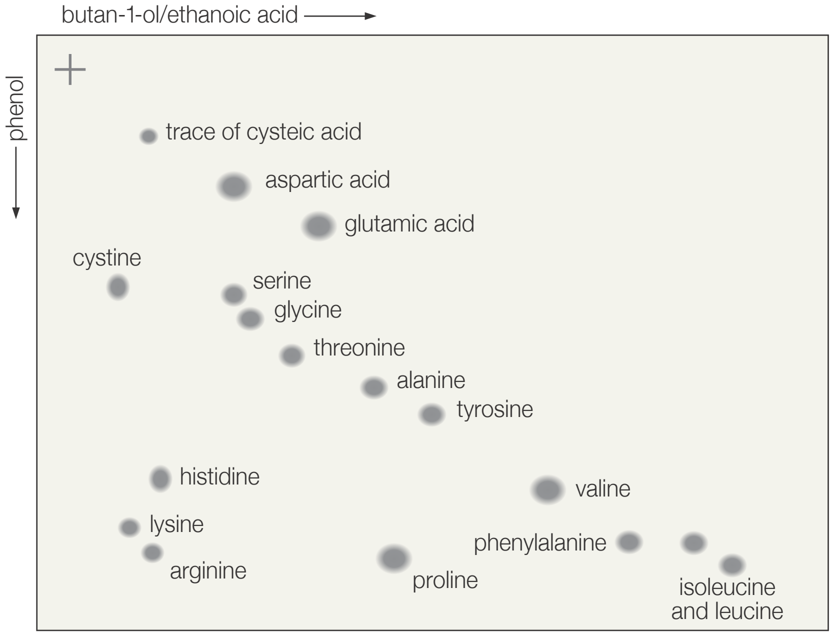

solvent front
x
y
starting point
$R_f = \frac{\text{distance moved by solute}}{\text{distance moved by solvent}}$
$= \frac{x}{y}$

■ **Figure 23.22** Calculating the R_f value from a chromatogram

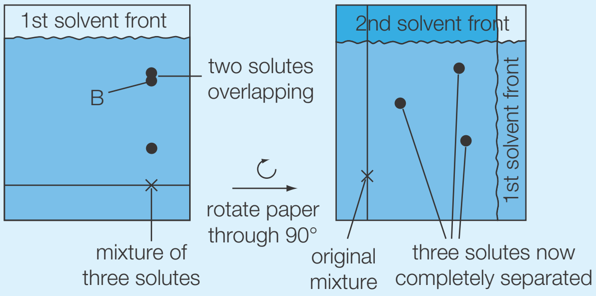
For each of the amino acids a **retention factor** (**R_f value**, or retardation factor) can be calculated (Figure 23.22). Standard R_f values are tabulated for amino acids using specific solvents or solvent mixtures under standard conditions. The amino acids in the hydrolysed protein can be identified by comparing their R_f values against the standards.

Further resolution of amino acids can be achieved by performing two-dimensional chromatography (Figure 23.23). After the initial separation process the paper is allowed to dry and then a different solvent or solvent mixture is used with the paper rotated through 90°. This approach allows the separation of complex mixtures of amino acids.

■ **Figure 23.23** Two-dimensional paper chromatography of hydrolysed protein



9 Calculate the R_f value in the **second solvent** of the spot labelled **B** in the following experiment.

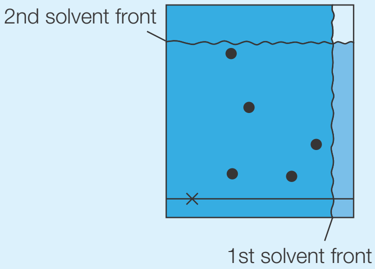


10 The diagram shows a two-way chromatogram carried out on the products of the hydrolysis of a polypeptide. The X marks the starting point of the sample.

a How many different amino acids were present in the sample?

b What could have been used to make the spots visible?

c Sketch the chromatogram you might expect if only the first solvent had been used.



Thin-layer chromatography (TLC) is similar to paper chromatography, but the stationary phase is a thin layer of a solid such as alumina or silica supported on an inert base such as glass, aluminium foil or insoluble plastic. The plate can be prepared in the laboratory by using a slurry of the powder and then carefully oven dried; alternatively, commercially prepared plates are available. TLC has advantages over paper chromatography in that its results are more reproducible, and separations are very efficient because of the much smaller particle size of the stationary phase.

The ways in which spots of colourless compounds on a TLC plate can be visualized are similar to those used for paper chromatography, but the visualization of compounds containing aromatic rings, or other systems, such as amino acids or peptides, that absorb UV radiation at 254 nm can be enhanced by impregnating the silica or alumina layer with a fluorescent insoluble compound that absorbs UV light and emits it as visible light. When placed under a UV lamp, the plate emits a bright white light except where a UV-absorbing compound is situated. Here, a dark spot is observed.

Analysis of a hydrolysate of aspartame

An interesting experiment in this context is the hydrolysis and chromatography of the artificial sweetener aspartame. Aspartame contains two amino acids, aspartic acid and phenylamine. In this activity, aspartame is hydrolysed by heating with hydrochloric acid ($6\,\text{mol dm}^{-3}$). The hydrolysed product is then analysed by paper chromatography using amino acid standards to demonstrate their identity. The amino acids are located using a UV lamp or ninhydrin spray.

Method

Stage 1: hydrolysis of the sweetener

12 g of aspartame is placed in a round-bottomed flask and $200\,\text{cm}^3$ of hydrochloric acid ($6\,\text{mol dm}^{-3}$) is added. The mixture is then refluxed for 30–60 minutes. After a short time, the mixture will begin to turn brown; by the time this stage is finished, it will be black.

Stage 2: paper chromatography of the amino acids

Care should be taken to touch the chromatography sheets only at the top corners as fingerprints contain traces of amino acids. It is best to wear thin surgical gloves.

A small sample of the black mixture is first decolourized by using activated charcoal. A dropping pipette is used to transfer about $5\,\text{cm}^3$ of the hydrolysed sample to a clean test tube. This is decolourized with about 100 mg of activated charcoal and then filtered to give a clear solution for spotting on to the chromatogram.

The solvent mixture (ethanol : water : ammonia / 80 : 10 : 10) is placed in the tank, which is covered to produce a saturated atmosphere. Cling film or aluminium foil can be used for this.

The paper is prepared and spots of each of the reference amino acids and also the sample are placed on the paper. This can be done using capillary melting point tubes. The spots should not be too large. This can be achieved by spotting several times for the one sample, and drying with a hairdryer or heat lamp in between applications. Pencil identification marks are made at the top of the paper.

The paper is formed into a cylinder and secured with clips. It is then placed in the tank with the spotted end down, taking care not to let the paper touch the glass walls. The tank is closed. No observations can be made while the chromatogram is running because the compounds used are colourless. The separation is run for a minimum of 1 hour, longer if possible (take note of where the solvent front is). It is then removed from the tank, the solvent front is marked with a pencil and the paper is allowed to dry.

The paper is then hung up in a fume cupboard and sprayed sparingly with the ninhydrin solution. It is then heated in an oven at 110 °C for 5–10 minutes, when the amino acids should appear as purple spots. The colour is stable for some weeks if kept in the dark and can be photocopied to give a permanent record.

Alternatively, the dried chromatogram can be viewed under a UV lamp to visualize the spots.

This experimental method can be used to analyse other standard mixtures of amino acids and introduce the idea of two-dimensional chromatography (using a different solvent such as a butan-1-ol/ethanoic acid/water mixture). Such a two-dimensional method gives a better separation of the amino acids.

Nature of Science

Scientific testing and reliability

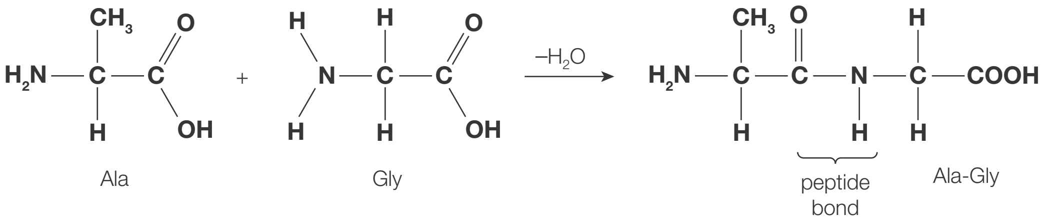
The artificial sweetener aspartame is a further example of an 'accidental' discovery; it is also a cautionary tale in how electronic media, un-edited as it largely is, can be misleading. In December, 1965, Jim Schlatter, a chemist at G.D. Searle, was working on a project to discover new treatments for gastric ulcers. To test new anti-ulcer drugs, the biologists used a tetrapeptide (four amino acids) normally produced in the stomach; Schlatter was synthesizing this tetrapeptide in the laboratory, and one of the steps in the process was to make a dipeptide intermediate, aspartyl-phenylalanine methyl ester. In the course of his work, Schlatter accidentally got a small amount of the compound on his hands without noticing it. This led to his discovering and testing the sweet taste he found on his fingers. Eventually this adventitious discovery led to the mass market sweetener that is available to us today.

The use of aspartame has hardly been uncontroversial and despite several reproducible trials by various food agencies internationally, all of which have shown that it is not harmful to health, it has been subjected to misleading campaigns and adverse publicity on the internet. Such campaigns can undermine correctly conducted science and need to be carefully appraised when making judgements regarding health issues. One recent salutary example of the danger of 'bad science' campaigning is the undermining of the child vaccination programme in the UK. The results of biased and ill-managed testing led to many parents avoiding the MMR vaccine for their children, resulting in an increase in the incidence of dangerous conditions affecting the health of children. Such instances are an important reminder to be wary of some internet material, and of the need for comprehensive and reliable product testing.

Formation of polypeptides

Within the cell, proteins are formed from amino acids inside **ribosomes**. Adjacent amino and carboxylic acid functional groups are joined together to form a peptide linkage. This reaction is a condensation process since it involves the formation of a water molecule. The reaction between two amino acids results in the formation of a dipeptide (Figure 23.24). The two **amino acid residues** are joined by a strong carbon–nitrogen bond. The process of condensation polymerization is repeated until a long chain of amino acids, known as a **polypeptide** (Figure 23.25), is formed. A protein results when the polypeptide chain leaves the ribosome and undergoes a folding process.

Figure 23.24 Peptide bond formation between two amino acids



By convention, when you are drawing peptide chains, the $-NH_2$ group which has not been converted into a peptide link is written at the left-hand end. The unchanged $-COOH$ group is written at the right-hand end. The end of the peptide chain with the $-NH_2$ group is known as the **N terminus**, and the end with the $-COOH$ group is the **C terminus**. This reflects the fact that the synthesis of a protein is directional, with the protein chain being synthesized from the N-terminal end to the C-terminal end.

Depending on which ends of the amino acids form the peptide linkage, two distinct dipeptides can be formed from two different amino acids. For example, the two dipeptides formed from alanine (Ala) and glycine (Gly) are glycyl-alanine, Gly-Ala, and alanyl-glycine, Ala-Gly (Figure 23.26). Note that both dipeptides are drawn according to the convention that the N terminus is on the left.

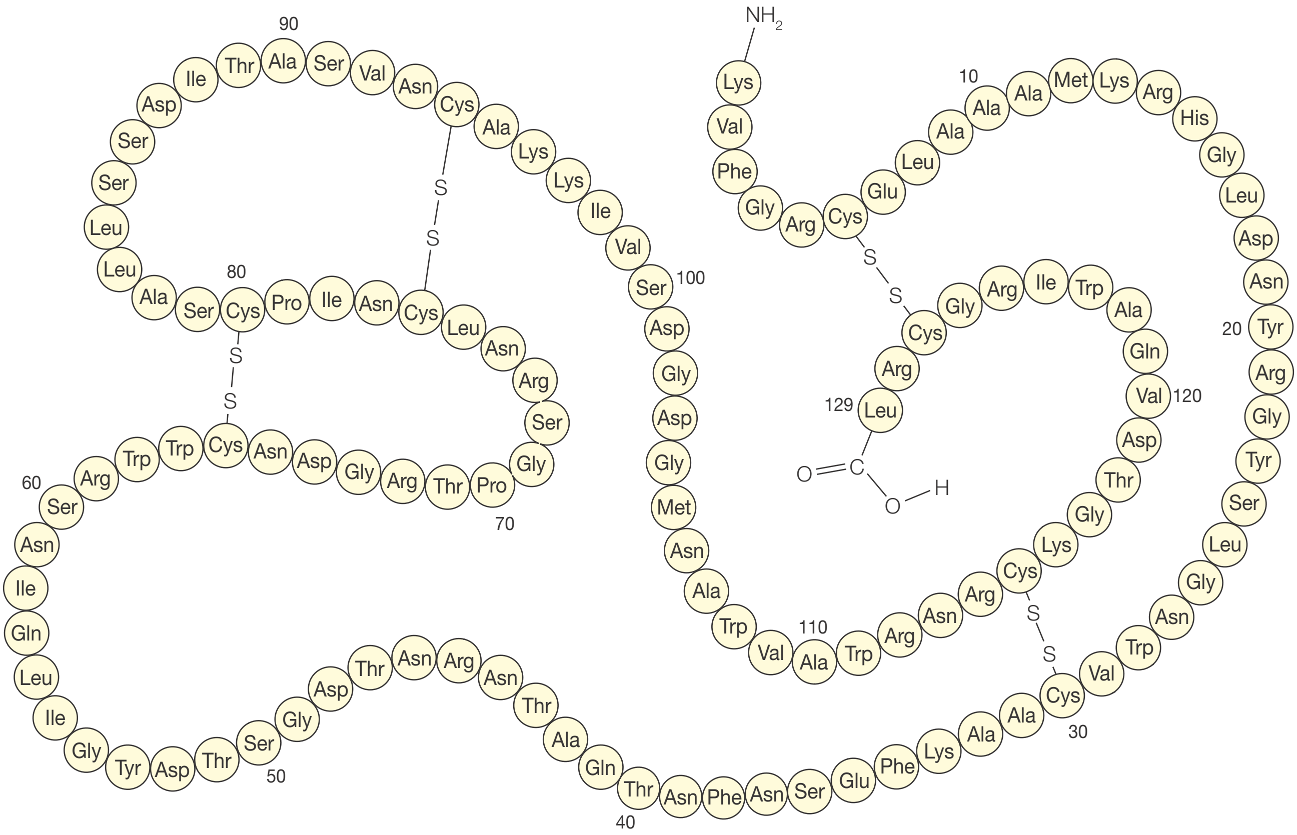
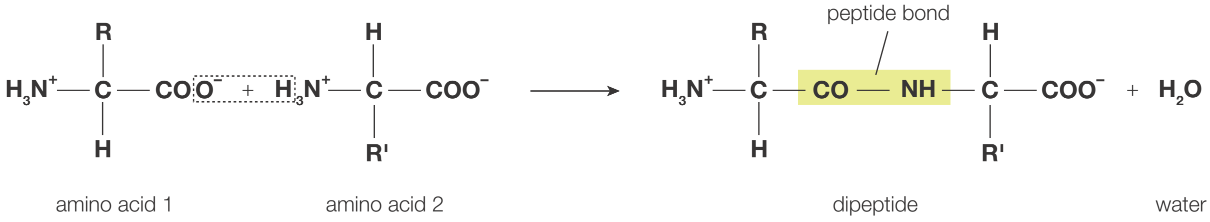
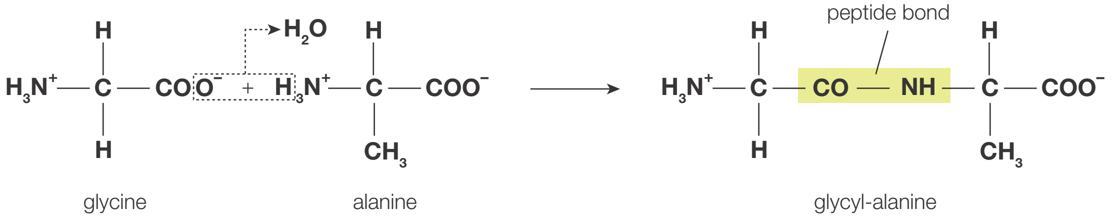


Figure 23.25 The structure of the human protein lysozyme


peptide bond
R
H
H3N+ — C — COO− + H3N+ — C — COO−
H
R'
H3N+ — C — CO — NH — C — COO− + H2O
amino acid 1
amino acid 2
dipeptide
water

For example, glycine and alanine can react like this:


peptide bond
H2O
H
H3N+ — C — COO− + H3N+ — C — COO−
H
CH3
H3N+ — C — CO — NH — C — COO−
glycine
alanine
glycyl-alanine

But if the amino group of glycine reacts with the carboxyl group of alanine, a different dipeptide, alanyl-glycine, is formed

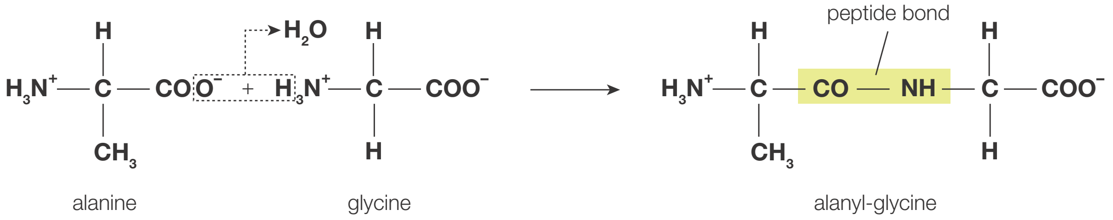

peptide bond
H2O
H
H3N+ — C — COO− + H3N+ — C — COO−
CH3
H
H3N+ — C — CO — NH — C — COO−
alanine
glycine
alanyl-glycine

Figure 23.26 The structures of the two dipeptides formed from alanine (Ala) and glycine (Gly)

A protein chain will have somewhere in the region of 50–2000 amino acid residues. The term amino acid residue has to be used since a peptide chain is not made up of amino acids. When the amino acids chemically combine together, a water molecule is lost. The peptide chain is made up from what is left after the water is lost – in other words, it is made up of amino acid residues.

The peptide bond

The peptide bond has some special properties due to π delocalization or resonance (Chapter 14), shown in Figures 23.27 and 23.28. This gives the peptide bond some double bond character, preventing rotation about the bond. The peptide bond is rigid and planar, and usually occurs in the *trans* conformation shown in Figure 23.28, as opposed to the *cis* conformation in which the hydrogen atom would be on the same side of the double bond as the oxygen atom. The *trans* conformation prevents steric hindrance between the oxygen and the hydrogen atom, whereas in the *cis* conformation, interactions between the two groups destabilize the conformation. In the *cis* conformation, steric clashes occur between the side-chains. This is a major reason why the peptide bond is usually found in the *trans* conformation.

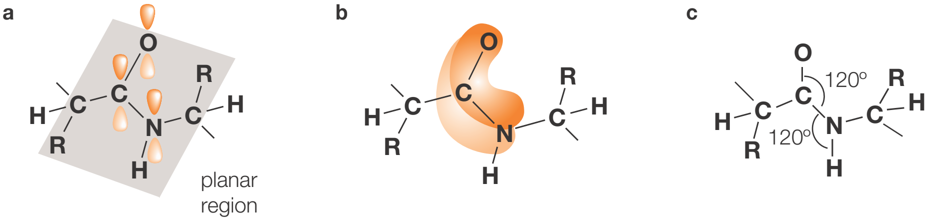
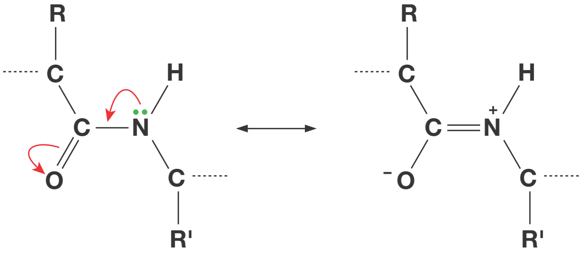


Figure 23.27 The structure of the peptide bond showing **a** the unused p orbitals following formation of the 'skeleton' of σ bonds; **b** the delocalized π bond; **c** the bond angles involved

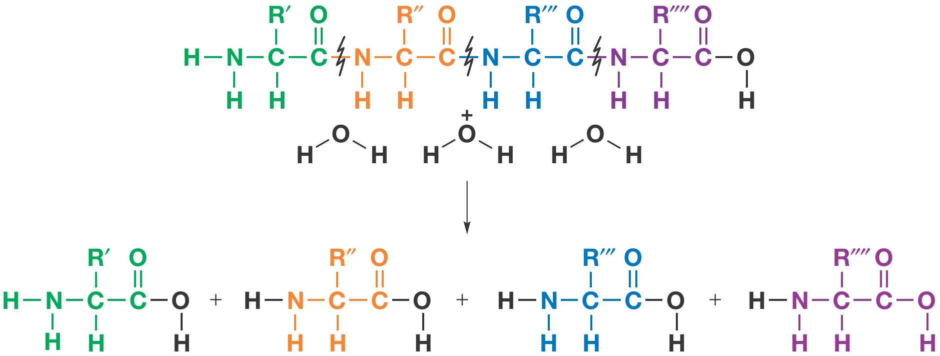
Figure 23.28 Resonance structures of the peptide bond (in the *trans* conformation)



Protein hydrolysis

As mentioned earlier proteins can be hydrolysed back to their amino acids by boiling with 6 mol dm^{-3} hydrochloric acid solution for 24 hours. The peptide bonds (amide links) are broken and free amino acids are released (Figure 23.29). In this acid hydrolysis, the bond between the C=O and the N of each amide link must be broken, and the elements of water added: –OH to the C=O and H to the N–H.

Figure 23.29 The acid hydrolysis of a tetrapeptide



The presence of peptide bonds in a protein can be demonstrated by the Biuret test. Biuret is a blue reagent and contains copper(II) sulfate dissolved in alkaline solution. Proteins give a purple colour with Biuret reagent (Figure 23.30), whereas Biuret remains blue in the presence of amino acids.

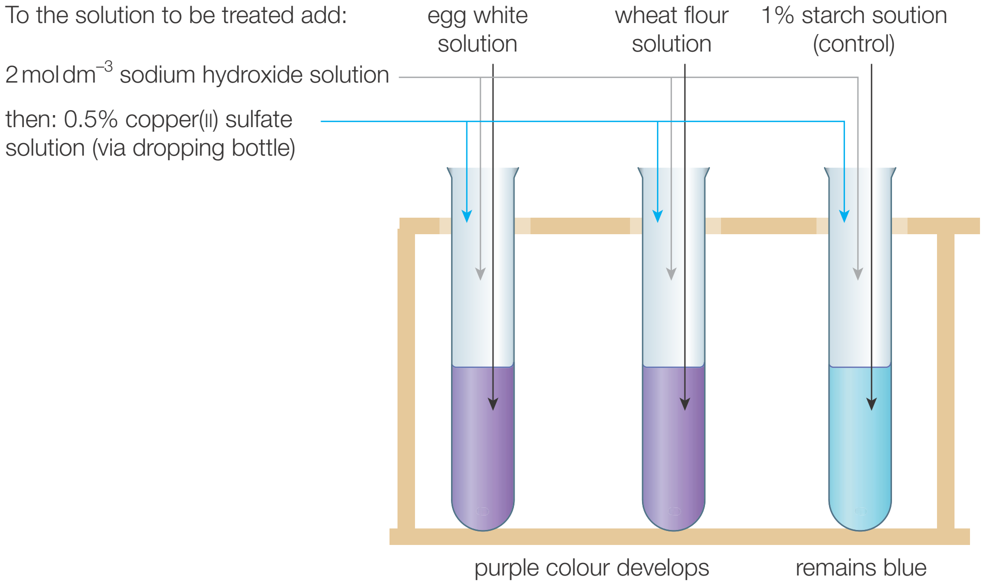

To the solution to be treated add:
egg white solution
wheat flour solution
1% starch soution (control)
2 mol dm−3 sodium hydroxide solution
then: 0.5% copper(II) sulfate solution (via dropping bottle)
purple colour develops
remains blue

Figure 23.30 The Biuret test for protein

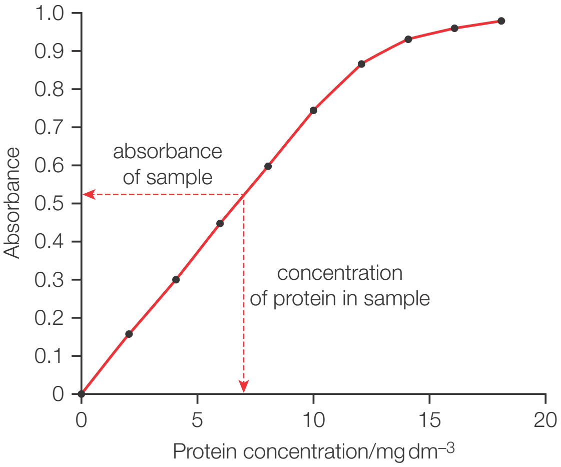


Figure 23.31 A calibration curve for the Biuret test for finding protein concentration

Protein analysis by UV–visible spectroscopy typically involves first reacting the sample with a reagent such as Biuret that generates a colour change which is dependent on the amount of protein present. This will promote absorption in the UV–visible range. Several different reagents can be used which react with different groups within the protein molecules, such as peptide bonds, aromatic side-groups or basic groups. Biuret reagent generates a purple colour by reaction with peptide bonds.

In order to find the protein concentration of an unknown sample a **calibration curve** (or standard curve) must first be established using protein solutions of known concentration. Such a calibration curve is based on standard solutions, those with a known concentration of protein, which are prepared to cover a range of concentrations on either side of the value being investigated. Usually a series of dilutions of a standard such as bovine serum albumin of known concentration are prepared. These solutions are then tested with the Biuret reagent and their absorbance is measured at the selected wavelength (660 nm for the Biuret test). The absorbance readings are then plotted to obtain the calibration curve (Figure 23.31).

The absorbance of the solution being analysed is then measured at the same wavelength and its protein concentration determined from the calibration curve as shown (Figure 23.31). Although referred to as a calibration curve, the useful portion of the graph is the linear region in which **Beer–Lambert's law** is being observed. In this region the absorbance is directly proportional to the amount of protein present in the sample.

Protein structure

In 1959, Perutz and Kendrew published an article on the three-dimensional structure of whale myoglobin, which is a small protein responsible for the transport of oxygen in whale cells. By investigating the protein's structure, the two scientists wanted to understand the oxygen-carrying mechanism at a molecular level. They grew crystals of the protein and determined its structure by analysing the X-ray diffraction pattern of the crystal. Crystallizing proteins is an awkward task, having to be carried out in a controlled way. However, it is possible to devise a suitable protocol to obtain crystals of lysozyme in a school laboratory and this makes for a useful practical exercise (Figure 23.32; for an article which can be downloaded from the online journal *Science in School* see www.scienceinschool.org/2009/issue11/lysozyme).

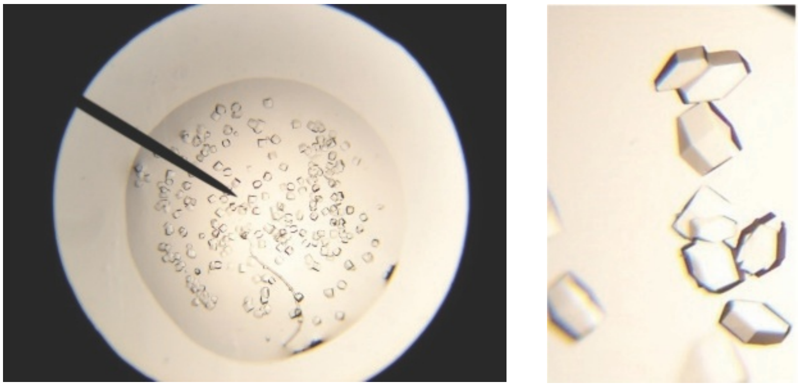


Figure 23.32 Crystals of the protein lysozyme grown in a school laboratory

Nature of Science

Exploiting a technique

Pioneering studies such as those of Perutz and Kendrew laid the foundation for the elucidation of the different levels of organization of protein chains and how structure relates to function for different types of protein. Their work on myoglobin and hemoglobin gained them the Nobel Prize in Chemistry in 1962. Their achievements were linked to, and developed from, the establishment of the principles of X-ray crystallography by scientists such as Lawrence Bragg. Bragg, to date the youngest ever winner of a Nobel Prize, had outlined the basis of the technique following on the work of his father, William Bragg. Together they are the one father and son combination to win the prize (for Physics), which they did in 1915.

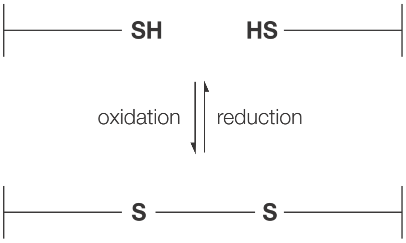
Perutz and Kendrew were just two of a series of crystallographers who participated in a 'golden era' of the use of X-ray crystallography in probing the structure of proteins and other biological macromolecules.

Dorothy Hodgkin, a British biochemist, was another credited with the development of protein crystallography. She was awarded the Nobel Prize in Chemistry in 1964. She advanced the technique of X-ray crystallography and among her most influential discoveries are the confirmation of the structure of penicillin, and then the structure of vitamin B_{12}, for which she became the third woman to win the Nobel Prize in Chemistry. In 1969, five years after winning the Nobel Prize, Hodgkin was able to decipher the structure of insulin. X-ray crystallography became a widely used tool and was critical in later determining the structures of many biological molecules where knowledge of structure is critical to an understanding of function. David Phillips also elucidated the structure of egg white lysozyme in 1965, an influential step in understanding the functioning of enzymes.

The technique was central to the work undertaken by Maurice Wilkins and Rosalind Franklin. Their crystallographic studies were instrumental in the discovery of the structure of DNA reported by James D. Watson and Francis Crick in February 1953. Such crystallography studies proved very influential in forwarding the understanding of molecular biology, particularly because the findings were reported alongside other crucial developments in other aspects of biochemistry including protein sequencing.

Frederick Sanger was an English biochemist awarded the Nobel Prize in Chemistry twice. He used the protease trypsin to partially hydrolyse insulin and then used electrophoresis to separate the fragments based on their charge and solubility. Sanger published the primary structure of insulin 1955 and was awarded the Nobel Prize in 1958. Later he developed methods of sequencing DNA and was awarded a second Nobel Prize in 1980 with Walter Gilbert.

This whole period of breakthrough in the structural analysis of biological macromolecules illustrates the significance of periods in science when a novel technique is developed and then creatively applied to a range of problems, producing an upsurge in understanding of structure and function.

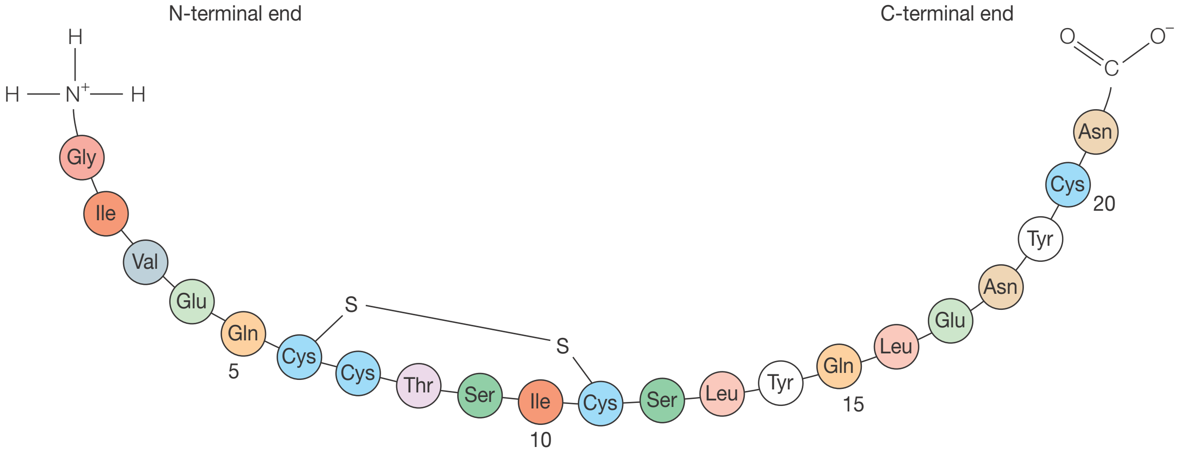


■ **Figure 23.33** The formation of a disulfide bridge

Primary structure

Each polypeptide chain is a linear polymer of amino acids and as such has an amino- (or N-) terminal end and a carboxyl- (or C-) terminal end. The **primary structure** (Figure 23.34) of a protein is the linear sequence of amino acid residues in the polypeptide chain including any covalent cross links. Covalent bonds, known as disulfide bridges, –S–S–, can form when two cysteine amino acid side-chains react (under enzyme control) (Figure 23.33). Proteins differ in the variety, number and order of their constituent amino acid residues.

The primary structure of a protein is determined by the sequence of bases in the gene that codes for the protein (see later). As mentioned earlier, in a cell polypeptide chains are always synthesized from the N-terminal end to the C-terminal end. Thus, when writing out the primary sequence of a polypeptide chain the amino acids are numbered from the N-terminal end. As an example, the primary structure of the insulin ‘A’ chain is shown in Figure 23.34. (The position in the primary structure of any cysteine residues is of particular significance, as these can form disulfide bridges that stabilize the three-dimensional structure of a protein.) The changing of a single amino acid (due to a mutation (change) in the DNA of its gene) will alter its properties, often drastically.



■ **Figure 23.34** The primary sequence of the insulin A chain (a short polypeptide of 21 amino acids). Note the cysteine residues at positions 6, 7, 11 and 20 in the chain. They form disulfide bridges (the one between residues 6 and 11 is shown here) that stabilize the three-dimensional structure of the insulin molecule. The Cys residues at positions 7 and 20 form inter-chain disulfide bridges to the insulin B chain

The amino acids present in a protein can be established by acid hydrolysis followed by paper chromatography or electrophoresis. However, these techniques do not indicate the primary structure. The primary sequence of a protein can be established by partially hydrolysing the protein (with a number of specific proteases) and then overlapping the fragments to give the complete sequence (Figure 23.35).

■ **Figure 23.35** Establishing the primary structure of a protein by partial hydrolysis

Fragments from first enzyme cleavage: Glu-Met-Leu-Gly-Arg
Ala-Gly
Tyr-Lys

Fragments from second enzyme cleavage: Tyr-Lys-Glu-Met
Leu-Gly-Arg-Ala-Gly

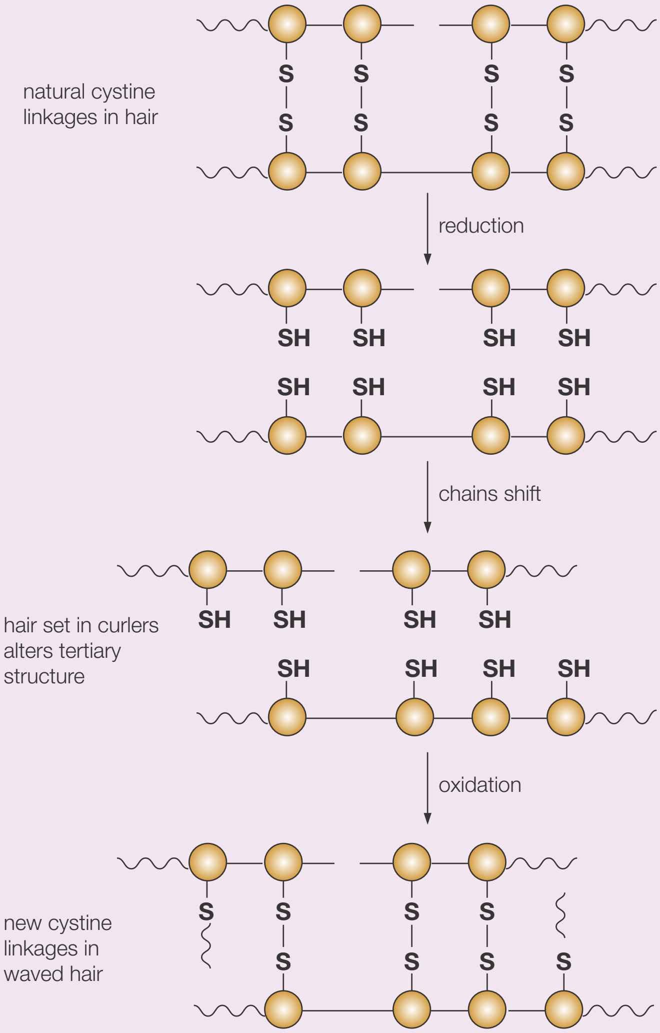
Deduced sequence: H_2N–Tyr-Lys-Glu-Met-Leu-Gly-Arg-Ala-Gly–COOH

Additional Perspectives

Permanent styling

Hair is composed of a structural protein named keratin whose long polypeptide chains fold up into an α-helix structure. One of the amino acids in keratin is cysteine, the side-chain of which contains a thiol (–SH group). Thiol groups can interact to form a covalent **disulfide bridge** (–S–S–). Perming the hair involves the breaking and re-forming of disulfide bonds. A two-stage process is used: first, the hair is treated with a solution of a thiol which reacts with the disulfide bonds. The hair is then set into the required style, usually with rollers. The final step is to form new disulfide bonds by oxidation of the thiol groups, which sets the new style permanently into position (Figure 23.36)

■ **Figure 23.36** The chemistry of hair styling

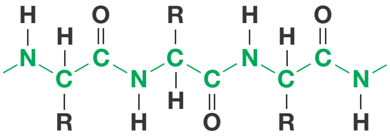


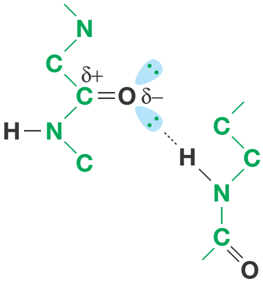
natural cystine linkages in hair
S
S
reduction
SH
chains shift
hair set in curlers alters tertiary structure
oxidation
new cystine linkages in waved hair

Secondary structure

The **secondary structure** of a protein refers to the regular and permanent arrangement of sections of the polypeptide chain. Every polypeptide has a 'backbone' that runs the length of the chain (Figure 23.37).

■ **Figure 23.37** The 'backbone' of a polypeptide chain





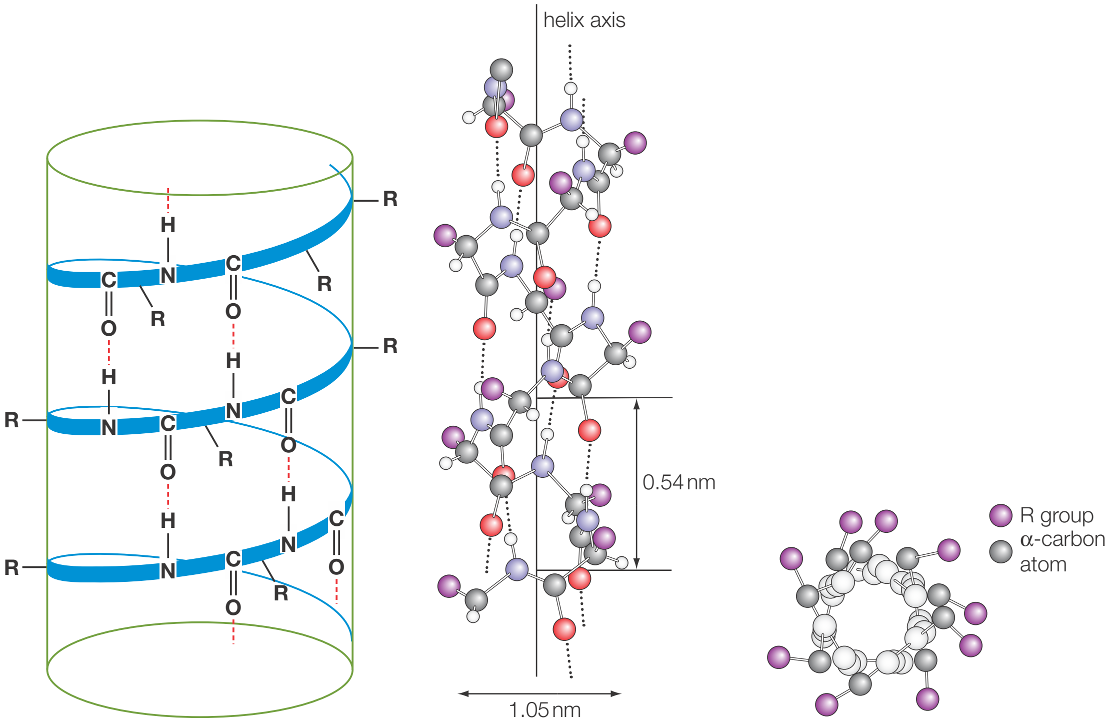
■ **Figure 23.38** Hydrogen bonds between peptide bond regions stabilize protein secondary structure

As the only difference between the amino acids lies in the R groups, this backbone is essentially the same for all proteins (–C–C–N–C–C–N–, etc.). Although the rigid planar structure of the peptide bond restricts some of the shapes that a polypeptide chain can adopt, the backbone is flexible and in certain sections can fold in a regular manner, and this is the secondary structure. The folding in the polypeptide backbone is stabilized by hydrogen bonding. The N–H of one peptide bond forms a hydrogen bond to the C═O of another peptide bond (Figure 23.38).

Two of the most stable types of secondary structure are the **α-helix** and **β-pleated sheet**. In both these types the polypeptide chain is folded in a very stable arrangement because of the many hydrogen bonds formed between adjacent peptide bond regions.

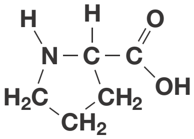
The α-helix

The α-helix (Figure 23.39) is a regular coiled conformation of the polypeptide chain, somewhat like a narrow tube. The polypeptide chain is coiled in a spiral, and the variable R groups of the amino acid residues project outwards from the spiral of the helix. There are 3.6 residues per turn (Figure 23.39b). Each peptide group is involved in two hydrogen bonds, with the C═O of a peptide group being hydrogen bonded to the N–H of the peptide group four units ahead in the primary structure. The hydrogen bonds run down the length of the α-helix and hold the structure in place. Alpha-helices are always right-handed.

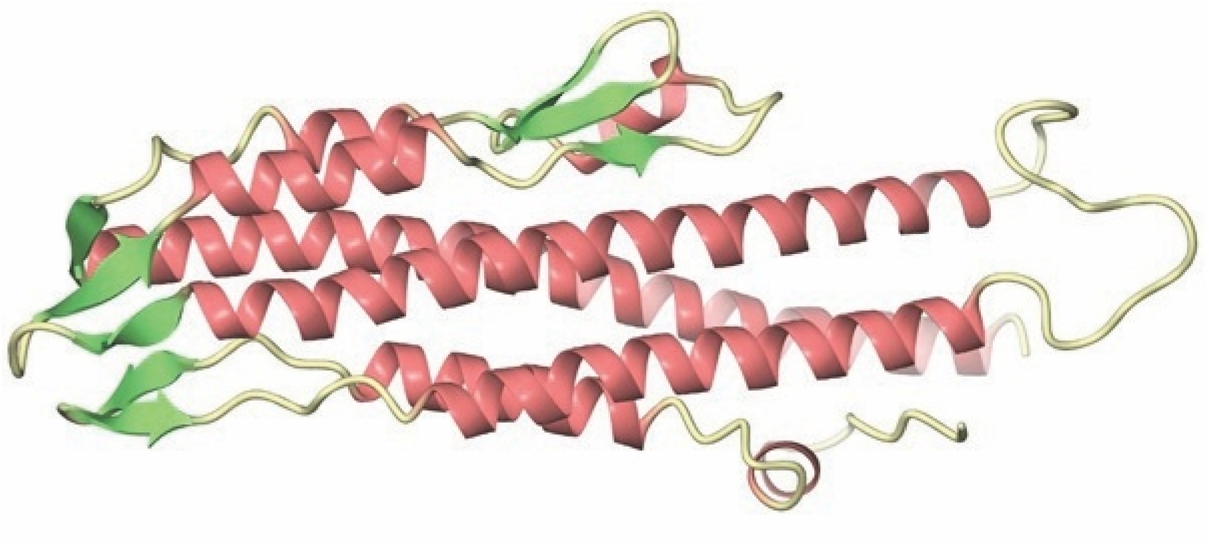


■ **Figure 23.39** The α-helix. **a** The structure of the α-helix. **b** The helix extends 0.54 nm for each complete turn (this is known as the pitch of the helix). **c** The end-on view shows that the R groups all point outwards from the helix

However, not all amino acids allow such structures to form; some R groups destabilize an α-helix. Of particular note here is the amino acid proline.



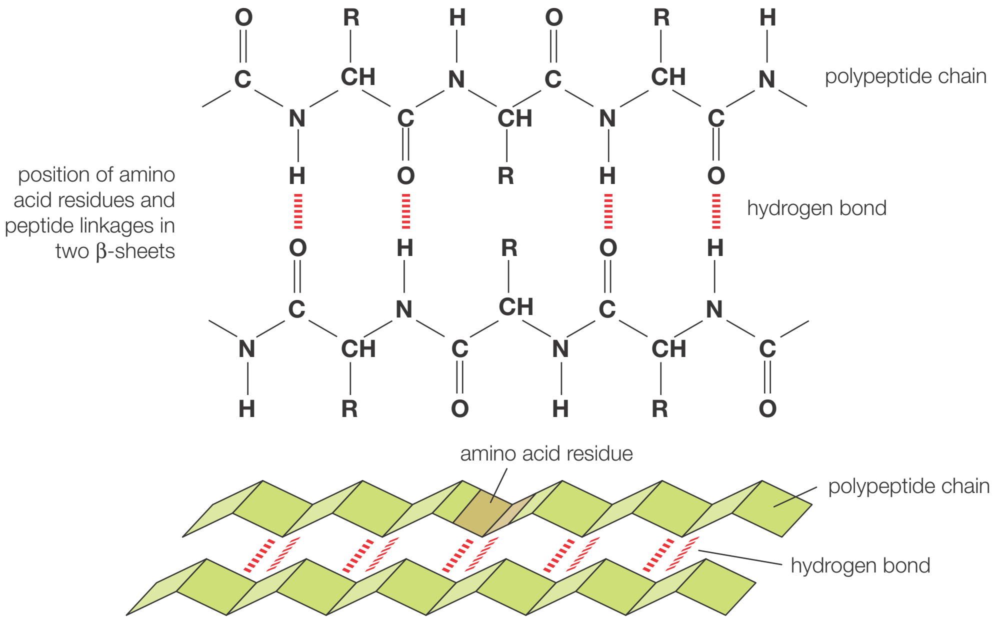
Because of its distinctive cyclic structure, proline breaks up a α-helical region by introducing a sharp kink in the chain. Many proteins have extensive stretches of α-helices located along their polypeptide chain (Figure 23.40). Hemoglobin and myoglobin are other examples (see Figure 23.47), while keratin, found in nails and hair, consists of a number of α-helices coiled around each other and held together by **disulfide linkages**.



■ **Figure 23.40** Some proteins show extensive regions of α-helical structure such as BipD, an invasion protein from the bacterium *Burkholderia pseudomallei* which causes the disease mellioidosis. Alpha helices are shown in red and beta strands in pale green, with the intervening loops coloured cream

The β-sheet

The β-sheet is the second form of protein secondary structure. It is composed of adjacent polypeptide chains (within the same protein) lying side-by-side and connected by intramolecular hydrogen bonding, forming a sheet (Figure 23.41). The R groups of the amino acid residues point above and below the plane of the sheet while the C═O and N–H groups of the peptide groups in adjacent sections point towards each other. Hydrogen bonds form between these adjacent sections and so stabilize the structure.



■ **Figure 23.41** Structure of a β-sheet

The protein fibroin found in the silk fibres produced by silkworms (the larvae of the silk moth, *Bombyx mori*) has extensive β-sheets. Figure 23.42 shows three antiparallel sections of a polypeptide chain in a β-sheet, illustrating the pleated sheet arrangement. All the variable side-chains extend above or below the sheet. It should be noted that the strands of polypeptide arranged in a sheet can run antiparallel (as here), parallel or as a mixture of the two, depending on the folding in the adjacent regions.

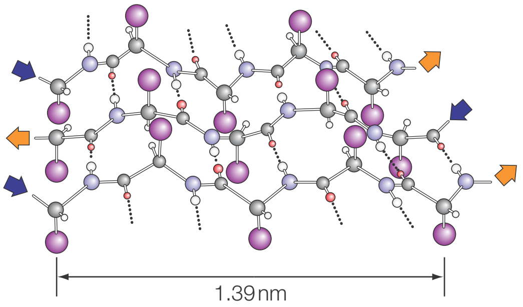
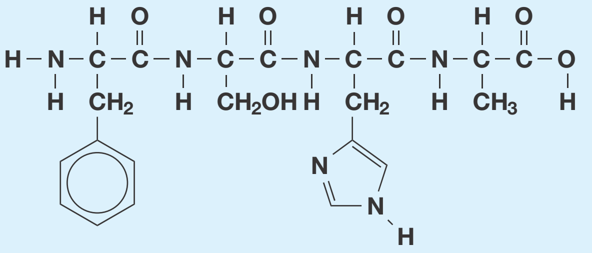


Figure 23.42 The β-pleated sheet structure – showing that the R groups point above and below the sheet

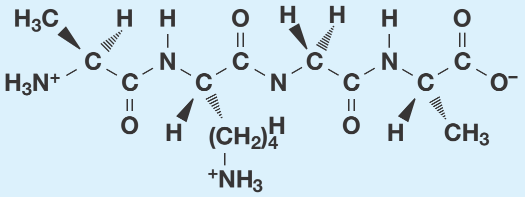
11 **a** What type of polymerization takes place when a protein chain is assembled from 2-amino acids?
b Draw a diagram showing the bond formed between two amino acids ($NH_2CH(R)COOH$) during formation of a protein chain. Label the approximate bond angles and describe the shape of the fragment that you have drawn,
c Explain (i) why the structure in this region is rigid, and (ii) why the >C=O group and the >N-H group are arranged *trans* to each other (consider the size of the R groups on each amino acid).

12 **a** What type of bonding is responsible for maintaining the primary structure of a protein chain?
b A polypeptide chain is said to have direction. How are the two ends of the chain referred to?

13 Name and give the structures of the amino acids that would be formed by the hydrolysis of the polypeptide below. Use the information on their structures given in Table 23.2 and Section 33 of the *IB Chemistry data booklet* to help identify them.



14 The structure of a tetrapeptide is shown below.

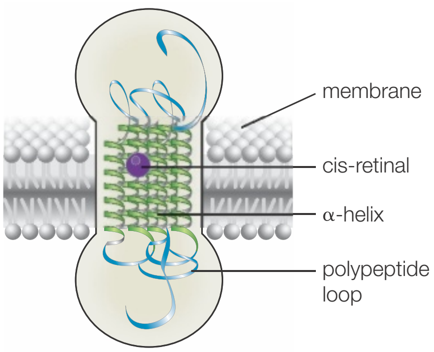


a State the name of the chemical bond linking the amino acids together in the tetrapeptide.
b Draw the structural formula of the amino acid at the N- and C-terminal ends of the peptide.
c State how many optically active molecules are formed when the tetrapeptide is completely hydrolysed.

Secondary structure and membrane proteins

Membrane proteins are an important group of proteins that play a part in biological processes as enzymes (phospholipase A, for instance), receptors for hormones and pharmaceutical drugs (see Chapter 25), as an integral part of photosystems in cells (see Chapter 24), and as part of the immune system (see the discussion on AIDS in Chapter 25). Here we mention them as a case study in how the secondary structure of a protein links to its role and position in cell structure.

Membrane proteins can be either intrinsic or extrinsic to the membrane. Extrinsic (or peripheral) membrane proteins, such as phospholipase A, are attached to one surface of the



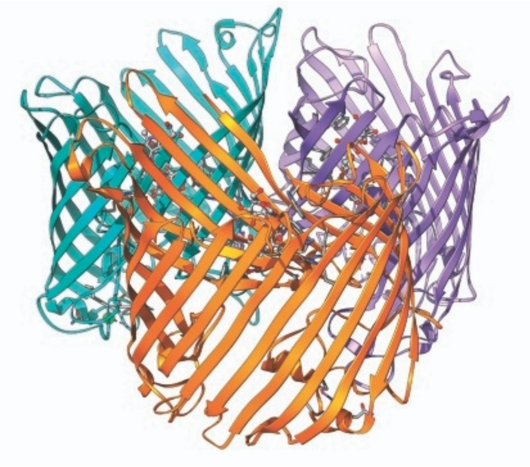
■ **Figure 23.43** The molecular structure of bacteriorhodopsin. This protein is found in primitive organisms known as Archaea. The central *cis*-retinal molecule is surrounded by α-helices (spirals) and polypeptide loops. The protein sits in the membranes of the cell (grey)

membrane whereas intrinsic proteins are integral to the structure of the membrane. Most of the latter span the whole of the phospholipid bilayer (i.e. they are transmembrane proteins). Membrane proteins have proved difficult to analyse structurally; of over 100 000 structures available in the Worldwide Protein Database only 494 are membrane proteins (http://blanco.biomol.uci.edu/mpstruc/ and http://www.wwpdb.org/stats.html). However, in all the transmembrane proteins examined to date, the membrane-spanning domains are α-helices or multiple β-strands (as in porins).

Proteins containing seven membrane-spanning α-helices form a major class that includes bacteriorhodopsin and many cell-surface receptors. Bacteriorhodopsin is a protein found in a photosynthetic bacterium (Figure 23.43). Absorption of light by the retinal group attached to bacteriorhodopsin causes a conformational change in the protein that results in pumping of protons from the cytosol across the bacterial membrane to the extracellular space. Other seven-spanning membrane proteins include the opsins (eye proteins that absorb light), cell-surface receptors for many hormones, and receptors associated with the sense of smell.

The porins are a class of transmembrane proteins whose structure differs radically from that of other integral proteins. All porins are trimeric transmembrane proteins. Each subunit is barrel-shaped with β-strands forming the wall and a transmembrane pore in the centre. Several types of porin are found in the outer membrane of gram-negative bacteria such as *E. coli* (see Figure 23.44). The outer membrane protects an intestinal bacterium from harmful agents (e.g. antibiotics, bile salts, proteases) but permits the uptake and disposal of small hydrophilic molecules including nutrients and waste products. The porins in the outer membrane of an *E. coli* cell provide channels for passage of disaccharides (sucrose, for instance), phosphate and similar molecules.

■ **Figure 23.44** Sucrose-specific porin – molecular model. Porins are proteins that span cell membranes and act as a channel through which specific molecules can diffuse

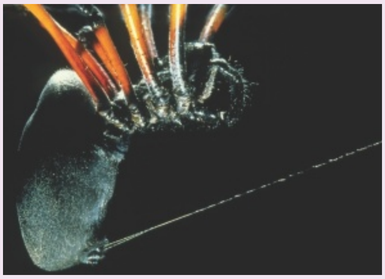


Additional Perspectives

Spider silk

Spider silk is a remarkable material. The purpose of this activity is to research, using the library or internet, the seven different forms of spider silk mentioned below, and their structure and function.

Based on weight, spider silk is five times stronger than steel wire of the same diameter. There are historical records that suggest that bullets have been unable to break through a silk handkerchief. George Emery Goodfellow (a doctor at Tombstone, Arizona, USA) wrote in the spring of 1881, ‘I was a few feet from two men quarrelling, they began shooting,



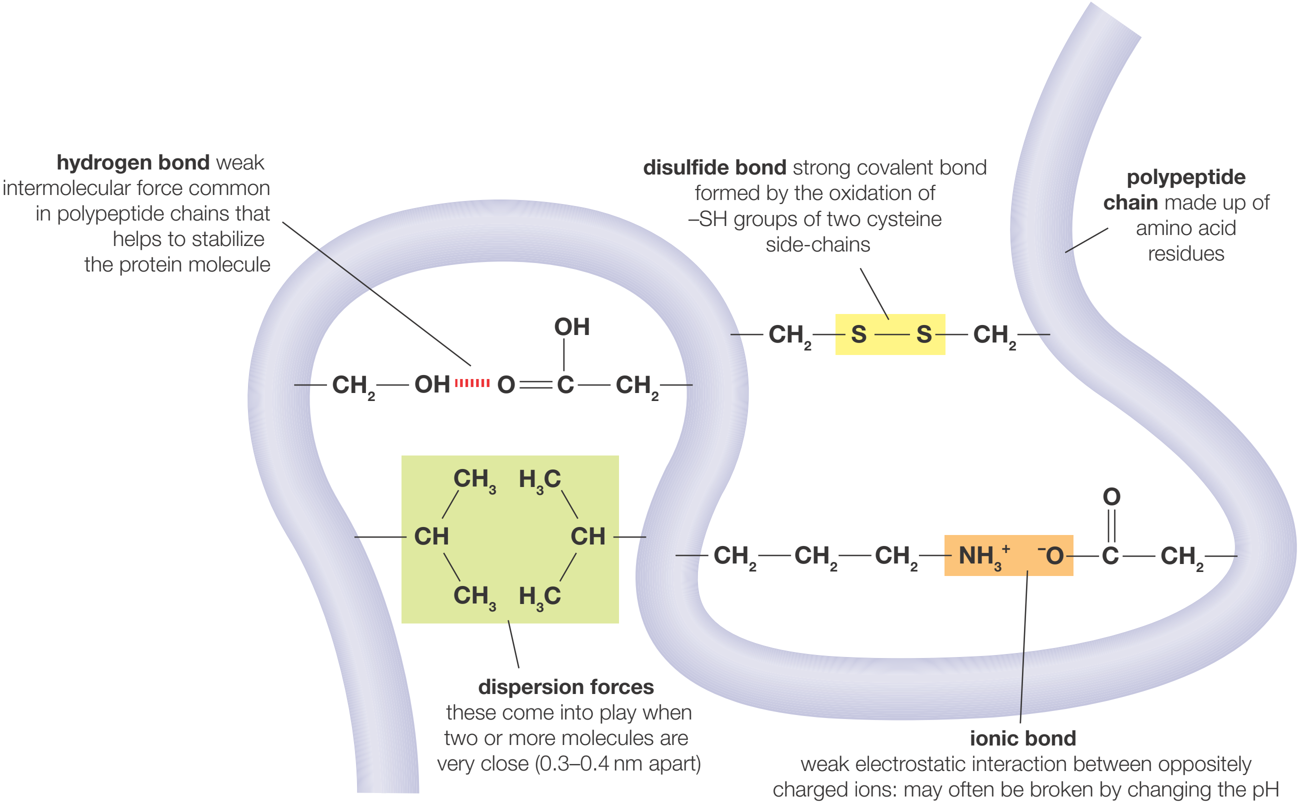
■ **Figure 23.45** Female golden orb spider (*Nephila* sp.) producing silk. The *Nephila* spiders are a tropical genus that spin enormous webs up to 2 m across. Their silk is among the strongest known, and their webs often catch small birds

two bullets pierced the breast of one man, who staggered, fired his pistol and crumpled on to his back. Despite fatal injuries, not a drop of blood had come from either of the two wounds.' Further investigation located a bullet wrapped within a silk handkerchief. It appears that the bullet had passed through clothes, flesh and bones but not his silk handkerchief. More recently, it has been suggested that a strand of spider silk as thick as a pencil would stop a jumbo jet in flight.

The golden orb-weaving spider (Figure 23.45) produces a dragline silk (a dragline connects a spider to its web and is also used for dynamic kiting – the means by which spiders move location) that is the strongest form of spider silk. The protein in dragline silk is fibroin. There are actually seven different kinds of spider silk, each being produced by different glands. Each type has a different function. Fibroin has a molecular mass of 200 000–300 000 and consists of large regions of β-pleated sheet structure and a distinctive amino acid composition of 42 per cent glycine and 25 per cent alanine, with the remainder coming from just seven other amino acids.

Tertiary structure

The **tertiary structure** is the overall three-dimensional shape of a single protein. A series of possible interactions between the R groups of different amino acid residues produces this third level in the hierarchy of protein folding. This is known as the tertiary structure and is crucially important to a protein's function. The protein molecule is held in a specific shape by hydrogen bonds and other intramolecular forces involving the side-chains. Disulfide bridges may also be present. At this level, the chemical nature of the different R groups becomes particularly significant. The different possible interactions responsible for maintaining the tertiary structure of a polypeptide chain are summarized in Figure 23.46.


hydrogen bond weak intermolecular force common in polypeptide chains that helps to stabilize the protein molecule
disulfide bond strong covalent bond formed by the oxidation of –SH groups of two cysteine side-chains
polypeptide chain made up of amino acid residues
dispersion forces these come into play when two or more molecules are very close (0.3–0.4 nm apart)
ionic bond weak electrostatic interaction between oppositely charged ions: may often be broken by changing the pH
CH2 — OH ····· O=C — CH2
OH
CH2 — S — S — CH2
CH3 H3C
CH CH
CH3 H3C
CH2 — CH2 — CH2 — NH3+ -O — C — CH2
O

■ **Figure 23.46** Interactions maintaining protein tertiary structure

These interactions are:

- ionic bonds between charged R groups, such as those of lysine and aspartic acid;
- hydrogen bonds between polar R groups;
- instantaneous dipole-induced dipole forces between non-polar side-chains, such as those of valine and phenylalanine;
- covalent disulfide bonds formed between cysteine residues at different locations in the primary sequence (see Figure 23.33).

The interactions between particular R groups in different regions of a polypeptide chain reinforce a specific folding arrangement. Some of these interactions are relatively easily disrupted, and others not so. The formation of disulfide bridges is of particular significance. Because of their covalent nature, disulfide bonds can have the effect of locking a particular tertiary structure in place.

Figure 23.47 shows the secondary and tertiary structures of myoglobin. This protein acts as a store of oxygen in muscle.

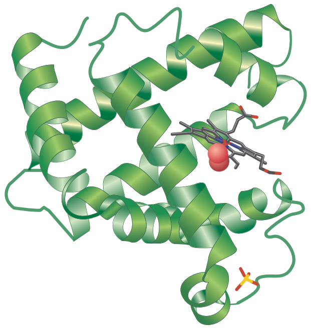


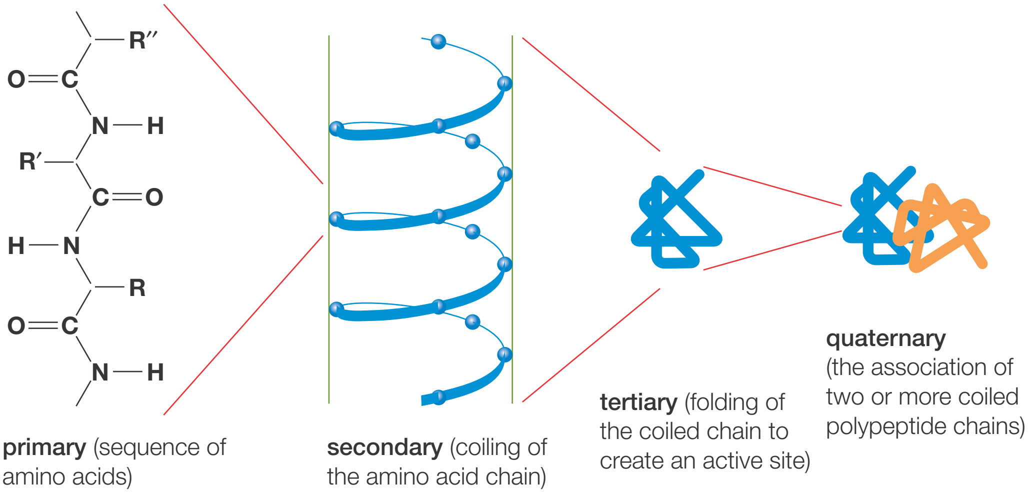
Figure 23.47 Three-dimensional structure of myoglobin – an oxygen-carrying heme protein in muscle

Quaternary structure

Some proteins, for example hemoglobin, consist of two or more tightly bound polypeptide chains, or subunits. The overall conformation or shape is referred to as the **quaternary structure**. The types of forces between the chains are the same as those that maintain the tertiary structure of the individual chains.

The different levels of the hierarchy of protein folding are summarized in Figure 23.48.

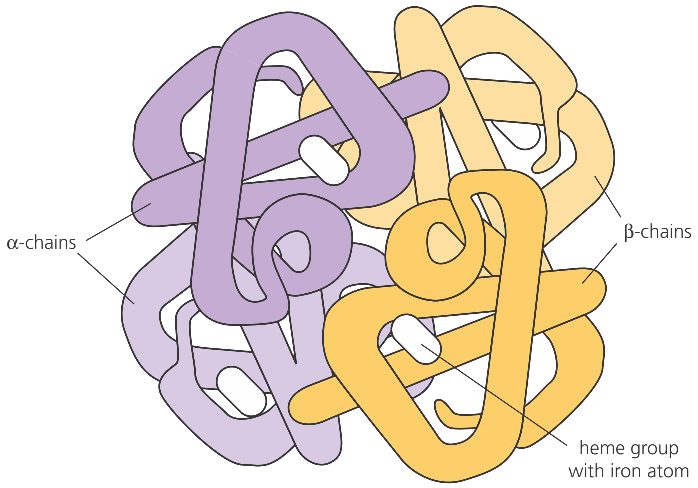
Figure 23.48 The different levels in the hierarchy of protein structure



R″
O=C
N—H
R′
C=O
H—N
R
O=C
N—H
primary (sequence of amino acids)
secondary (coiling of the amino acid chain)
tertiary (folding of the coiled chain to create an active site)
quaternary (the association of two or more coiled polypeptide chains)

Hemoglobin (Figure 23.49) is composed of four polypeptide chains: two α-chains and two β-chains. (The use of the terms α- and β- as applied by biochemists in this context might seem confusing! The terms α- and β- are used here just to differentiate between the two different chains, and this use is not at all related to how we used them before, to describe the α-helix and the β-pleated sheet secondary structures of proteins.) The α- and β-chains both have very similar primary structures to each other and to myoglobin. Each of these is bound to a heme unit, at the centre of which is an iron(II) ion that binds reversibly to oxygen. There is a high percentage of α-helices in the α- and β-chains. The four chains are held together by a variety of non-covalent interactions.

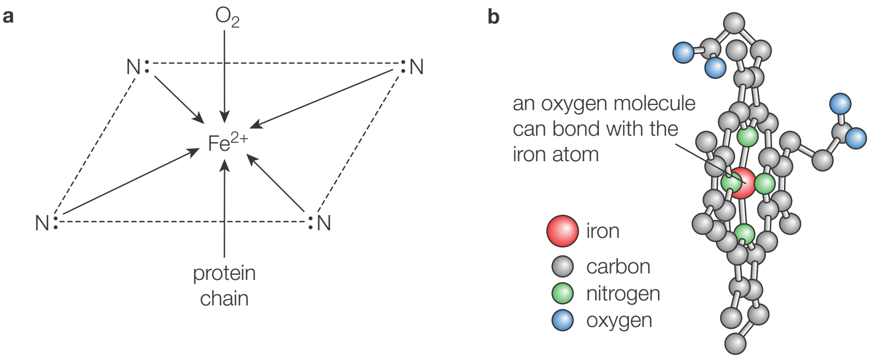
Figure 23.49 Structure of hemoglobin



Overall, hemoglobin is described as $\alpha_2\beta_2$. Each of the four protein chains is also bound to a non-protein heme group that contains an iron(II) ion (Fe^{2+}). It is the Fe^{2+} ions in the heme groups that bind oxygen to hemoglobin. Each heme group can bind one oxygen molecule (Figure 23.50) and each of the four heme groups binds oxygen simultaneously, so the overall reaction is:

$$\underset{\text{hemoglobin}}{Hb} + 4O_2 \rightleftharpoons \underset{\text{oxyhemoglobin}}{HbO_8}$$

Figure 23.50 **a** The complex ion in hemoglobin. **b** The heme group contains an Fe^{2+} ion which can bond reversibly with an oxygen molecule



The Fe^{2+} ions act as the centres of complex ions; the ligands are the heme group, the protein chain and a molecule of oxygen. The heme group binds to the Fe^{2+} ion via four nitrogen atoms and the protein chain also binds via a nitrogen atom. The elucidation of the structures of the related proteins, myoglobin and hemoglobin, explained why the Fe^{2+} ion in the heme group is not oxidized in the process of binding oxygen.

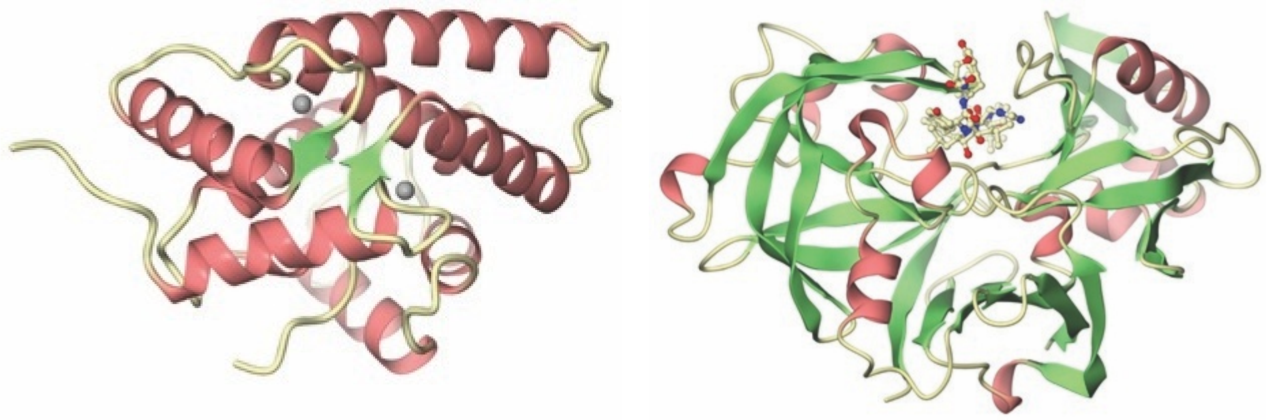
15 Describe how a section of a protein containing valine, glutamic acid and cysteine (see the *IB Chemistry data booklet* for structures of these amino acids) can contribute to the ordered secondary and tertiary structure of a protein with an α-helix. Include diagrams in your answer and discuss the relevant bonds and forces that stabilize the structure.

16 Tertiary structure in proteins is stabilized in part by hydrogen bonds between the R groups of amino acid residues. Consider the R groups present in the amino acids lysine, asparagine, valine and aspartic acid. For each amino acid, predict whether the R groups can contribute to a hydrogen bond.

Nature of Science

Computer graphic imagery of protein structures

The interaction between studies of the structure of a wide range of proteins and the onset of sophisticated methods of computer graphics and imaging has proved immensely productive and resulted in a new 'language' of visualization of protein structure. Such developments vastly aid understanding and we have seen in the images represented throughout this chapter the way in which the two key types of secondary structure are represented in these computer graphics (Figure 23.51).



■ **Figure 23.51 a** The protein calexcitin with the bound calcium ions in grey (the secondary structure is coloured according to the same scheme as Figure 23.40). **b** The pepsin-like protease endothiapepsin with a drug molecule bound in the active site cleft shown in ball-and-stick representation

The assembly of databanks of computer-generated images of a wide range of protein molecules, together with precise information on the 'mapping' of their structures, has been a notable development in international cooperation within science and the development of the field of proteomics.

UniProt is a central resource for storing and interconnecting biological information for storing and interconnecting information from large and disparate (very different) sources. It is a comprehensive catalogue of protein sequence (primary structure of amino acids) and functional annotation, which gives information about the function of the protein. UniProt is built upon the bioinformatics infrastructure and scientific expertise at the European Bioinformatics Institute (EBI), Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB). UniProt has three different protein databases optimized for different uses. UniProt is updated and distributed every month and can be accessed online for searches or downloaded at www.uniprot.org.

The Worldwide Protein Data Bank (wwPDB; www.wwpdb.org/) consists of organizations that act as deposition, data processing and distribution centres for PDB data. Members are RCSB PDB (USA: www.rcsb.org/pdb/), PDBe (Europe: www.ebi.ac.uk/pdbe/), PDBj (Japan: http://pdbj.org/) and BMRB (USA: www.bmrb.wisc.edu/). The Worldwide Protein Data Bank's aim is to maintain a single archive of macromolecular structural data that is freely and publicly available to the global community.

Protein sequencing is now a routine operation in biochemical research, with machines able to deduce the entire sequence of a large protein very quickly. This is a major part of **proteomics**, the study that explores the relationship between structure and function of proteins. The synthesis of new so-called designer proteins through **protein engineering** has many applications in pharmaceutical and environmental research. The databases mentioned above also allow for extensive comparative studies of protein sequences in studies of biochemical evolution to determine the relationships between organisms.

Globular and fibrous proteins

Proteins are classified into **fibrous** and **globular proteins**. Fibrous proteins consist of long molecules arranged to form fibres, such as keratin of hair and nails, and collagen, which is present in skin, bones, teeth and tendons (Figure 23.52). Fibrous proteins are insoluble.

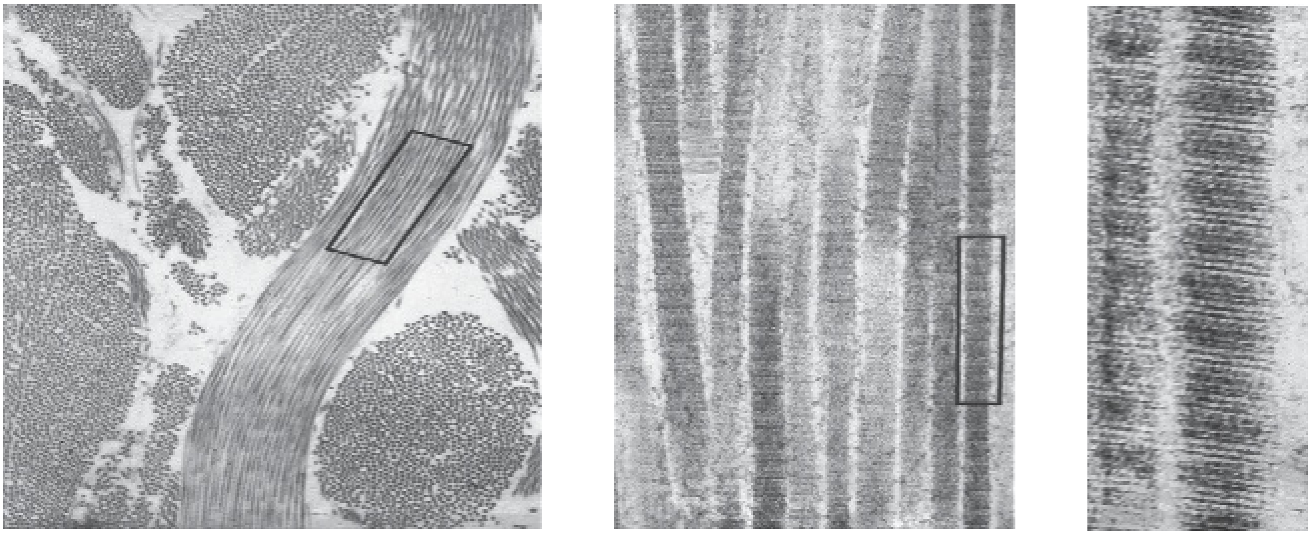
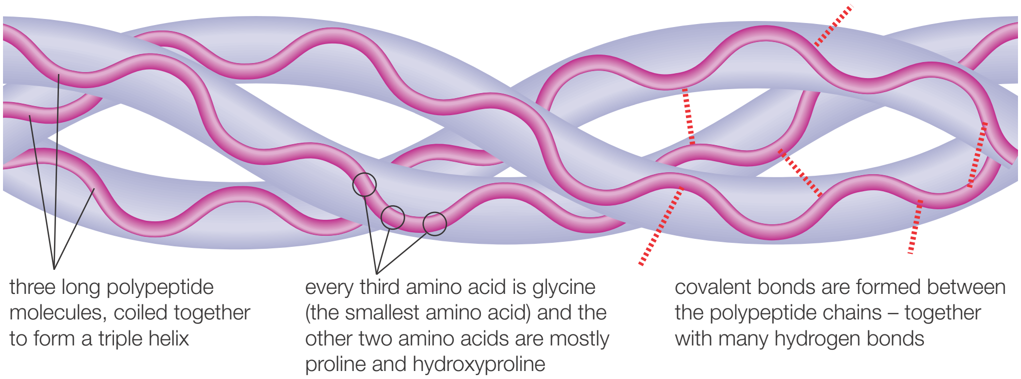


Figure 23.52 Electron micrographs of collagen fibres. **a** The collagen fibres are arranged in bundles in the body. Note some are cut in cross section. **b, c** Stained fibres: the repeating pattern of the stain shows the regular, repetitive amino acid sequence of the collagen chains

Collagen is the most abundant fibrous protein in the human body. It consists of three polypeptide chains, each about 1000 amino acid residues long. The three polypeptide chains are wound together as a unique triple helix (Figure 23.53).

Figure 23.53 Collagen – an example of a fibrous protein

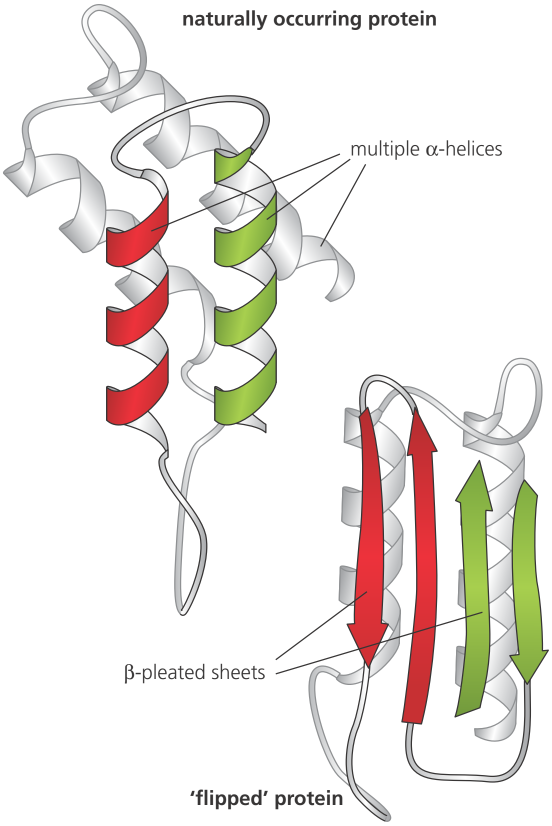

three long polypeptide molecules, coiled together to form a triple helix
every third amino acid is glycine (the smallest amino acid) and the other two amino acids are mostly proline and hydroxyproline
covalent bonds are formed between the polypeptide chains – together with many hydrogen bonds

In a globular protein the polypeptide chain(s) are folded into a compact but precise shape. Enzymes, hemoglobin and protein-based hormones, for example insulin, are globular. They tend to be soluble in water. The mechanisms by which globular proteins fold are still being actively researched. However, it is known that they fold compactly, leaving little room for water in the interior. The polar side-chains are on the outer surface of the molecule and help to make the protein water soluble. In contrast, the non-polar side-chains point to the interior of the molecule. When a protein folds there is a loss of entropy (Chapter 15) in the protein chain, but the removal of water from the interior increases the entropy in the surroundings. There is a fine balance between these processes and there is often only a small difference in free energies (Chapter 15) between the folded and unfolded states.

Protein folding disorders

The structure of a protein and its ability to carry out its biological function are so strongly correlated that very small structural defects can lead to a number of protein folding diseases. These include genetic diseases such as sickle cell anemia, which is caused by a single residue mutation. A number of diseases have been linked to protein folding problems which lead to the build-up of insoluble protein plaques in the brain. These diseases include diseases such as bovine spongiform encephalopathy (BSE) ('mad cow disease') and its human equivalent Creutzfeld–Jakob disease (CJD). Researchers have found that normal prion proteins in the brain consist of many α-helices but in CJD these prion proteins 'flip', unfold and refold into a protein with β-sheets – these then cause other normal prion proteins to 'flip' and insoluble fibrils are formed inside brain cells (Figure 23.54).

Figure 23.54
Prion protein 'flipping'



Gel electrophoresis

We saw earlier the usefulness of separating amino acids on the basis of their charge at a particular pH. That technique of electrophoresis can be applied to proteins and has proved extremely useful in analysing complex mixtures of proteins and obtaining data on the molecular size of protein chains. Electrophoresis of proteins is usually performed in a polyacrylamide gel rather than on paper. The polyacrylamide gel is an inert supporting material through which molecules will move when a potential difference is applied. The gel acts a molecular sieve, slowing larger molecules and allowing smaller molecules to move more quickly. Therefore in this form of electrophoresis separation is based upon both molecular size and charge (Figure 23.55) – the electrophoresis here is carried out on native proteins.

Proteins and other biological molecules are often treated with a strong detergent known as sodium dodecyl sulfate (SDS) which attaches to the molecules, making them negatively charged. This form of electrophoresis is known as SDS-PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis; Figure 23.56) and separation is based solely on molecular size or mass.

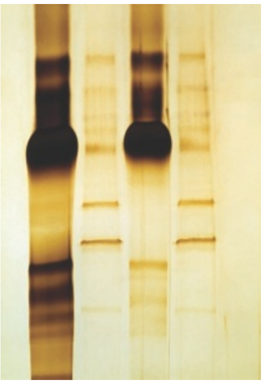
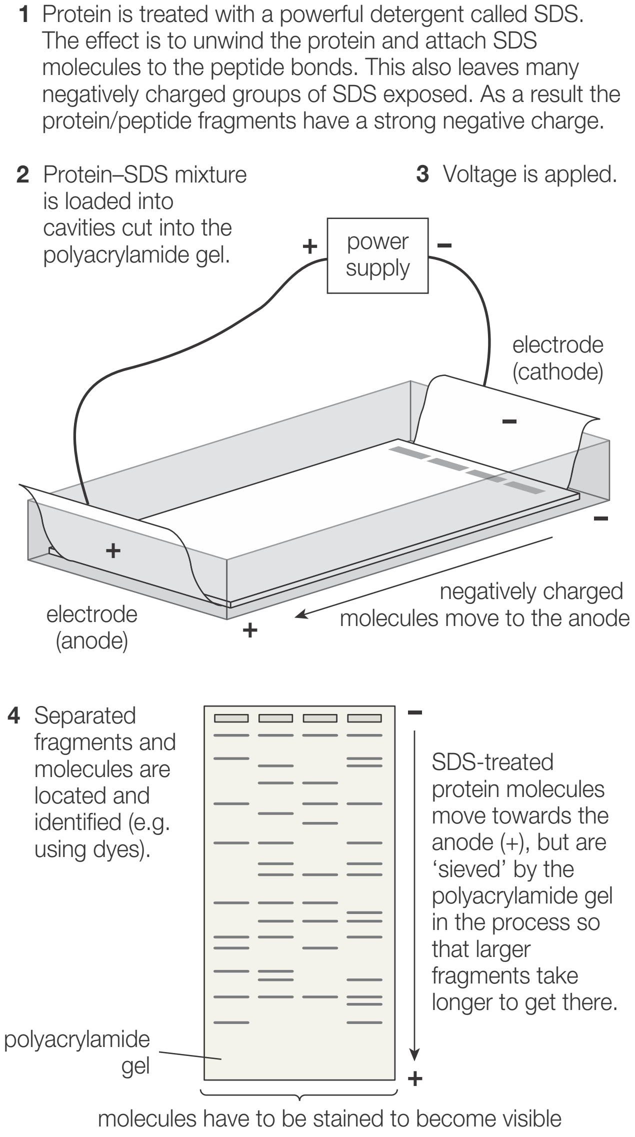


Figure 23.55
Close-up of bands produced by protein samples on a polyacrylamide gel electrophoresis (PAGE) plate



1 Protein is treated with a powerful detergent called SDS. The effect is to unwind the protein and attach SDS molecules to the peptide bonds. This also leaves many negatively charged groups of SDS exposed. As a result the protein/peptide fragments have a strong negative charge.

2 Protein–SDS mixture is loaded into cavities cut into the polyacrylamide gel.

3 Voltage is appled.

4 Separated fragments and molecules are located and identified (e.g. using dyes).

SDS-treated protein molecules move towards the anode (+), but are 'sieved' by the polyacrylamide gel in the process so that larger fragments take longer to get there.

molecules have to be stained to become visible

■ **Figure 23.56** Separation of amino acids and polypeptide fragments by SDS-PAGE

Functions of proteins

Proteins are crucial components for basic life processes (Figure 23.57). They are responsible for transport throughout a cell or organism, for maintaining cellular structures and for basic metabolism, among other processes.

Essential functions, such as transport of oxygen throughout blood and its storage in muscle cells, are carried out by proteins such as hemoglobin and myoglobin.

Structural proteins such as elastin are included in the walls of arteries and veins and of the bronchioles (airways) in the lungs. Collagen is the main structural protein of connective tissue in animals and the most abundant protein in mammals. It is the major component of cartilage, ligaments (joining bone to bone) and tendons (joining bones to muscle). Muscle is also composed of proteins, for example actin and myosin form the fibres.

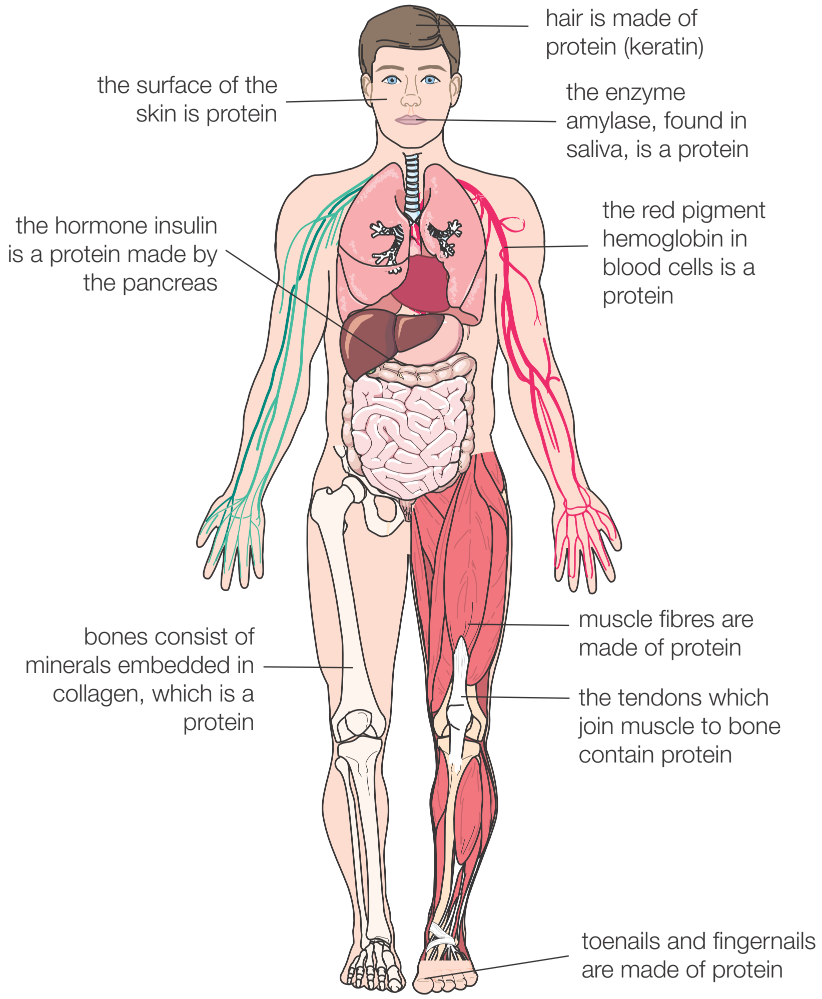

hair is made of protein (keratin)
the surface of the skin is protein
the enzyme amylase, found in saliva, is a protein
the hormone insulin is a protein made by the pancreas
the red pigment hemoglobin in blood cells is a protein
muscle fibres are made of protein
bones consist of minerals embedded in collagen, which is a protein
the tendons which join muscle to bone contain protein
toenails and fingernails are made of protein

Figure 23.57 Examples of proteins in the human body

Pore proteins form channels that transport ions, water and other molecules through the cell membrane. Many proteins have enzymatic activity, for example catalysing the digestion of protein, starch and lipids in the gut. Digestive enzymes are extracellular enzymes which operate outside cells. However, there are also many intracellular enzymes that operate inside cells. They control respiration, photosynthesis (in plants), DNA repair and replication, and many other biochemical reactions.

Proteins are also involved in the immune system by acting as immunoproteins (antibodies). Antibodies are present in blood and identify and neutralize harmful bacteria and viruses. Antibodies are produced by a group of white blood cells known as B cells. HIV infects a variety of cells in the body, most notably T cells which regulate the B cells of the immune system.

Hormones are chemical messengers that carry a signal from a cell or a group of cells to another via the bloodstream. All animals and plants produce a variety of hormones. In general hormones control the function of their target cells. Insulin is a protein-based hormone produced by the pancreas of mammals. It is released by cells within the pancreas when blood glucose levels are low. Its main target cells are liver cells and muscle cells, which absorb the excess glucose and convert it to glycogen.

Protein denaturation

So far we have studied the formation and structure of proteins. However, it is also important to consider the disruption of protein structure, which is known as **denaturation**. Many enzymes only function within a narrow range of pH and temperature values. This is because high or low values of pH and temperature cause a physical change known as denaturation. This involves loss of tertiary and quaternary structures, but is often reversible. In denaturation no covalent bonds are broken.

The peptide bonds between the individual amino acids in a polypeptide chain are strong covalent bonds. However, we have seen that the overall shape and function of a protein depends on weaker interactions that are more easily disrupted. When a protein loses its unique three-dimensional shape it has been denatured. This denaturation can be temporary or permanent. Everyday examples of denaturation include:

- the curdling of milk proteins when milk turns sour or when milk is mixed with vinegar;
- the thermal denaturation of egg or meat protein when it is cooked;
- the mechanical denaturation of egg white when it is whisked.

In all these cases the process is irreversible; the protein cannot be reassembled. Proteins may change shape in response to a variety of factors, such as:

- changes in pH;
- changes in temperature;
- disruption of instantaneous dipole-induced dipole forces by certain metal ions, including those of heavy metals;
- the addition of a small polar molecule such as urea ($CO(NH_2)_2$) in concentrated solution – urea causes complete denaturation by specifically disrupting hydrogen bonds;
- the presence of mild reducing agents capable of breaking disulfide bridges.

Enzymes

Nature can be considered as a chemical industry: it turns out billions of tonnes of a vast range of products every year using the simplest starting materials. The catalysts that make all this possible are enzymes – all large protein molecules. Enzymes take part in the homogeneous catalysis of reactions in an aqueous environment which otherwise would be so slow that life would be impossible. The digestion of food, the release of energy for movement, the copying of genes in reproduction – all of these processes and more rely on enzymes.

Enzymes are globular proteins specialized to catalyse biochemical reactions, and exist in compact spherical shapes when in aqueous solution in cells. As biological catalysts, enzymes increase the rate of a chemical reaction without undergoing a permanent chemical change. The enzyme performs its catalytic function by providing an alternative mechanism with a lower activation energy for the reaction. The molecule whose reaction is catalysed by an enzyme is referred to as its **substrate**. A relatively small part of the protein is called its **active site**, to which the substrate can bind. Enzymes combine temporarily with the substrate to produce a transition state having a lower free energy than the transition state of the uncatalysed reaction (Figure 23.58). When the reaction products are formed, the free enzyme is regenerated.

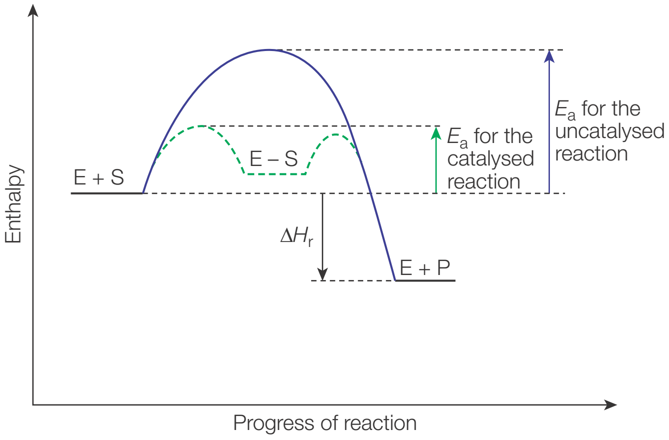


Figure 23.58 The enthalpy profile of an enzyme-catalysed reaction (ΔH_r, E_a, S, E and P represent enthalpy change of reaction, activation, substrate, enzyme and product, respectively)

Enzymes have a well-defined tertiary structure that gives them a specific three-dimensional shape (Figure 23.59), which is essential for enzyme activity. Enzymes are typically relatively large molecules, containing several hundred amino acids, and some also have a quaternary structure. For example, many of the enzymes involved in the first stage of respiration are dimeric proteins, meaning they contain two polypeptide chains. In addition, some enzymes require non-protein molecules to be bound for activity. These are known as co-factors and may be organic, when they are known as **co-enzymes**, or inorganic, such as metal ions (note the comment on the activity of carbonic anhydrase in Figure 23.59). Common examples include vitamins, many of which act as precursors for co-enzymes.

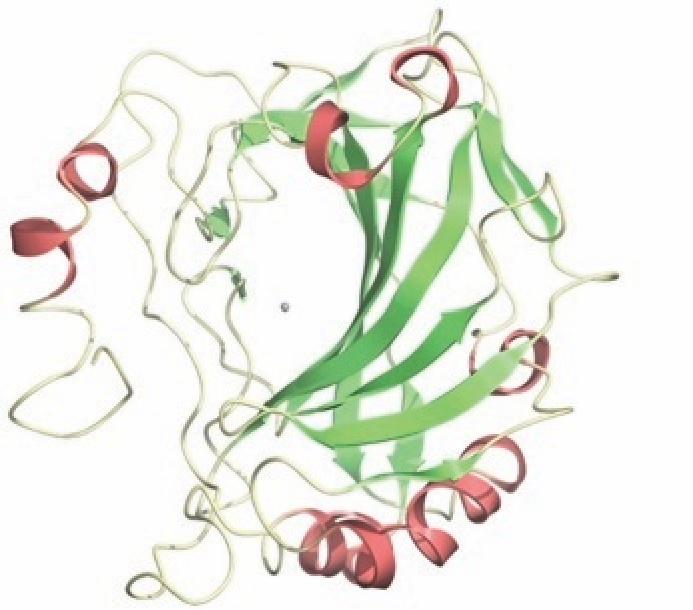


Figure 23.59 A computer-generated image of human carbonic anhydrase. Carbonic anhydrase catalyses the reaction $CO_2 + H_2O \rightarrow H_2CO_3$ (carbonic acid) within red blood cells and speeds the reaction up by a factor of about 10^9 time. However, if the zinc ion (Zn^{2+}, depicted as the grey sphere) in the centre of the enzyme is removed the enzyme has no activity

Enzyme activity is the rate at which a biochemical reaction takes place in the presence of an enzyme. It is measured in terms of the rate of appearance of a product or consumption of the reactant. Enzymes generally have very high specificity, meaning that only certain substrates are acted upon and only a single type of reaction takes place, without side reactions or by-products. This high specificity occurs because the active site has a very close fit to the substrate, and enzyme and substrate have complementary structures whereby all the charged hydrophilic and hydrophobic amino acid residues are paired. Although very specific compared to inorganic catalysts, enzymes do vary considerably in their degree of specificity. Some are absolutely specific for a particular substrate and will not attack even the enantiomer, whereas others will react with a whole class of molecules but at widely differing rates.

Table 23.4 The turnover numbers of certain enzymes

Enzyme	Turnover number[a]/s^{-1}
Carbonic anhydrase	600000
Catalase	93333
β-Amylase	18333
β-Galactosidase	208
Phosphoglucose isomerase	21
Succinate dehydrogenase	19

[a] The turnover number is defined as the maximum number of molecules of substrate that an enzyme can convert to product per catalytic site per unit time. Each carbonic anhydrase molecule can produce up to 600000 molecules of carbon dioxide product *per second.*

Enzymes are very efficient catalysts and function in dilute aqueous solution at biological pH and moderate temperature, in contrast to the rather extreme conditions often used with inorganic industrial catalysts. Although enzymes function within the rules that define catalytic activity, they differ from inorganic catalysts in several important respects:

- Higher reaction rates: the rates of enzyme-catalysed reactions are typically increased by factors of 10^6–10^{12} compared to the uncatalysed reaction (see Figure 23.60 later). These reaction rates are several orders of magnitude greater than those for inorganic catalysts (Table 23.4).
- Milder conditions: enzyme-catalysed reactions occur under relatively mild conditions: temperatures below 100 °C, atmospheric pressure and usually at pH values around pH 7.
- Greater reaction specificity: enzymes are highly selective in their actions; they are usually capable of catalysing the reaction of just one molecule or class of molecules.
- The reactions are 'clean', with very few side products.
- Ease of control: the catalytic activities of many enzymes can be varied by altering the concentrations of substances other than the reactant: the way the reaction is controlled can be complex.

The activity and specificity of enzymes depends on their conformation or three-dimensional shape (tertiary and quaternary structures). Small changes in the conformation of a protein will lead to loss of activity and specificity. The differences between inorganic catalysts and enzymes are summarized in Table 23.5.

Table 23.5 Differences between enzymes and inorganic catalysts

Enzymes	Inorganic catalysts
Enzymes are complex globular proteins	Inorganic catalysts are generally ions or simple molecules
Enzymes are synthesized by living cells	Inorganic catalysts are not produced by living cells
Enzymes are usually highly specific in action	Inorganic catalysts are usually less specific in action
Enzymes are sensitive to changes in pH and temperature	Inorganic catalysts are usually less sensitive to changes in pH and temperature
Enzymes only function in aqueous solution	Some inorganic catalysts function in aqueous solution

Categorizing enzymes

The digestive enzymes of the stomach and small intestine were among the first enzymes to be discovered. These enzymes were given names ending with '-in': hence the names pepsin, trypsin and chymotrypsin. These enzymes are all proteases, meaning that they break proteins down by hydrolysis. The first enzyme to be successfully isolated and crystallized was urease. James Sumner (1887–1955) was the American chemist who achieved this and showed that urease was a protein. He shared the Nobel Prize in Chemistry in 1946. Enzymes are now categorized and named on the basis of six main reaction types (Table 23.6). As well as ending in '-ase', an enzyme's name also indicates the type of reaction it catalyses and the substrate involved.

Table 23.6 The categorizing and naming of enzymes

Type of enzyme	Reaction	Example(s)
Oxidoreductase	Redox reactions, e.g. removal of hydrogen	Succinate dehydrogenase, cytochrome oxidase
Transferase	Transfer of groups, e.g. transfer of a phosphate from one molecule to another	Phosphofructokinase, hexokinase
Hydrolase	Breakdown of molecules by hydrolysis	Urease, trypsin, lipase, ribonuclease, amylase
Lyase	Removal of a group from, or addition to, a double bond	Pyruvate decarboxylate
Isomerase	Isomerization, moving groups within a molecule	Phosphoglucose isomerase, maleate isomerase
Ligase	Synthetic reactions, joining molecules together	Glycogen synthase

What is perhaps rather surprising is that there are only six categories of enzyme-catalysed reactions. This illustrates an important point about metabolic systems in cells. The changes achieved by these metabolic pathways are often large, but they are achieved in a series of simple reactions, each catalysed by a specific enzyme. It is the cumulative effect of these small changes that produces the overall large change.

Genetically inherited disorders

The importance of shape and structure to the functional activity of enzymes is highlighted by the range of genetically inherited conditions where mutation produces a defective or reduced function for a specific enzyme, resulting in a health condition. Many serious or fatal illnesses are the product of genetically inherited disorders that result in the failure of a single enzyme. For example, the condition phenylketonuria (PKU) that can lead to intellectual disability is the consequence of a malfunction in the enzyme responsible for the breakdown of the amino acid phenylalanine in the liver. This condition is the reason why many food and drinks that contain aspartame are labelled ‘contains a source of phenylalanine’. In many parts of the world babies are screened for PKU soon after birth. Patients who are diagnosed early and maintain a strict diet can have a normal lifespan with normal cognitive development.

Ehlers–Danlos syndrome (EDS) is an inherited connective tissue disorder with different presentations that have been classified into six primary types. EDS is caused by a defect of genetic origin in the structure, production or processing of collagen, or proteins that interact with collagen, including the enzymes involved in post-translational processing of collagen (e.g. lysyl hydroxylase).

There are a range of such genetic disorders and their incidence can show ethnic and racial differences in distribution. The development and extension of antenatal and postnatal screening for such conditions, together with treatment regimes, is needed internationally to help deal with the consequences of these disorders.

Measuring rates of enzyme-controlled reactions

The rate of an enzyme-catalysed reaction is the amount of substrate that has been consumed from a reaction mixture, or the amount of product that has been formed. The SI units of rate (Chapter 6) are moles per cubic decimetre per second, $\mathrm{mol\,dm^{-3}\,s^{-1}}$, but other units such as moles per minute ($\mathrm{mol\,min^{-1}}$) or cubic centimetres per second ($\mathrm{cm^3\,s^{-1}}$) are also used.

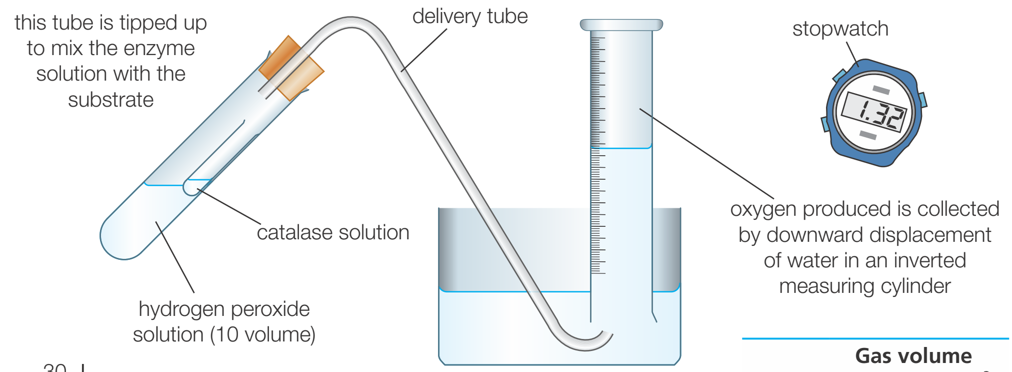
One well-studied enzyme is catalase, which catalyses the breakdown of hydrogen peroxide to water:

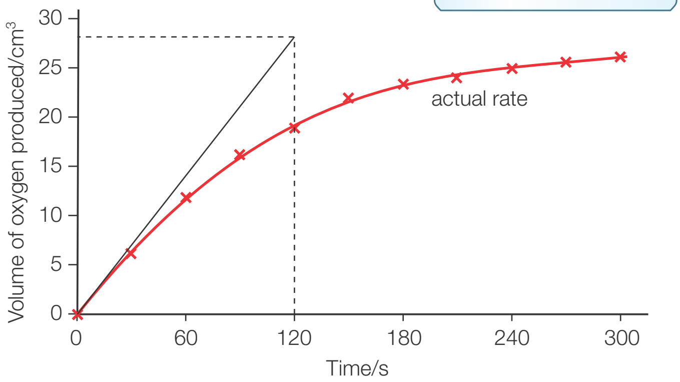
$$2H_2O_2 \rightarrow 2H_2O + O_2$$

Catalase occurs in all cells (liver cells are an especially rich source) and protects them from hydrogen peroxide, a highly oxidizing minor by-product of respiration and a chemical employed by natural killer cells in the immune system.

The rates of enzyme-controlled reactions can be followed by measuring any variable that varies with time during the reaction, for example pH, absorbance, turbidity (cloudiness) and, in the case of catalase, total gas volume. The results are plotted on a graph. Tangents can be drawn to the curve obtained to calculate initial rates (Figure 23.60).

Figure 23.60 Measuring the rate of reaction, using catalase (from liver)


this tube is tipped up to mix the enzyme solution with the substrate
delivery tube
stopwatch
1.32
catalase solution
oxygen produced is collected by downward displacement of water in an inverted measuring cylinder
hydrogen peroxide solution (10 volume)



Time/s	Gas volume collected/cm^3
30	6
60	12
90	16
120	19
150	22
180	23
210	24
240	25
270	25.5
300	26

If the initial rate of O_2 production continued for 120 s, then 28 cm^3 of O_2 would be produced

Therefore the initial rate = $\frac{28}{120}$ cm^3 s^{-1} = 0.23 cm^3 s^{-1}

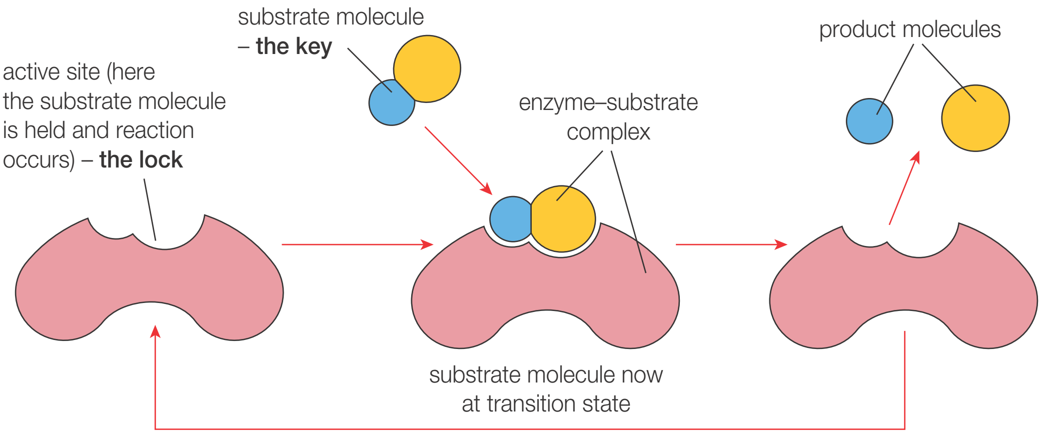
Mechanisms of enzyme action

Enzymes function as catalysts by binding to their substrate molecule(s) at a specific pocket or cleft in the enzyme. The binding site is known as the active site and is where catalysis occurs. The active site contains specific amino acid residues which are responsible for the substrate specificity and catalysis, often acting as proton donors or acceptors. The active site is also the site of inhibition of enzymes.

The activity and specificity of many enzymes can be explained by the **lock and key hypothesis** (Figure 23.61). As the enzyme (E) and substrate (S) interact they form an enzyme–substrate complex (ES), which forms a transition state that breaks down to form products (P) and unchanged enzyme (E).

Enzyme + Substrate ⇋ Enzyme–Substrate → Enzyme + Products
(lock) (key) (key in lock)
$$E + S \leftrightharpoons ES \rightarrow E + P$$

Figure 23.61 The lock and key hypothesis


substrate molecule – **the key**
product molecules
active site (here the substrate molecule is held and reaction occurs) – **the lock**
enzyme–substrate complex
substrate molecule now at transition state

Induced fit model

Enzymes were originally considered to be a rigid template in which the substrate had to fit like a key in a lock. However, it became apparent that a rigid fit between molecular structures cannot explain all aspects of enzyme catalysis. The lock and key model of enzyme activity does not fully account for the combined events of binding and simultaneous chemical change observed in some enzyme-catalysed reactions. It also fails to account for the broad specificity of some enzymes – that is, the ability of enzymes to bind to several related substrates.

In 1958, Daniel Koshland postulated that the substrate may cause an appreciable change in the three-dimensional relationship of the amino acids at the active site. The idea of a precise fit was retained from Fischer's lock and key model, but the fit occurred only after the changes induced by the substrate itself. In the active sites of some enzymes a small but essential change of shape is induced in the enzyme molecule when the substrate binds. This change in shape is critical in converting the substrate to resemble the transition state. An analogy for the induced fit hypothesis is that of a hand slightly changing the shape of a glove as the glove is put on (Figure 23.62).

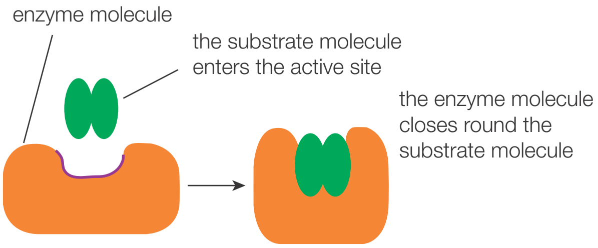

enzyme molecule
the substrate molecule enters the active site
the enzyme molecule closes round the substrate molecule

Figure 23.62 The induced fit model of enzyme action

Once the transition state is formed, other amino acids residues of the active site catalyse the breaking of specific bonds in the substrate molecule. The **induced fit model** (Figure 23.63) is based on experimental data that suggests the active sites of some enzymes are relatively 'flexible' structures.

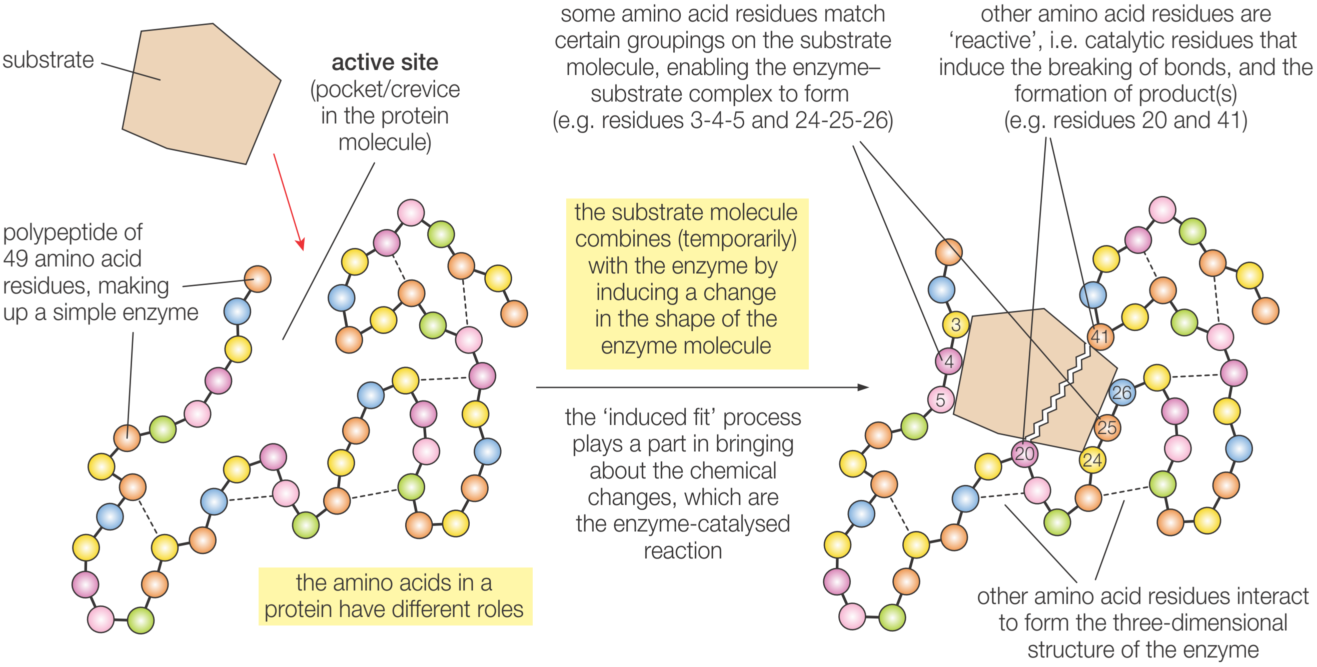

substrate
active site (pocket/crevice in the protein molecule)
polypeptide of 49 amino acid residues, making up a simple enzyme
the amino acids in a protein have different roles
the substrate molecule combines (temporarily) with the enzyme by inducing a change in the shape of the enzyme molecule
the 'induced fit' process plays a part in bringing about the chemical changes, which are the enzyme-catalysed reaction
some amino acid residues match certain groupings on the substrate molecule, enabling the enzyme–substrate complex to form (e.g. residues 3-4-5 and 24-25-26)
other amino acid residues are 'reactive', i.e. catalytic residues that induce the breaking of bonds, and the formation of product(s) (e.g. residues 20 and 41)
other amino acid residues interact to form the three-dimensional structure of the enzyme
3
4
5
41
26
25
24
20

Figure 23.63 A detailed view of the induced fit model of enzyme action

Induced fit has the advantage that it allows the possibility of ordered binding to occur. For example, with some enzymes that have two substrates (say, A and B), one of them (A, for example) binds and induces a conformational change such that the enzyme closes partly to form the binding pocket for the second substrate (in this case, B). This is an example of how induced fit would enhance the subtlety and rate of catalysis of an enzyme compared to the lock and key model. This is because, with the lock and key model, the two binding sites for A and B would have to be present all the time and, if B were to bind first, it might physically block A from binding, and the reaction would not occur.

Nature of Science

Refining a model

Science involves an ever-changing and developing body of knowledge – ideas build on previous theories and extend them as new information and greater experimental sophistication enhance understanding. The induced fit theory introduced by Koshland in the 1950s has superseded the lock and key theory of enzyme activity but it is important to see that the basis of the current model, or analogy, is embedded in the original. The induced fit model is a subtle development from the lock and key analogy. It is not that the earlier model is now totally erroneous, but simply that we now have a more comprehensive and developed view of enzyme activity. Both models emphasize the key importance of the need for the integrity of an enzyme's structure to be maintained for it to function effectively.

Indeed although the induced fit model provides a reasonably good description of how enzymes work it is now clear that the model needs further refinement. Rather than the binding of the substrate 'forcing' or inducing a change to the enzyme, it is actually much more accurate to think of the free enzyme as being in a dynamic state in which most of the time it looks much like the 'free state' (as seen in crystal structures of free enzyme, for example) but it also undergoes motional processes in which it can adopt other conformations, some of which look much more like the conformation of the active enzyme.

The substrate is then thought to bind to one of these more active conformations, altering the shape to that of the enzyme–substrate complex (ES). Thus, in this model, there is a pre-existing conformational equilibrium which is disturbed by the binding of the substrate. This model has been called the **conformational selection** or **pre-existing equilibrium model** and is widely agreed to be in general a more accurate model than induced fit.

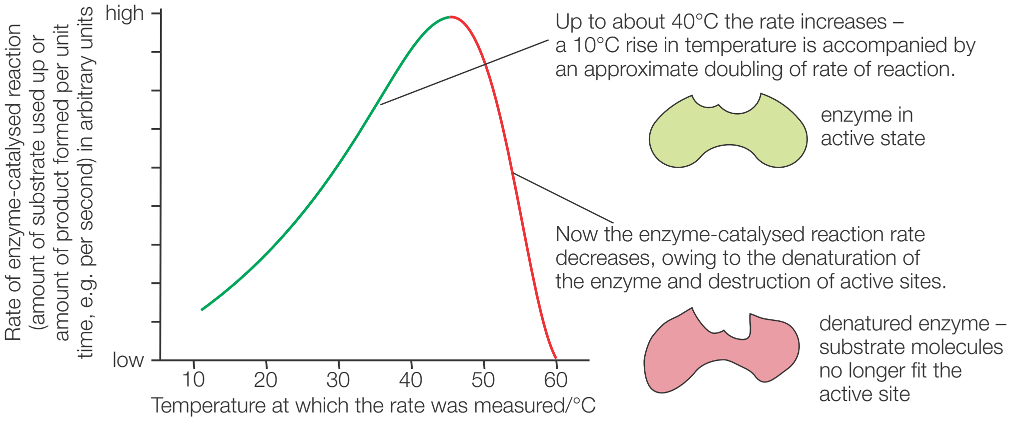
While beyond the scope of the current IB syllabus, the development of this model serves to illustrate how scientific ideas develop – in this case, around the concepts of the importance of subtle changes of shape to the functioning of a protein. This is an idea we will meet again in discussion of the function of hemoglobin in a later section.

Factors that affect enzyme activity

Temperature

Temperature has two effects on the rate of an enzyme-catalysed reaction. An increase in temperature always increases the number of effective collisions – collisions that have sufficient combined kinetic energy to bring about the reaction (Chapter 6). Initially the rate of the reaction increases exponentially with increasing temperature (Chapter 16) until a maximum rate is achieved. However, beyond this temperature the rate of reaction decreases, often rapidly, and this loss of activity is often irreversible (Figure 23.64).

Figure 23.64 The effect of temperature on a typical human enzyme



Up to about 40°C the rate increases – a 10°C rise in temperature is accompanied by an approximate doubling of rate of reaction.

Now the enzyme-catalysed reaction rate decreases, owing to the denaturation of the enzyme and destruction of active sites.

The activity of an enzyme depends on its precise three-dimensional shape (conformation). Many of the intermolecular forces maintaining that structure are relatively weak. As the temperature rises, the weak inter-molecular forces between and within the polypeptide chains break as molecular movement increases. At quite moderate temperatures, the molecules start to unravel and become disordered (a so-called random coil). Some enzymes are much more susceptible to temperature change than others; each enzyme is said to have a temperature optimum at which it works best.

There are bacteria, known as archaebacteria, living in the hot springs of the Yellowstone National Park; one bacterium has been isolated which grows best at about 105 °C and can survive in superheated water at 113 °C. Enzymes which have unusually high optimum temperatures are known as thermostable enzymes and are being used increasingly in industry. One example is *Taq* polymerase; this can be used at 60–90 °C to increase the amount of DNA available for DNA profiling.

Heavy metal ions

The **heavy metals** are metals with a relatively high relative atomic mass. Examples of heavy metals include mercury, cadmium, zinc and silver. Heavy metals and their ions can act as irreversible inhibitors of some enzymes at very low concentrations. They form bonds with free −H groups present in the amino acid cysteine. The free −SH groups, if present in the active site, may be essential to the activity of the enzyme (Figure 23.65).

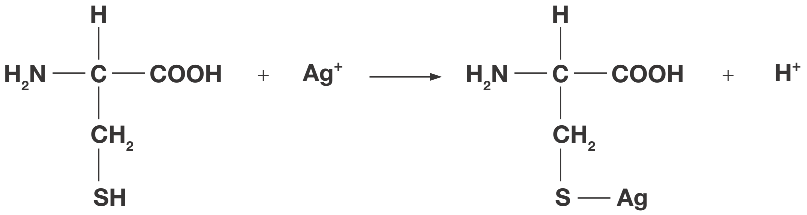


Figure 23.65 The action of silver ions on the –SH functional group of cysteine

pH

Many enzymes work efficiently over a narrow range of pH values. The optimum pH is the pH value at which the maximum rate of reaction occurs. For many enzymes their optimum pH is close to neutrality (pH 7). When the pH value is above or below this value, the rate of enzyme activity is significantly decreased. Most enzymes exhibit a characteristic bell-shaped curve of enzyme activity against pH (Figure 23.66).

Changes in pH alter the charge of the acidic groups (–COO^-) and basic groups (–NH_3^+) present in the amino acid residues of the enzyme's active site. This leads to a change in the enzyme's shape, particularly at the active site. The effects of a small change in pH are usually reversible, and if the pH is restored to the optimum for the enzyme its activity may be restored. Buffer solutions (Chapter 18) are often used during investigations involving enzymes to maintain a constant pH.

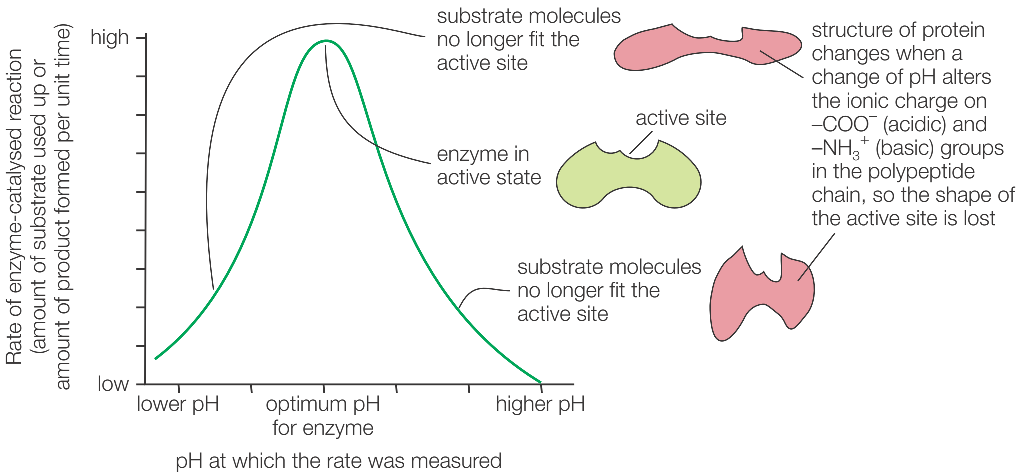


Figure 23.66 The effect of pH on enzyme shape and activity

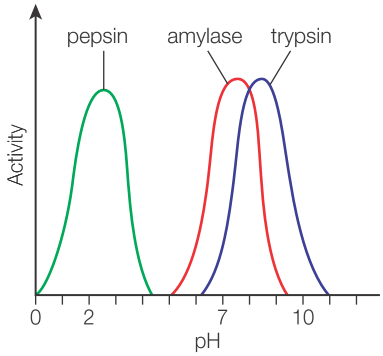


Figure 23.67 The pH profiles for pepsin, amylase and trypsin

Digestive enzymes show clearly how each enzyme has its own distinct optimum pH, reflecting the environment in which it is operating (Figure 23.67):

- Pepsin hydrolyses proteins to peptides in the very acidic conditions of the stomach.
- Amylase, found in saliva, hydrolyses starch to a mixture of glucose and maltose. Saliva is approximately neutral.
- Trypsin hydrolyses peptides to amino acids in the mildly alkaline conditions of the small intestine.

17 **a** Sketch the energy profile of an uncatalysed exothermic reaction, showing:
i the activation energy (E_a)
ii the enthalpy change of reaction (ΔH_r).
b Sketch a similar energy profile for the reaction in part a when the reaction is enzyme-catalysed.

18 The turnover number of an enzyme is the number of substrate molecules that one molecule of an enzyme can convert to product in unit time under optimum conditions. The turnover numbers of several enzymes are given in Table 23.4 on p. 37.
a In measuring the turnover number of a particular enzyme, there must be an excess of the substrate present. Why is this so?
b What effect would a fall in temperature of 10°C have on the turnover number of the enzymes? Explain your answer.

23.3 Lipids *– lipids are a broad group of biomolecules that are largely non-polar and therefore insoluble in water*

Lipids are a varied group of biochemical compounds that contain the elements carbon, hydrogen and oxygen, but the proportion of oxygen is less than in carbohydrates. They are grouped together because of their non-polar nature. Lipids are largely insoluble in water, but are soluble in non-polar solvents such as hexane. Lipid molecules are not polymers but their non-polar nature means that they tend to group together when placed in water.

Lipids play essential roles in cell structure and metabolism. Biochemically important lipids include the following: triglycerides, phospholipids and steroids.

- **Triglycerides** occur in animal fats, which are semi-solid at room temperature, and vegetable oils, which are liquid at room temperature. Triglycerides are the major storage form of the energy needed to drive reactions in plants and animals.
- **Phospholipids** (glycerophospholipids or phosphoglycerides) are important constituents of cell membranes.
- **Steroids** are another series of biologically important molecules.

An epidemic of obesity

Fat is an important part of the human diet. Among tribal communities in Africa a normal cycle of accumulating and using body fat can be demonstrated. Immediately after harvest each year everyone in a rural village gains weight. Most adults double their reserves of body fat at this time. This fat is then used to prevent starvation when food stores diminish later in the year. This natural cycle reminds us that the ability to lay down stores of energy in the form of fat is an evolutionary advantage – fat people are more likely to survive a famine.

Obesity is often stigmatized in the modern Western world. However, it has been perceived as a symbol of wealth and fertility at other times in history, and still is in many parts of Africa.

Obesity is a condition in which excess body fat has accumulated to an extent that health may be negatively affected. Excessive body weight is associated with various diseases, particularly heart disease, diabetes mellitus and certain types of cancer. Hence, obesity reduces life expectancy. Obesity is generally due to lack of exercise and a diet that supplies energy in excess of the body's requirements. Obesity is usually treated with dieting and physical exercise. The risk of obesity is greater when the diet is high in fat. In this context fats are divided into 'good fats' and 'bad fats'. Good fats include mono-unsaturated and poly-unsaturated fats. 'Bad fats' include saturated fats and *trans* fats.

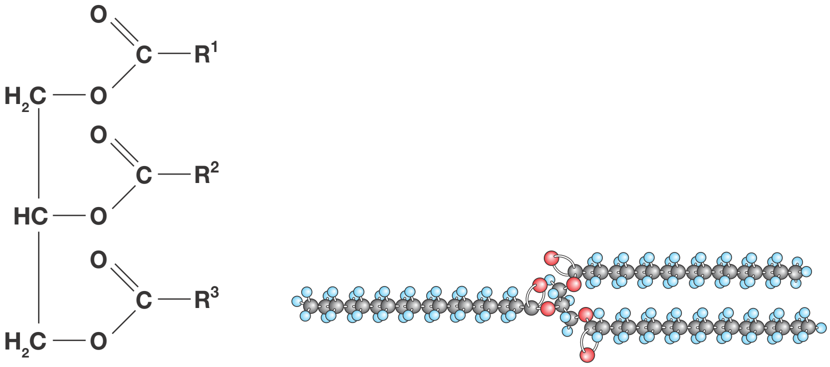
With supermarkets and restaurants open long hours and much food now highly processed, most modern societies have eliminated famine. However, the penalty for this is a rise in obesity. Obesity now ranks with asthma as one of the fastest-growing medical epidemics afflicting the West. The proportion of obese adults in the UK doubled during the 1980s and is continuing to rise.

In 1995, newspapers in the USA reported the death of the fattest man in the world. He died weighing 465 kg. The tragedy of his situation was emphasized by the fact that it once required a fork-lift truck to transport him to hospital. And yet, his over-eating need not have been particularly excessive. If he had started at a weight of 70 kg when he was 16 years old, it has been estimated that he merely needed to eat slightly less than a bar of chocolate too much each day in order to reach his final weight. His was an extreme case of a loss of regulatory control.

Most of us regulate our body weight more tightly. Even so, the growing number of people becoming obese indicates that our sedentary lifestyle and high-fat diet are taking their toll. The need for scientists to develop a greater understanding of how the body balances energy input with its energy requirements is going to be of great importance in averting major health problems.

Triglycerides

Triglycerides are the most common lipids in living organisms and are classified as fats (semi-solid at room temperature) or oils (liquid at room temperature), depending upon their physical state at 20 °C. Triglycerides are formed from the condensation reactions of propane-1,2,3-triol (glycerol) and fatty acids (long-chain carbon carboxylic acids) (Figure 23.68). They are non-polar and hence do not dissolve in water. The fatty acid chains of a triglyceride can be classified as saturated or unsaturated, depending on whether or not it contains only carbon–carbon single bonds.



■ **Figure 23.68 a** The generalized formula of a triglyceride, where R^1, R^2 and R^3 represent fatty acid chains, which may be the same or different. **b** A model of tristearin, a triglyceride found in animal fat

Simple triglycerides contain three molecules of one particular fatty acid bonded to one molecule of propane-1,2,3-triol. For example, tristearin, which is found in the fatty tissue of animals, is formed from propane-1,2,3-triol and three molecules of stearic acid.

Most naturally occurring triglyceride molecules are mixed triglycerides: they have two or three different fatty acids attached to the propane-1,2,3-triol molecule. The fatty acid composition varies with the organism that produces them.

Structure of fatty acids

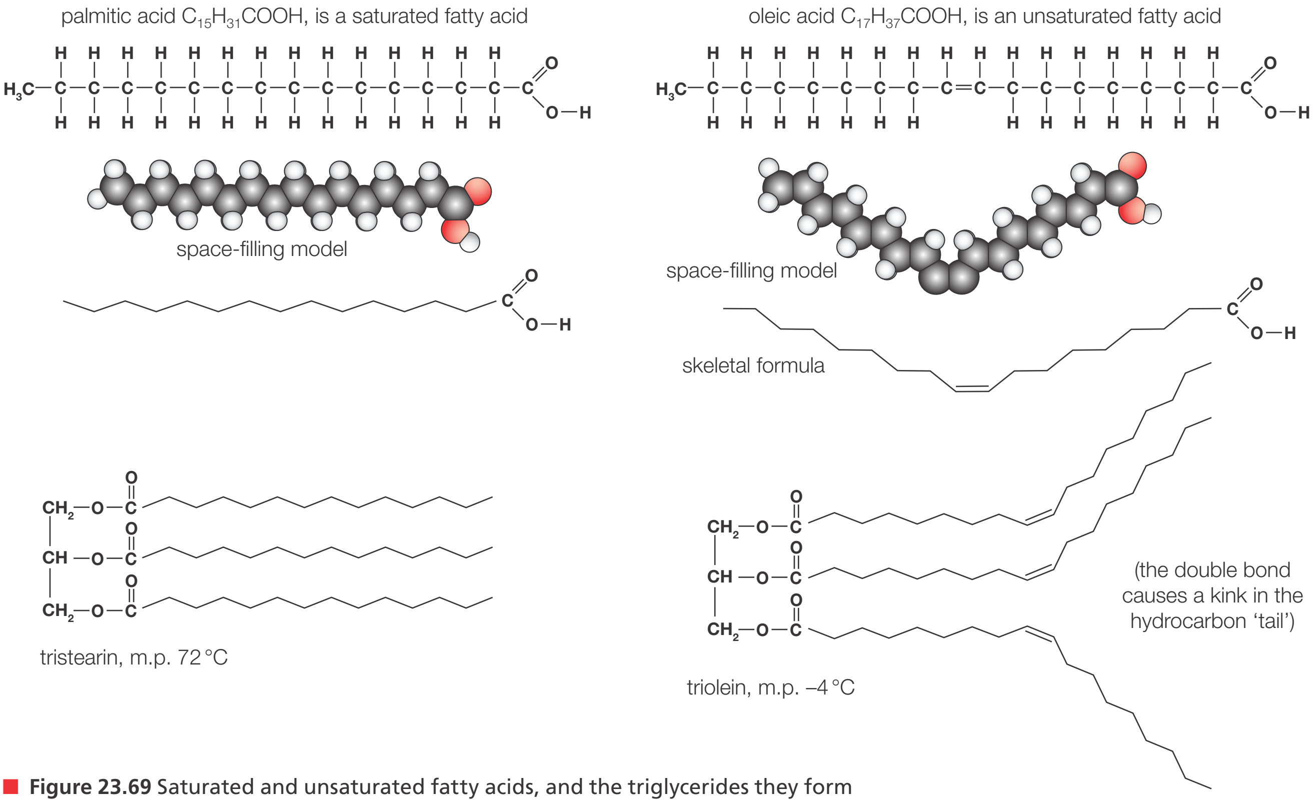
The major distinction in the fatty acid composition of triglycerides is between saturated and unsaturated fatty acids. **Saturated** fats and oils have no carbon–carbon double bonds between the carbon atoms in their fatty acids. They contain fatty acids such as palmitic acid and stearic acid. **Mono-unsaturated** or **poly-unsaturated fats** have one or more carbon–carbon double bonds in their fatty acids; they involve fatty acids such as oleic acid and linoleic acid (Figure 23.69).

Table 23.7 The melting points of selected saturated and unsaturated fatty acids

Name of fatty acid	Number of carbon atoms	Number of C=C double bonds	Melting point/°C
Saturated fatty acids			
Lauric acid $CH_3(CH_2)_{10}COOH$	12	0	44.2
Myristic acid $CH_3(CH_2)_{12}COOH$	14	0	54.1
Palmitic acid $CH_3(CH_2)_{14}COOH$	16	0	62.7
Stearic acid $CH_3(CH_2)_{16}COOH$	18	0	69.6
Arachidic acid $CH_3(CH_2)_{18}COOH$	20	0	75.5
Unsaturated fatty acids			
Palmitoleic acid $CH_3(CH_2)_5CH{=}CH(CH_2)_7COOH$	16	1	−0.1
Oleic acid $CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$	18	1	10.5
Linoleic acid $CH_3(CH_2)_4CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$	18	2	−5.0
Linolenic acid $CH_3CH_2CH{=}CHCH_2CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$	18	3	−11.0
Arachidonic acid $CH_3(CH_2)_5(CH{=}CHCH2)_4(CH_2)COOH$	20	4	−49.0

Unsaturated fatty acids tend to have lower melting points than saturated fatty acids (see Table 23.7). This means that fats containing unsaturated fatty acids melt at lower temperatures than those with saturated fatty acids. This trend in melting points is the consequence of a steric effect and occurs because the introduction of a double bond prevents the triglyceride molecules from approaching each other closely and hence interacting via London dispersion forces.

As can be seen from Figure 23.68, long-chain saturated fatty acids have a regular tetrahedral arrangement of carbon atoms. This means that they can pack closely together and the dispersion forces between chains are strong because of their extended surface area. In unsaturated fatty acids the bond angle in the chains changes around the double bond and the structure becomes rigid at that point. This introduces a kink in the chain (Figure 23.69) and they are unable to pack so closely together. The dispersion forces become weaker which results in these acids having lower melting points. This packing arrangement is similar in the triglycerides and explains why unsaturated fats (oils) have lower melting points.


palmitic acid $C_{15}H_{31}COOH$, is a saturated fatty acid
space-filling model
oleic acid $C_{17}H_{37}COOH$, is an unsaturated fatty acid
space-filling model
skeletal formula
tristearin, m.p. 72 °C
triolein, m.p. –4 °C
(the double bond causes a kink in the hydrocarbon 'tail')

■ **Figure 23.69** Saturated and unsaturated fatty acids, and the triglycerides they form

This melting point trend seems to be a significant factor in influencing the triglyceride composition produced by different organisms. Animal fats are usually richer in saturated fatty acids than plant oils (Table 23.8). In warm-blooded animals, body temperature is high enough to allow most fats to be above their melting temperature. This eases their transport around the body in liposomes. These fats tend to be highly saturated, with a high content of palmitic acid and stearic acid.

Plants have no means of keeping themselves warm when faced with cold conditions. Plant oils must therefore have lower melting points and are often highly unsaturated. They contain high proportions of oleic acid and linoleic acid.

Fish oils are highly unsaturated for similar reasons. They can contain long fatty acid groups with as many as six C=C bonds.

■ **Table 23.8** Approximate composition of selected naturally occurring lipids

Fat or oil	Main fatty acids
Palm oil	Oleic (45%), palmitic (40%)
Olive oil	Oleic (80%), linoleic acid (10%)
Lard	Oleic (56%). palmitic (28%), stearic (8%)
Butter fat	Oleic (30%), palmitic (30%), stearic (11%), myristic (10%)
Ground nut	Oleic (57%), linoleic (23%), palmtic (12%)

Most naturally occurring fats and oils contain a mixture of saturated, mono-unsaturated and poly-unsaturated fatty acids and are classified according to the predominant type of unsaturation present (Figure 23.70).

For example, linseed soil (from the flax plant) has a relatively low percentage of saturated fatty acid residues and hence is classified as an unsaturated fat. In contrast, beef tallow – extracted from beef fat – is high in saturated fat and low in unsaturated fatty acids and hence is classified as a saturated fat. Animal lipids are generally saturated and vegetable lipids tend to be unsaturated.

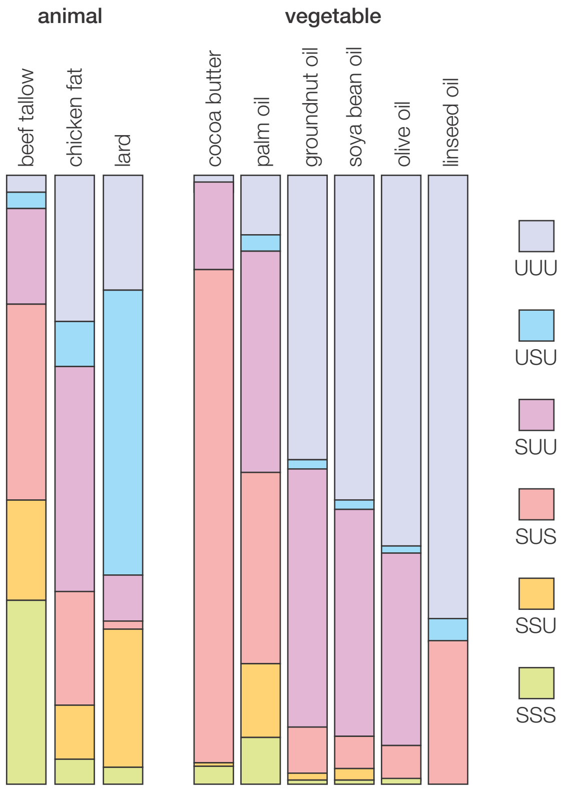


Figure 23.70 Triacylglycerol components of lipids, classified according to the position and degree of saturation of their fatty acid residues. For example, UUU indicates a triacylglycerol in which the three fatty acid residues are unsaturated

Essential fatty acids

Just as we saw with the proteogenic 2-amino acids, there are certain poly-unsaturated fatty acids which cannot be synthesized in the human body and therefore must be supplied from the diet. Linoleic acid (Figure 23.71) is a member of a group of **essential fatty acids** called the omega-6 fatty acids. Chemically, linoleic acid is a carboxylic acid with an 18-carbon chain and two *cis* carbon–carbon double bonds; the first double bond is located at the sixth carbon from the furthest (or *omega*, ω) end from the acid group. The omega end refers to the methyl end of the fatty acid chain. Linoleic acid is an essential fatty acid found in many vegetable oils, especially safflower and sunflower oils. Linoleic acid is used in the synthesis of prostaglandins.

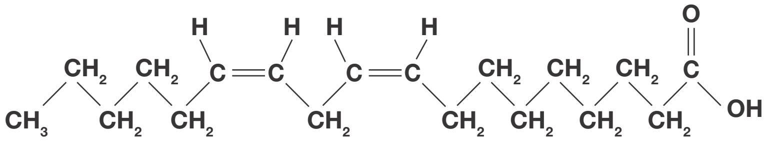


Figure 23.71 The structure of linoleic acid (*cis,cis*-9,12-octadecadieonic acid)

Linolenic acid (Figure 23.72) is an essential fatty acid found in rapeseed, soya beans, walnuts and hemp. Green-leaved vegetables are also good sources of linolenic acid. Linolenic acid is an omega-3 fatty acid.

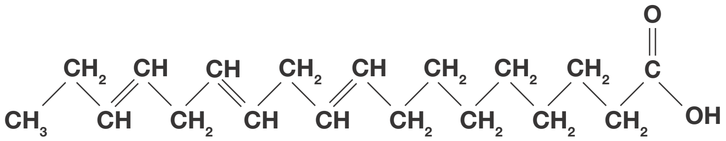


Figure 23.72 The structure of linolenic acid

■ **Figure 23.73** Flaxseed (linseed) oil is rich in omega-3 fatty acids

Linolenic acid is a carboxylic acid with an 18-carbon chain and three *cis* carbon–carbon double bonds. The first double bond is located at the third carbon from the omega end. Studies have found evidence that linolenic acid is related to a lower risk of cardiovascular disease.

Note that oleic acid, $CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$, can be made in human metabolism and is a **non-essential fatty acid**. Using the omega notation, it is an ω-9 (omega-9) fatty acid. Plants, seeds and vegetable oils are good dietary sources of omega-6 fatty acids, while fish, shellfish and flaxseed oil are notably rich in omega-3 fatty acids (Figure 23.73).

■ The hydrogenation of oils

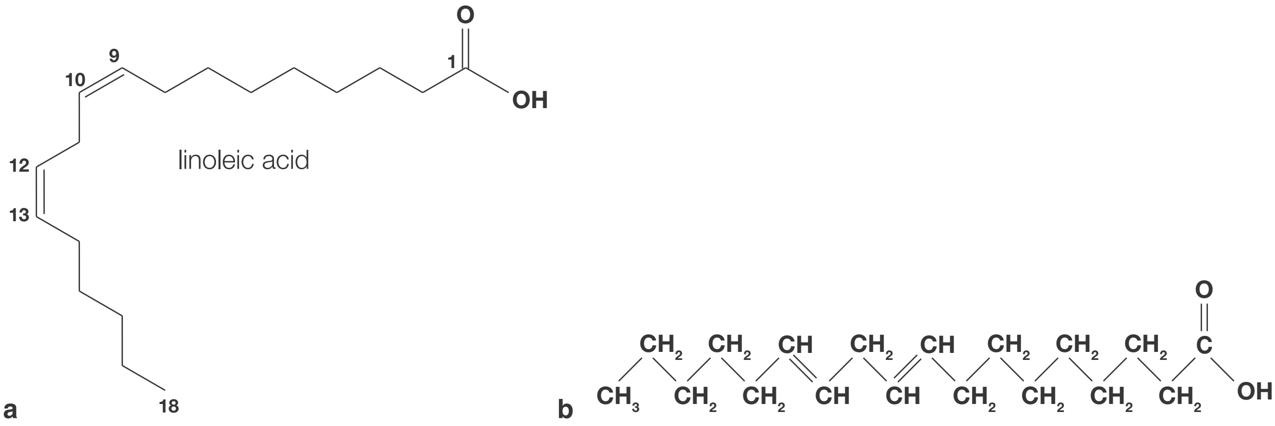
Unsaturated oils can be hydrogenated to convert them to semi-solid fats with a lower degree of unsaturation. Margarine is made by the hydrogenation of corn oil or sunflower oil. The liquid component of margarine is made from milk and water. This process is called 'hardening' and the principal oil involved is oleic acid, which may be hardened by converting it to stearic acid:

$$\underset{\text{oleic acid}}{CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH(l)} + H_2(g) \rightarrow \underset{\text{stearic acid}}{CH_3(CH_2)_{16}COOH(s)}$$

The finely divided metal catalyst (nickel, copper or zinc) is mixed with the oil and this is heated to a temperature of about 180 °C. The mixture is stirred and the hydrogen bubbled through the reaction mixture. After hydrogenation, the catalyst and the oil are separated by simple filtration. In addition to hardening the oil, hydrogenation increases its chemical stability.

The process of catalytic hydrogenation is intended to add hydrogen atoms to *cis*-unsaturated fats, eliminating a double bond and making them more saturated. These saturated fats have a higher melting point, which makes them attractive for baking and solid but spreadable from the refrigerator, and extends their shelf life.

However, a partial process is often used that has a side effect that converts some *cis*-isomers into *trans*-unsaturated fats instead of hydrogenating them completely. *Trans*-unsaturated fats (Figure 23.74) have a straight, rather than kinked, shape for the carbon chain, more like the straight chain of a fully saturated fat. A margarine may contain up to 20 per cent of fatty acids with *trans* double bonds.



■ **Figure 23.74 a** The structures of the *cis,cis* form of linoleic acid and **b** the straightening of the *trans,trans* form of linoleic acid

Trans fats occur when the *cis*-double bonds in the fatty acid chains are not completely saturated during the hydrogenation process. The catalysts used to aid the addition of hydrogen appear to cause the remaining double bonds to isomerize to their *trans* configuration. These unnatural *trans* fats appear to be associated with increased heart disease, cancer, diabetes and obesity, and to give rise to immune response and reproductive problems.

Several large studies have indicated a link between consumption of high amounts of *trans* fats and coronary heart disease. This is mainly because *trans* fats increase the amount of LDL cholesterol and decrease the amount of HDL cholesterol in the bloodstream. *Trans* fatty acids are used by the body like saturated fats, mainly in respiration, but tend to block the use of omega-3 and omega-6 fatty acids for vital bodily functions.

Trans fats are only present in small amounts in natural oils and fats of plants and animals. The following products contain *trans* fats: margarine and other spreads, cake mixes, fast foods (fries and fried chicken), baked products (biscuits and cakes), salad dressings and crisps.

Although there seems compelling evidence for a correlation between the consumption of *trans* fats and adverse effects on human health, a definitive link has not been established.

Nature of Science

A causal relationship?

There has been evidence put forward that saturated fatty acids, particularly lauric (C_{12}), myristic (C_{14}) and palmitic (C_{16}) acids, increase the levels of low-density lipoprotein cholesterol (LDL cholesterol) as do *trans* unsaturated fats. Conversely omega-3 poly-unsaturated fatty acids, such as in natural unsaturated oils (olive oil, for instance), are thought to lower LDL cholesterol and consequently be beneficial. The hydrogenation of unsaturated oils to produce semi-solid fats has had unintended consequences, increasing the level of *trans* fats in the diet in parts of the world. Although hydrogenation imparts desirable features – such as spreadability, texture, 'mouth feel' and increased shelf life – to naturally liquid vegetable oils, it may introduce some serious health problems.

Such considerations can even affect our liking for chocolate – particularly if we live in a warm climate. Cocoa butter is a pale-yellow vegetable fat extracted from ground cocoa beans. It is used to make chocolate, ointments and suppositories. Cocoa butter is one of the most stable fats known, containing natural antioxidants that prevent rancidity and give it a storage life of 2–5 years, making it a good choice for non-food products. Cocoa butter has a narrow melting range that is close to body temperature. This accounts for its characteristic 'melt-in-your mouth' property. About 80% of its triglycerides in cocoa butter are of one class, which contain palmitic or stearic acids and oleic acid. The triglycerides of this class all have similar shapes and pack together relatively well, giving a relatively sharp melting range. However, the chocolate produced in warm countries needs to have a higher melting point and to achieve this the cocoa butter used is partially hydrogenated to decrease its unsaturation level. This increases the level of *trans* fats in the product.

There have been a number of studies on the influence of the lipid content of a regional diet on human longevity – with reference made to the Mediterranean diet, rich in olive oil, and the Japanese diet, with its emphasis on fish oils. The aboriginal population of the Arctic has a heavily meat-based diet high in fats and proteins. The metabolism of the Inuit peoples has adapted to their diet and they are able to synthesize large amounts of glucose from fat and protein metabolites. This process of neoglucogenesis requires large amounts of energy derived from the high fat consumption. In addition, the fats that predominate in the Inuit diet are naturally rich in mono- and poly-unsaturated fatty acids and so do not impose the same health risk as the typical Western diet.

The scientific evidence links the consumption of saturated and *trans* fats with elevated LDL cholesterol and the incidence of coronary heart disease. The hypothesis is that by raising LDL cholesterol the *trans* fats contribute to the formation of atherosclerotic plaques in arteries. There is a strong correlation between the consumption of *trans* and saturated fatty acids and the incidence of coronary heart disease, but a causal relationship is very difficult to prove because other factors also play a role, including genetics, other foods, consumption of alcohol and lifestyle.

Iodine number and unsaturation

We have seen that the degree of unsaturation of a fat or oil is a significant property of lipids. Saturated lipids do not react with bromine or iodine, but unsaturated lipids will undergo an addition reaction with halogens.

The **iodine number** is the mass of iodine in grams that reacts with 100 g of a chemical substance, such as unsaturated lipid. An iodine solution is yellow–brown in colour and any carbon–carbon double bonds in the lipid that reacts with iodine will make the colour disappear at a precise concentration. The amount of iodine solution thus required to keep the solution yellow–brown is a measure of the amount of the unsaturation in the lipid.

19 In an experimental determination of iodine number, 0.01 moles of linoleic acid reacts with 1.5 grams of iodine.

a Determine the number of carbon–carbon double bonds present in the fatty acid.
b Calculate the iodine number of linoleic acid, $C_{17}H_{31}COOH$.

20 Fats and vegetable oils can be saturated or unsaturated. A simple experiment can be carried out to measure the degree of unsaturation in these compounds.

Five drops of a liquid, or a similar volume of a solid, are dissolved in 4 cm^3 of ethanol. A dilute iodine solution is added a drop at a time. At first, the brown colour disappears. When enough iodine has been added to react with all the double bonds in the sample, the brown colour remains. The number of drops of iodine solution needed to produce a permanent brown colour is recorded. This test was carried out on a variety of cooking products. The results are shown in the following table:

Cooking product	Mass of saturated fat in 100 g of cooking product/g	Mass of unsaturated fat in 100 g of cooking product/g	Number of drops of iodine solution used
Olive oil	11	84	14
Peanut oil	20	72	12
Butter	45	30	5
Soft margarine	35	40	
Poly-unsaturated margarine	11	66	

a Why was ethanol, and not water, used as a solvent in this experiment?
b Suggest values to complete the table.
c One medical theory is that using unsaturated fats in the diet, instead of saturated fats, will reduce the number of cases of heart disease. Which of the cooking products in the table is the least likely to cause heart disease?

21 A sample of vegetable oil (2.5 g) reacted completely with 19 cm^3 of a 0.50 $mol\,dm^{-3}$ solution of iodine.
a What is the iodine number of the oil?
b Estimate the average number of carbon-carbon double bonds per molecule of this oil if its average molecular mass is 865 $g\,mol^{-1}$.
c Why is the figure in **b** necessarily an average figure for the number of double bonds per molecule?

22 **a** Predict and explain which fatty acid in each group has the highest melting point.
i Butanoic acid, palmitic acid and stearic acid.
ii Linoleic acid, oleic acid and linolenic acid.
b Chocolate is a food made from cocoa, sugars, unsaturated vegetable fats, milk whey and emulsifiers. Chocolate bars sold in hot climates are made with a different blend of vegetable fats from those sold in colder climates.
i Explain why fats with a different physical properties are used for making chocolate in different climates.
ii Suggest how the fat molecules used in a hot climate differ chemically from those used in a cold climate.

Fat or oil	Iodine number
Soya bean oil	122–134
Olive oil	80–90
Bacon fat	47–67
Beef fat	35–45

■ **Table 23.9** Typical iodine number values of certain fats and oils

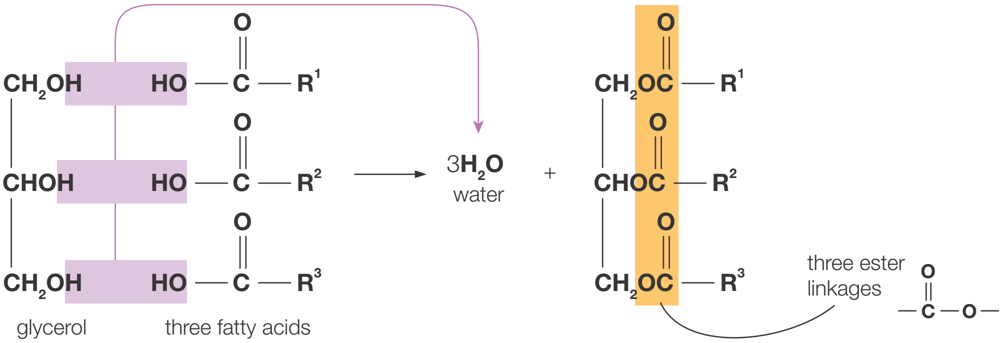
Animal fats contain relatively few carbon–carbon double bonds and thus have low iodine numbers, typically in the region of 40–70 (see Table 23.9). Vegetable and fish oils have a greater degree of unsaturation, and hence higher values of iodine number, often lying between 80 and 140, though they can be as low as 10 for coconut oil and as high as 200 for linseed and fish oils.

Triglyceride formation and hydrolysis

Condensation

Glycerol (propane-1,2,3-triol) has three hydroxyl groups (−OH), all of which can undergo a condensation reaction with a fatty acid molecule to form an ester. Usually all three hydroxyl groups undergo an enzyme-controlled condensation reaction, as shown in Figure 23.75, and the lipid formed is a triester known as a triglyceride.

Figure 23.75 Formation of a trigylceride from fatty acids and glycerol molecules by condensation reaction



The fatty acid hydrocarbon chains, R^1, R^2 and R^3, may be identical. For example. tristearin has three stearic acid molecules and triolein has three oleic acid molecules. However, R^1, R^2 and R^3 are usually different.

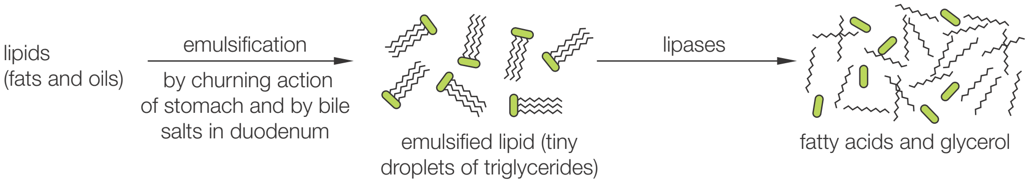
Enzyme-catalysed hydrolysis

Lipids are poorly soluble in water and hence do not undergo significant hydrolysis in water. Lipases are a group of digestive enzymes that break lipids down chemically. The major component of lipids in the human diet is triglycerides. Although a small amount of lipase is secreted by the tongue and the stomach, these digestive actions are not significant, as almost no breakdown of lipids occurs until they reach the first part of the small intestine.

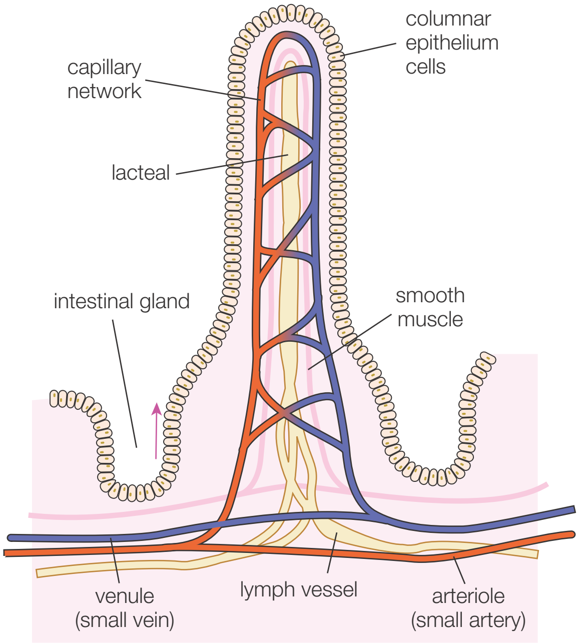
Lipid digestion and absorption require that the lipid molecules be broken down into smaller and hence more soluble molecules. The lipids mix with the lipase, which enters the small intestine from the pancreas – the main source of enzymes for digesting lipids and proteins.

Lipase hydrolyses triglyceride molecules into fatty acid molecules and glycerol molecules (Figure 23.76). However, because lipids do not dissolve in water, the lipid molecules enter the small intestine in a congealed mass. This makes it impossible for the pancreatic lipase enzymes to attack them, since lipase is a water-soluble enzyme and can only attack the surface of the lipid molecules.

Figure 23.76 Summary of the digestion of lipids



To overcome this problem the digestive system uses a substance called bile, produced in the liver but stored in the gallbladder, which enters the small intestine via the bile duct. Bile emulsifies fats – it disperses them into small droplets which then become suspended in the watery contents of the digestive tract. Emulsification allows lipase to gain easier access to the fat molecules and thus accelerates their digestion.



■ **Figure 23.77** The internal structure of a villus in the small intestine

Absorption of fatty acids and glycerol occurs in the villi – the finger-like projections which cover the walls of the small intestine. Inside each villus (Figure 23.77) is a series of lymph vessels (lacteals) and blood vessels (capillaries). The lacteals absorb the fatty acids and glycerol into the lymphatic system which eventually drains into the bloodstream. The fatty acids are transported via the bloodstream to the membranes of adipose cells or muscle cells, where they are either stored or respired for energy. The glycerol enters the liver.

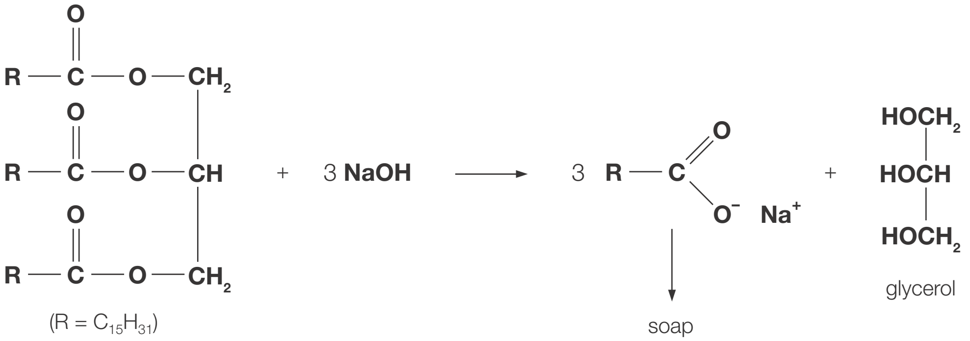
In addition to the biological function, non-enzymatic hydrolysis also has important analytical and commercial uses. Since it is the fatty acid component of triglycerides that varies from one source to another, isolating the fatty acids by hydrolysis is the first stage in the analysis of triglycerides. The fatty acids present in the different oils are separated and analysed by a variety of chromatography techniques.

The hydrolysis of triglycerides is commercially important in the manufacture of soaps. Soft soaps are generally made by hydrolysing the triglyceryl esters present in a blend of animal fats and vegetable oils by heating them with a solution of sodium hydroxide.

Soap production

Triglycerides (fats and oils) are the glyceryl triesters of long-chain carboxylic acids. Heating triglycerides with a concentrated solution of aqueous sodium hydroxide (a strong alkali) causes hydrolysis (Figure 23.78). The sodium salts of the long-chain carboxylic acids are precipitated by adding salt to the mixture, and this solid is then washed and compressed into bars of soap. Perfume and dyes are also added.

■ **Figure 23.78** Production of soap



The alkaline hydrolysis of fats and oils is known as saponification. The name saponification literally means ‘soap making’. The root word, *sapo*, is Latin for soap. Lipids can be classified into saponifiable lipids, for example triglycerides and phospholipids (as diglycerides), and non-saponifiable lipids, for example cholesterol.

Fats and oils are hydrolysed commercially in alkaline conditions. The products are glycerol and the salts of the fatty acids. The salts of the long-chain fatty acids are known as soaps. They are extremely useful as they help oil and water mix together by reducing the surface tension. Glycerol is a useful by-product of this reaction as it can be used to make pharmaceuticals and cosmetics. Soap molecules have a polar carboxylate functional group, which attracts water and other hydrophilic groups. They also have a non-polar end, which is the long-chain hydrocarbon, which attracts oils and other hydrophobic species. The same molecule is thus attracted to aqueous species and oily species, thus enabling them to mix. When added to water, the soap molecules lower the surface tension of the water, so that it wets objects more easily. The molecules also interact with the grease present (Figure 23.79). The hydrophobic hydrocarbon chain is attracted to the grease and becomes embedded in it.

The hydrophilic ionic head of the molecule sits outside the grease, in contact with water. When the water is agitated, the grease is released from the cloth fibre or dish and is completely surrounded by soap molecules. Rinsing with fresh water removes these grease–detergent droplets.

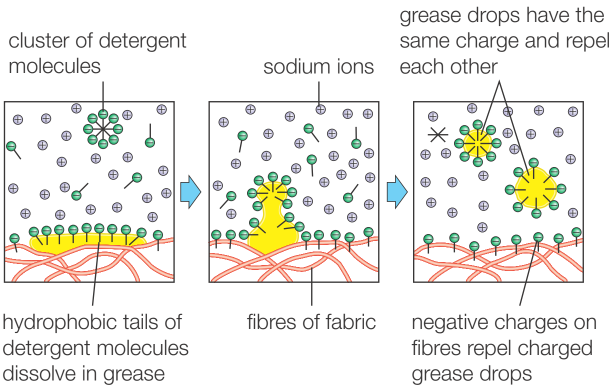

cluster of detergent molecules
sodium ions
grease drops have the same charge and repel each other
hydrophobic tails of detergent molecules dissolve in grease
fibres of fabric
negative charges on fibres repel charged grease drops

Figure 23.79 Interactions between soap and grease lift the grease from the fabric

Ghanaian black soap

Although a massive worldwide industry, soap making also has a long artisanal history and the technology for making it stretches back in history in different parts of the world. African black soap production is an integral part of village life and uses readily available raw materials. Black soap is a handmade soap, that has been used for centuries throughout western Africa, known for being gentle and alleviating skin ailments (Figure 23.80). It consists of a naturally derived emollient (unrefined shea butter) combined with the nutrient-rich ashes of native African plant materials – cocoa pod or plantain ashes. The ashes supply a natural, local and income-generating source of potassium hydroxide. Local coconut oil or palm kernel oil complete the saponification process and generate a moisturizing soap.

Figure 23.80 Making black soap using ashes as a source of alkali

Rancidity of fats

Vegetable oils and animal fats can develop an unpleasant smell if they are kept for too long. They are said to go rancid. Fatty acids formed through the hydrolysis of triglycerides are the cause of this rancidity. This type of rancidity is known as **hydrolytic rancidity**. The rancid smell and flavour is due to the release of free fatty acids such as butanoic and octanoic acids, which are released from rancid milk and dairy products. As hydrolytic rancidity is favoured by higher temperatures, it can be substantially reduced by refrigeration. This hydrolysis is speeded up by the presence of certain microorganisms (**microbial rancidity**).

Oxidative rancidity occurs when unsaturated fats react with oxygen from the air. The products responsible for the rancidity are volatile aldehydes and ketones and are a result of the reactivity in the carbon–carbon double bonds in unsaturated triglycerides. The process, known as auto-oxidation, is often accelerated by light and enzymes or metal ions. The reactions proceed via a free radical mechanism and yield a mixture of products. Oxidative rancidity is characteristic of fats and oils that have a high proportion of carbon–carbon double bonds, such as those from oily fish like herring. It can be controlled by the use of light-proof packaging, a protective (oxygen-free) atmosphere and the addition of natural or synthetic antioxidants.

Food antioxidants

Food antioxidants are compounds that increase the resistance of fats to oxidation and consequent deterioration or rancidity. They are common additives in developing countries. Only specifically approved antioxidants are able to be added to foods susceptible to rancidity. Antioxidant inclusion is restricted to specific limits and must be declared on product labels. Some of the approved antioxidants are butylated hydroxyanisole (BHA), butylated hydroxytoluene (BHT), propyl gallate and tocopherols (Chapter 9). These primary antioxidants are often used in combination with citric or phosphoric(V) acid. Use of one or more of the primary antioxidants in combination with one of the acids is common because combinations are much more effective than single antioxidants. Freezing and vacuum packaging are also used to slow down rancidity in meat sold by supermarkets in developed countries.

Natural antioxidants, such as those contained in spices, are used in developing countries for the slowing down of rancidity in meat products. These and other natural antioxidants not only slow down the rancidity in pre-cooked meat products, but also provide a pleasant smell (aroma) and flavour. Some spice extracts, particularly rosemary, are prepared primarily for their antioxidant activity and do not include strong flavour components.

Energy value of fats

Fats and oils are efficient long-term stores of chemical energy. Typical fats and oils provide approximately 38 kilojoules of energy per gram, while typical carbohydrates provide only 17 kilojoules of energy per gram (Table 23.10). This is because lipids are a more highly reduced form of biomolecule than a sugar. In other words, lipid molecules contain a higher proportion by mass of hydrogen and carbon than sugars.

Food constituent	Available energy/ kJ g^{-1} dry weight
Fat	38
Carbohydrates	17
Proteins	17
Ethanol	30
Dietary fibre	0–8

Table 23.10 The energy content of the major food components

A person's supply of glycogen acts as a relatively short-term energy store: it can provide energy for less than 24 hours. Triglycerides are the longer-term energy store in the human body. The fat content of an average person in the West (21 per cent for men, 26 per cent for women) enables them to survive starvation for 2–3 months.

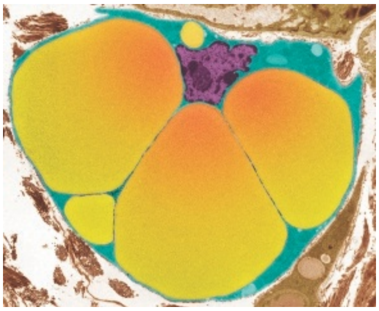
Table 23.11 summarizes the main features of glycogen and fats as energy stores.

In animals, triglycerides are synthesized and stored in specialized fat cells (Figure 23.81). These cells can be almost entirely filled with fat globules, unlike other cell types that contain only a few droplets of fat. Triglycerides, being weakly polar molecules, are stored in anhydrous form. In contrast, glycogen binds about twice its weight of water. Because of this, under the conditions found in the body, fats provide about six times the metabolic energy of an equal weight of hydrated glycogen.

Fats are a highly efficient form in which to store energy because they can be regarded essentially as long hydrocarbon chains made up of $-CH_2-$ units. Carbohydrates, in contrast, contain a much higher proportion of oxygen. They can be regarded as molecules made up of $-CH(OH)-$ units. Carbohydrates are already partly oxidized and therefore give less energy when converted to carbon dioxide (CO_2) and water (H_2O).

■ **Table 23.11** A comparison of glycogen and triglycerides as energy stores

Factor	Glycogen	Triglycerides (fats)
Efficiency	Efficient, relatively short-term, concentrated energy store	Highly efficient, long-term, concentrated energy store
Storage	Compact, insoluble polymer molecules, stored in granules	Triglyceride molecules are completely insoluble, aggregate together in droplets
Need for hydrolysis	Glycogen chains readily hydrolysed to glucose prior to oxidation	Triglycerides readily hydrolysed to free, long-chain acids prior to oxidation
Energy per mole	Glucose molecules from glycogen are already partially oxidized, containing oxygen atoms – they can be regarded as having the formula $(CHOH)_n$ and so: • provide more instant access to energy than fats • produce less energy per mole than fats	Long-chain acid molecules represent carbon and hydrogen in a highly reduced form – they can be regarded as having the formula $(CH_2)_n$ and so: • produce more energy per mole than glucose • also release more water when metabolized (known as metabolic water; this is important in dry climates)
Need for oxygen	Glucose molecules are already partially oxidized, so can be metabolized under anaerobic conditions to produce some energy	Triglycerides cannot be metabolized under anaerobic conditions – energy cannot be released in the absence of oxygen



■ **Figure 23.81** Coloured transmission electron micrograph (TEM) of lipid droplets (yellow) in a developing fat cell (adipocyte). The cell's nucleus is purple. Adipocytes form adipose tissue, which stores energy as an insulating layer of fat

This difference can be illustrated by comparing the energy diagrams for the oxidation of the two basic units (Figure 23.82). Using bond enthalpies it is possible to estimate the difference between the energy given out in the two cases. Do note that this is only an illustration as the values we use are for changes in the gas phase.

Triglycerides store energy in a highly efficient form. However, unlike glycogen, they cannot produce energy when muscles are starved of oxygen (anaerobic conditions), for example during extreme exercise such as sprinting. This means that both glycogen and triglycerides are required for humans to function properly.

The respiratory quotient

The respiratory quotient (RQ) is the ratio of the amount of carbon dioxide produced to the amount of oxygen taken in by an organism in a given time:

$$RQ = \frac{CO_2 \text{ produced}}{O_2 \text{ taken in}}$$

The RQ value is useful because it indicates which substrate is being oxidized during respiration. For example, when glucose is respired aerobically the reaction is:

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$$

Hence, the RQ is:

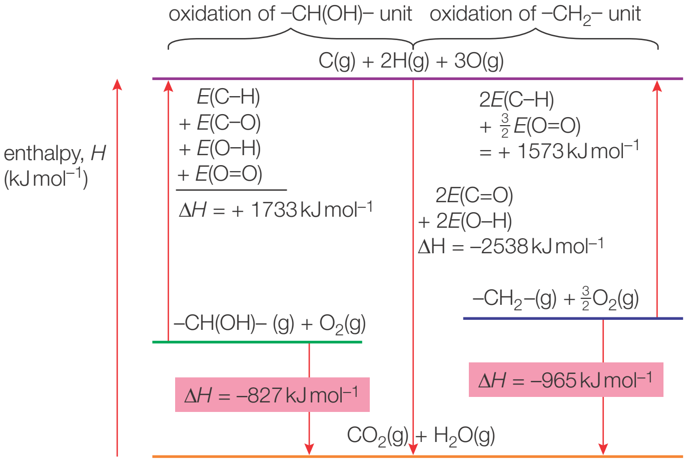
$$\frac{6CO_2}{6O_2} = 1.0$$

However, when fatty acids are respired aerobically (during starvation), the reaction is:

$$C_{18}H_{36}O_2 + 26O_2 \rightarrow 18CO_2 + 18H_2O$$

and the RQ value is:

$$\frac{18CO_2}{26O_2} = 0.7$$



■ **Figure 23.82** Comparison of the energy released from glucose and triglycerides

The RQ due to fatty acid respiration is significantly lower than that due to the respiration of carbohydrates. This is because fatty acids have a greater proportion of hydrogen atoms relative to oxygen atoms. Therefore the metabolism of fat consumes a great deal more oxygen for each carbon dioxide molecule produced than the metabolism of carbohydrates.

23 **a** Explain the meaning of the following terms: lipid; triglyceride; saturated fatty acid; and polyunsaturated fatty acid.
b If too much glucose is absorbed by organisms, some of the excess is converted into fatty acids. One such fatty acid is palmitic acid, which has the formula $C_{15}H_{31}COOH$.
i Write down the formula of the triglyceride derived from palmitic acid and propane-1,2,3-triol.
ii What name is given to the type of link that joins the fatty acid to propane-1,2,3-triol?
iii Is this compound likely to be solid or liquid at room temperature?
c **i** Write a structural formula for the mixed triglyceride molecule formed from propane-1,2,3-triol, hexadecanoic acid, octadecanoic acid and octadeca-9,12-dienoic acid.
ii Write the equation for the saponification of this triglyceride.

The role of lipids in the body – an overview

A major function of lipids is to act as a long-term energy store – they are a more reduced energy source than carbohydrates. Consequently they have a higher calorific value than carbohydrates and are aerobically respired when glycogen levels in the muscles and liver run low. Fat is also respired after exercise so that glycogen levels can be restored.

Animals, such as bears, store extra fat when hibernating during the winter, and fat is also found below the dermis of the skin of mammals where it serves as an insulator to prevent heat loss. It is most extensive in aquatic mammals living in cold climates, such as whales and seals, where it takes the form of blubber. Fat also protects a number of organs, including the kidneys and intestines.

■ **Figure 23.83** Oilseed rape: the oil contains both omega-6 and omega-3 fatty acids in a ratio of 2 : 1

Plants usually store oils rather than fats. Seeds and fruits are often rich in oils, for example coconuts, soya beans, peanuts, and flax and sunflower seeds. When fat is respired (oxidized), water is a product. This is known as metabolic water and is essential to animals that live in hot deserts, such as the kangaroo rat. Camels store fat in their humps primarily as a water source rather than as an energy source.

Phospholipids are lipids with a covalently bonded phosphate group. They are the major constituents of cell membranes in plant, animal and bacterial cells. Lipoproteins are associations of lipids with proteins. They are present in cell membranes and play an important role in the transport of cholesterol in the blood.

Steroids are classified as lipids, but they are not formed from fatty acids. They have similar physical properties to triglycerides. Steroids are common in both animals and plants and have a wide range of functions. Steroids act as precursors for the synthesis of the sex hormones (progesterone, oestrogen and testosterone) and aldosterone. Steroids are also involved in the synthesis of bile, which emulsifies lipids during digestion. Vitamin D is a steroid derivative (see Section 23.5).

Omega-3 poly-unsaturated fatty acids are found in oil from certain types of fish, vegetables and other plant sources (Figure 23.83). These fatty acids are not made by the body and must be consumed in the diet. Omega-3 poly-unsaturated fatty acids work by lowering the body's production of triglycerides. High levels of triglycerides can lead to coronary heart diseases and strokes. There is preliminary evidence that supplementation might be helpful in cases of depression and anxiety.

Mono-unsaturated fat is the primary fat source found in olive oil. Research shows that mono-unsaturated fat may have an LDL cholesterol-lowering effect when substituted for equal amounts of saturated fat, and can help reduce the risk of heart disease. Mono-unsaturated fat may also help control blood sugar levels. Poly-unsaturated fat is found in plant oils, nuts and fish. Poly-unsaturated fat also helps to maintain heart health and lower blood cholesterol levels.

All animal fats (meat, poultry and dairy) contain saturated fat. These fats can raise blood cholesterol levels and increase the risk of heart disease. The major triglycerides present in saturated fat are lauric, myristic and palmitic acids. Lauric acid (C_{12}) is the main fatty acid in coconut milk and palm kernel oil. Myrisitic acid (C_{14}) is present in palm oil, coconut oil and butter fat. Palmitic acid (C_{16}) is one of the most common saturated fatty acids found in animals and plants.

Trans fats are also naturally present in meat and dairy products, though in small amounts. Most *trans* fats are created through hydrogenation. *Trans* fats remain solid at room temperature, like saturated fats. *Trans* fats may raise LDL cholesterol levels while decreasing HDL cholesterol levels.

ToK Link

Food labelling and dietary information

Food labels are important, both practically and ethically. Reading the label is a key way to make sure the food item you are buying meets your needs. Labels can help inform consumers about what they are buying, reducing what economists call information asymmetries between buyer and seller. Where substantial information asymmetries exist, voluntary exchanges can fail to live up to the promise of mutual benefit, and society as a whole suffers from the resulting reduction in market efficiency. The recent experience regarding the labelling of meat products in the UK has shown that deviation from compliance with labelling standards can involve serious criminality.

It can be argued that any substance in a food that is scientifically proven to be a hazard should be a mandatory label. For example, a label that a product contains nuts is justified by severe allergic reactions, even though the additional label may add to the cost of a product for people who do not have allergies.

Any label that does not have a proven hazard is simply a label of preference so should not be mandatory. Instead, voluntary labels are appropriate. For example, producers may choose to label products as free from animal products if they think the cost of sourcing non-animal ingredients, testing and labelling will be rewarded by additional purchases of their products by vegetarians and vegans. Non-vegetarians should not have to pay for a label that is based on preference, not science.

Practical concerns are not the only reason to label or not label foods, however. Ethics definitely comes into play. Do people have a right to labels, such as labels that indicate a product contains ingredients derived from genetically modified organisms (GMO)?

Some consumers think that GMOs are a 'like to know' issue and that a 'Contains GMOs' type label would simply be confusing to consumers, possibly interpreted as a warning. Other consumers might argue that GMOs are a 'right to know' issue and hence an ethical issue. Do consumers who want to know if products contain products of genetic engineering have rights that are upheld to be more important than the rights of consumers who do not care? What about the rights of farmers, distributors, grocers and supermarkets?

The use and labelling of food additives is regulated by national and international law. In Europe a system of 'E' numbers is used whereby each additive has its own unique number. The same numbers without the E prefix are used in many non-European countries. In the USA food additives are regulated by the US Food and Drugs Administration (FDA). Worldwide though it is not uncommon for food additives approved in one country to be listed as harmful and banned in another. The International Organization for Standardization has drawn up a set of universal standards but as yet these have not been adopted by many countries.

Phospholipids

All cells are surrounded by a membrane which controls the exchange of chemicals, such as food and waste products, between the cell and its environment. Membranes are also present within cells, where they surround the various internal compartments. The major lipid component of these cell membranes is **phospholipids** (phosphoglycerides).

Phospholipids have a structure very similar to that of a trigylceride, except that one of the three fatty acid groups is replaced by a phosphate group (Figure 23.84) – hence, phospholipids are diglycerides. The phosphate group is ionized and negatively charged. Water molecules will therefore be attracted to this polar part of the molecule, making this end of the molecule soluble in water (this part of the molecule is **hydrophilic**). However, the other two fatty acid chains are long hydrocarbon chains and hence non-polar. These parts of the molecule will be **hydrophobic** (they will not attract water molecules). Note the meaning of the important terms hydrophilic and hydrophobic: literally hydrophilic means 'water liking' and hydrophobic means 'water hating'. Since these molecules contain both hydrophilic and hydrophobic regions they can be referred to as **amphiphilic**.

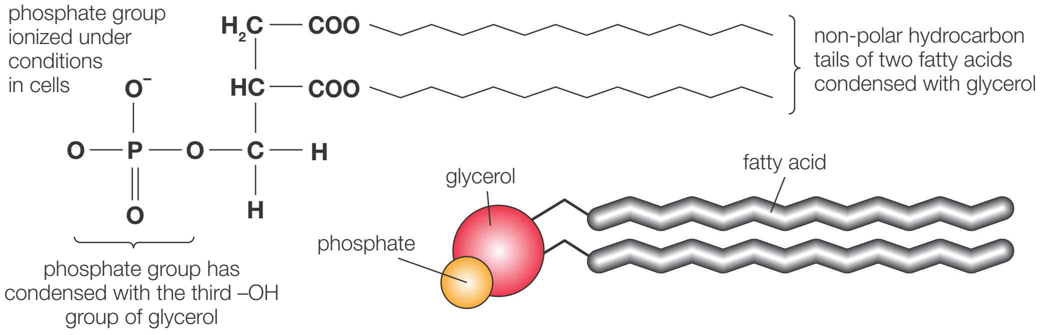
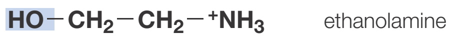
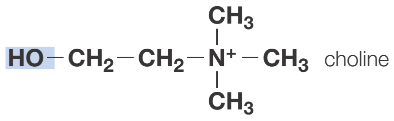
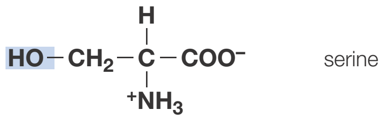

phosphate group ionized under conditions in cells
non-polar hydrocarbon tails of two fatty acids condensed with glycerol
glycerol
fatty acid
phosphate
phosphate group has condensed with the third –OH group of glycerol

Figure 23.84 The generalized structure of a simple phospholipid

In some phospholipids the phosphate group can also be attached, also via an ester link, to the −OH group of a small polar molecule such as choline, ethanolamine or serine.







Different phospholipids vary in their fatty acid chains and in the group attached to the phosphate. One of the most common phospholipids is lecithin, or phosphatidylcholine; this belongs to a group of phospholipids which incorporate choline as a head group (Figure 23.85). They are a major component of biological membranes and can be isolated from egg yolk or soya beans, from which they are mechanically extracted or chemically extracted using hexane. Lecithin is used as a natural emulsifier.

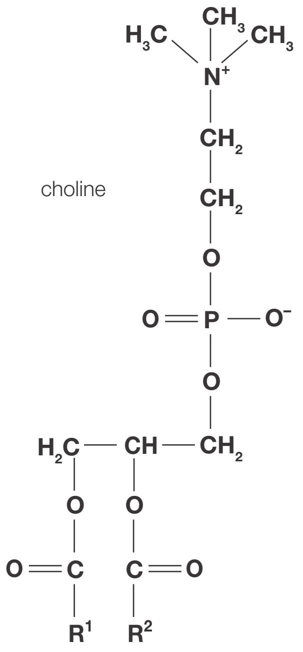


Figure 23.85 The generalized structure of lecithin (phosphatidylcholine)

Cell membranes

In cells, phospholipids often occur with proteins and cholesterol in a bilayer in the form of a cell membrane (Figure 23.86). Lipid bilayers occur when hydrophobic 'tails' line up against one another, forming a membrane with hydrophilic 'heads' on the outside facing the water. This phospholipid bilayer provides the basis of membrane structure.

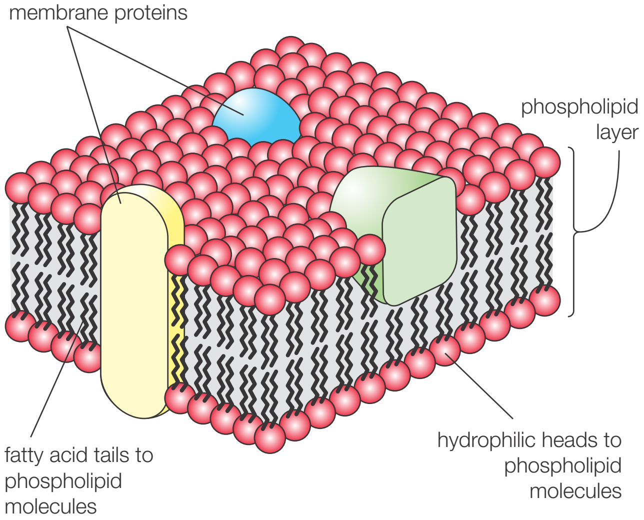
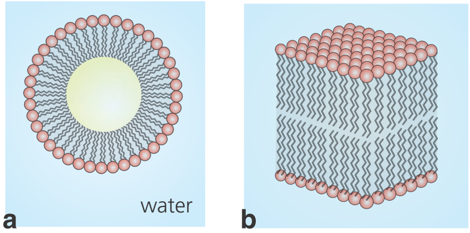

membrane proteins
phospholipid layer
fatty acid tails to phospholipid molecules
hydrophilic heads to phospholipid molecules

Figure 23.86 The structure of a cell membrane



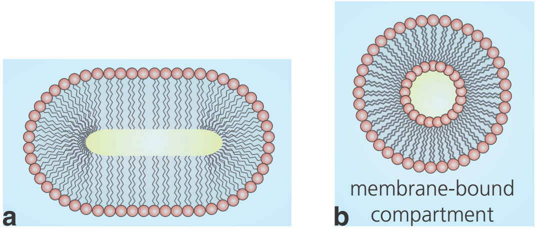
■ **Figure 23.87 a** The structure of a micelle formed from triglycerides. **b** A bimolecular layer formed by phospholipids

Self-assembly – monolayers, micelles and membranes

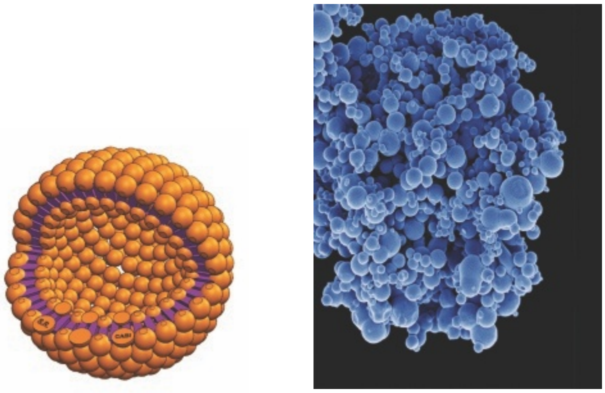
Triglyceride molecules are insoluble in water. If enough triglyceride molecules are present, they can group together to form spherical structures known as micelles. In these structures, the polar heads of the triglyceride molecules face out into the water, and the non-polar tails group together in the centre of the sphere, away from the water (Figure 23.87a). This process is a simple example of self-assembly.

Triglycerides do not form structures more complicated than the simple micelles shown in Figure 23.87a. This is because of the bulk of the three hydrocarbon chains in a triglyceride molecule. However, the structure of phospholipids, with two hydrocarbon chains forming the non-polar 'tail', is more rectangular. Structures formed from phospholipids can be enlarged to contain sections where the molecules are arranged in a bimolecular layer (Figure 23.87b). These bimolecular layers can take different forms:

- The packing together of phospholipids in aqueous suspension gives rise to disc-shaped micelles that are really extended bimolecular layers (Figure 23.88a).
- If a concentrated suspension of phosphoglycerides in water is subjected to ultrasonic vibrations then structures are formed that contain water inside, bounded by only a single bimolecular layer (Figure 23.88b). This type of structure, often referred to as a liposome, has been used as a model of cell membranes (Figure 23.89).



■ **Figure 23.88 a** Structure of a disc-shaped micelle formed from phospholipids in water can be considered an extended bimolecular layer. **b** Phospholipids can form structures that contain water within the bimolecular layer – a liposome



■ **Figure 23.89 a** Computer artwork of a liposome. **b** Coloured scanning electron micrograph (SEM) of liposome vesicles

Our current model for the structure of membranes is the **fluid mosaic model**. The phospholipid bimolecular layer forms the core of the structure. The phospholipids are able to move from side to side, giving the cell flexibility, while cholesterol molecules are present to give the membrane greater rigidity. The phospholipids also give the cells high electrical resistance and an impermeability to highly polar molecules. A 'mosaic' of proteins is embedded in the bimolecular layer. Proteins can be located in one or other of the membrane surfaces or can span the whole membrane from one surface to the other (see Figure 23.86 and Section 23.2).

Nature of Science

Liposomes

The development of liposomes as a model of cell membranes has illustrated how a technique involved in 'pure' research can impact on practical technology and applied science. These artificially constructed spherical vesicles possess a selectively permeable wall that closely resembles the membrane of a living cell. The membrane consists of a dual layer of phospholipids. Liposomes can be used to carry drugs or genes to target cells. This is particularly useful for highly toxic cancer drugs as it reduces unpleasant side effects.

Liposomes are used in biotechnology research to investigate the functioning of the cell membrane and, since they can be incorporated into living cells, are used to deliver drugs of high toxicity to specific cells in the body, such as cancer cells. Liposomes are also used in the cosmetics industry.

24 **a** What distinctive features of phospholipid molecules enable them to fulfil their role as a major component of cell membranes?

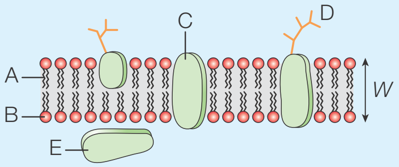
b The diagram represents the structure of a cell membrane.

i What term is applied to this model of cell plasma membrane structure?

ii State the names of components A–E.

iii Research an approximate figure for the width (*W*) of the plasma membrane.

iv What role does cholesterol play in plasma membranes?



Steroids

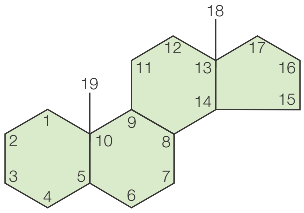


Figure 23.90 The steroid 'nucleus' of four fused rings

Steroids are classified as lipids, although they do not contain fatty acids. They are classified as lipids because they have similar physical properties to triglycerides and are synthesized using common intermediates.

Steroids all contain a 17-carbon atom skeleton (or 'nucleus') consisting of four fused rings (Figure 23.90). Methyl groups ($-CH_3$) are usually attached to carbon atoms 18 and 19 and a side-chain usually occupies position 17. Steroids vary by the functional groups attached to these rings and the oxidation state of the rings.

Steroids are found in animals and plants and have many important biochemical roles. Steroids form bile acids, the constituents of bile that emulsify and solubilize lipids during the physical digestions of fats and oils. The sex hormones, for example progesterone, oestrogen and testosterone, are all steroid-based hormones. Aldosterone, secreted by the adrenal glands, is a member of another family of steroid-based hormones concerned with controlling the concentration of sodium and potassium ions. Vitamin D (calciferol) is a steroid-based vitamin (see Section 23.5).

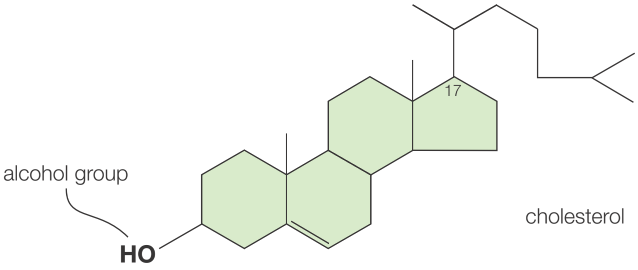


Figure 23.91 The skeletal structure of cholesterol. This is the most abundant steroid in humans

Anabolic steroids

Anabolic steroids are synthetically produced variants of the naturally occurring male hormone testosterone. They include compounds such as dianabol and nandrolone. Anabolic steroids are primarily used by body builders and athletes who claim steroids give them a competitive advantage or improve their physical performance. Steroids increase lean body mass, strength and aggressiveness. Steroids also reduce recovery time between workouts, which makes it possible to train harder and thereby further improve strength and endurance.

Apart from giving unfair advantage to athletes and sports participants in international competition, anabolic steroids present significant health risks ranging from acne to high blood pressure and liver damage. Additionally many of these anabolic steroids suppress the normal production of sex hormones in the body and increase the level of LDL cholesterol.

Such steroids are banned by most sports governing bodies and regular testing of participants is carried out, both in competition and during training periods. The methods used to detect these steroids and their metabolites in urine and blood samples include gas-liquid and high-performance liquid chromatography and mass spectrometry.

Cholesterol

Cholesterol is found in all tissues since it is a component of cell membranes (see Figure 23.86). High concentrations of cholesterol are found in the blood, brain and spinal cord. Some cholesterol enters the body via the diet.

Since cholesterol, like other lipids, is almost insoluble in blood, it is transported in the plasma of the blood within **lipoproteins** known as apoproteins. The outer surface of the apoprotein is water soluble (polar) and the inward-facing surface is fat soluble (non-polar). Some of the cholesterol within the lipoprotein is in the form of cholesteryl ester: an ester bond is formed between the carboxylic acid group of a fatty and the hydroxyl group of cholesterol.

There are two main types of lipoproteins in the blood: high-density lipoprotein (HDL) and low-density lipoprotein (LDL). The cholesterol within the two types of lipoproteins is identical.

High-density lipoproteins (HDL) are composed mainly of proteins, with only small amounts of cholesterol. HDLs are often referred to as 'good cholesterol' because they help remove cholesterol from artery walls and transport it to the liver for removal from the body. In healthy individuals, about 30 per cent of blood cholesterol is carried by HDL.

Low-density lipoproteins (LDL) (Figure 23.92) are composed mainly of cholesterol and have very little protein. They are often referred to as 'bad cholesterol' because they are primarily responsible for depositing cholesterol within arteries.

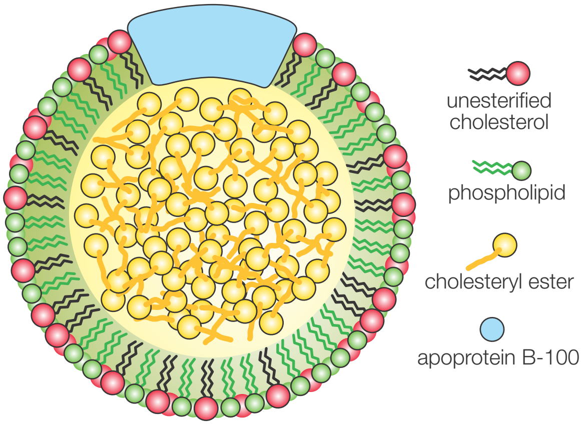


Figure 23.92 Structure of LDL

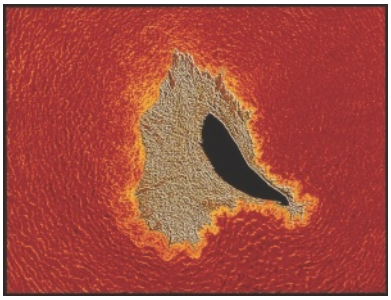


Figure 23.93 Coloured light micrograph of a cross section through an artery obstructed with atheroma plaque. This arterial disease is known as atherosclerosis. The muscular wall of the artery (orange) takes up much of the image. At the centre, fatty deposits of plaque (grey) are seen on the inner arterial wall; the lumen (black) has been severely reduced for the flow of blood

Lipoproteins and health

Excess lipids in the diet are increasingly linked to negative effects on health. These arise largely due to their low solubility that causes some lipids to be deposited in the walls of the main blood vessels (Figure 23.93). This can restrict blood flow, a condition known as atherosclerosis. It is usually associated with high blood pressure and can lead to heart disease. In addition, because of the body's ability to convert excess fats into adipose tissue, a diet too rich in lipids can lead to the excess accumulation of body fat known as obesity.

This is linked to many other health issues including type 2 diabetes and a variety of cancers. The molecule that is often the main culprit in circulatory diseases is cholesterol. It is present in the human diet, particularly from animal fat, and is also synthesized in the body. Because cholesterol is insoluble in blood, it is transported bound in different lipoproteins as we have seen earlier. LDL cholesterol is associated with increased deposition in the walls of the arteries, while high levels of HDL cholesterol seem to protect against heart attack.

HDL is thought to carry cholesterol away from the arteries and back to the liver, where it can be metabolized, so slowing its build-up. The main sources of LDL cholesterol are saturated fats and *trans* fats, the chemical nature of which we discussed earlier.

Clearly the type of fat consumed is as important as the total amount. In general an intake of poly-unsaturated fats, such as those found in fish, many nuts and corn oil, is considered beneficial in lowering levels of LDL cholesterol. Also, omega-3 poly-unsaturated fatty acid, found for example in fish oils and flax seeds, has been shown to be linked with reduced risk of cardiovascular disease as well as with optimum neurological development. These fatty acids must be taken in as part of the diet.

Nature of Science

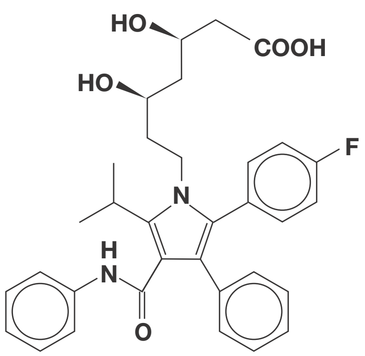
Scientific interventions

The public explanation of scientific developments in terms of understanding the impact of the negative effects of diets high in saturated and *trans* fats and cholesterol has led to discussion about possible interventions and new food products.

Statins

Atorvastatin (Figure 23.94) is a member of the drug class known as statins, used for lowering levels of blood cholesterol and preventing strokes. Marketed by Pfizer under the trade name Lipitor, atorvastatin became the world's largest-selling drug of all time. Pfizer's patent on atorvastatin expired in November 2012, so various generic forms have been available under a range of brand names since May 2013. Atorvastatin works by inhibiting a liver enzyme involved in the synthesis of a key intermediate in the production of cholesterol.

Recent discussion in the UK has centred on the feasibility and ethics of prescribing statins to individuals over a certain age en masse in order to lower the incidence of strokes and heart disease.



■ **Figure 23.94** Structure of atorvastatin (Lipitor)

Fat replacers and mimetics

A large number of substances have been suggested as fat replacers. They include modified starches, fibre, gums, emulsifiers and restructured proteins. Fat substitutes are lipid-like substances which replace fats and oils on a one-to-one basis. Fat mimetics are proteins or carbohydrates which imitate the texture, taste and mouth feel of real fats and oils.

Olestra, also known as Olean, is an emulsifier produced by reacting sucrose with six to eight moles of C_{12}–C_{22} fatty acids in the presence of a catalyst. Olestra was approved in 1996 for use in savoury foods. Less completely esterified sucrose esters (two to three moles of fatty acids) have been produced which are hydrophobic and also more digestible.

The fat mimetics are non-lipid compounds that are able to simulate the physical properties of lipids, such as creaminess and smoothness. Carbohydrate fat mimetics, such as Avicel and Methocel, include micro-particulate cellulose. These materials provide the mouth feel and flow properties of fat but lack the flavour characteristics of edible fats.

ToK Link

Health issues

Dealing with poor health arising from personal lifestyle choices is a major public health challenge. Doctors see the consequences of poor diet, smoking and alcohol abuse every day in clinics, emergency departments and hospital wards. Doctors and nurses have a clear role to play in treating illness and in providing support and education to patients. By their nature, therapeutic treatments, such as surgery, tend to treat the consequences rather than the causes of these health problems.

How far the government or state can legitimately interfere in the private choices of its citizens is controversial. Most people accept that the state has a role to play in health promotion. Dealing successfully with health threats, such as infectious diseases, and the close link between health and happy lives provides clear justification. But liberal societies also value freedom of choice. Although it is widely accepted that the government is justified in restricting the freedom of individuals to prevent harm to others, the extent to which the state should intervene in the decisions of individuals for their own benefit is far more controversial.

23.4 Carbohydrates – *carbohydrates are oxygen-rich biomolecules which play a central role in metabolic reactions of energy transfer*

Carbohydrates, as the name implies, are composed of the three elements carbon, hydrogen and oxygen, with the hydrogen and oxygen always in the same ratio as they are in water (i.e. 2 : 1). Note that this higher ratio of oxygen represents a more oxidized state for the carbon atoms than occurs in lipids and this is significant when comparing carbohydrates and lipids as fuels (see Section 23.3)

■ 'Non-stop fuel' – carbohydrate loading and endurance

Modern sports training has developed the link between correct diet and physical performance. Such interest has shown that carbohydrate-rich foods provide the most appropriate fuel for prolonged heavy exercise. Studies using samples from the muscles of athletes have shown a clear link between fatigue and reduced levels of the carbohydrate glycogen (glycogen is the main short-term energy store in the muscles).

If an athlete in training eats a high-carbohydrate diet after exercising to exhaustion, then their muscle glycogen levels recover to higher values than those existing before the exercise. This 'carbo-loading' increases the length of time that an athlete can exercise by up to 50 per cent (Figure 23.95). Drinking a well-formulated sports drink containing 5–8 per cent carbohydrate also aids performance during sport. Such sports performance studies emphasize the role of carbohydrates as energy sources.

In this section we will consider the different roles of carbohydrates – as a source of energy, as a store of energy and as a component of physical structures in organisms. This last role is also linked to the usefulness of some carbohydrate polymers as a source of dietary fibre for humans. Carbohydrates are classified according to their structure:

- monosaccharides – simple sugars with the empirical formula $(CH_2O)_n$
- disaccharides – dimers of monosaccharides
- polysaccharides – polymers of monosaccharides or disaccharides.

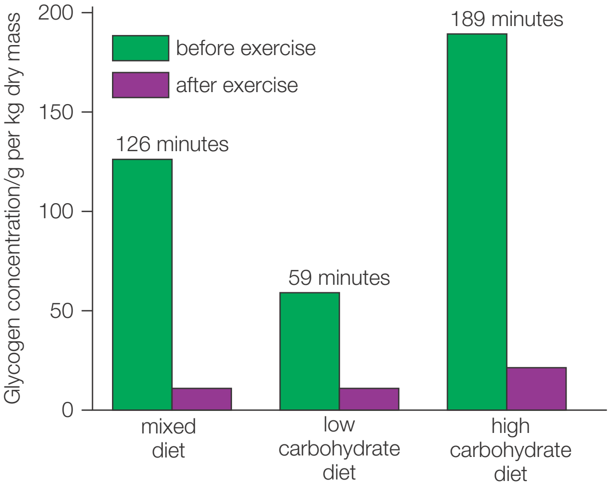


Figure 23.95 The results of carbo-loading. The graph shows the concentration of glycogen in muscle before and immediately after strenuous exercise for athletes following three different diets. The time recorded on the graph is the time taken to reach exhaustion. A high-carbohydrate diet leads to higher glycogen levels both before and after exercise and enables athletes to exercise for longer

■ Structures of monosaccharides

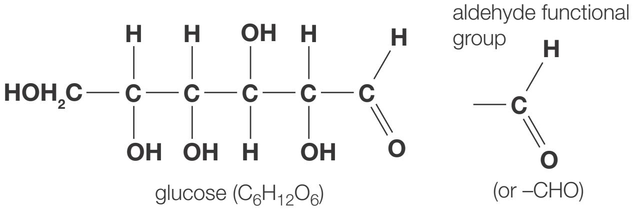
Monosaccharides are simple sugar molecules that provide instant-access energy to plants and animals, and act as structural units of polysaccharides. They are simple sugars that contain a carbonyl group (>C=O) and two or more hydroxyl groups (–OH). Glucose, $C_6H_{12}O_6$, is the most common monosaccharide and is the monomer for the **polysaccharides** starch, glycogen and cellulose. Many monosaccharides, such as glucose, have the general formula $(CH_2O)_n$. The value of n ranges from 3 to 9; monosaccharides can be categorized by the value of n (Table 23.12).

Table 23.12 The classification of simple sugars according to chain length

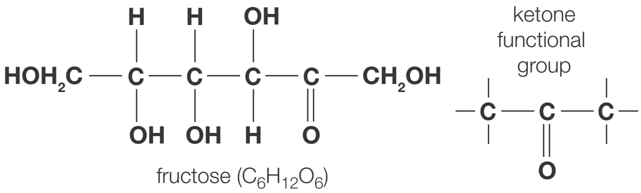
Value of n	Example	Formula	Type of sugar
3	Glyceraldehyde (2,3-dihydroxpropanal)	$C_3H_6O_3$	Triose
4	Erythrose	$C_4H_8O_4$	Tetrose
5	Ribose	$C_5H_{10}O_5$	Pentose
6	Glucose	$C_6H_{12}O_6$	Hexose

Monosaccharides are aldehydes or ketones, depending on the position of the carbonyl group (>C=O) in the carbon chain. If the carbonyl group is at the end of the molecule then the monosaccharide is an **aldose**, for example glucose (Figure 23.96); if the carbonyl group is in any other position, the monosaccharide is a **ketose**, for example fructose (Figure 23.97). All simple monosaccharides are white, crystalline solids soluble in water due to the ability of the polar −OH groups to hydrogen bond with water.

■ **Figure 23.96** The structure of D-glucose, an aldose



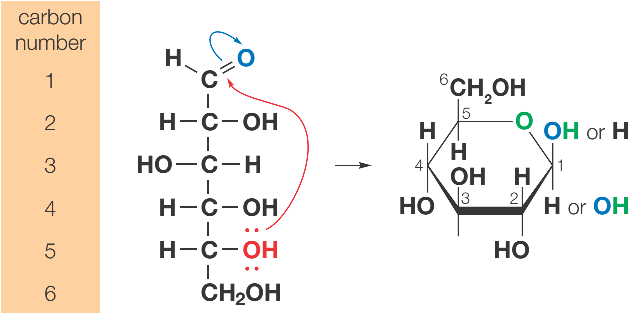
■ **Figure 23.97** The structure of D-fructose, a ketose



Glucose and fructose are both simple monosaccharides that have the molecular formula $C_6H_{12}O_6$. Sometimes the number of carbon atoms and the functional group are combined in a single designation of the structure. Thus glucose is an aldohexose whereas fructose is a ketohexose.

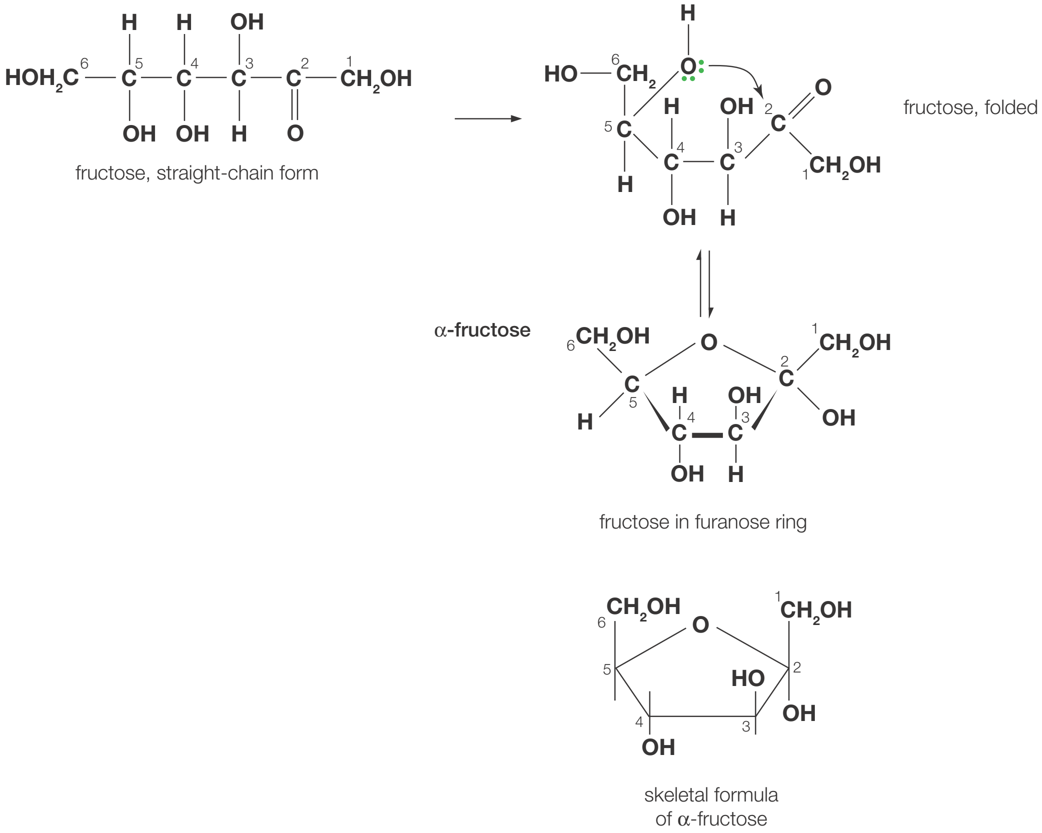
Due to the presence of a carbonyl group and several hydroxyl groups in the same molecule, straight-chain forms of monosaccharides are unstable and undergo intramolecular nucleophilic addition reactions to form cyclic structures. This reaction occurs internally within a monosaccharide molecule due to the shape and flexibility of the open-chain structure. For example, in glucose a lone pair of electrons on the C-5 hydroxyl group can attack the C-1 aldehyde group. This nucleophilic addition reaction results in the formation of a cyclic structure consisting of five carbon atoms and an oxygen atom (Figure 23.98). The six-membered ring formed from glucose is sometimes referred to as a pyranose ring.

■ **Figure 23.98** The formation of the pyranose ring structure of glucose. The cyclic structure shows the two possible isomers of glucose, differing in the positions of the H and OH on carbon 1



Fructose, a ketohexose, also forms a cyclic structure, though this time it is a five-membered (furanose) ring (Figure 23.99)

■ **Figure 23.99** The structures of fructose in aqueous solution


fructose, straight-chain form

fructose, folded

α-fructose

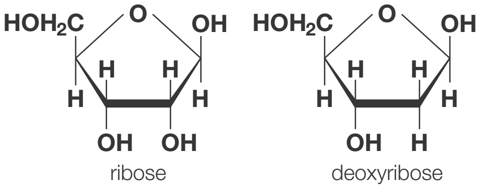
fructose in furanose ring

skeletal formula of α-fructose

Each cyclic form of a monosaccharide can exist as two stereoisomers known as α- and β-forms. The stereoisomerism of monosaccharides is covered in Section 23.10. The straight-chain and α-ring forms of glucose and fructose are given in Section 34 of the *IB Chemistry data booklet* so they do not have to be learnt.

Crystalline glucose consists only of the cyclic form of the sugar. However, when dissolved in water, it is converted into an equilibrium mixture of the α- and β-forms. A very small, but significant, amount of the open-chain structure is also present. The −OH group attached to carbon-1 is particularly reactive – it is the group involved in forming bonds to other glucose molecules during the condensation reactions that produce disaccharides and polysaccharides.

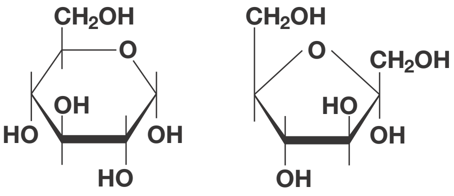
Aldopentoses such as ribose and deoxyribose predominantly exist in the five-membered (furanose) cyclic form (Figure 23.100). These sugars are highly significant as they are constituents of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), respectively. You will note from Figure 23.100 that deoxysugars, such as deoxyribose, have one less oxygen atom than a normal monosaccharide of the same carbon chain length.


ribose

deoxyribose

■ **Figure 23.100** The structures of ribose and deoxyribose – the sugar components of RNA and DNA

Glucose is the most important monosaccharide found in nature. As the product of photosynthesis, glucose has played a pivotal role in the development of life on Earth. The nature of its reactivity means that the energy from the Sun trapped by photosynthesis can be easily released by the metabolic processes of respiration (see Section 23.1). Because glucose can be polymerized, its energy can be stored for later use. The human brain requires the energy from two small spoonfuls of glucose per hour. In the diet, glucose can come from the monosaccharide itself, from some disaccharides or from starch-based foodstuffs.



skeletal formula of α-glucose

skeletal formula of α-fructose

■ **Figure 23.101** The Haworth projections of the cyclic structures of α-glucose and α-fructose

Haworth projections

The skeletal representations of the cyclic forms of sugars are often drawn as Haworth projections. The edge of the ring nearest the reader is represented by bold lines, and the letter C for the carbons in the ring are usually omitted from the structure (Figure 23.101).

Haworth projections are useful in that they emphasize the nature of the ring and the positions of the functional groups attached to it. The cyclic forms adopted by monosaccharides result in space-efficient molecules and this is important in their role as components of disaccharides and polysaccharides.

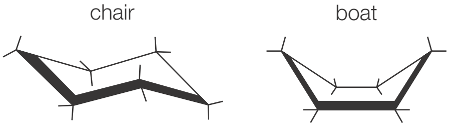
Nature of Science

Representing structure

The use of simplified structural diagrams (Haworth projections) of molecules such as monosaccharides allows biochemists to visually represent the key features of the stereochemistry and three-dimensional arrangement of the units in biopolymers such as starch or DNA. It, and other agreed 'shorthand forms', allows us to focus on the overall features and shape of the molecules concerned.

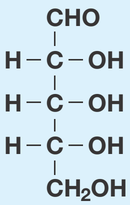
Haworth formulas can, however, be misleading as they suggest that the five-membered ring of fructose and six-membered ring of glucose are planar, which is not the case. In reality these exist as puckered rings ('chair' and 'boat' forms) in different conformations which can interconvert by rotation around single bonds (Figure 23.102). It is an interesting exercise to construct models of these structures and see how they interchange.

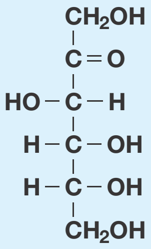
Models and visualizations are important in science to communicate knowledge, but often any single representation is inadequate. Understanding can therefore be enhanced by the use of a variety of models, including computer simulations and mathematical models.

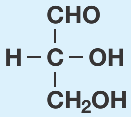


■ **Figure 23.102** 'Chair' and 'boat' forms of a single-bonded six-membered ring structure such as occurs in glucose

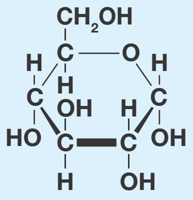
25 Which of the following is or are a ketone, an aldehyde, a triose, a pentose and a hexose?

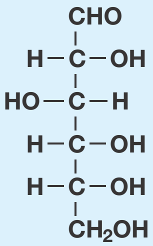
A


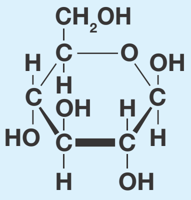
B


C


26 Glucose is an example of an aldohexose sugar which dissolves in water. In aqueous solution, glucose exists in three different structures which are in equilibrium with each other. The three structures are shown below:

A


B


C


a What chemical term is used to describe the different forms shown above?
b Explain why glucose is very soluble in water.
c Which of the two cyclic forms of glucose can be described as α-glucose?

Health concerns – diabetes

Diabetes is a group of metabolic diseases in which there are high blood glucose levels over a prolonged period. This high blood sugar produces the symptoms of frequent urination, increased thirst and increased hunger. Untreated, diabetes can cause many complications. Acute complications include diabetic ketoacidosis and induced coma. Serious long-term complications include heart disease, stroke, kidney failure and damage to the eyes.

Approximately 180 million people worldwide have diabetes, and one significant aspect of the picture is that it is thought that up to one in five people with diabetes are unaware that they have the condition. Worldwide, in 2012 and 2013, diabetes resulted in 1.5 of 5.1 million deaths per year, making it the eighth leading cause of death. Of particular concern is the increase in the incidence type 2 diabetes which is linked to the health concerns surrounding raised levels of obesity in the adult population discussed earlier.

Diabetes can be successfully managed, but it is a chronic disorder which currently does not have a cure. It has several long-term health effects. This is particularly true if an individual's blood glucose levels are poorly controlled. There are two types of diabetes affecting the adult population: type 1 and type 2 diabetes.

Type 1 diabetes is an autoimmune disease and accounts for up to 10 per cent of diabetes cases in a population (in the UK, for instance). It typically develops before the age of 40 and occurs when the pancreas can no longer produce insulin.

Type 1 diabetes develops when the person's own immune system destroys the beta cells of the islets of Langerhans. As a result insulin is no longer produced and blood sugar levels rise. This leads to the rapid onset of the symptoms of diabetes, including fatigue, unquenchable thirst, weight loss and the production of large volumes of urine. The risk of developing type 1 diabetes has recently been linked with genetic factors and may be associated with lifestyle factors such as diet and exercise.

Type 1 diabetes is treated by insulin injections alongside a healthy diet and regular exercise. People with type 1 diabetes are usually required to have either two or four injections of insulin every day. These injections of insulin are vital to keep these people alive.

Type 2 diabetes is a disorder that is increasing in both developed and developing nations as unhealthy diets and lifestyles become more common. It develops when the body can still make some insulin but not enough, or when the insulin that is produced does not work properly (known as insulin resistance). In most cases this is linked with the person being overweight. Type 2 diabetes usually occurs in people over the age of 40. In South Asian and African-Caribbean people it often appears after the age of 25. Recently more children are being diagnosed with the condition. Type 2 diabetes is more common than type 1 diabetes accounting for 85–95 per cent of people with diabetes.

Many factors influence the development of type 2 diabetes, such as an inherited predisposition to diabetes and a diet high in saturated fats and sugar and low in fibre. Being overweight also increases the chance of developing type 2 diabetes.

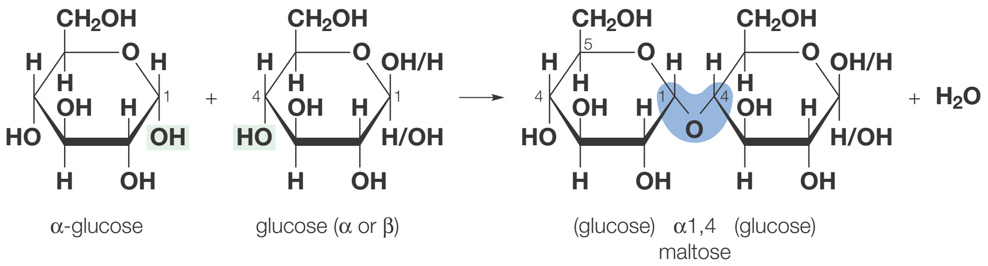
This increase in the level of type 2 diabetes, and the extension of its occurrence to younger groups in the population, is of concern to health professionals and authorities in a number of countries and emphasizes the importance of effective dietary and lifestyle advice to the public.

Disaccharides

The **disaccharides** consist of two monosaccharides joined by a **glycosidic linkage** (−O−). The bond is formed by the condensation reaction between the −OH on the carbon-1 atom with a hydroxyl group, −OH, of another monosaccharide. The reaction is a condensation reaction as it involves the elimination of water. Disaccharides are all soluble molecules that can be hydrolysed into two monosaccharides by acid hydrolysis or by enzyme-catalysed reaction. Combining different monosaccharides will produce different disaccharides.

Maltose is a disaccharide composed of two α-glucose residues and is found in germinating seeds (Figure 23.103).

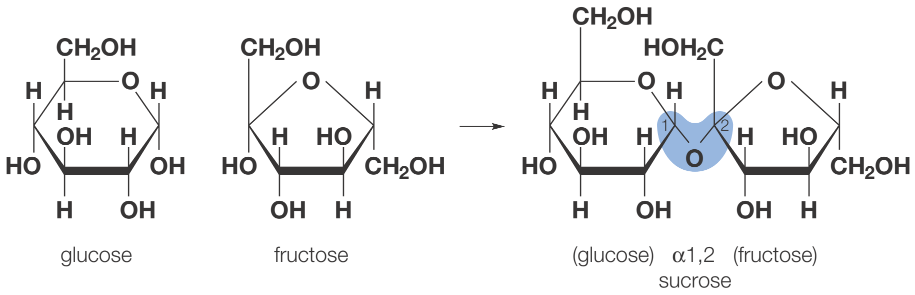
Figure 23.103 The formation of maltose



The oxygen bridge in maltose connects the carbon-1 of the first glucose residue with the carbon-4 of the second glucose unit, so it is a 1,4-glycosidic link. The stereochemistry of glycosidic links is discussed in Section 23.10.

The most common disaccharides are maltose, lactose and sucrose. Sucrose, table sugar, is composed of an α-glucose and a β-fructose residue and is found in fruit, sugar cane and sugar beet.

Figure 23.104 The formation of sucrose



In this case, the six-membered (pyranose) ring of glucose is linked to the five-membered (furanose) ring of fructose. As can be seen in Figure 23.104, the bonding in this case is by a 1,2-glycosidic link.

Sugar production

Sugar is produced in 121 countries and global production now exceeds 120 million tons a year. Approximately 70 per cent is produced from sugar cane, a very tall grass with big stems which is largely grown in the tropical countries. Sugar cane is the world's largest crop and is exported largely from tropical and subtropical regions of Brazil, India, China and Thailand.

The remaining 30 per cent is produced from sugar beet-producing countries in more temperate areas of the world. Sugar beets are grown mainly in cooler climates, such as the northern USA, countries in the European Union and Russia. The plant grows a bushy leaf structure above the ground and is grown in rows. The root of the plant provides the sugar, which has about 15 per cent sugar content. At harvest, the roots are dug up and sent to a processor for refining of the sugar. Other parts of the plant can be used to feed to livestock.

Figure 23.105 A field of sugar cane

Sugar cane (Figure 23.105) is a genus of tropical grasses which requires strong sunlight and high levels of water for growth. Binomial names of the species include *Saccharum officinarum*, *S. spontaneum*, *S. barberi* and *S. sinense*. Farmers usually grow hybrid species which can reach heights of 5 m. A typical sugar content for mature cane would be 10 per cent by weight but the figure depends on the variety and varies from season to season and location to location. Ethanol is generally available as a by-product of sugar production. It can be used as a biofuel alternative to gasoline and is widely used in cars in Brazil. It may become the primary product of sugar cane processing, rather than sugar.

Lactose is the primary carbohydrate of human and cow's milk, providing about 40 per cent of their total energy values. It is composed of a glucose residue and a β-galactose residue (Figure 23.106).

It should be noted that galactose is an aldohexose isomer of glucose, differing in the orientation of the hydroxyl (–OH) group on the carbon-4 of the structure (Figure 23.107).

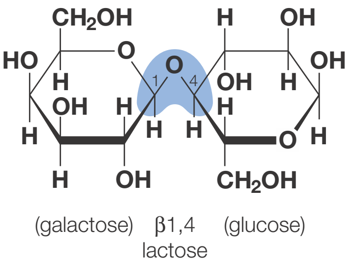


Figure 23.106 The structure of lactose is that of galactose and glucose residues bonded by a 1,4-glycosidic link

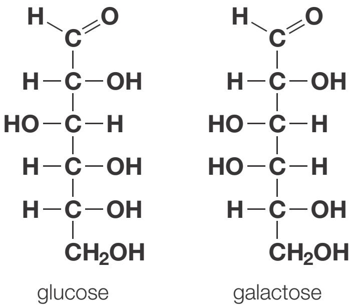


Figure 23.107 Galactose is an isomer of glucose

Lactose, sucrose and another important disaccharide, trehalose, are used to transport energy.

- Carbohydrates are transferred from a mother to her infant as the disaccharide lactose in milk.
- Plants usually transport sugars between tissues as concentrated solutions of sucrose. When extracted and refined from sugar cane or sugar beet, sucrose is sold as table sugar and is used as the sweetener in many food products and drinks.
- Trehalose is important for the transport of sugars in insects and certain fungi.

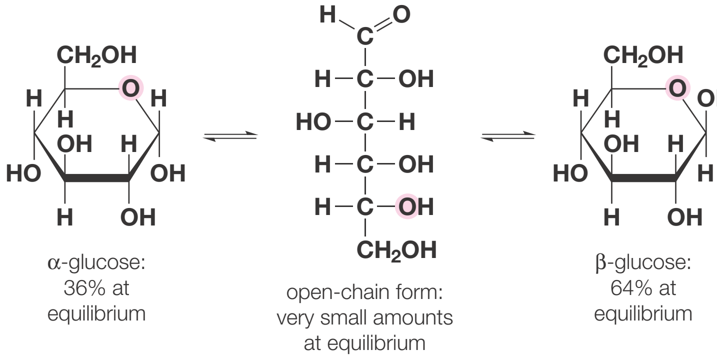
However, higher animals invariably use the monosaccharide glucose to transport energy in the bloodstream as we have seen in our earlier discussion of diabetes.

Reducing sugars

The redox properties of monosaccharides disaccharides depend on the position of the carbonyl in their molecules. Aldoses, including glucose, are **reducing sugars** in solution because they contain a terminal carbonyl (aldehyde) group and are readily oxidized under relatively mild conditions.

Reducing sugars can be detected in the laboratory by Fehling's solution – an alkaline solution containing sodium potassium tartrate, copper(II) sulfate and sodium hydroxide. When heated with an aldose solution the deep blue colour of the reagent solution disappears and a red precipitate of copper(I) oxide is produced as the Cu^{2+} ions are reduced to Cu^{+} ions. Benedict's reagent can be used as a less hazardous alternative – being less strongly alkaline – to Fehling's solution.

Figure 23.108 The equilibrium composition of an aqueous solution of glucose



The reaction takes place because, as mentioned earlier, the glucose solution contains a small proportion of the straight-chain form of the structure (Figure 23.108). As these molecules react the equilibrium is disturbed and more of the straight-chain molecules are formed (Le Chatelier's principle). The process continues until all of the glucose is oxidized.

Disaccharides such as maltose and lactose will also give a positive test with Fehling's or Benedict's reagent on heating. This indicates the presence of an aldehyde grouping in the straight-chain form of the molecules present in equilibrium in the aqueous solutions of these sugars. Sucrose, however, does not rearrange into a straight-chain form in solution and so is a non-reducing sugar, giving a negative test with these reagents. Thus sucrose can be distinguished from reducing monosaccharides and disaccharides.

Lactose intolerance

Lactose intolerance is the inability to digest the lactose found in milk and other dairy products into its constituent monosaccharides, glucose and galactose. Lactose is a disaccharide and cannot pass through the plasma membrane of the cells of the gut epithelium. It must first be hydrolysed by enzymes bound to the membrane of microvilli. Digestion of lactose is carried out by the enzyme lactase present in all young mammals. The monosaccharides produced are transported across the cell membranes of the gut epithelial cells into the bloodstream and then metabolized by tissues to yield ATP or are stored as glycogen in the liver and muscle for later use.

Milk provides carbohydrate, protein, fat, mineral and vitamins and is produced by the mammary glands of all female mammals after giving birth. When mammals stop breast feeding their young, their young lose the ability to make lactase; after weaning most mammals never drink milk again so lactase is not required – continuing to synthesize an enzyme in the gut and hydrolyse a nutrient that will not appear in the diet would be a waste of energy.

Lactose intolerance can be considered to be a physiological response to the intake of lactose in the diet by an individual who has undergone a genetically programmed loss of the enzyme lactase after weaning. The lactase gene plays a critical role in directing the synthesis of lactase at birth, as a major source of carbohydrate in the baby's nutrition is the lactose in the mother's milk. In many developed countries due to modern storage and distribution methods, adults now consume milk and milk products in greater quantities. In other parts of the world, such as Asia, milk products do not form a part of the adult diet and switching off the lactase gene is a normal phenomenon that goes unnoticed, so the lactose remains undigested and cannot be absorbed into the bloodstream. Lactose therefore passes unchanged into the last part of the large intestine, and bacteria living here switch their metabolism and begin to ferment lactose, producing lactic acids and large amounts of gases.

A geographical bias can be seen in the distribution of lactose intolerance worldwide linked to the dietary pressures of availability and preference. Lactose-free milk has become available and commercially viable as Western dietary preferences have spread into other regions. The situation has become more complex as the condition of lactose intolerance has tended to become confused with a 'milk intolerance' caused by a mutation in the protein content of cow's milk. A major study by scientists at Curtin University (Perth, Australia) has suggested that millions of people who believed that they were lactose intolerant could return to consuming commercial cow's milk using the A2 version that is now becoming more readily available.

Originally all cows produced only the A2 beta-type protein in their milk, but a genetic mutation in European herds produced the A1 protein. This spread throughout many countries and milk with this A1 protein now makes up the majority of the milk in our fridges. It is this A1 protein that causes many people the digestive discomfort that may be mistaken for lactose intolerance. The increasing commercial availability of A2 milk which does not cause this problem could mean that fewer people have an adverse reaction to milk and dairy products, but there are intriguing conflicts of commercial interest that could play out in various countries over the next few years.

Polysaccharides

Polysaccharides are condensation polymers of monosaccharide. The three most common and important polysaccharides are starch, cellulose and glycogen. All three are formed from glucose residues. **Starch** (Figure 23.109) and **glycogen** are both polymers of α-glucose; cellulose is a polymer of β-glucose.

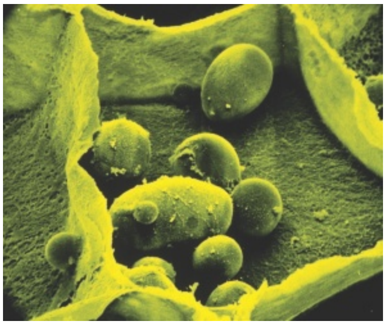


Figure 23.109 False-colour scanning electron micrograph (SEM) of a single, broken cell from a potato, showing several large starch granules

Polysaccharides are a further example, together with proteins and nucleic acids, of biologically important condensation polymers. You will remember that condensation polymers are built from monomers that each contain two functional groups capable of reacting to produce water. Each time a bond is formed between monomers, a water molecule is eliminated. In forming a polysaccharide chain, it is the −OH groups of the glucose molecules which react, and glycosidic links are formed between the monomers.

Starch and glycogen are used to store glucose in plants and animals, respectively. Glycogen is present in the liver and muscles, while starch is formed in leaves but stored in seeds and in roots. **Cellulose** is the major component of plant cell walls and, together with lignin, provides the structure to the cell walls. It is the major component of wood and cotton.

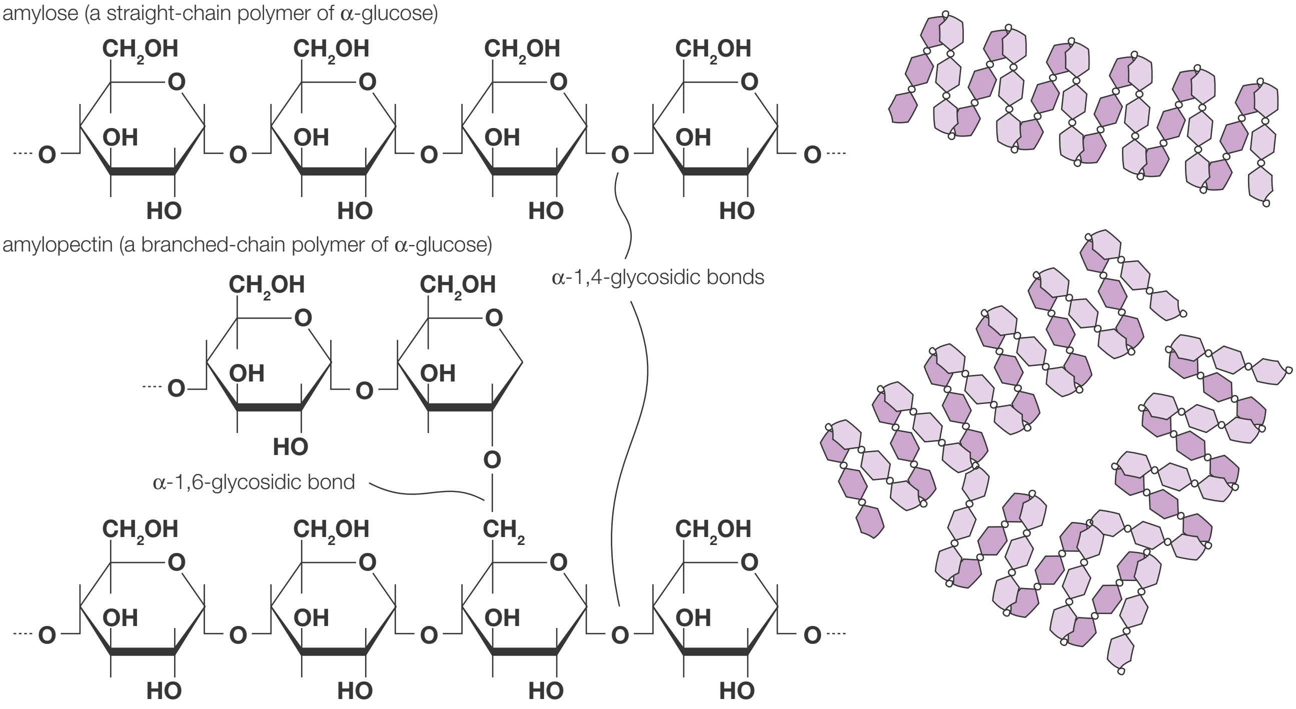
The differences in the structure and function of these three important polysaccharides, starch, cellulose and glycogen, depend largely on two features:

- the type of glycosidic link between the monomer units
- the isomer of glucose involved in their construction.

Comparing starch and cellulose

Starch occurs in two forms: **amylose** and **amylopectin** (Figure 23.110). Amylose consists of long unbranched chains of around 300 residues in which all the glucose units are bonded via α-1,4-glycosidic linkages. The orientation of the rings around the α-1-4 links produces a helical molecule with six glucose units per turn. This structure is stabilized by intra-molecular hydrogen bonding involving the hydroxyl groups on carbon-2 and carbon-3 in each ring and the O atoms of the link and ring respectively.

The other form of starch, amylopectin, is a branched molecule with the length of each branch being on average between 24 and 30 glucose residues. The backbone of amylopectin contains α-1,4 glycosidic linkages. However, the branch points are α-1,6-glycosidic linkages. Glycogen has a structure very similar to that of amylopectin. However, it is more highly branched and has a higher molar mass. The polymer molecules are very large, involving up to a million glucose units. This makes amylopectin molecules among the largest in nature.


amylose (a straight-chain polymer of α-glucose)
amylopectin (a branched-chain polymer of α-glucose)
α-1,4-glycosidic bonds
α-1,6-glycosidic bond
CH₂OH
CH₂
OH
HO
O

■ **Figure 23.110** Structures of amylose and amylopectin

Starch constitutes up to 80 per cent of the dry mass of staple foods such as wheat, corn, rice and potato, which makes it the commonest carbohydrate in the human diet. Starch is the storage polymer of glucose in plants. It is used as a carbohydrate store in roots, tubers, seeds and fruits. Plant cells store starch in the form of insoluble starch grains (see Figure 23.109), which contain variable amounts of the two polysaccharides, amylose (25–30 per cent) and amylopectin (70–75 per cent).

The structures of amylose and amylopectin are well suited to their storage function because:

- They are compact and do not take up much space, their helical structure enabling this.
- They are insoluble and so cannot move out of the cells in which they are stored.
- They enable a very large number of glucose units to be stored within cells without generating a high osmotic pressure (osmotic pressure is a property dependent on the number of solute particles present in a solution – if a large number of glucose monomers are present inside the cell, the net flow of water molecules into the cell would burst it).
- They do not become involved in the immediate metabolic processes of the cells.
- They are easily hydrolysed by enzymes to soluble sugars when required.

Worldwide, about 30 million tonnes of starch are obtained from plants each year. Perhaps surprisingly, most of this starch is used for industrial purposes. These uses include making glues, wallpaper paste, paper and card coatings, carbon paper, corrugated board, textiles, paints, packaging and insulating materials, biodegradable plastics and rubber. Various different forms of starch are used to make high-value products such as cosmetics and medicines, and even a form of concrete.

Cellulose (Figure 23.111), like amylose, is an unbranched polymer of glucose. Up to 15 000 units of glucose are present in each chain. Unlike amylose, the glycosidic linkages are β-1,4 linkages. These β-1,4 linkages are bonds that connect carbon-1 of one glucose to carbon-4 of the other glucose, where the oxygen is in the β position on carbon-1. The change in bonding results in significant differences in the properties of starch and cellulose. Cellulose is insoluble in water; starch is slightly soluble. Cellulose forms fibres; starch is a powder.

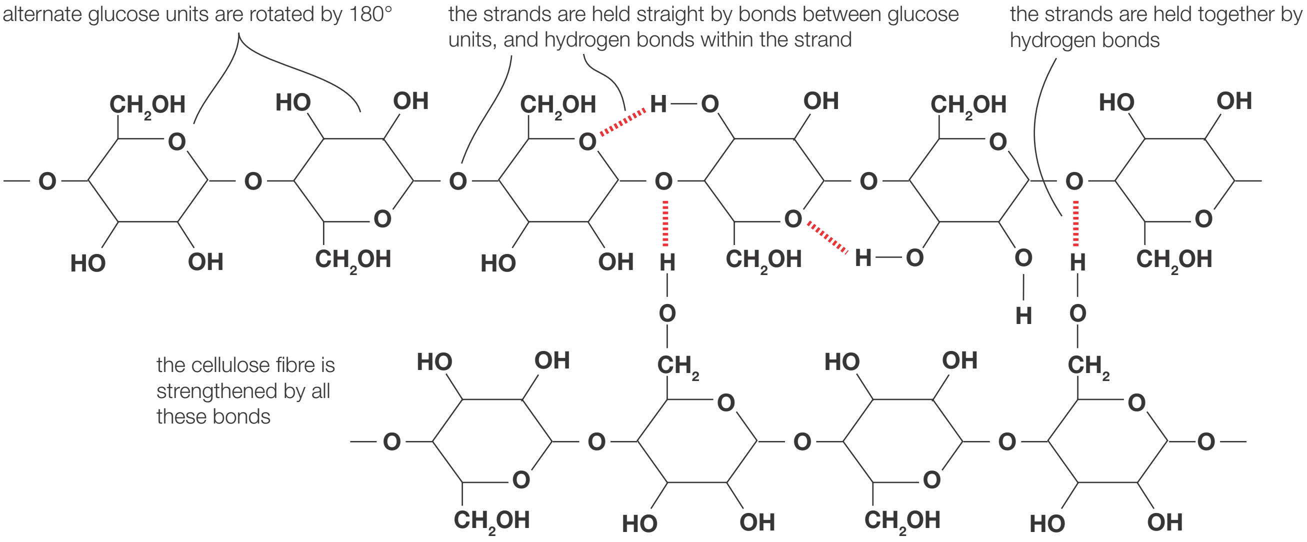
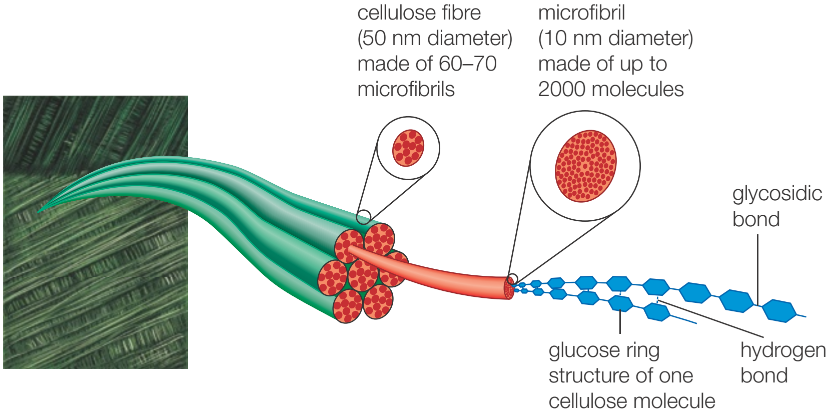
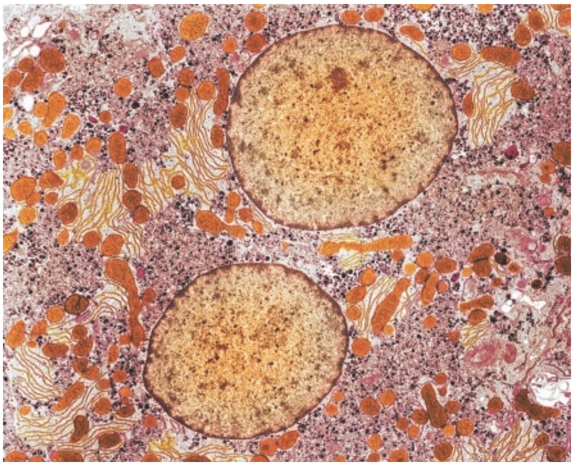

alternate glucose units are rotated by 180°
the strands are held straight by bonds between glucose units, and hydrogen bonds within the strand
the strands are held together by hydrogen bonds
the cellulose fibre is strengthened by all these bonds
CH₂OH
HO
OH
O
H
CH₂

Figure 23.111 Structure of cellulose

Due to the orientation of the glucose units the molecules do not coil into a helical shape but form a linear molecule. Parallel neighbouring chains are able to interact with each other by cross-linked hydrogen bonds. Up to 60 or 70 individual chains can be held together to form a column known as a microfibril (Figure 23.112). Bundles consisting of many microfibrils are assembled together to form the cellulose fibres which give tensile strength to the cell wall.

Figure 23.112 Cellulose chains can be held together by hydrogen bonds to form bundles. Between 60 and 70 of these bundles can pack together to produce a cellulose microfibril and then a fibre for use in structures such as plant cell walls


cellulose fibre (50 nm diameter) made of 60–70 microfibrils
microfibril (10 nm diameter) made of up to 2000 molecules
glycosidic bond
glucose ring structure of one cellulose molecule
hydrogen bond



■ **Figure 23.113** Coloured transmission electron micrograph (TEM) of a liver cell. Glycogen granules are seen as black dots. Glycogen is also stored in skeletal muscle

Animals have enzymes that degrade amylose and amylopectin into glucose. However, these enzymes do not bind cellulose and they are unable to hydrolyse cellulose. It therefore passes through the digestive system as dietary fibre (see below). Some animals, such as termites, secrete cellulase which breaks down the wood they eat. Cows contain bacteria in their stomachs which secrete cellulase and allow the cow to digest grass. Fungi are also able to break down cellulose.

Glycogen ('animal starch') is an energy storage molecule in animals, where it occurs in liver cells and muscle tissue (Figure 23.113). The structure of glycogen is very similar to amylopectin (see Figure 23.110), but glycogen is more extensively branched. The unbranched sections are formed from glucose molecules linked by α-1,4 glycosidic bonds. At the branch points, α-1,6 glycosidic bonds are found. In glycogen the branching points occur every 8 to 12 residues and the branches are shorter. Glycogen forms a very compact structure.

The structures of the three main polysaccharides, cellulose, starch and glycogen, are closely related to their function. Table 23.13 summarizes the features of the three main polysaccharides.

■ **Table 23.13** A summary of the features of the major polysaccharides

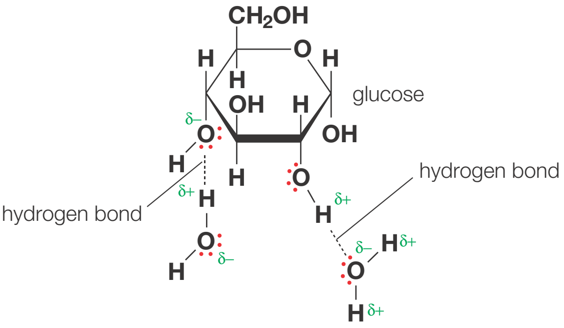
Polysaccharide		Monomer	Type of glycosidic link	Shape of macromolecule	Function
Starch	Amylose	α-Glucose	α-1,4	Unbranched chains wound into a helix	Carbohydrate storage in plants
	Amylopectin	α-Glucose	α-1,4 and also α-1,6 at branches	Tightly packed branched chains	Carbohydrate storage in plants
Glycogen		α-Glucose	α-1,4 and also α-1,6 at branches	Very branched, compact molecules	Carbohydrate storage in animals
Cellulose		β-Glucose	β-1,4	Linear	Structural component in plant cell walls

Carbohydrate solubility

All the different carbohydrates contain a significant number of −OH groups. For the monosaccharides and disaccharides, the −OH groups are an important factor in their solubility since they form hydrogen bonds with water molecules (Figure 23.114). This solubility enables small sugar molecules to be transported and metabolized readily.

In contrast with monosaccharides and disaccharides, an essential part of the function of polysaccharides is that they are not readily soluble in water:

- In cellulose, the linear molecules interact strongly with one another, forming fibrous structures that are insoluble in water. Although there are many −OH groups, they tend to be on the inside of the fibres and are involved in internal hydrogen bonding between the chains. As a result, the interactions of the polymeric chains with water are restricted.
- Neither amylose nor amylopectin is truly soluble in water. The −OH groups in the unbranched, coiled chains of amylose are more accessible to water than those of the more compact,



■ **Figure 23.114** The hydrogen bonding between glucose and water enables the sugar to be soluble branched amylopectin chains. Thus there is more scope for hydrogen bonding with water in the case of amylose. As a result, amylose is the more soluble polymer of the two.

- Glycogen is even more branched than amylopectin and is likewise only sparingly soluble in water.

Cooking a rice or pasta dish illustrates the effect of water on carbohydrate polymers. As the pasta or rice is heated in water, it swells as the polymer chains become hydrated. However, although swollen, the polymer chains do not dissolve away.

Hydrolysis of carbohydrates

Hydrolysis is the reverse of the condensation reaction that forms disaccharides and polysaccharides. Covalent bonds are broken during hydrolysis, with water (as H and OH) being added to the fragments.

$$\text{monosaccharides} \underset{\text{hydrolysis}}{\overset{\substack{\text{condensation}\\ \text{polymerization}}}{\rightleftharpoons}} \text{disaccharides or polysaccharides} + \text{water}$$

Acid hydrolysis

Hydrolysis of disaccharides or polysaccharides can be brought about by warming (70 °C) with dilute hydrochloric acid ($1\,\text{mol}\,\text{dm}^{-3}$). Acid hydrolysis of polysaccharides can be used in analysing the structure of complex carbohydrates. The hydrolysis is usually followed up by chemical testing and then paper chromatography is used to find the nature of the monomers present in the structure. Acid hydrolysis of starch can also be used as a source of glucose in the food industry. An example of this is the production of sweetening syrup. Glucose itself is not as sweet as fructose and so the glucose produced by acid hydrolysis is converted to fructose using immobilized glucose isomerase.

Enzyme hydrolysis

Biological systems require the hydrolysis of specific glycosidic bonds. Plants and animals have enzyme systems for the controlled hydrolysis of stored starch or glycogen, respectively, while starch taken in by animals as food can also be hydrolysed.

In humans, the digestion of starch begins in the mouth. Saliva contains an α-amylase, an enzyme that specifically hydrolyses α-l,4 glycosidic bonds. Digestion of starch with this enzyme produces a mixture of glucose and maltose. This enzyme cannot break bonds in the region of the branching points in amylopectin. A pancreatic α-amylase continues to digest starch in the small intestine. This enzyme is again specific for α-l,4 links.

Starch and cellulose are isomers because they are different compounds with the same molecular formula $(C_6H_{10}O_5)_n$. They are stereoisomers because only the three-dimensional arrangement of atoms is different. How the six-membered rings are joined together has an enormous effect on the shape and properties of these carbohydrate molecules.

Human amylases cannot hydrolyse β-glycosidic links. Because of this, we cannot digest cellulose. Any cellulose in our diet acts as dietary fibre. This has the beneficial effect of providing bulk to our food and helping its passage through the digestive system. Despite the abundance of cellulose in nature, cellulase enzymes are rare and most animals cannot utilize cellulose as food. However, some bacteria and fungi are able to synthesize such an enzyme. Animals such as cows and sheep have bacteria in their stomachs that produce a cellulase to break down cellulose. These mammals can use grass and hay as food; they then absorb the monosaccharides and disaccharides released by the bacteria.

The difference between them can be described using slightly different terminology relating to the orientation of the linking bonds in the molecule. Cellulose and starch are both composed of the same repeating unit, α-glucose, but they differ in the position of the oxygen atom joining the rings together. In cellulose the oxygen atom joins two rings using two equatorial bonds, but in starch the oxygen atom joins two rings using one equatorial and one axial bond.

Cellulose is composed of long chains held together by inter-molecular hydrogen bonds, thus forming sheets that stack in an extensive three-dimensional network. The axial–equatorial ring junction in starch creates chains that fold into a helix. The human digestive system contains the enzyme necessary to hydrolyse starch by cleaving its axial C–O bond, but not an enzyme to hydrolyse the equatorial C–O bond in cellulose (Figure 23.115). Hence an apparently minor difference in the three-dimensional arrangement of atoms gives very different properties to starch and cellulose.

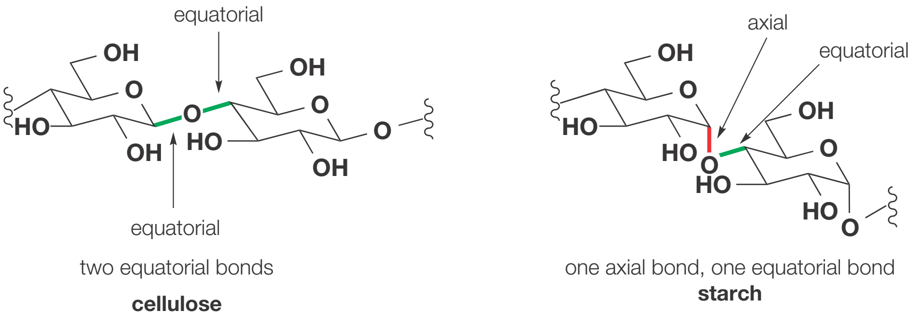
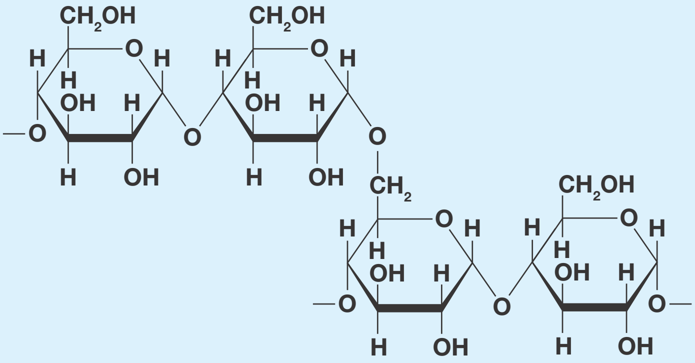


Figure 23.115 The different orientations of the linking bonds in cellulose and starch

27 The diagram below shows a part of the structure of a carbohydrate storage polymer which can be broken down by enzyme action.



a What types of glycosidic linkages are present in this structure?
b Name this type of breakdown reaction.
c Draw the structure of the monomer unit produced
d Suggest which carbohydrate polymer this drawing represents.
e Suggest why the polymer functions as an energy store but the monomer does not.

28 Describe the difference in structure between starch and cellulose. How does this difference affect human nutrition?

Figure 23.116 Probiotic cleanse: coated capsules that enable probiotic bacteria to pass through the highly acidic environment of the stomach

Dietary fibre

Dietary fibre (or roughage) is the indigestible portion of plant-based food that retains water and so aids motility and makes defecation easier. Dietary fibre contains cellulose, hemicellulose, lignin and pectins. Lignin is present in wood and plant stems and is an important component of mature plant cell walls. It is a complex alcohol-based polymer and cross-linked with other cell wall components. Hemicellulose is similar to cellulose but is formed from a variety of monosaccharides. Pectin is another complex polysaccharide present in the material that binds adjacent plant cells together.

Sources of dietary fibre are usually divided according to whether they are water soluble or water insoluble. Both types of fibre are present in all plant foods. Insoluble fibre possesses water-attracting properties that help to shorten transit time through the intestine. Soluble fibre undergoes fermentation by the gut microflora (probiotic bacteria – Figure 23.116) with healthy effects. Prunes have a thick skin covering a juicy pulp. The prune's skin is an example of an insoluble fibre source, whereas soluble fibre sources are inside the pulp.

The importance of dietary fibre in the diet

As the dietary fibre passes through the small intestine it undergoes a process of fermentation (to varying degrees). A variety of short fatty acids are produced that protect and promote the environment in the large intestine and stimulate the immune system. Dietary fibre (Figure 23.117) may be helpful in the prevention of conditions such as diverticulosis, irritable bowel syndrome, constipation, obesity, Crohn's disease, hemorrhoids and diabetes mellitus, and in lowering cholesterol levels.

Figure 23.117 A high-fibre breakfast cereal

Diverticulosis is characterized by the presence of pockets of tissue in the lining of the colon (large intestine) due to weakness in the muscle layers. It is thought to be caused by pressure in the colon which may result from a diet low in fibre. Symptoms include bleeding, bloating and abdominal cramps after eating.

Irritable bowel syndrome (IBS) is a bowel disorder characterized by abdominal pain relieved by defecation. The underlying cause of IBS is not known, but it may be caused by the body's immune system as a response to the presence of excessive microflora in the colon. IBS may also be triggered by certain foods in the diet.

Crohn's disease has similar symptoms to IBS. It affects men and women equally and appears to be genetically determined. It is especially common in Jewish people and African-Americans. It is believed to be an autoimmune disease involving the body's immune system responding to food and its own bacteria as 'foreign'. A common complication with Crohn's disease is blockage of the intestine.

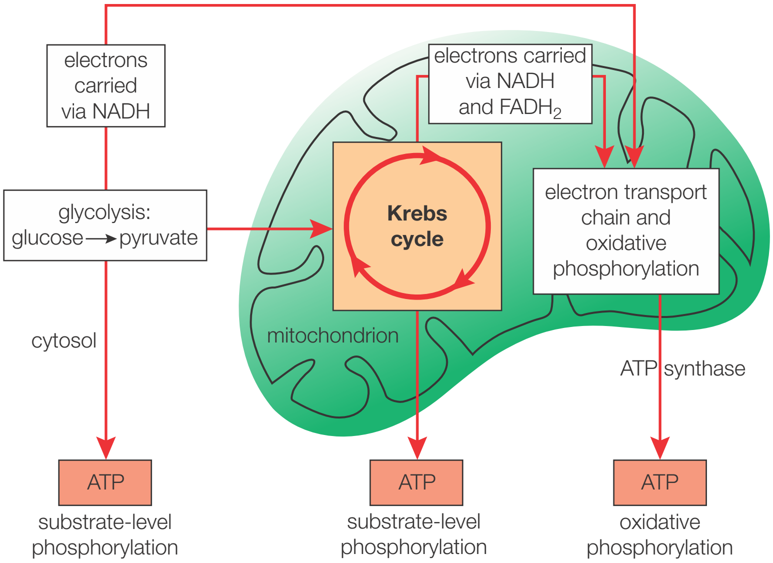
Hemorrhoids or piles occur when the veins in the rectum or anus become swollen and inflamed. Consequently, defecation is painful and blood is present in the stools. The general cause of hemorrhoids is pressure on the rectal veins. Increased straining caused by constipation or diarrhoea may lead to hemorrhoids. Severe cases of hemorrhoids are treated by surgery. A common approach is to cut the blood supply so the hemorrhoid dies and enters the stools.

Uses and functions of carbohydrates

Metabolic involvement

The themes of energy and energy storage have persisted through the aspects of carbohydrates that we have considered in this section. It is worth reiterating that carbohydrates often act as energy sources for respiration (Figure 23.118), with glucose being the major substrate for respiration. Starch and glycogen can be broken down by appropriate enzymes to release glucose. Fructose can also enter the respiratory process (via phosphorylation). The intermediates of sugars formed during respiration can act

Figure 23.118 Overview of cellular respiration


electrons carried via NADH
electrons carried via NADH and FADH$_2$
glycolysis: glucose → pyruvate
Krebs cycle
electron transport chain and oxidative phosphorylation
cytosol
mitochondrion
ATP synthase
ATP
ATP
ATP
substrate-level phosphorylation
substrate-level phosphorylation
oxidative phosphorylation

as precursors for a range of molecules, for example amino acids and porphyrins such as chlorophyll and hemoglobin. The nitrogenous bases of DNA are also synthesized from amino acids, which in turn are synthesized from the intermediates of aerobic respiration.

Carbohydrate fillers

We have mentioned a range of major uses for glucose, sucrose and various polysaccharides as we have progressed through this section. It is worth adding that many carbohydrates and their derivatives are used in the pharmaceutical industry to bind preparations into tablets or to act as 'inert' bulking agents for small quantities of drugs. Glucose, starch, sucrose and mannitol are all used for their adhesive properties. Mannitol is an alcohol sugar with the formula $C_6H_8(OH)_6$. Dextran is used as a plasma substitute, which increases blood volume in patients who have lost large amounts of blood. Dextran is a complicated branched polysaccharide of glucose.

23.5 Vitamins – *vitamins are organic micronutrients with diverse functions that must be obtained from the diet*

Vitamins as micronutrients

A **nutrient** is a substance required by an organism. Nutrients are needed for metabolism – the chemical reactions that occur inside cells and ensure growth and survival. Nutrients include proteins, carbohydrates, lipids, vitamins, minerals and water. Nutrients are divided into **micronutrients** and **macronutrients** (Figure 23.119).

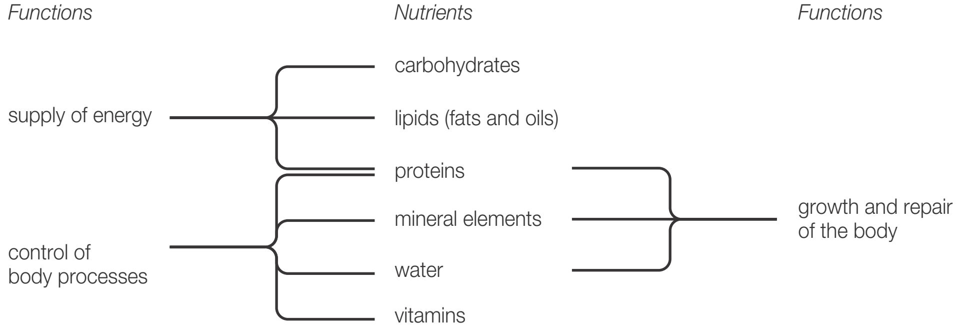


Figure 23.119 The importance of different macronutrients and micronutrients to the body

So far in this chapter we have discussed proteins, carbohydrates and lipids. All these are required in relatively large amounts in the human diet – they are known as macronutrients. Vitamins are a group of complex organic molecules that are classified as micronutrients. These are essential in small amounts (<0.005 per cent body weight) for the normal functioning of the body. They are required in milligram (mg) or microgram (μg) amounts. Micronutrients include the vitamins and many so-called trace minerals such as iron (Fe), copper (Cu), zinc (Zn), iodine (I), selenium (Se), manganese (Mn), molybdenum (Mo), chromium (Cr) and cobalt (Co).

Vitamins are categorized according to their biological function rather than chemical structure. Such functions include acting as prosthetic groups or co-factors by binding to enzymes, acting as hormones or antioxidants, or facilitating the transfer of functional groups or electrons. The name vitamin C refers to a single compound, ascorbic acid, but the name vitamin A refers to a group of compounds that include retinol (an alcohol), retinal (an aldehyde), retinoic acid and several poly-unsaturated hydrocarbons (the carotenes, can be converted to retinol and retinal in the body). Vitamin B comprises a diverse group of compounds (with molecular masses ranging from 123 to 1580). Vitamin D consists of four structurally similar compounds produced from the same precursor, 7-dehydrocholesterol, by different metabolic pathways (note that 7-dehydrocholesterol is also a precursor of cholesterol).

Nature of Science

The importance of vitamins

Careful observation is a cornerstone of science and the idea of vitamins originated when it was observed that certain diseases arose when people consumed diets deficient in certain foods. For instance, sailors often developed scurvy when they went on long journeys and did not have access to citrus fruit or other fresh fruit and vegetables; and people who ate white rice often developed beriberi, but those who ate brown rice did not. In these cases the treatment preceded a clear understanding of any underlying cause. These and similar observations led to the hypothesis of deficiency diseases – the idea that a lack of something in the diet could cause disease. This was formulated in 1912 and is now firmly established in healthcare practice. Vitamins were originally named 'vitamines' (vital amines). However, it was later found that not all vitamins are amines, so the 'e' was dropped and the term vitamins adopted.

Vitamins and their solubility

Vitamins cannot generally be made in the body (with the exceptions of vitamin D, which can be synthesized in the skin, and niacin, which can be formed from the essential amino acid tryptophan) so they must be obtained from the ingestion of suitable foodstuffs as part of a healthy diet. They are present in very small quantities in food and absorbed into the body during digestion. They have no energy value, but are essential for a healthy body and for maintaining body metabolism. They are also sometimes taken in the form of food supplements – vitamin tablets. The structures of vitamin A (retinol), vitamin D (calciferol) and vitamin C (ascorbic acid) are shown in Figure 23.120. Note that the names and structural formulas of these three vitamins are given in the *IB Chemistry data booklet* which is available during examinations.

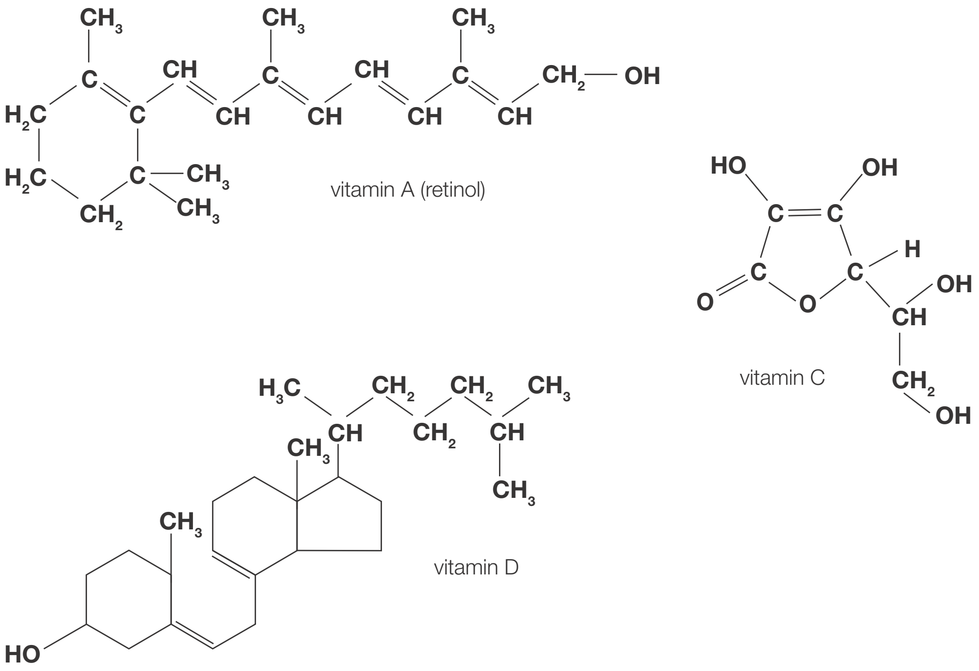
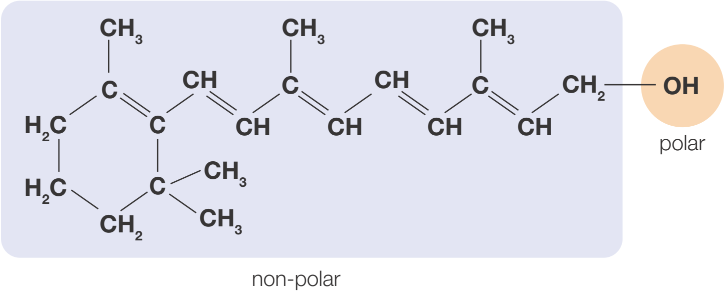


Figure 23.120 Structures of vitamins A, C and D

The vitamins are classified into two groups: water-soluble vitamins and fat-soluble vitamins. Vitamins A, D, E and K are fat-soluble vitamins and accumulate in the liver and adipose (fat) tissue where they can be stored for prolonged periods of up to several months. Such vitamins can be consumed less frequently than others without any detrimental health effects. Indeed, if excessive amounts are taken, levels can build up, resulting in toxicity.

Fat-soluble vitamins

The solubility of a vitamin can be readily deduced from its structure. Consider the structure of vitamin A (retinol) (Figure 23.121). Although the molecule contains a polar hydroxyl group, −OH, it is essentially non-polar due to the presence of a large hydrocarbon 'skeleton'. Vitamin A is therefore predicted to be predominantly hydrophobic, fat soluble and largely insoluble in water.



■ **Figure 23.121** The structure of vitamin A and classification into polar and non-polar regions

Retinol is a long-chain alcohol with an extensive system of alternating single and double carbon–carbon bonds. The π-electron clouds of the adjacent double bonds overlap with each other to produce an extended cloud of delocalized electrons – a conjugated system which, in retinol, involves ten carbon atoms (including two in the six-membered ring). Carotenes, another group of vitamin A compounds, have even longer conjugated systems (up to 22 carbon atoms). The longer chain of β-carotene can be cleaved to produce retinol. This electron conjugation makes the vitamin A compounds efficient antioxidants that readily react with molecular oxygen and free radicals. The conjugated systems also mean that all these compounds absorb visible light and therefore have bright colours.

The absorption of these hydrophobic vitamin A compounds into the intestinal tract and their transport depends on certain lipids and lipoproteins. Consequently, low-fat diets may lead to secondary vitamin A deficiencies that cannot be corrected simply by increased intake of the vitamins and may require a change in dietary habits.

Vitamin D is a collective name referring to cholecalciferol (vitamin D_3), ergocalciferol (vitamin D_2) and two other related compounds (vitamins D_4 and D_5), each also with a partly broken steroidal structure. Vitamin D without a subscript usually refers to either D_2 or D_3 or both. These are known collectively as calciferol.

In the human body cholecalciferol can be synthesized in the skin and requires ultraviolet (UV) light. The human body is normally able to synthesize enough vitamin D to meet metabolic requirements. However, when circumstances limit exposure to sunlight – winter months at high latitudes, for instance – vitamin D becomes an essential micronutrient that must be obtained from the diet or by supplement.

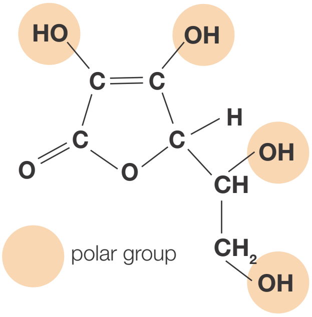
Vitamin D is fat soluble and transported in body fluids in the form of lipoprotein complexes. Natural sources and food supplements that contain vitamin D are usually rich in lipids and so do not require any additional fat intake as is the case with vitamin A. There are only a few dietary sources of vitamin D: oily fish such as salmon and herring, liver and eggs. Vitamin D stimulates the uptake of calcium ions by cells and is important in the health of bones and teeth

Water-soluble vitamins

In contrast to vitamins A and D, consider the structure of vitamin C (Figure 23.122). The molecule has a much smaller hydrocarbon 'skeleton', but this contains four polar hydroxyl groups, −OH, which can form hydrogen bonds with water molecules. Vitamin C (ascorbic acid) is therefore expected to be soluble in water.

Both vitamin C and the B group of vitamins (consisting of eight different vitamins) are water soluble. Water-soluble vitamins are transported directly in the blood, and excesses are filtered out by the kidneys. They are excreted readily in the urine and stores can become depleted, so a dietary intake is required. Unlike fat-soluble vitamins where there may be large body stores, reserves of water-soluble vitamins are small. Note, however, that it may take up to 6 months to deplete body reserves of vitamin C enough for deficiency signs to appear. By contrast, deficiency of vitamin B_1 may occur within about a week of depletion.

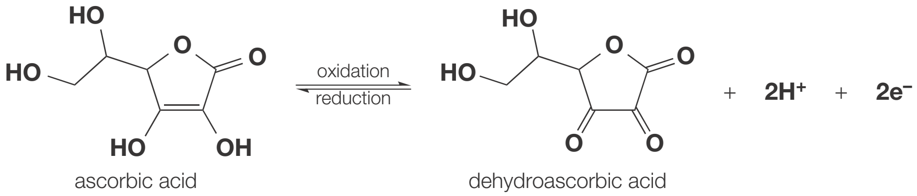
Vitamin C participates in a broad range of metabolic processes, including the biosynthesis of collagen. It is a powerful antioxidant and reducing agent capable of donating electrons in redox reactions.



■ **Figure 23.122** The structure of vitamin C showing its polar groups

The half-equation for its conversion to dehydroascorbic acid is:

$$C_6H_8O_6 \leftrightharpoons C_6H_6O_6 + 2H^+ + 2e^-$$



The concentration of vitamin C in solution can be determined by redox titration using DCPIP (2,6-dichlorophenol-indophenol, $C_{12}H_7NCl_2O_2$) as an indicator. In the presence of ascorbic acid the blue solution becomes colourless as DCPIP is reduced:

$$\underset{\text{blue}}{C_{12}H_7NCl_2O_2} + C_6H_8O_6 \rightarrow \underset{\text{colourless}}{C_{12}H_9NCl_2O_2} + C_6H_6O_6$$

When all the ascorbic acid has reacted the blue colour of the DCPIP will persist. Note that some plant materials may be very acid, in which case the colour of the DCPIP becomes pink–magenta. In this case, look for the appearance of a permanent pink colour during the titration, rather than a blue colour. Note that when using the DCPIP assay for vitamin C in vegetables or fruit, the food must be homogenized in acid (commonly phosphoric acid). The problem is that plant cells contain ascorbate oxidase, which destroys the vitamin C at neutral pH. It is inactivated when the food is homogenized in acid.

Ascorbic acid is a food additive (E300) used, along with other antioxidants, to prevent oxidative rancidity.

The solubility and key properties of vitamins A, C and D are summarized in Table 23.14.

Table 23.14 The solubility and key properties of vitamins A, C and D

Vitamin	Solubility and properties
A, retinol	Fat soluble Hydrocarbon chain and ring are non-polar and influence the solubility more than the single –OH group does Involved in the visual cycle in the eye, and particularly important for vision at low light intensity
C, ascorbic acid	Water soluble Several –OH groups enable hydrogen bonds to form with water Acts as co-factor in some enzyme reactions (hydroxylation of proline in collagen synthesis, for example); important in tissue regeneration during wound healing and following injury It can also act as an antioxidant, protecting the body from damage by free radicals produced naturally during normal metabolic processes
D, calciferol	Fat soluble Predominantly a hydrocarbon molecule with four non-polar rings and only one –OH group; chemically similar to cholesterol Stimulates the absorption of calcium from the gut; important in whole body calcium homeostasis and the health of bones and teeth

Vitamin C is an important factor in the functioning of enzymes involved in collagen synthesis. More generally, a number of enzymes function only in the presence of a **co-enzyme**. These are small organic molecules and are temporary parts of an enzyme's structure. Co-enzyme molecules are often vitamins or are made from vitamins. Examples of vitamins acting as co-enzymes include vitamin B_1 (thiamin), vitamin B_6 (pyridoxine) and vitamin B_2 (riboflavin).

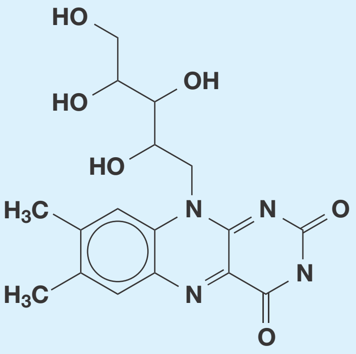
Heat sensitivity of vitamins

Some vitamins, particularly those that are water soluble, such as vitamin C and vitamin B_1 (thiamin), are sensitive to heat. This can be important as many foods that we consume are heat processed in some way (the pasteurization of milk, for instance) and this can cause a reduction in their vitamin content.

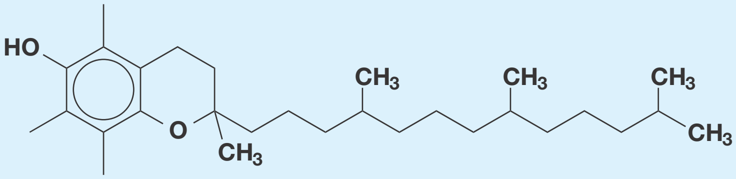
Vitamin C will be lost when vegetables such as broccoli are boiled in water. This is due to the effect of the heat, which results in the oxidation of the vitamin to dehydroascorbic acid, but also because the vitamin C is soluble in water and will dissolve in the cooking water. The hydrocarbon backbones of the fat-soluble vitamins A and D are relatively stable to heat and do not decompose significantly when food is steamed or boiled. Over-cooked or fried food can lose more than 50 per cent of its fat-soluble vitamin content and virtually all its vitamin C.

Vitamins A and C, containing carbon–carbon double bonds and –OH groups, are more sensitive to light and air than vitamin D as they are susceptible to free radical and redox reactions.

29 Consider the structures of two vitamins given below:



I



II

a Name two oxygen-containing functional groups in the structure of I.
b Name two oxygen-containing functional groups in the structure of II.
c Classify each of these vitamins as either water soluble or lipid soluble.

Measuring vitamin C content

It is possible to measure the vitamin C (ascorbic acid) content of a sample of fruit juice by measuring the volume of the sample required to decolourize a solution of DCPIP (2,6-dichlorophenol-indophenol).

Calibrate the results by comparison with a known concentration of vitamin C. A standard 1% solution of vitamin C with 1 g of vitamin C dissolved in 100 cm^3 is suggested; this is 10 mg cm^{-3}.

1 Pipette 2 cm^3 of the juice or vitamin C solution into a test tube.
2 Using a graduated pipette or a burette, add 1% DCPIP drop by drop to the vitamin solution. Shake the tube gently after adding each drop. Add DCPIP until the blue colour just disappears.
3 Record the exact amount of DCPIP solution that was added.
4 Repeat the procedure and calculate an average result.
5 Repeat with the juice to be tested. If the juice decolourizes a large volume of DCPIP, dilute the fruit juice and repeat the test. If the juice has a strong colour that will interfere with determining the end-point, dilute the juice before testing.
6 Calculate the amount of vitamin C in the standard solution in mg cm^{-3}. Calculate how much vitamin C there is in each of the fruit juices in mg cm^{-3}.

Having established the reliability of the procedure, the following statements could be tested:

- Fresh juices have more vitamin C than long-life juices.
- Juice 'not from concentrate' is best in terms of vitamin content.
- Fruit squashes have less vitamin C than fruit juices.
- If heat destroys vitamin C, then heat-treated long-life juices will have lower concentrations.
- If heat destroys vitamin C, then boiled fruit juice will have lower concentrations than unboiled juice.
- Manufacturers generally provide reliable information about their products.
- Vitamin C degrades in vitamin tablets, and old tablets will have less vitamin C than fresh ones.

Note that:

- The method can be adapted to measure the vitamin C content of fruit and vegetables directly – see earlier note on homogenization in acid (www.saps.org.uk/secondary/teaching-resources/191-measuring-changes-in-ascorbic-acid-vitamin-c-concentration-in-ripening-fruit-and-vegetables)
- Similar estimates of ascorbic acid content can be carried out using redox titrations involving potassium iodate(V) or iodine (www.outreach.canterbury.ac.nz/chemistry/documents/vitaminc_iodine.pdf).

Nature of Science

Vitamin C and the common cold

Science is not without controversy and it is definitely not the case that all scientists believe the same things, or indeed that one scientist always backs the correct ideas. One of the greatest chemists of the 20th century was the US scientist Linus Pauling (the only person to win two unshared Nobel prizes – one for chemistry and one for peace). He elucidated the structure of the α-helix in protein secondary structure and we will refer to him again when discussing the structure of DNA. Later in life he heavily backed the idea that high doses of vitamin C could prevent a number of diseases, including the common cold and could even be a treatment for cancer. His claims and the scientific studies associated with them were very controversial. There were flaws in the methodology of the studies – a lack of blind trials – and in the logic of the arguments used. He appears to have taken the vitamin C intake of primates in a zoo, where they were fed large amounts of fruit, as their requirement for the vitamin, which is not so. The debate still carries on today as to whether high doses of vitamin C have any beneficial effect.

In the context of such controversies it is important to note that the 'authority of science' lies in its methodology and not in individual or groups of scientists. It is important also to realize that this 'authority' is confined to areas which the scientific method is competent to address.

ToK Link

Authority and science

The involvement of well-known and prestigious scientists in backing controversial ideas does raise issues around how scientists use their authority and their responsibility to speak carefully on issues of the day. There is a social contract between science and other institutions in society, and following World War II, an almost unquestioned cultural belief in the value and material benefits of science. This meant that scientists have enjoyed considerable social and political freedom within many developed countries.

In this period the 'authority of science' was based on its distinctive understanding and 'way of knowing' as a valid and, to some, a seemingly better approach compared to competing understandings (for example, an emotional intuition or a religious understanding that treats as known matters revealed by the word of God, etc.).

However, beginning in the late 1960s, the heavy reliance on science and technology for sustained economic growth combined with the growing technical and moral risks of scientific development (for example, the use of embryonic stem cells) resulted in constraints on the autonomy of science, and increased demands from various societal interests for greater input on science-related policy.

There is also a growing reaction to what some see as a growing 'absolutism' in the approach of some scientists in communicating the findings of science and opinion generated from them. Professor Brian Cox, one of Britain's best-known scientists and communicators on science, became involved in a debate on this issue while commenting on the issues involved in assessing climate change. In proposing the validity of the computer models of the progress of climate change, he made a particular statement that the subsequent discussion latched on to, generating significant online reaction. That statement was that 'it's clearly a bad thing for knowledge to be controversial'. To what extent is this valid? His critics argued to the contrary, stating that from the Enlightenment onwards the reasoned believed that actually it was good – essential, in fact – for knowledge to be treated as controversial and open to the most stringent questioning. For instance, John Stuart Mill suggested that controversy is the lifeblood of knowledge: 'complete liberty of contradicting and disproving our opinion is the very condition which justifies us in assuming its truth for the purposes of action'.

Does the general public possess adequate knowledge to weigh and assess science and technology claims that appear in the mass media? Has the public followed and reflected on the issues? What role does public trust in institutions play in the public understanding of science, and can institutions promote understanding via new forms of public engagement? What is the inter-relationships between trust and knowledge in shaping public perceptions and subsequent action?

Nutrient deficiencies

A disorder occurs if the diet does not contain enough of a vitamin, which is termed a deficiency disease. When the intake in the diet is insufficient, the deficiency disease can be avoided by supplementing the diet with the necessary vitamin. Table 23.15 indicates some of the sources and biochemical functions of selected vitamins required in the human diet and the associated deficiency disease.

Table 23.15 The sources, functions and deficiency diseases of the major human vitamins

Name of vitamin	Important sources	Biological function	Deficiency disease
A (retinol)	Milk, butter, eggs, liver, cod liver oil, green and yellow vegetables	The aldehyde form known as retinal is essential for the formation of the visual pigment rhodopsin	Skin and cornea in eye become dry (not connected with the function of retinal in rhodopsin); poor 'night vision'
D (calciferol)	Cod liver oil, eggs, margarine (if fortified), milk (if fortified); also synthesized by the body in the presence of sunlight	Controls calcium absorption; important in bone and tooth formation	Rickets in children: failure of bones to calcify and become hard; osteomalacia in adults: spontaneous fractures
C (ascorbic acid)	Citrus fruits, green vegetables, potatoes, tomatoes	Essential for collagen synthesis	Scurvy – skin of gums becomes weak and bleeds; wounds fail to heal; connective tissue fibres fail to form
B_1 (thiamin)	Unprocessed rice, whole cereals, egg yolk, liver, milk, green vegetables, fruits	Acts as a co-enzyme in aerobic respiration	Beriberi – nervous system affected; muscles become painful and weak; loss of appetite
Niacin	Meat, wholemeal bread, yeast extract, liver	Essential part of several co-enzymes	Pellagra – skin lesions, rashes, diarrhoea

Vitamin A deficiency

Mammals, including humans, will die if their diet is deficient in vitamin A, but they do not die because of problems with their visual **pigments**. Vitamin A has a much more important function in maintaining life, growth and general health. The active form of A in this systemic mode of action is retinoic acid. Rats given a diet containing retinoic acid as the sole source of vitamin A become blind because they cannot reduce retinoic acid to retinal to serve as visual pigments, but they grow normally and are outwardly healthy, illustrating the two separate functions. Note that retinoic acid has hormone-like actions, controlling gene expression, and is also essential for the functions of vitamin D and thyroid hormone.

Deficiency of vitamin A can result in a condition called xerophthalmia, which is a leading cause of blindness in many developing countries. Fortification of foods with vitamin A (adding vitamin A to foods) has proved a successful strategy for combating this deficiency.

Programs also exist to encourage farmers to grow varieties of foods richer in β-carotene (provitamin A; which can be converted into vitamin A in the body) – this is called biofortification. For instance, the introduction of orange-fleshed sweet potato into Uganda to replace the indigenous white-fleshed variety has met with some success in reducing vitamin A deficiency. The complexity of the situation is illustrated by the fact that many trials to increase consumption of dark green leafy vegetables and other sources of carotene have failed to prevent vitamin A deficiency in developing countries where the intake of fat is below about 20 per cent of energy requirement, meaning that the carotene is not absorbed.

Vitamin B group deficiency

Niacin is converted to a co-enzyme that plays a key role in oxidation–reduction processes in the cell. Deficiency results in a condition called pellagra, characterized by diarrhoea, dermatitis and dementia. Vitamin B1 (thiamin) is converted to a co-enzyme that is essential for energy production within cells. Deficiency leads to beriberi, characterized by muscle weakness. Many foods, such as breakfast cereals, are fortified with niacin and thiamin, and deficiency is rare in developed countries.

Vitamin C deficiency

Vitamin C deficiency results in scurvy – a condition with symptoms of bleeding gums, poor wound healing, poor resistance to infection and dark spots on the skin. This used to be a common problem among sailors who spent long periods at sea without fresh fruit and vegetables – then it was recognized that a regular intake of citrus and other fruits and vegetables would prevent this disease.

James Cook succeeded in circumnavigating the world (1768–1771) in HM *Bark Endeavour* without losing a single man to scurvy, but his suggested methods, including a diet of sauerkraut, are now thought to be of limited value. Sauerkraut was the only vegetable food that retained a reasonable amount of ascorbic acid in a pickled state, but it was boiled to prepare it for preservation and much of the vitamin C content would have been lost. In Cook's time it was impractical to preserve citrus fruit for long sea voyages. More important was Cook's regime of shipboard cleanliness, enforced by strict discipline, as well as frequent replenishment of fresh food. But it is interesting to note that while 16th and 17th century British and other sailors suffered from scurvy during long sea voyages, the Dutch who consumed relatively large amounts of sauerkraut, rarely did.

Nowadays, scurvy is rarely present in adults, although infants and elderly people can be affected. Vitamin C is destroyed by the process of pasteurization, so babies fed with ordinary bottled milk sometimes develop scurvy if they are not provided with adequate vitamin supplements. Virtually all commercially available baby formulas contain added vitamin C for this reason, but heat and storage destroy vitamin C. Human breast milk contains sufficient vitamin C if the mother has an adequate intake herself.

Scurvy is one of the accompanying diseases of malnutrition and thus is still widespread in areas of the world depending on external food aid. Though rare, there are also documented cases of scurvy due to poor dietary choices by people living in industrialized nations.

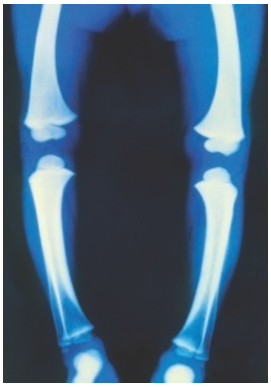


Figure 23.123 Coloured X-ray of the weakened bones and bowed legs of a child suffering from rickets. The skeleton of the legs appears deformed, with the long bones of the limbs severely curved. Rickets is caused by a nutritional deficiency of vitamin D

Vitamin D deficiency

Vitamin D deficiency can result in rickets in children (Figure 23.123). As explained above, it is a condition in which softening and deformity of the bones occur due to a reduction in the uptake of calcium and phosphate from food. Fortification of dairy products with vitamin D means that deficiency is now rare in industrialized countries. However, it is still a problem in some developing countries where intake of dairy products may be low or where religious or social customs (wearing clothes that cover virtually all the body) or climatic conditions prevent an adequate exposure to sunlight, and there is still a problem of (preclinical) rickets in developed countries with poor sunlight exposure.

A recognized consequence of low vitamin D is osteomalacia in adults. It also contributes to osteoporosis in elderly people that results in fragility fractures, partly through increased risk of falls. Vitamin D deficiency has also been implicated in cardiovascular disease, increased cancer risk and mortality, falls, diabetes, multiple sclerosis, osteoarthritis and epilepsy.

The 'sunshine vitamin' and supplements

The main source of vitamin D comes from exposure of the skin to sunlight. Hence there is considerable seasonal variation, with concentrations higher at the end of summer compared to other seasons. Vitamin D is found in fatty fish such as herring, salmon and mackerel. Other sources include eggs, meat and fortified foods such as margarine. Adequate vitamin D is unlikely to be achieved through dietary sources alone without fortification. Exposure to sunlight must be balanced with the risks of skin aging and skin cancer.

Urban lifestyles and the more widespread use of sunscreen lotions during the summer months seem to have resulted in greater concern over an increase in vitamin D deficiency even in affluent populations. Even a sunscreen with a minimal SPF (sun protection factor) of 15 is sufficient to block a significant proportion of ultraviolet radiation and reduce the synthesis of cholecalciferol in the skin by 98 per cent. Higher SPF sunscreens can effectively prevent any synthesis of vitamin D and make an individual totally dependent on dietary supplements.

The supplementation of vitamin D in the diet using 'vitamin pills' is probably one of the most frequent use of such supplements. There is increasing evidence that higher intakes of vitamin D than are required to prevent rickets and osteomalacia may be beneficial, especially with respect to some cancers and type 2 diabetes.

The nutritional supplement industry, especially the sale of vitamin pills, is in a time of continual global growth. The demand for vitamins and nutritional supplements continues to increase at a steady pace, with sales rising to $84 billion in 2011, up 6.1 per cent from $80 billion in 2010, with the USA holding top market share. Global sales in dietary supplements are estimated to continue growing by at least 4% each year up to 2018.

Malnutrition

Malnutrition is a general term for a medical condition caused by an improper or inadequate diet. An individual will experience malnutrition if the appropriate amount or quality of nutrients comprising a healthy diet is not consumed. An extended period of malnutrition can result in starvation, vitamin or mineral deficiency diseases and infection (since the immune system is affected). Protein–energy malnutrition refers to a form of malnutrition where there is inadequate protein intake. There are two forms: **marasmus** and **kwashiorkor** (Figure 23.124).

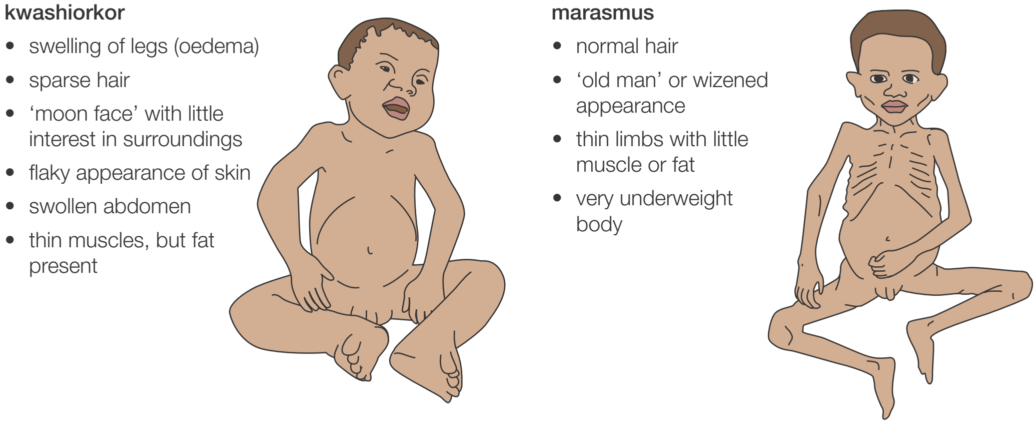

kwashiorkor
• swelling of legs (oedema)
• sparse hair
• 'moon face' with little interest in surroundings
• flaky appearance of skin
• swollen abdomen
• thin muscles, but fat present
marasmus
• normal hair
• 'old man' or wizened appearance
• thin limbs with little muscle or fat
• very underweight body

Figure 23.124 Kwashiorkor and marasmus

Marasmus is a form of severe malnutrition characterized by energy deficiency. The malnutrition associated with marasmus leads to extensive tissue and muscle wasting and oedema (accumulation of fluid beneath the skin or in a body cavity). Other common characteristics include dry skin and loose skin folds hanging over the buttocks and armpits. There is also drastic loss of adipose (fat) tissue from normal areas of fat deposits, such as the buttocks and the thighs. People with marasmus are often irritable and extremely hungry. In children body mass may be reduced to less than 80 per cent of the normal mass for that height.

Kwashiorkor is a type of malnutrition that was commonly believed in part to be caused by insufficient protein consumption. However, all the recent evidence (over the past 20 years) suggests that the problem of kwashiorkor is overall lack of food, not specifically of protein. Children aged 1–4 years are most often affected, although it also occurs in older children and adults. Kwashiorkor is a Ga word meaning 'displaced child', referring to the fact that the condition often develops after breast feeding has stopped – the disease was first described and named in what was the Gold Coast, now Ghana. Symptoms of kwashiorkor include a swollen abdomen (pot belly, due to fatty infiltration of the liver because of failure to synthesize transport proteins for export of fat from the liver), alternating bands of pale and dark hair, and weight loss. Common skin symptoms include a sunburn-like dermatitis, similar to that seen in pellagra, often called sooty dermatitis.

Kwashiorkor is likely due to deficiency of protein in combination with one of several types of micronutrients (e.g. iron, folic acid, iodine, selenium and vitamin C), particularly those involved with antioxidant properties.

Food fortification

Some foods are 'fortified' to ensure that a normal diet can provide sufficient vitamins or minerals to maintain health. For example, the milling of wheat removes part of the grain richest in vitamin B_1, so white flour has this vitamin added to it. Margarine has vitamins A and D added (to match that found in butter), and vitamin C is often added to fruit juices and dehydrated mashed potato. Potassium iodate(V) is commonly added to table salt in regions where iodine deficiency is a problem.

A more controversial form of biofortification involves genetic modification (GM) of foodstuffs to make them richer in a particular vitamin. Genetic modification has been used to produce 'golden rice' – a variety of rice rich in provitamin A (β-carotene). It is hoped that the use of golden rice will make a significant difference to vitamin A deficiency in countries such as India, Bangladesh and Vietnam.

Biofortification and genetically modified crops

A number of scientists are claiming that GM plants can significantly reduce malnutrition, especially in the developing world, through the development of plants that are resistant to pest-derived disease, adverse soil pH or drought conditions. In addition, plants can be engineered to have high levels of specific nutrients.

The advantages of using GM crops include a potential reduction in input in terms of labour and machinery costs (of particular importance to resource-poor farmers), a reduction in the use of potentially harmful fertilizers and insecticides, and a potential reduction in the amount of land required for cultivation due to an increase in yield (or rather a decrease in the numbers of diseased plants).

Golden rice is a variety of rice (*Oryza sativa*) produced through genetic engineering to biosynthesize β-carotene, a precursor of vitamin A (retinol) in the edible parts of rice. Golden rice was developed as a fortified food to be used in areas where there is a shortage of dietary vitamin A. However, it is worth noting that an adequate fat intake is essential for the β-carotene from golden rice to be absorbed. Approximately 24 million people, in 118 countries, were estimated to be affected by vitamin A deficiency. It is responsible for 1–2 million deaths, 500 000 cases of irreversible blindness and millions of cases of xerophthalmia every year. Vitamin A deficiency is currently treated orally or by injection.

ToK Link

Fluoridation and iodization

The practice of food fortification increases the nutritional values of dietary products and avoids widespread deficiencies caused by geographical or cultural factors. However, while beneficial for the majority of a population, it limits the freedom of an individual to choose their diet and, in rare cases, can lead to vitamin poisoning and allergic reactions. The practice raises issues concerning the balance between the rights of an individual, the interests of society and the level at which government should provide for and protect society. Illustrative cases that highlight these issues are two additions to public supply carried out in the interests of public health:

- fluoridation of the drinking water supply to prevent tooth decay
- the iodization of salt to counteract iodine deficiency disorders.

For almost 100 years an inverse correlation between the levels of fluoride and occurrence of dental decay has been noted. This resulted in a preventative policy in many countries of fluoridation of water in areas with a low fluoride level. There were objections on the basis of suspected undesirable side effects of fluoride such as Down's syndrome and various cancers. The ethical objection is that adding fluoride to the water supply is a compulsory medication and, therefore, a violation of individual rights.

The issues involved include:

- Beneficence (action for the benefit of others): iodine deficiency disorders comprise a spectrum ranging from abortion to hypothyroidism. The most serious and irreversible consequences of iodine deficiency are abortions, stillbirths, congenital malformations and mental retardation. Fortification of salt with iodine can be considered as a 'vaccine' for ensuring the proper physical and mental development of a child.

- Non-maleficence: no harmful effects of excess iodine consumption have been reported from countries that were previously iodine sufficient. Allergy to iodine in the fortified salt has not been reported, though it is possible.
- Justice (equity): this means that priority should be given to those who are in need, in proportion to their need. In the case of iodine deficiency, the communities that are deficient in iodine are the 'needy' communities.
- Autonomy (freedom to make choices): universal salt iodization can be looked upon as a compulsory medication and, therefore, a violation of an individual's rights.

Important minerals

Other nutrients that are highly significant components of a balanced diet include minerals. Minerals act as metabolites for various cell processes, raw materials for body tissue formation and components of enzymes and provide the correct chemical environment for cells.

A selection of minerals essential to human health are listed in Tables 23.16 and 23.17. Plants and animals lacking an essential mineral will develop a characteristic deficiency disease or disorder.

Minerals have four functions in the body:

- They act as raw materials for the formation of body tissues.
- They provide the necessary chemical environment for cells.
- They act as metabolic intermediates for cell processes.
- They act as co-enzymes.

The minerals in the first two categories above are required in relatively large amounts and are macronutrients (Table 23.16). The quantities required will be in excess of 0.005 per cent of body mass.

Minerals in the last two categories are required in much smaller amounts and are micronutrients (Table 23.17). They are required in milligram or microgram quantities (10^3 mg = 1 g; 10^6 μg = 1 g).

Table 23.16
Macronutrient minerals

Mineral	Major food source	Function
Calcium	Milk, cheese, bread	Bone and teeth formation, muscle contraction, nerve action, blood clotting, blood formation
Phosphorus	Cheese, eggs	Bone and tooth formation, respiration, ATP and nucleic acid formation
Sulfur	Dairy products, meat, eggs, broccoli	Formation of keratin and extracellular matrix
Potassium	Potatoes, meat, chocolate	Muscle contraction, nerve action, active transport across cell membranes
Magnesium	Meat, green vegetables	Formation of bones and co-enzymes for respiration
Chlorine	Salted foods, e.g. crisps, sea food	Maintaining the anion/cation balance in cells, formation of gastric juice (hydrochloric acid)
Sodium	Any salted food, meat, eggs, milk	Muscle contraction, nerve action, active transport across cell membranes

Table 23.17
Micronutrient minerals

Mineral	Major food source	Function
Iron	Liver, red meat, some vegetables, e.g. spinach	Heme group in hemoglobin (oxygen carrier in blood)
Copper	Most foods	Cytochrome c oxidase (electron transport chain), many other enzymes
Fluorine	Milk, drinking water in some areas	Component of tooth enamel and bone
Zinc	Most foods	Co-factor for many enzymes
Iodine	Seafood and iodized salt	Thyroxine (thyroid hormone) synthesis (control of basal metabolic rate)
Selenium	Plants, seafood, meat, mushrooms	Antioxidant enzymes, reduces risk of cancer and heart disease, boosts selenoproteins in immune system
Manganese	Most foods	Phosphatase enzymes (transfer phosphate groups)
Molybdenum	Most foods	Enzyme co-factor
Chromium	Most foods	Uptake of glucose
Cobalt	Most foods	Red blood cell development

Just as for vitamins, deficiency diseases can also arise when a person's diet lacks a specific mineral (Table 23.18).

Table 23.18 The sources, functions and deficiency diseases of the major minerals in the human diet

Mineral	Deficiency disease	Symptoms of deficiency disease
Calcium and phosphorus	Rickets in children and osteomalacia (softening of bones) in adults, osteoporosis in adults	Bones and teeth are affected: twisted limbs or unformed teeth
Sulfur		Skin problems or disorders, muscle pain, nerve disorders, circulatory trouble, stress, infection
Potassium	Rarely deficient	Heart disease
Chlorine		Muscular cramps
Sodium		Muscular cramps, heart disease
Iron	Anaemia: low level of red blood cells	Fatigue, dizziness, rapid heart beat
Iodine	Goitre (Figure 23.125)	Enlarged thyroid gland, protruding eyes

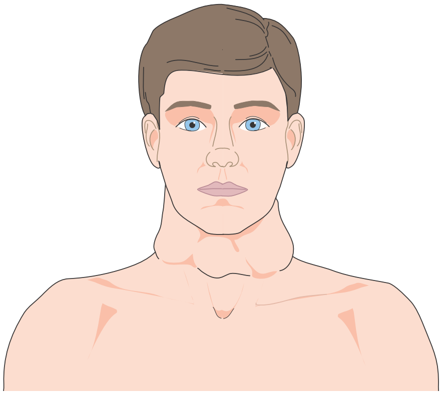


Figure 23.125 An adult untreated with goitre

23.6 Biochemistry and the environment – *our increasing knowledge of biochemistry has led to several environmental problems, while also helping to solve others*

Biochemistry is a multidisciplinary science that has expanded significantly in recent years and greatly increased our understanding of biochemical processes and our ability to control biological systems. However, that development, together with massive technological expansion and development, has created and revealed serious ecological problems. It has also increased our awareness of environmental impact and the ethical implications of scientific and technological development. This topic focuses on the use of biochemical techniques in industrial, agricultural and household applications and their effects on worldwide and local ecosystems. It also indicates aspects of the role of biochemistry in alleviating the environmental impact of human activities.

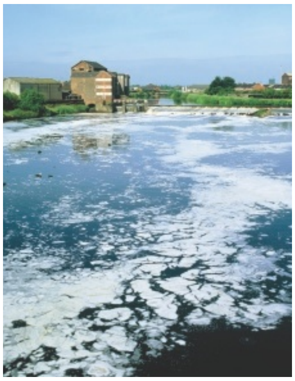


Figure 23.126 River pollution in northern England. The frothy scum of an industrial pollutant formed on the surface of a river. Although the pollutant may have reached the river many miles upstream, it is not until the river passes over a weir that the scum forms

Xenobiotics – foreign compounds

Xenobiotics are chemical compounds that are found in a living organism but are foreign to that organism (from the Greek *xenos*, 'stranger', and biotic, 'related to living organisms'). The term can also be applied to chemicals found in higher-than-normal concentrations in an organism, or to compounds that are not produced naturally but only by synthetic processes – in other words, chemicals that are foreign to the biosphere.

Specifically, drugs such as antibiotics are xenobiotics in humans because the human body does not produce them itself and they are not part of a normal diet. Natural compounds can also become xenobiotics if they are taken up by another organism, such as the uptake of natural human hormones by fish found downstream of sewerage treatment plant outfalls.

The presence of pharmaceuticals, including antibiotics, chemotherapy drugs and hormones, in waste water is becoming a big problem. These pharmaceuticals are then carried to sewerage treatment plants. These pharmaceuticals may be discharged from industries or hospitals, or passed through the human body and released unmodified or partially metabolized in urine. This situation is a relatively new problem which sewerage treatment plants were not specifically designed to deal with. Sewerage treatment plants may break down the xenobiotics through bacterial action, but too often this process is incomplete. The diverse nature of the chemicals involved compounds the problem. These issues taken together mean that these extraneous substances are not effectively removed from the waste water and some are released into the environment. The result of this is that water containing a variety of pharmaceuticals can be taken up by fish living downstream and is used for drinking and the irrigation of agricultural land. Although these pharmaceuticals are found in drinking water in only very small amounts (at concentrations of the order of ng dm^{-3}), there are concerns that long-term exposure could result in damage to human health.

One particular problem is the release of antibiotics into the environment in waste water. This is regarded as a particular problem because not only can they cause damage to aquatic organisms, but they can also result in increased bacterial resistance to antibiotics. Antibiotics are used to treat a variety of conditions but if bacteria develop resistance to antibiotics such as penicillin these diseases can become much more difficult to cure. This is an environmental aspect to the problem of antibiotic resistance that is reportedly resulting from over-prescription in parts of the developed world.

The waste water from sewerage treatment may also contain sex hormones such as the female oestrogens that have been released in human urine, particularly due to use of the synthetic contraceptive pill. There is some concern that male fish may absorb sufficient quantities of oestrogens to become 'feminized' and unable to breed.

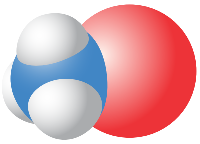
The following xenobiotics are causing current concern in the environment:

- household and industrial detergents
- pharmaceutical drugs, including antibiotics such as penicillin
- food additives
- pollutants, such as PCBs and dioxins
- insecticides, such as DDT
- heavy metals, such as mercury and lead compounds
- hormones, such as oestrogens
- plasticizers from polymer manufacture

Depending on their chemical nature, some xenobiotics can be completely digested by microorganisms, plants and animals. The body removes xenobiotics by **xenobiotic metabolism**. Xenobiotic metabolism is the set of metabolic pathways that modify the chemical structure of xenobiotics. In general, drugs are metabolized more slowly in fetal, neonatal and elderly humans and animals than in adults.

Polar molecules are often soluble in water and are quickly metabolized within the organism or undergo photochemical oxidation. Non-polar molecules, on the other hand, pass relatively easily across the hydrophobic cell membranes. They enter cells where they may be modified by enzymes and then detoxified. This consists of the deactivation and the excretion of xenobiotics and happens mostly in the liver. This is how many drugs are broken down in the body. Excretion routes are urine, faeces, breath and sweat. However, many synthetic chemicals produce toxic metabolites, affect the metabolism of other compounds or affect the development of the organism. Certain xenobiotics cannot be metabolized by existing pathways and remain within the organism or are excreted unmodified.

If the xenobiotic cannot be modified in the organism it may build up in the cells. The increasing concentration of the substance in an organism is known as **bioaccumulation**. For example, mercury compounds in the form of methylmercury, which is non-polar, cross into the brain, where they build up, causing mercury poisoning (Figure 23.127).



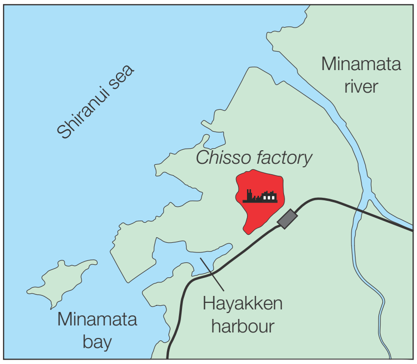
$[CH_3Hg]^+X^-$

Figure 23.127 Methylmercury(I), a toxic xenobiotic that accumulates in the brain causing mercury poisoning

Minamata disease

Outbreaks of methylmercury poisoning occurred in several places in Japan during the 1950s due to industrial discharges of mercury into rivers and coastal waters. The best-known instances were in Minamata and Niigata. In Minamata alone, more than 600 people died due to what became known as Minamata disease. More than 21 000 people filed claims with the Japanese government, and almost 3000 became certified as having the disease. In 22 documented cases, pregnant women who consumed contaminated fish showed mild or no symptoms but gave birth to infants with severe developmental disabilities.

The Chisso Minamata factory first started ethanal production in 1932, and output increased progressively until 1960. The chemical reaction used to produce the ethanal used mercury sulfate as one of the catalysts. Starting in August 1951, following a change of co-catalyst, a side reaction of the



■ **Figure 23.128** The location of the Chisso Minamata factory showing the proximity of the factory to Minamata Bay

catalytic cycle led to the production of a small amount of methylmercury. This highly toxic compound was released into Minamata Bay from the change of the co-catalyst in 1951 until 1968, when this production method was discontinued.

In February 1959, the mercury distribution in Minamata Bay (Figure 23.128) was investigated. The results shocked the researchers involved. Large quantities of mercury were detected in fish, shellfish and sludge from the bay. The highest concentrations centred around the Chisso factory waste-water canal in Hyakken Harbour and decreased going out to sea, clearly identifying the plant as the source of contamination. Pollution was so heavy at the mouth of the waste-water canal that a figure of 2 kg of mercury per ton of sediment was measured: a level that would be economically viable to mine.

The incidence of this form of mercury poisoning was not confined to Japan. Ontario Minamata disease is a neurological syndrome caused by severe mercury poisoning. It occurred in the Canadian province of Ontario in 1970, and severely affected two First Nation communities in north-western Ontario following consumption of local fish contaminated with mercury, and another First Nation community in southern Ontario (Sarnia) due to illegal disposal of industrial chemical waste.

The first of these Ontario cases concerned a chlor-alkali plant manufacturing sodium hydroxide using the mercury electrolysis cell technology (Castner–Kellner cell). Such cells use a flowing mercury cathode. Current-day mercury cell plant operations are criticized on environmental grounds and as a result of these concerns mercury cell plants are being phased out. They are being replaced by plants using semi-permeable membrane cell technology, which eliminates use of mercury. This is a successful example of environmental concern driving the development of new industrial technology for a key economic process.

Biomagnification

Although biological processes can produce many harmful substances, snake venom and plant toxins for instance, natural toxins do not build up in the environment as they are broken down by enzymes. By contrast, some synthetic chemicals are not broken down naturally as there are no enzymes to achieve this. Consequently, these synthetic compounds build up in the air, water, soil and living cells, and in some cases their concentration can increase in food webs to potentially harmful levels.

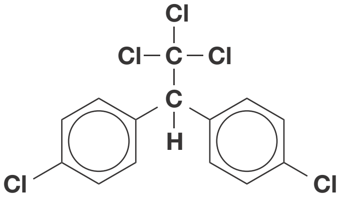
Biomagnification is a term that refers to this increase in concentration of a xenobiotic substance as it passes up a food chain. It occurs when a xenobiotic cannot be metabolized, and so is taken up directly when one organism feeds on another, causing the greatest effects in animals that feed at the top of a food chain. For biomagnification to occur the organic substance involved should:

- not be broken down in the environment
- not be broken down naturally
- be a lipid- or fat-soluble compound so that it is not readily excreted but is stored in fatty tissue instead.

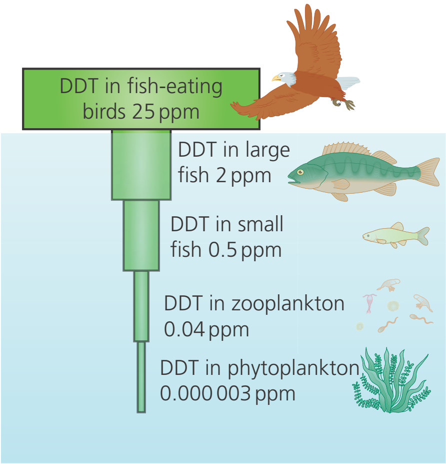
The terms bioaccumulation and biomagnification have become important ideas in discussing xenobiotics and it is important to distinguish the two. Bioaccumulation refers to the build-up of a substance within an organism; biomagnification refers to the build-up of a substance at different levels in a food chain.

Probably the most noted and well-studied example of this is the insecticide DDT, dichlorodiphenyltrichloroethane (Figure 23.129), the first synthetic pesticide.

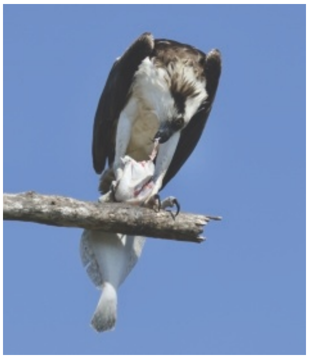
DDT is a complex aromatic molecule which was used to great effect starting during World War II to control the mosquitoes that are responsible for the spread of diseases such as malaria, dengue fever and typhus. The World Health Organization (WHO) suggests that 5 million lives were saved in the early years of its use. DDT is readily soluble in fat, and does not undergo metabolic breakdown. It therefore accumulates in tissues and passes unchanged through food chains. Some of the DDT used to spray areas of land found its way into rivers and lakes. It was taken up by microscopic plants, which were eaten by microscopic animals, which were in turn eaten by fish. The accumulated DDT progressed up the food chain and at each level, the concentration of DDT in the organism increased (Figure 23.130).



■ **Figure 23.129** The structure of DDT, dichlorodiphenyltrichloroethane


DDT in fish-eating birds 25 ppm
DDT in large fish 2 ppm
DDT in small fish 0.5 ppm
DDT in zooplankton 0.04 ppm
DDT in phytoplankton 0.000 003 ppm

■ **Figure 23.130** The biomagnification of DDT concentration as it passes up the food chain



■ **Figure 23.131** An osprey eating fish

In the 1960s it was noticed that birds of prey, such as ospreys (Figure 23.131), bald eagles and peregrine falcons, suffered a serious decline in their numbers. The cause was traced to a thinning of their eggshells. These birds could not build shells strong enough to last through incubation – the eggs broke under their parents' weight – because they could not process calcium properly.

Investigation showed that this thinning was due to the toxic effects of the high levels of DDT in their tissues. This and other negative environmental impacts eventually led to a ban of the use of indiscriminate spraying of DDT in many countries by the 1970s. Nonetheless, some continued use of DDT persists in vector control, particularly in countries where malaria is a serious health risk, and this remains controversial.

Nature of Science

Significant publications

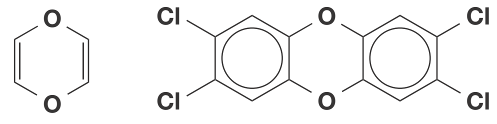
Awareness of the environmental impact of DDT and other pesticides, particularly on bird life, came to public attention through the publication in 1962 of the book *Silent Spring* by US biologist Rachel Carson. This book made a highly significant impact and is now widely credited as launching the global environmental movement.

An understanding of science is essential if the public and politicians are to make informed judgements about the advantages and disadvantages of using substances such as DDT, and this is precisely what the book *Silent Spring* did. It is essential that scientists provide the evidence in a way that is as complete as possible, but also objective so that people can make their own decisions. In the case of the re-introduction of DDT into Africa, the scientists provide the evidence but it is the politicians that ultimately make the decisions.

Another example of a publication that generated and informed public and political debate is Al Gore's *An Inconvenient Truth*, dealing with the issues around global warming and climate change. Significant in the development of environmental awareness in the past half-century have been the images of the Earth that have come into the shared domain through space exploration (Figure 23.132). This is in addition to the valuable hard data that space probes have provided on problems such as ozone depletion. Many people in recent generations have now seen the Earth from a totally different perspective and this has altered the psychology of our relationship with the planet.

When DDT was first introduced scientists did not consider possible negative effects on the environment but nowadays we are much more aware of such issues. When substances are made the environmental impact of the synthesis and use of the substance are often major considerations. There are two major approaches to the environmental issues and these are remedial action and prevention. The areas of hazard awareness, risk assessment and baseline evaluation have developed significantly over recent years, meaning that prevention is now becoming a more important factor.

■ **Figure 23.132** The Earth from the Moon – North and South America are visible



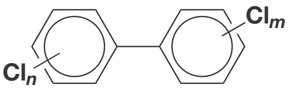
1,4-dioxin 2,3,7,8-tetrachlorodibenzodioxin

Figure 23.133 The structure of 1,4-dioxin and 2,3,4,8-tetrachlorobenzodioxin (an example of a PCCD). The general formulas of PCCDs and PCBs are given in Table 31 in the *IB Chemistry data booklet*

DDT, mercury and other heavy metals can become more concentrated as they move up a food chain. There continue to be worries about the level of mercury present in tuna consumed by humans, especially pregnant women. Sources of mercury include waste incineration, gold mining and coal combustion.

There are also health worries about biomagnification of other organic substances – for example, dioxins and dioxin-like substances such as polychlorinated dibenzodioxins (PCCDs) (Figure 23.133) and polychlorinated biphenyls (PCBs).

Dioxins are produced as by-products in the manufacture of some chlorinated organic compounds and the incineration of polymer waste materials. They are highly carcinogenic, particularly the chlorinated dioxins, and they can disrupt the endocrine system (hormone action) and lead to cellular and genetic damage.



polychlorinated biphenyl

Figure 23.134 The generalized formula of PCBs

PCBs are not strictly dioxins as they contain no oxygen. They contain one to ten chlorine atoms attached to a biphenyl molecule (Figure 23.134). They are highly stable, with high electrical resistance, and so were widely used in the 20th century as coolants, plasticizers, lubricants and insulating liquids. Their production in most countries was banned in the 1970s, but they can still persist in the environment.

Persistent organic chemicals (POPS)

It is the responsibility of scientists and industrial companies to consider the ways in which their research and findings impact on the environment, and to find ways of amelioration where damage has already been done. This involves baseline evaluation, risk assessment and long-term data collection. The issues presented raise ethical issues which cross national boundaries and demand international collaboration by scientists from different disciplines.

The Stockholm Convention on Persistent Organic Pollutants (POPs) in 2001 identified 12 chemicals, 'the dirty dozen', whose use should be banned. Although legislation to this effect now exists in many countries, many of these chemicals are particularly stable and they continue to persist in the environment. There is concern that climate change causing melting of the polar ice is re-mobilizing some of these banned chemicals into the Arctic atmosphere.

Bisphenol A (BPA) is an organic compound widely used in making polymers such as the polycarbonate plastics used in food packaging, reusable water bottles and water pipes. BPA has become controversial because it may mimic hormones, especially oestrogens, and so give rise to a range of health problems, including a lowering of the sperm count in men. Studies have shown that the risk of the chemical leaching from the plastic is increased when it is heated, and so particular concern has been expressed about its use in babies' bottles that are routinely heated during sterilization. While debate continues about safe levels, many industries have withdrawn these products and several governments have legislation pending to limit their use.

30 The information in the following table represents the energy flow in a hypothetical spring which was sprayed with DDT. The concentration of DDT found in the organisms at each trophic level is also given.

Trophic (feeding) level	Productivity/ $kcal/m^2/yr$	Ecological efficiency	DDT present/ ppm	Increase in DDT
Producers (plants)	9000	–	0.04	–
Primary consumers	1500		0.23	
Secondary consumers	120		2.07	
Tertiary consumers	12		13.8	
Percentage of energy from producers to tertiary consumers: ___% Concentration of DDT in the producer as compared to the water: ___ times more				

a Calculate the percentage ecological efficiency of energy transfer from:
  i producers to primary consumers
  ii primary consumers to secondary consumers
  iii secondary consumers to tertiary consumers
b Calculate the percentage of the energy from the producers that is transferred to the tertiary consumers.
c The concentration of DDT in the water was $1.0 \times 10^{-8}\,mg\,dm^{-3}$.
  i Deduce how many times more concentrated is the DDT in the producers as compared to the water.
  ii Calculate the increase of DDT between trophic levels, as it accumulates from producers to primary consumers, primary consumers to secondary consumers, and secondary consumers to tertiary consumers. (This will be expressed as how many times more concentrated, rather than a percentage.)

Host–guest chemistry

There is a growing focus of environmental concern on remedial action to ameliorate some of the situations that have arisen in ecosystems. Host–guest chemistry is one possible approach to amelioration. Host–guest chemistry is a form of supramolecular chemistry – it deals with systems bigger than a single molecule. The interaction between an enzyme and its substrate is an example of host–guest chemistry. The substrate (guest) does not form covalent bonds with the groups in the active site of the enzyme (the host) but is held in place by interactions such as London forces, ionic interactions and hydrogen bonding.

When host–guest chemistry is applied to environmental remediation the xenobiotic is the guest, and its chemical features determine the synthesis of the host, which is designed to bind to it. Many host molecules have a cage-like or tube-like structure which traps the guest molecule.

$$\underset{\text{host}}{H} + \underset{\text{guest}}{G} \rightleftharpoons \underset{\text{host–guest complex}}{H\text{–}G}$$

There are many types of host–guest complexes that have applications in removing toxic materials from the environment. One class that has been used for removal of heavy metal ions from solutions is the calixarenes. These have been used in the removal of radioactive ions such as caesium-137 from nuclear waste and to extract uranium ions from water. Calixarenes have the basic structure shown in Figure 23.135a and form a cup-like structure (Figure 23.135b) into which the metal ion is 'captured'. The cup-like shape, which is similar to the active site of an enzyme, makes it size-selective and ion–dipole interactions can form between the caesium ions and the oxygen atoms of the −OH groups.

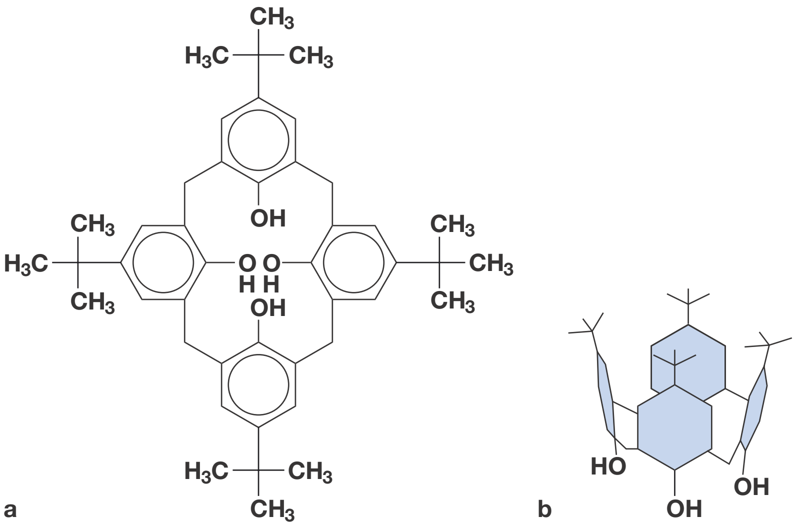


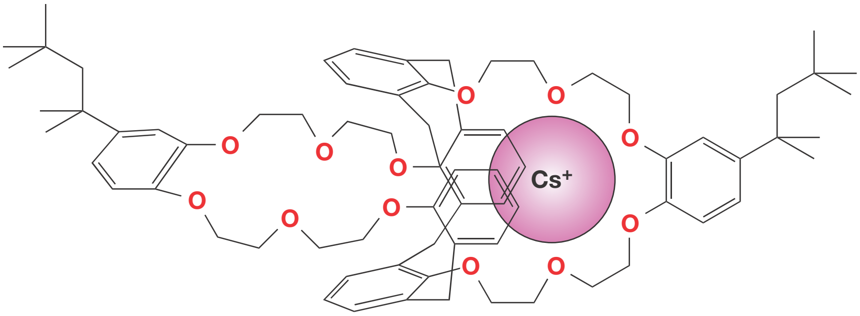
Figure 23.135 a A calixarene molecule. **b** The cup-like structure of a calixarene showing the rings in different colours to make the structure clearer

The word calixarene is derived from calix (or chalice), because this type of molecule resembles a vase, and arene, which refers to the aromatic building block. Calixarenes have hydrophobic cavities that can hold smaller molecules or ions and belong to the class of cavitands known in host–guest chemistry; others include crown ethers and zeolites. Calixarene nomenclature is straightforward and involves counting the number of repeating units in the ring and include it in the name. A calix[4]arene, such as the one in Figure 23.135, has four units in the ring, while a calix[6]arene has six units.

BOBCalix6 (Figure 23.136) is an example of a host molecule developed for a specific purpose: to bind the caesium-137 ions in nuclear waste.

Another environmental example of the use of host–guest chemistry is the removal of carcinogenic aromatic amines from polluted water. Drugs and drug metabolites are also targets of research for suitable host molecules for their removal.

Figure 23.136 Host molecule (BOBCalix6) shown with a positively charged caesium ion held inside one of its cavities



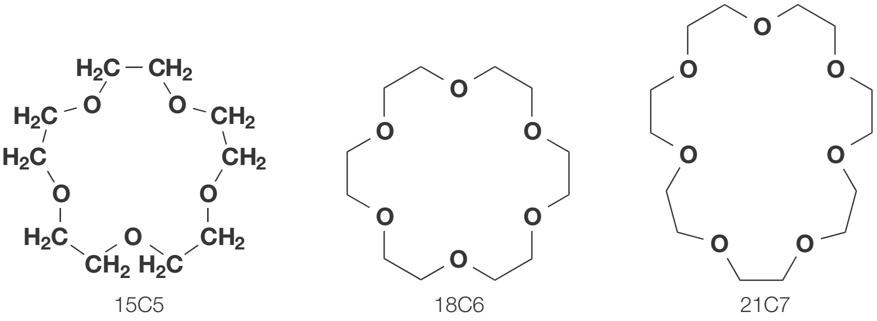
Nature of Science

Transferable ideas – coordination chemistry

It is important for the development of ideas in chemistry that information from one application feeds ideas into another area. The use of crown ethers in coordination chemistry provided useful information that fed into the above ideas for xenobiotic remediation. Metal cations are Lewis acids (electron pair acceptors) and so can interact with Lewis bases (electron pair donors). This is the basis of coordination chemistry (Chapter 13). A molecule or anion that reacts with a metal ion or atom is a ligand and forms a coordinate bond to a metal. A compound containing ligands coordinated to a metal is known as a complex.

The group 1 cations have a single positive charge and a large ionic radius and so have a low charge density. They are weak Lewis acids, but can form stable complexes with ligands known as crown ethers (in non-aqueous solvents). A crown ether (Figure 23.137) is an example of a macrocyclic ligand where all the donor atoms are contained within a ring.

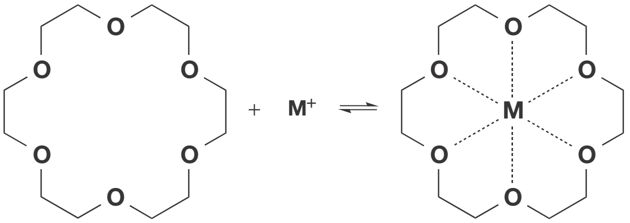
Figure 23.137 Various different crown ethers



31 Research, using the internet and school library, the structure and action of the antibiotic vancomycin.

The size of the ring is crucial in determining which metal ion forms the most stable complex. For example, 18[crown]-6 forms more stable complexes with K^+ than with either Na^+ or Cs^+. Na^+ is too small to coordinate with all of the oxygen atoms at the same time and Cs^+ is too large to fit into the cavity inside the ring. The larger crown ether [21]crown-7 forms its most stable complexes with Cs^+.

■ **Figure 23.138** The coordination of a metal ion by a crown ether



The complexes are known as supramolecules or host–guest complexes (Figure 23.138) and mimic the structures of enzyme–substrate complexes in the active sites of enzymes.

■ Biodegradable plastics

Substances that cannot be broken down by natural processes, which mostly involve microbial action, are said to be non-biodegradable. Compounds in this category often contain carbon–halogen bonds or stable aromatic structures, which enzymes are not able to break. This is why many plastics, such as PVC and polystyrene, and compounds like DDT persist in the environment indefinitely.

Alternatively, a compound is biodegradable if it can undergo bacterial degradation into end-products that are found in nature and therefore are not harmful to the environment.

This plastic bag is 100% degradable.
This bag will fully degrade to carbon dioxide, water and a small amount of inorganic residue after approximately 3 years in the domestic waste stream.
This compares to conventional plastic taking over 100 years to degrade.
SAFETY FIRST To avoid danger of suffocation, keep this bag away from babies and young children.
Please dispose of this bag carefully.
COMMENTS? QUESTIONS?
FREEPHONE 0800 0686 727
Manufactured using d2w supplied by Symphony Environmental Ltd

■ **Figure 23.139** A biodegradable plastic bag

Biodegradable and photodegradable plastics

Much research attention has been given to the development of biodegradable plastics. By modifying the structure of the polymer, it is possible to produce a plastic which can be decomposed when buried in a landfill. Other plastics are designed to undergo photochemical reactions, so the action of sunlight gradually breaks down the plastic. Biodegradable (Figure 23.139) and photodegradable plastics must be carefully sorted from plastic waste, because if they are incorporated into recycled plastic they will weaken the resulting items.

Two main types of biodegradable plastic bag currently exist.

Plant-based hydro-biodegradable plastic

This has a high starch content and is often obtained from corn. Genetic modification of grasses may help to produce similar plastics. The breakdown is initiated by hydrolysis and produces carbon dioxide and water. Swelling of starch grains can help to break up the plastic. At high temperatures it decomposes relatively quickly, but when buried in a landfill it may take much longer to decompose and may produce methane when decomposed anaerobically. In theory this type of starch-based bioplastic produced as biomass could be almost carbon neutral but there remain problems with using land that could serve more productive purposes, and the release of methane if the plastics are disposed of in landfill sites.

Polylactic acid (PLA) is a polyester derived from lactic acid (2-hydroxypropanoic acid). The lactic acid can be obtained from corn starch by microbial fermentation. PLA has found use as a packaging material and for making plastic cups and surgical thread. Again there are issues regarding land use to grow the corn for this purpose; the corn is a GM crop, and PLA will only degrade at a measurable rate in an industrial composter. (This highlights the difference between the terms biodegradable and compostable – the first simply means the product can be broken down by bacteria; the second means that the breakdown takes place at a rate comparable to that of naturally occurring polymers.)

Biopol is a polyester composed of hydroxybutyrate units with hydroxyvalerate units randomly distributed along the polymer chain. Biopol is produced by fermenting sugar with *Alcaligenes eutrophus*, a bacterium found in water and soil. Biopol biodegrades to water and carbon dioxide, in both aerobic and anaerobic conditions. Genetically modified plants which produce Biopol may make the plastic commercially viable.

Petroleum-based oxo-biodegradable plastic

This is derived from a by-product of the oil industry. Additives, often cobalt, are used to act as catalysts for the breakdown process, which can be programmed for different times depending on the use of the plastic. The plastic degrades into microfragments which are dispersed and eventually broken down by bacteria.

Making polylactic acid

This activity allows you to join about 10–30 lactic acid molecules together to begin to make a polymer. In industry several hundred molecules are joined together and so the properties of the polymer product are different from those of the polymer you will make.

Apparatus and chemicals

Test tube
Test-tube holders
Bunsen burner and heat-proof mat
Anti-bumping granules
Lactic acid (*Irritant*)
Hydrochloric acid 2 mol dm^{-3} (*Irritant*)
Petri dish or white tile
Eye protection

Safety: The boiling point of lactic acid is 122 °C. It will get very hot during the experiment. Be careful not to get it on your skin. If you do, put your skin under the cold tap immediately.

Method

1 Fill a test tube ⅕ full with lactic acid.
2 Add five drops of hydrochloric acid and two anti-bumping granules.
3 Put the test-tube holders around the top of the test tube and begin to heat the tube.
4 Be careful not to point the open end of the test tube at anyone in the room.
5 Keep the mixture gently boiling and stir or gently shake the tube occasionally to mix the contents.
6 After about 10 or 15 minutes the mixture will begin to go a yellowish colour. Leave it for another minute or two and then quickly pour the contents of the tube on to a petri dish or a white tile.
7 Leave the mixture to cool.

32 The Great Pacific garbage patch, also described as the Pacific trash vortex, is a gyre of marine debris particles in the central North Pacific.

Use the internet, magazines and journals, and the library to research the impact of pollution by plastic waste on the Earth's oceans. Consider the issues raised below and others that come to your attention:

- whether the above area of the Pacific is the only such accumulation of waste in the oceans, and how do they form
- the impact on this pollution on the natural environment and marine life
- the longevity of this form of pollution in the oceans – consider the role of photodegradation, and the possible leaching out of harmful organic chemicals such as bisphenol A and PCBs.

Enzymes and bioremediation

Although crude oil is a natural product, it can be present in sufficient concentrations to be considered xenobiotic. Such a situation arises, for example, when oil spills occur on land or at sea, depositing millions of litres into the surrounding environment (Figure 23.140). Organisms such as sea birds and marine mammals are then at risk from physical damage and from the toxic effects of the crude oil spill. Ways to ameliorate the impact of oil spills include the use of microorganisms which are able to break down the oil by using it as a food source and oxidize it in respiration. This is known as **bioremediation**.

Figure 23.140 Oil from the Deepwater Horizon oil spill

As crude oil contains a large number of different compounds, many different chemical reactions are involved in its breakdown. Microorganisms have evolved different enzymes that are specific for the degradation of different hydrocarbons in the oil so breakdown of the oil takes the combined action of a community of bacteria and fungi. Most of these are found naturally in the environment.

Over time, microbes are a dependable means of breaking down oil in the environment, although some of the larger and more highly branched molecules seem resistant to break down by enzymes. Environmental conditions such as temperature, supply of nutrients and availability of oxygen all influence the efficiency of the process, and in many cases the process may be too slow to prevent ecological damage. Research is ongoing to find ways to enhance the enzymatic processes involved. It has been suggested that the naturally occurring bacteria should be augmented with extra bacteria capable of breaking down the oil, or a mixture of enzymes and added nutrients to aid the breakdown of the oil spill. Oil spill Eater II (OSE II) is a commercial product that is able to convert crude oil into feedstock for naturally occurring bacteria and has been used in this type of situation.

Immobilized enzymes (enzymes attached to a solid support – see Section 23.7) have been used in the clean-up of industrial waste such as effluents from paper mills, textile industries and the leather industry. They can break down specific chemicals, such as pesticides or cyanide, to prevent their release into the environment.

Biological detergents

Another type of common environmental pollutant is industrial and household detergents containing amphiphilic molecules that facilitate the cleansing of fabrics and surfaces. Many detergents are branched-chain alkylbenzenesulfonates (ABSs) and as such have a very limited degree of biodegradability. They accumulate in sewerage treatment plants, producing persistent foam and altering the bacterial composition of recycled water. Many developed countries have phased out branched-chain detergents in favour of biodegradable linear alkylbenzenesulfonates (LASs).

'Biological' washing powders are those that contain enzymes. The first biological washing powders to be produced contained a protease called subtilisin. (Proteases are enzymes that break down proteins.) Subtilisin removed protein stains such as blood and egg, was stable at the alkaline pH of the detergent and retained its activity at temperatures up to 60 °C. Subtilisin has since been adapted to be effective at different, and lower, temperatures. Modern biological washing powders are also likely to contain an amylase to digest starch-based stains and a lipase to digest fat-based stains. The most recent powders also contain a cellulase (an enzyme that breaks down cellulose) to 'condition' the fabric. The cellulase enzyme removes stray ends of fibres produced by wear on the fabric. Such biological detergents are now available in powder, tablet or liquid form and for use in dishwashers too.

Biological detergents allow for use at lower temperatures than non-biological ones and so save energy. The only known side effect of biological detergents in humans is the possibility of an adverse allergic response in certain individuals with increased skin sensitivity.

Environmental sustainability

Whether carrying out an experiment in a laboratory, planning the protocols around the handling of any chemical material or setting up an industrial plant, the emphasis has shifted to the need to assess the hazards involved and the risk that such hazards may come into play. The need is for evaluation of the risks involved so as to ensure minimal harm, whether to personnel or the environment.

Concerns over health and environmental impact have put the chemical industry under a lot of pressure for greater accountability. This has become formalized to some extent with the establishment of various protocols to guide the baseline testing (Figure 23.141) and setting up of a project, and to assess and guide its sustainability. The following is just one

■ **Figure 23.141** Biomonitoring the health of the fish population as part of a baseline study prior to a mining project in northern British Columbia

example of such a protocol, produced in this case to guide the implementation of hydroelectric power projects but illustrating the complex range of issues that must guide such assessments.

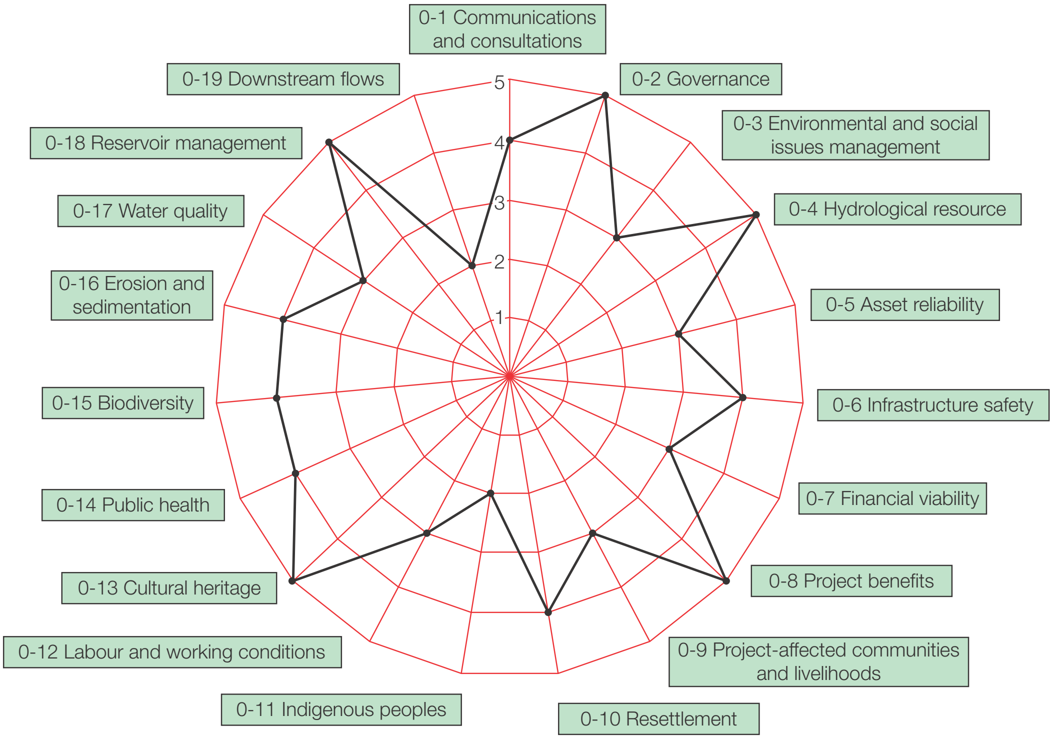
The Hydropower Sustainability Assessment Protocol is an international enhanced sustainability assessment tool used to measure and guide performance in the hydropower sector. The protocol assesses the four main stages of hydropower development:

- early stage
- preparation
- implementation
- operation.

Assessments rely on objective evidence to create a sustainability profile against some 20 topics depending on the relevant stage, covering all aspects of sustainability (Figure 23.142).
In summary the protocol is:

- a method for the assessment of individual projects against globally applicable criteria
- a series of assessment tools applicable to all stages of hydropower development in all global contexts
- an evidence-based objective assessment of a project's performance, prepared by an accredited assessor
- a project which involves leading NGOs (e.g. WWF, the Nature Conservancy, Transparency International)
- developed and governed by a multi-stakeholder, consensus-based structure.

■ **Figure 23.142** Profiling the sustainability of a hydropower development using the Hydropower Sustainability Assessment Protocol



Sustainable/green chemistry

Green chemistry is chemists' response to the challenges of sustainable development and to the awareness of the impacts that human activities have on the environment, and of the implications of these impacts on humanity.

The scope of green chemistry is embedded in the concept of chemistry as the science of substances. The principles of green chemistry focus on substances, on the characteristics that are relevant for human health and safety (non-hazardous, non-toxic) and on the way in which they are produced. Green chemistry aims to protect human health and the environment by selecting the types of substances preferably to be produced, the best (safer and more environmentally compatible) ways of producing them and the best ways of using them.

Its emergence in the USA in 1991 was prompted by the awareness of the damages of environmental pollution. Green chemistry aims to prevent such damages at the source, by selecting less hazardous substances and designing less hazardous ways of producing them. In the USA new regulations restricted the use of chemicals and increased testing of chemical substances to determine their hazard provided powerful incentives for industry to find replacements, substitutes or alternatives. The toxicity testing required by many of these statutes (laws) generated new knowledge and a new awareness about the type and degree of hazard associated with numerous chemicals. As the collective knowledge grew in scientific and industrial circles, there was a corresponding growth in the public's demand for more information about chemicals that are present in their communities.

The spreading of green chemistry research, and also of the basic information about what it is, has followed different patterns in different countries and resulted in developments in education and legislation (passing of laws) in different countries.

By focusing on substances, their production and their usages, it relates to contextual realities. Therefore, some features of its spreading and development are different for different contexts, although the principles and the chemical background constitute a common basis for all contexts.

Twelve principles have been developed which cover concepts that seek to reduce the footprint of chemical manufacturing processes while improving product and environmental safety.

1. *Prevention* – it is better to prevent waste than to treat or clean up waste after it is formed.
2. *Atom economy* – synthetic methods should be designed to maximize the incorporation of all materials used in the process into the final product.
3. *Less hazardous chemical syntheses* – wherever practicable, synthetic methodologies should be designed to use and generate substances that possess little or no toxicity to human health and the environment.
4. *Designing safer chemicals* – chemical products should be designed to preserve efficacy of function while reducing toxicity.
5. *Safer solvents and auxiliaries* – the use of auxiliary substances (solvents, separation agents, etc.) should be made unnecessary whenever possible and innocuous when used.
6. *Design for energy efficiency* – energy requirements should be recognized for their environmental and economic impacts and should be minimized. Synthetic methods should be conducted at ambient temperature and pressure.
7. *Use of renewable feedstocks* – a raw material or feedstock should be renewable rather than depleting whenever technically and economically practicable.
8. *Reduce derivatives* – unnecessary privatization (blocking group, protection/deprotection, temporary modification of physical/chemical processes) should be avoided whenever possible.
9. *Catalysis* – catalytic reagents (as selective as possible) are superior to stoichiometric reagents.
10. *Design for degradation* – chemical products should be designed so that at the end of their function they do not persist in the environment and do break down into innocuous degradation products.
11. *Real-time analysis for pollution prevention* – analytical methodologies need to be further developed to allow for real time, in-process monitoring and control before the formation of hazardous substances.
12. *Inherently safer chemistry for accident prevention* – substances and the form of a substance used in a chemical process should be chosen to minimize the potential for chemical accidents, including releases, explosions and fires.

Different aspects of green chemistry have been referred to throughout this book. One such key concept is atom economy (see Chapter 1). It is important to remember that atom economy is not the same as the yield of a reaction. Atom economy is a theoretical quantity based on a chemical equation and allows evaluation of how much waste (products) will be produced. The yield of a reaction is an experimental quantity worked out from how much of the desired product is actually made in a chemical reaction. When evaluating how green or environmentally friendly a particular process is, both atom economy and yield must be considered.

23.7 Proteins and enzymes (AHL) – *analyses of protein activity and concentration are key areas of biochemical research*

Enzyme kinetics

The general principles of reaction kinetics apply to enzyme-catalysed reactions, but with one very important feature not usually observed in non-enzymatic reactions (except in surface chemistry) – that of saturation with substrate (Figure 23.143). (The graph takes the form of a right rectangular hyperbola which is asymptotic to the maximum enzyme rate, V_{max}.)

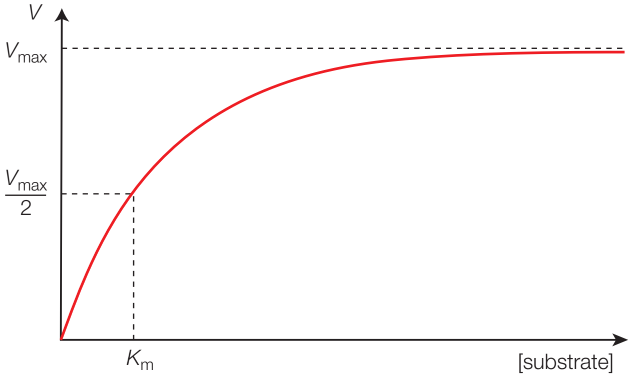
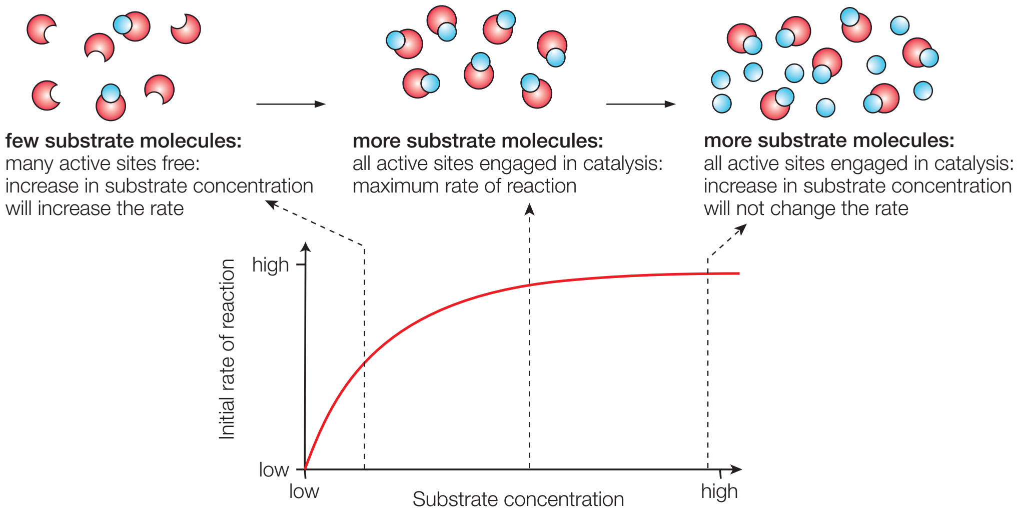


Figure 23.143 The effect of substrate concentration on the rate of an enzyme-catalysed reaction

At low substrate concentration, the enzyme activity, V (reaction rate), is nearly proportional to the substrate concentration, and the reaction is, therefore, approximately first order (Chapter 16) with respect to the substrate. As the substrate concentration is increased, the activity (reaction rate) increases less and is no longer nearly proportional to the substrate concentration: the reaction is now mixed order. With a further increase in the substrate concentration, the activity (reaction rate) tends to become independent of substrate concentration and approaches a constant rate, V_{max}. In this region the reaction is essentially zero order with respect to the substrate and the enzyme is said to be saturated with its substrate. All enzymes show this saturation (if the substrate is soluble enough) but there is variation in the substrate concentration required to produce it.

This saturation behaviour (Figure 23.144) suggests that the enzyme and substrate react reversibly to form a complex as an essential step of the enzyme-catalysed reaction, and also suggests that enzymes possess active sites where the substrate binds and chemical reaction occurs. Michaelis and Menten were the first researchers to develop a general theory of enzyme-catalysed reactions and kinetics.

Figure 23.144 Interpreting the change in initial reaction rate with substrate concentration of an enzyme-catalysed reaction


few substrate molecules:
many active sites free:
increase in substrate concentration
will increase the rate
more substrate molecules:
all active sites engaged in catalysis:
maximum rate of reaction
more substrate molecules:
all active sites engaged in catalysis:
increase in substrate concentration
will not change the rate
Initial rate of reaction
high
low
low
Substrate concentration
high

The Michaelis and Menten theory assumes that the enzyme, E, first binds with the substrate, S, to form an enzyme–substrate complex. This then breaks down to form the free enzyme and the product P. The first reaction is reversible and the enzyme–substrate concentration is assumed to be constant during the reaction.

$$E + S \leftrightharpoons ES$$

$$ES \leftrightharpoons E + P$$

For reactions of this type, at a very low substrate concentration [S], the activity (rate of reaction) V increases almost linearly as a function of [S]. As [S] increases further, V increases less rapidly. Eventually, V reaches a limiting value called V_{max} at saturating [S]. V_{max} is the maximum activity (rate of reaction) at 'infinite' substrate concentration. The [S] at which V equals $V_{max}/2$ is called the **Michaelis constant**, K_m (units of mol dm^{-3}) (shown in Figure 23.143).

An approximate value of the Michaelis constant for any given enzyme can be easily determined from a series of simple experiments in which the enzyme's activity (rate) is measured at different initial concentrations of the substrate with a fixed concentration of enzyme, and values plotted.

If the value of the Michaelis constant, K_m, is low, then a low concentration of substrate is sufficient to reach the maximum activity, V_{max}. If the value of the Michaelis constant, K_m, is high, then a high concentration of substrate is sufficient to reach the maximum activity, V_{max}. The Michaelis constant, K_m, is therefore a measure of the affinity of an enzyme for its substrate.

The Michaelis constant is not a fixed value but may vary with the structure of the substrate, with pH and with temperature. Inside cells, enzymes are not necessarily saturated with their substrates. The maximum activity (rate of reaction), V_{max}, also varies widely from one enzyme to another for a given enzyme concentration. V_{max} also varies with the structure of the substrate, with pH and with temperature.

The catalytic properties and specificity of an enzyme are determined by the functional groups in a small region of the protein surface called the active site. The active site is always found in a cleft or crevice in the enzyme structure and has two distinct functions:

- binding of the substrate
- catalysis.

Enzymes exhibit remarkable specificity because of the precise fit between their binding site and the substrate. This model of enzyme action is known as the **lock and key model**. This is the (lock/enzyme) and (key/substrate) analogy. The binding between substrate and enzyme can involve ionic bonding, hydrogen bonds and the range of van der Waals' interactions. We looked at this model, and its refinement in the **induced fit model**, in Section 23.2.

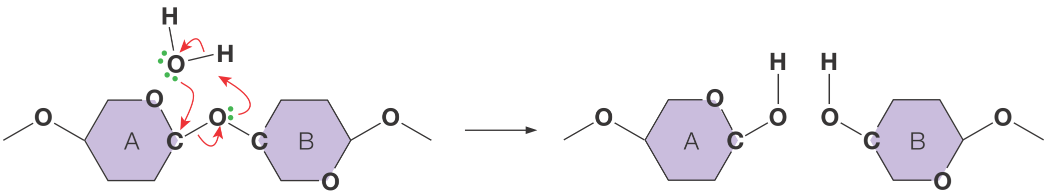
Enzymes bind the substrates so that the atoms participating in the bond to be made or broken are orientated properly with respect to catalytic groups in the enzyme's active site. The active site is stereospecific so that even enantiomers (optical isomers) of a substrate will not fit.

Case studies of enzyme action

Lysozyme

Lysozyme is an enzyme found in nasal mucus and tears. It was first named and studied by Alexander Fleming, discoverer of penicillin. Lysozyme is an antibacterial agent which catalyses the hydrolysis of specific polysaccharides. (Figure 23.145). The water molecule acts as a nucleophile. One of its lone pairs attacks the carbon atom of sugar A. The carbon–oxygen bond joining the two sugar residues is cleaved. The carbon–oxygen bridge is replaced by two hydroxyl groups.

Figure 23.145 Hydrolysis of a polysaccharide



Lysozyme contains 129 amino acids and four disulfide bridges. It is a globular protein with only small lengths of α-helix (Figure 23.146).

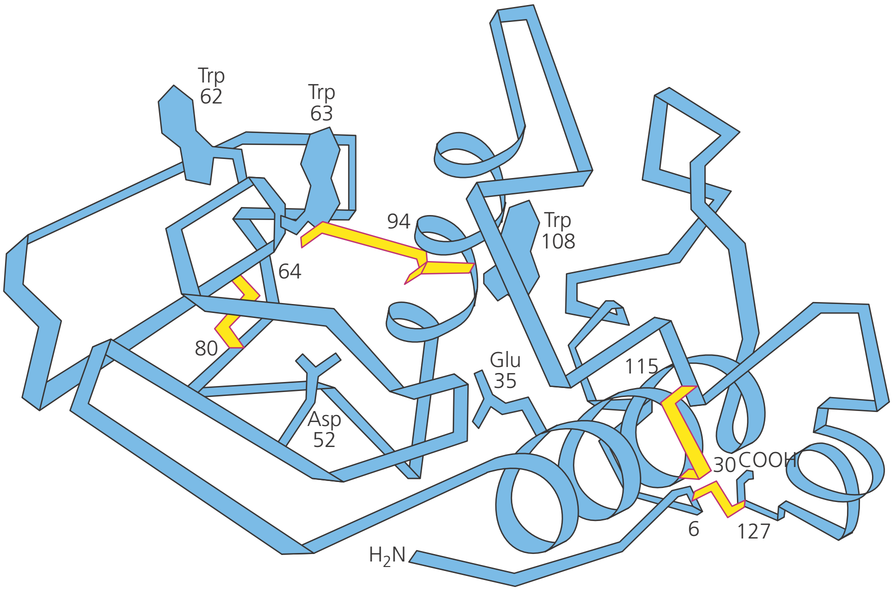


Figure 23.146 Computer model of the tertiary structure of lysozyme (amino acid residues Glu35 and Asp52 have crucial roles); disulfide bridges are shown in yellow

Enzymes bind their substrates in a cleft-shaped active site. The active site of lysozyme can hold six sugar (monosaccharide) residues. Five of the six sugars are bonded without any strain, but one of the sugars is stretched and bent, imposing stress on one of the glycosidic linkages (Figure 23.147).

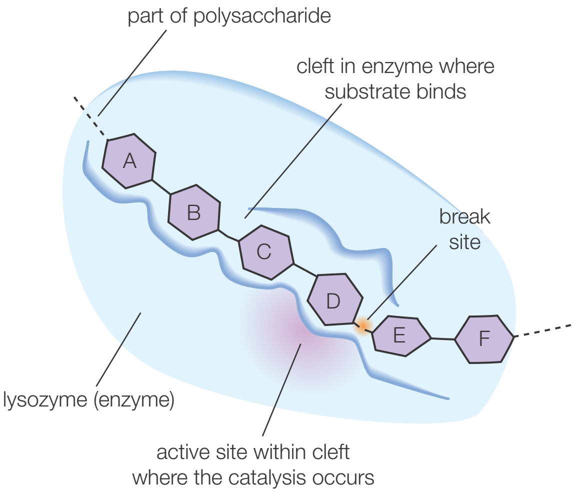
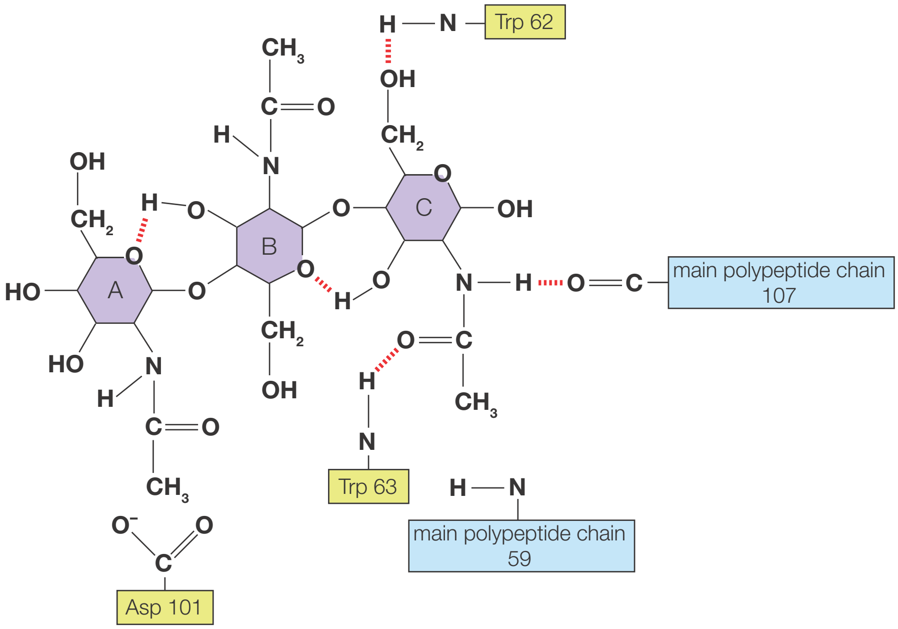

part of polysaccharide
cleft in enzyme where substrate binds
A
B
C
D
E
F
break site
lysozyme (enzyme)
active site within cleft where the catalysis occurs

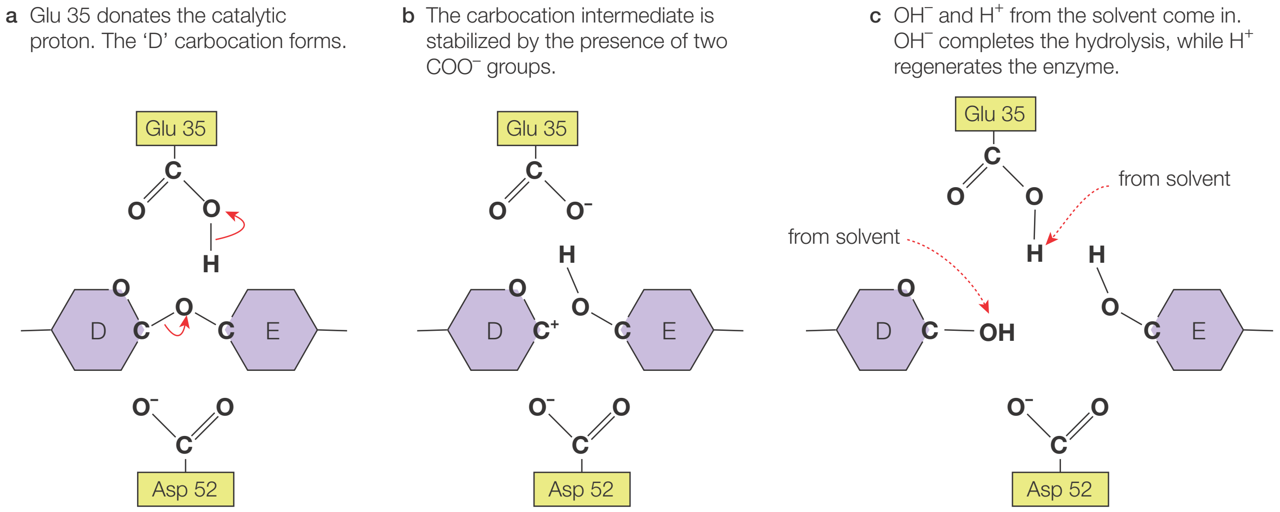
Figure 23.147 A simplified picture of the enzyme–substrate complex of lysozyme. Sugar residues A, B, C, E and F fit into the active site without distortion or strain while D, where the hydrolysis occurs, is strained and distorted

Part of the enzyme's active site is shown as a structural formula in Figure 23.148. The diagram shows three sugar residues A, B and C, and some key amino acid functional groups.



■ **Figure 23.148** Part of the lysozyme–polysaccharide complex

The enzyme-catalysed reaction of polysaccharides by lysozyme is similar to the acid hydrolysis of starch. The glutamic acid (glutamate) 35 residue donates a catalytic proton (H^+). Glutamine is an acidic amino acid with a pK_a value of 6, and hence is able to donate hydrogen ions at pH 7. The resulting carbocation intermediate is electrostatically stabilized by the presence of two charged carboxylate groups: glutamine 35 and aspartate 52. Monosaccharide D is in a strained conformation and the bridge carbon atom in ring D has been forced from a tetrahedral configuration into a planar configuration. Hence, it has the configuration it will adopt in the intermediate. The reaction is completed in a third step in which hydroxide ions (OH^-) and protons (H^+) from the dissociation of water complete the hydrolysis and 'regenerate' the catalytic hydrogen on glutamine 35. The cleaved polysaccharide molecule diffuses out of the active site (Figure 23.149).


a Glu 35 donates the catalytic proton. The 'D' carbocation forms.

b The carbocation intermediate is stabilized by the presence of two COO^- groups.

c OH^- and H^+ from the solvent come in. OH^- completes the hydrolysis, while H^+ regenerates the enzyme.

■ **Figure 23.149** The enzyme-catalysed hydrolysis of a polysaccharide

Chymotrypsin

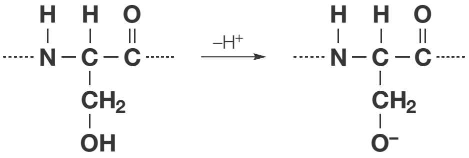
In humans, two of the proteases secreted by the pancreas into the small intestine neatly illustrate the subtleties of enzyme activity. Chymotrypsin and trypsin are both enzymes that break down proteins by hydrolysing peptide bonds, but the two enzymes break polypeptide chains in different places:

- Chymotrypsin hydrolyses the peptide bond immediately after a non-polar aromatic amino acid (phenylalanine, for example).
- Trypsin hydrolyses the peptide bond immediately after a basic residue (arginine, for example).

The substrate-binding regions of the two active sites must be different. That of chymotrypsin must consist of non-polar R groups that will be able to interact with aromatic amino acids, whereas that of trypsin will need to involve acidic R groups.

The catalytic action of conversion of substrates into products at an active site is carried out by specific R groups. It is important to note that when looking at the primary structure of an enzyme, these R groups are often found to be very far apart in the polypeptide chain. For example, the three R groups involved in the catalytic action of chymotrypsin are those of histidine (residue 57), aspartic acid (residue 102) and serine (residue 195). It is the folding involved in the tertiary structure of the enzyme that brings these groups into position to carry out their catalytic function. The three R groups involved in the active site of chymotrypsin act together by transferring protons (H^+ ions) between themselves. They act in such a way as to leave the R group of serine with an ionized $-O^-$ group.

This is something that would not normally happen in aqueous solution at the weakly alkaline pH of the small intestine. The negatively charged serine residue attacks the substrate as a nucleophile, bringing about the hydrolysis of a peptide bond. This particular example illustrates a feature of enzyme catalysis. An active site is a microenvironment in which the properties of chemical groups are altered so they can take part in reactions that would not occur in free aqueous solution.



serine-195

Enzyme inhibition

Inhibitors are chemicals that bind to enzymes and reduce their activity. Reversible inhibitors of enzymes are divided into two groups: competitive and non-competitive inhibitors (Figure 23.150). These can be recognized experimentally by their effects on the reaction kinetics of the enzyme.

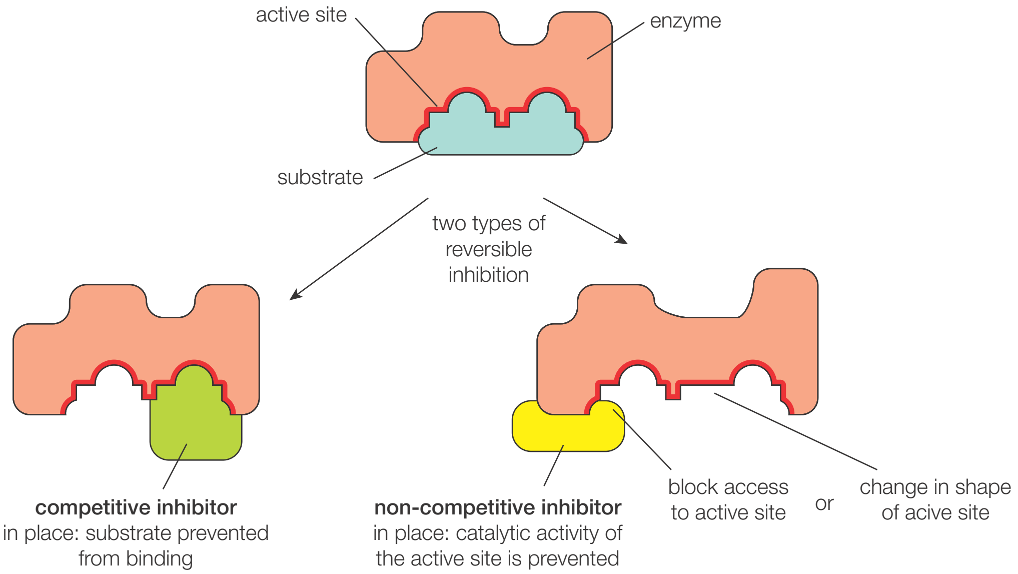
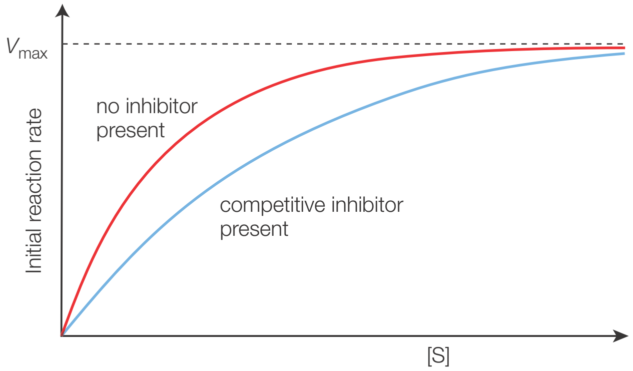

active site
enzyme
substrate
two types of reversible inhibition
competitive inhibitor
in place: substrate prevented from binding
non-competitive inhibitor
in place: catalytic activity of the active site is prevented
block access to active site
or
change in shape of acive site

Figure 23.150 The principles of competitive and non-competitive inhibitions



■ **Figure 23.151** The kinetic behaviour of a competitive inhibitor

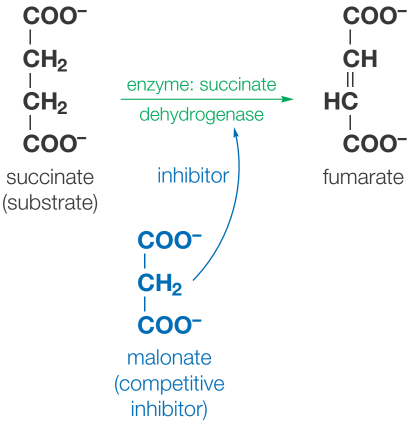
Competitive inhibitors

A **competitive inhibitor** will combine with the free enzyme in such a way that it competes with the substrate for binding at the active site. The inhibitor resembles the substrate molecules sufficiently well to form some of the proper interactions of the binding site but is not sufficiently similar to take part in the reaction and be released. The percentage of competitive inhibition at fixed inhibitor concentration can be decreased by increasing the substrate concentration.

At high concentrations of the substrate it is possible to reach V_{max} even in the presence of the inhibitor; however, the effect is to increase the value of K_m (Figure 23.151).

The extent of competitive inhibition will, therefore, depend on:

- the concentration of an inhibitor
- the concentration of the substrate
- the relative affinity of the active site for the inhibitor and substrate.



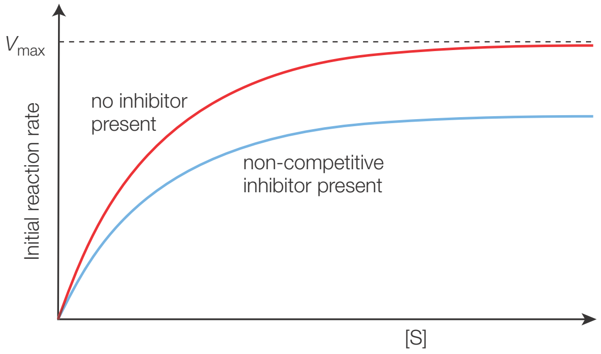
■ **Figure 23.152** The role of malonate as a competitive inhibitor of the enzyme succinate dehydrogenase

An example of competitive inhibition is the inhibition of the enzyme succinate dehydrogenase by molecules such as malonic acid which structurally resemble the substrate, succinic acid (Figure 23.152).

Non-competitive inhibitors

A **non-competitive inhibitor** will bind to a site on the enzyme other than the active site. The term **allosteric site** refers to the binding site of the non-competitive inhibitor on the enzyme. This often deforms the enzyme so that it does not form the enzyme–substrate complex readily. The non-competitive inhibitor may also combine directly with the enzyme–substrate complex.

The inhibitor in this case is not shaped like the substrate and combines with some region near or in the active site, preventing access by the substrate; in this case there is no competition between the substrate and the inhibitor. Non-competitive inhibition can be recognized from plots of V_0 (initial reaction rate) against [S] (substrate concentration) (Figure 23.153).



■ **Figure 23.153** The kinetic behaviour of a non-competitive inhibitor

V_{max} is decreased by the inhibitor and cannot be restored by increasing the substrate concentration.

Non-competitive inhibition depends on:

- the concentration of the inhibitor
- the affinity of enzyme for the inhibitor.

The most common type of non-competitive inhibition is given by reagents that can combine reversibly with some functional group of the enzyme that is essential for maintaining the catalytically active three-dimensional conformation (shape) of the enzyme molecule.

Table 23.19 summarizes some of the main differences between competitive and non-competitive inhibitors.

Table 23.19 A comparison of the main features of competitive and non-competitive inhibition

	Competitive inhibition	Non-competitive inhibition
Binding site on enzyme	Binds at active site	Binds at allosteric site
Effect on V_{max}	Not affected	Decreased
Effect on K_m	Increased	Not affected

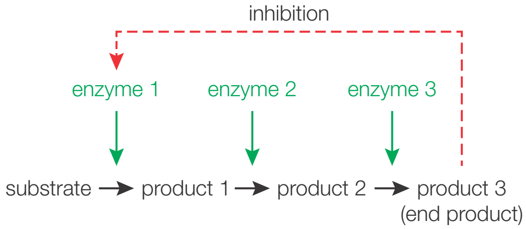


Figure 23.154 Feedback control of a metabolic pathway. The rate of reaction of the whole sequence is controlled by the concentration of the end-product

Feedback control by non-competitive inhibition

Feedback control by reversible non-competitive inhibition plays an important role in the delicate control of biochemical processes. This occurs when the end-product of a metabolic pathway inhibits an enzyme early in the pathway and so lowers the activity of the enzyme. This prevents too much of the end-product being formed.

When demand for the end-product increases, its concentration falls. The end-product molecules no longer inhibit the enzyme and activity is restored. This enables metabolic processes to respond quickly to immediate demands. Feedback control is summarized in Figure 23.154.

A simple example of feedback control can be seen in the use of a thermostat to control the temperature of a room. Here the end product is the heat produced by a radiator. When more heat is needed (as, for example, when cold air comes in through an open door), then the thermostat switches the radiator on. When the room is hot enough and no more heat is needed, the thermostat switches the radiator off.

One significant example of feedback control is that exerted by ATP (adenosine triphosphate) on its own synthesis in the sequence of reactions involved in glycolysis (the initial sequence of reactions involved in respiration; see Section 23.1). ATP is an inhibitor of the enzyme phosphofructokinase in that metabolic sequence, and so exerts control over its own synthesis. ATP is the short-term energy source for cells and so this example of feedback control is particularly important.

Nature of Science

Biotechnology and 'green' chemistry

Enzyme action is the basis of the brewing and cheese-making industries, possibly the oldest forms of biotechnology. Cheeses and fermented drinks are often associated with specific place names, according to the different flavours produced by enzymes in local microorganisms. Cheese making illustrates not just the use of enzymes but also the precipitation of proteins at their isoelectric point.

Beer is one of the world's oldest prepared beverages, possibly dating back to the early Neolithic or 9500 BCE, when cereal was first farmed, and is recorded in the written history of ancient Iraq and ancient Egypt. The earliest known chemical evidence of barley beer dates to around 3500–3100 BCE from the site of Godin Tepe in the Zagros Mountains of western Iran. Some of the earliest Sumerian writings contain references to beer; examples include a prayer to the goddess Ninkasi, known as 'The Hymn to Ninkasi', which served as both a prayer and a method of remembering, and passing on, the recipe for beer within the culture. Almost any substance containing sugar can naturally undergo alcoholic fermentation. It is likely that many cultures, on observing that a sweet liquid could be obtained from a source of starch, independently invented beer. Bread and beer increased prosperity to a level that allowed time for development of other technologies and so, as some have argued, contributed to the building of civilizations (Figure 23.155).

Figure 23.155 A 16th-century engraving of a brewer at work

Bread and beer making are both traditional examples of biotechnology that are reliant on the process of anaerobic respiration (fermentation) in yeast (Figure 23.156).

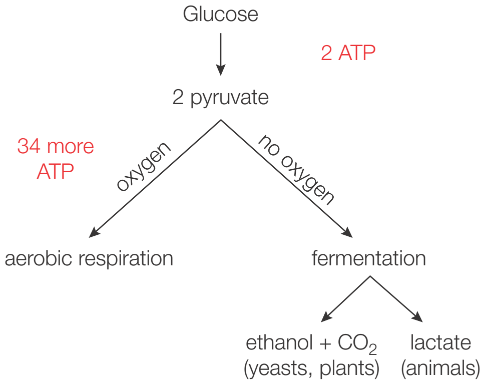


Figure 23.156 Anaerobic versus aerobic cell respiration. Anaerobic fermentation in yeast is the basis of bread and beer making

33 The addition of a phosphate group to glucose is the first step in the breakdown of glucose and is universally catalysed by an enzyme called hexokinase.

a **i** Sketch a graph to show how the rate of this reaction changes as a fixed amount of the enzyme is reacted with increasing concentrations of glucose (label the line **S**).

ii Explain the shape of your graph.

b Some compounds can inhibit the rate of an enzyme-catalysed reaction.

i On your graph from part **a** draw a second line to show the effect of a competitive inhibitor (label this line C).

ii Explain how a competitive inhibitor functions.

c On your graph from part **a**

i Draw a third line to show the effect of a non-competitive inhibitor (label this line **N**).

ii Explain how a non-competitive inhibitor functions.

Application	Enzymes employed	Uses
Biological washing powder	Primarily proteases, produced in an extracellular form from bacteria	Used for pre-soak conditions and direct liquid applications, helping with removal of protein stains from clothes
Dishwasher detergent	Amylase enzymes	Detergents for machine dishwashing to remove resistant starch residues
Baby foods	Trypsin (a protease)	To pre-digest baby foods
Fruit juices	Cellulase, pectinase (act on cell walls)	Clarify fruit juices, e.g. apple juice
Baking industry	Fungal α-amylase enzymes: normally inactivated at about 50 °C; destroyed during baking process	Catalyse breakdown of starch in the flour to glucose; yeast action on sugar produces carbon dioxide
Starch industry	Glucose isomerase	Converts glucose into fructose (high-fructose syrups derived from starchy materials have enhanced sweetening properties and lower calorific values)
Photographic industry (though largely replaced by digital now)	Protease	Dissolve gelatin off the scrap film, allowing recovery of silver present
Brewing industry	Enzymes from barley are released during the mashing stage of beer production	Degrade starch and proteins to produce simple sugars, amino acids and peptides that are used by yeast to enhance fermentation
Mouthwash	Glucose oxidase, lactoperoxidase, lactoferrin and lysozyme	Contains four natural antibacterial enzymes that kill bacteria found in oral infections and gingivitis without side effects

Table 23.20 Industrial uses of enzymes

Brewing and bread making represent the historical use of enzymes in biotechnology, but there are a range of important new technologies that involve enzyme biochemistry. Some of these are summarized in Table 23.20. The traditional uses of enzymes involve intact microorganisms (yeast and bacteria) and have the advantage that all the necessary reactants are present within the microorganisms. Perfect conditions for enzyme activity can be produced relatively easily.

A newer industry of enzyme technology, using enzymes that have been extracted and purified from organisms, now has many commercial applications. Enzymes are catalysts, and so remain unaltered by the reaction process; as a result, it should be possible to reuse them. However, it is often difficult to remove enzymes from the reaction mixture because they are soluble in water. This problem has been overcome by the development of immobilized enzymes.

Immobilized enzymes

For prolonged use in industrial processes, an enzyme can be immobilized by attachment to a solid surface. The overall advantages of this technology are that:

- the enzyme can be easily removed from a reaction mixture by centrifuging or filtration
- the enzyme can be packed into columns and used continuously for long periods
- the products can be easily removed from the reaction mixture so that inhibition of the reaction by the end-product can be avoided
- the stability of the enzyme to thermal denaturation can be increased, enabling the enzyme to be used for longer periods
- the optimum temperature of the enzyme may be increased, also allowing the reaction to be carried out at higher temperatures, so increasing the rate of reaction.

The initial costs of preparing the immobilized enzyme may be significantly greater than simply using the free enzyme. However, the ability to use the enzyme for longer under more productive conditions makes the use of the immobilized enzyme more economic overall.

The industrial importance of this idea has led to several different methods of immobilization being explored. The enzyme can be bound to a solid support either by weak intermolecular forces or by covalent bonding. The various methods of immobilization currently in use are summarized in Figure 23.157:

- adsorption on to an insoluble matrix
- covalent binding on to a solid support
- entrapment within a gel
- encapsulation behind a selectively permeable membrane.

Each of the methods shown in Figure 23.157 has been successfully used in a variety of applications. The key to success is that the interaction with the support must not interfere with the shape of the active site.

For example, the production of high-fructose syrup, which is used as a sweetener in the food industry, requires the conversion of glucose into fructose. Glucose, from the hydrolysis of starch, is converted into the much sweeter compound fructose using the enzyme glucose isomerase. This enzyme has proved easy to immobilize and can be used in a continuous reactor at 60–65 °C and pH 7 for over 1000 hours. The use of immobilized enzyme has dramatically reduced the cost of producing high-fructose syrup.

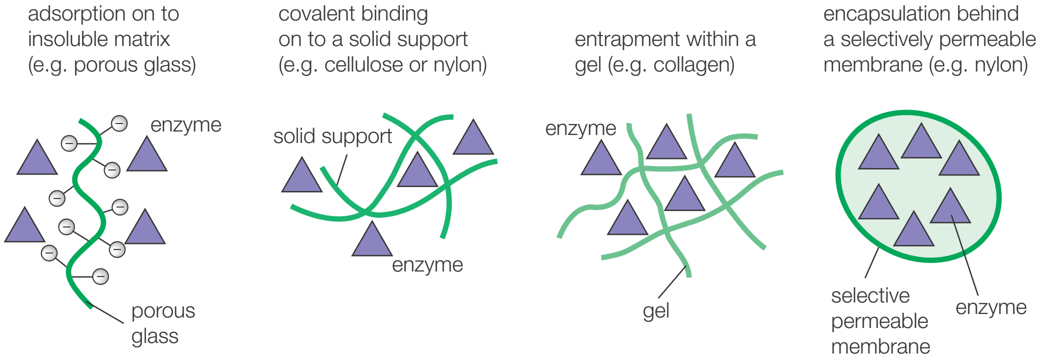


Figure 23.157 Different techniques for preparing immobilized enzymes for use in bioreactors

The benefits of using enzymes to bring about useful transformations under milder conditions links well with some of the key themes of sustainable, or 'green', chemistry (see Section 23.6 for an outline of the 12 principles of green chemistry). Table 23.21 summarizes the advantages of using enzymes to achieve chemical change using biological processing.

■ **Table 23.21** The advantages of the use of enzymes in industrial chemistry

Factor	Chemical/non-biological processing	Biological processing
Temperature	Very high or low temperatures often required to make the reaction occur – low temperatures are sometimes required to maintain stability of the catalyst or ensure solvents do not boil	Moderate temperatures usually employed – heating or refrigeration costs are lower
Solvent	Organic solvents often needed – may be both expensive and toxic	Water is the normal solvent for the enzymic reaction mixture or growth medium
Catalyst	Inorganic catalysts often used; often toxic, expensive and lacking in specificity; unwanted by-products may be produced	Enzymes are highly specific, therefore less wasteful of substrate and fewer by-products formed
Conditions	Extremely acidic or alkaline conditions often used – requires expensive corrosion-resistant vessels	Moderate pH usually used
Cost	Raw materials often expensive	Raw materials usually inexpensive
Length of reaction	Lengthy preparation time often required	Short fermentation or reaction times employed – higher production rate

Buffer action

A buffer solution is one that resists changes in pH when small amounts of acid or alkali are added to the solution. A buffer solution consists of two components – a conjugate acid–base pair such as ethanoic acid–ethanoate ion. The action of buffer solutions in maintaining a controlled pH is covered in Chapter 8, but consider the rationale behind the ethanoate buffer solution mentioned above.

Since ethanoic acid is only slightly dissociated and sodium ethanoate is completely dissociated, a mixture of the two contains a relatively low concentration of hydrogen ions, but a large proportion of ethanoic acid molecules and ethanoate ions:

$$CH_3COONa(aq) \rightarrow CH_3COO^-(aq) + Na^+(aq)$$
$$CH_3COOH(aq) \leftrightharpoons CH_3COO^-(aq) + H^+(aq)$$

If an acid is added to the buffer, the additional hydrogen ions will be removed by combination with the ethanoate ions to form undissociated acid molecules. The presence of sodium ethanoate ensures there is a large 'reservoir' of ethanoate ions to 'mop up' the additional hydrogen ions from an acid.

If an alkali is added, the hydroxide ions combine with the hydrogen ions to form water molecules:

$$H^+(aq) + OH^-(aq) \rightarrow H_2O(l)$$

The removal of hydrogen ions via neutralization results in the dissociation of ethanoic acid molecules to replenish the hydrogen ions removed. The presence of ethanoic acid ensures that there is a large 'reservoir' of undissociated ethanoic acid molecules that will dissociate following the addition of an alkali.

Other buffering systems are available such as that using the ammonia/ammonium ion conjugate pair. The use in experiments is determined by the pH that they are meant to maintain.

The use of buffer systems is crucial to much of the experimental work in biochemistry. We have seen that enzymes are sensitive to changes in pH and can be denatured by significant deviation from their pH optima. Analytical techniques such as electrophoresis and protein assay also involve the use of buffers.

The maintenance of the pH in the blood and tissues of the body is of great importance and amino acids and proteins contribute to the buffering that sustains these critical conditions.

Acid–base properties of 2-amino acids and proteins

In aqueous solution the amino and carboxylic acid functional groups both ionize, or dissociate. The carboxylic functional group releases hydrogen ions and hence acts as a Brønsted–Lowry acid (see Chapter 8):

$$-COOH(aq) \leftrightharpoons -COO^{-}(aq) + H^{+}(aq)$$

The amino functional group can accept hydrogen ions from solution and so acts as a Brønsted–Lowry base:

$$-NH_2(aq) + H^{+}(aq) \leftrightharpoons -NH_3^{+}(aq)$$

In neutral solution both the amino and carboxylic functional groups are ionized or dissociated. In an acidic solution (low pH), the amino acid accepts hydrogen ions and becomes positively charged. This reaction occurs because the lone pair of electrons on the nitrogen of the amino group can form a coordinate bond (Chapter 4) with an electron-deficient hydrogen ion. In an alkaline solution (high pH) the carboxylic acid functional group donates hydrogen ions to the hydroxide ions and forms a negatively charged carboxylate functional group.

Consequently, amino acids act to maintain the pH of an aqueous solution, because they remove excess hydrogen ions (H^+) or hydroxide ions (OH^-), forming water molecules. However, zwitterions of amino acids have pK_a values of approximately 2 and 10, and hence buffer effectively only at high and low values of pH.

This buffering capacity is retained when amino acids are polymerized to form proteins. Proteins are able to act as buffers because a number of amino acids contain either a basic amino group or an acidic carboxylic acid functional group in their variable side-chain (R group). Proteins play important roles as buffers (Figure 23.158) both inside and outside cells.

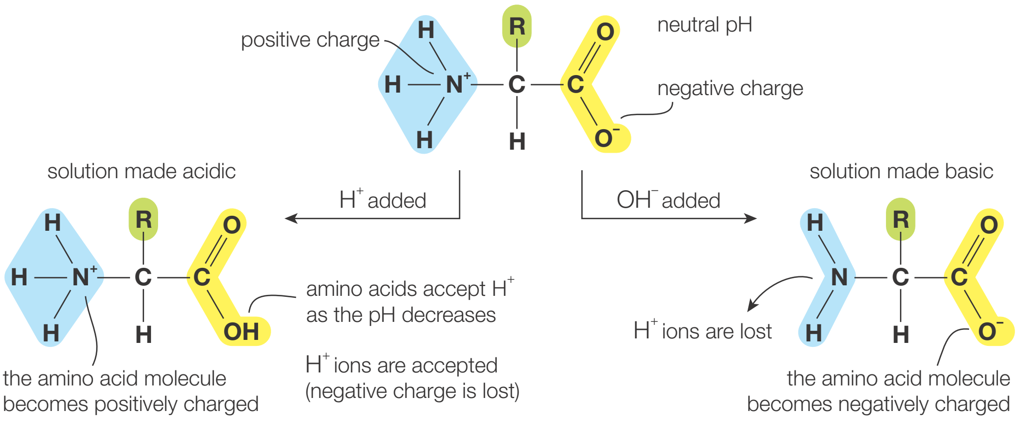

positive charge
neutral pH
negative charge
solution made acidic
H⁺ added
OH⁻ added
solution made basic
amino acids accept H⁺ as the pH decreases
H⁺ ions are lost
H⁺ ions are accepted (negative charge is lost)
the amino acid molecule becomes positively charged
the amino acid molecule becomes negatively charged
H
N⁺
C
R
O
O⁻
OH
N

Figure 23.158 The buffering action of amino acids in aqueous solution

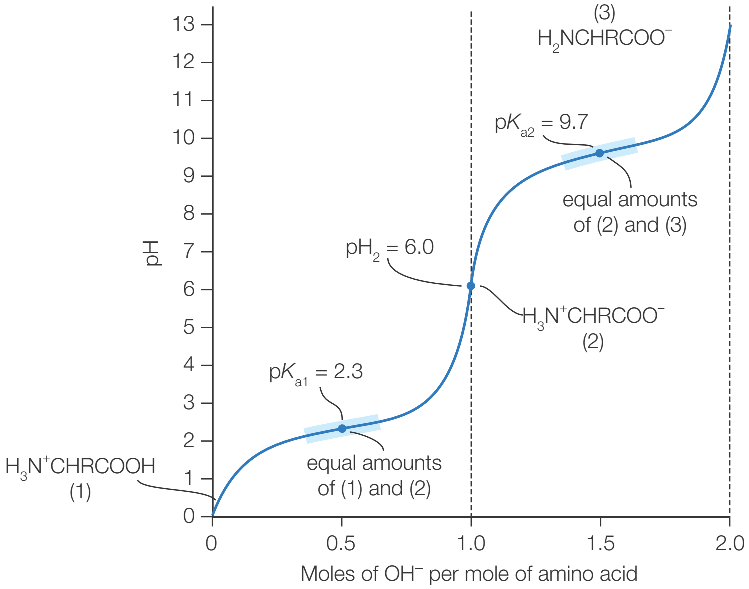
The buffering role of amino acids is important in helping to maintain a constant pH in cells, a crucial need for biological systems. Many of the protein components, especially enzymes, are extremely sensitive to changes in pH and could be made inactive by any significant fluctuation in pH. For example, human blood has a pH of 7.4, and an increase or a decrease of more than 0.5 pH units can be fatal. Clearly, effective buffering is important. There are several different buffer systems at work in the human body and these include those involving amino acids and proteins.

It is important to realize that an amino acid does not act as a buffer around its isoelectric point because there is only one species present. The titration curve for alanine (isoelectric point = 6.0) is shown in Figure 23.159, with the buffering regions being the almost horizontal regions of the curve (around pK_{a1} and pK_{a2}).

An amino acid such as alanine (where R is a methyl group) is dibasic when it exists in its fully protonated form. It can donate two protons (H^+) during its titration with a strong base:

$$H_3N^+CH(CH_3)COOH(aq) + OH^-(aq) \rightleftharpoons H_3N^+CH(CH_3)COO^-(aq) + H_2O(l)$$
$$H_3N^+CH(CH_3)COO^-(aq) + OH^-(aq) \rightleftharpoons H_2NCH(CH_3)COO^-(aq) + H_2O(l)$$

The first part of the curve corresponds to the first reaction and the second part of the curve corresponds to the second reaction. The flatter portions of the curves correspond to buffering regions, where the pH does not vary significantly with the concentration of base.



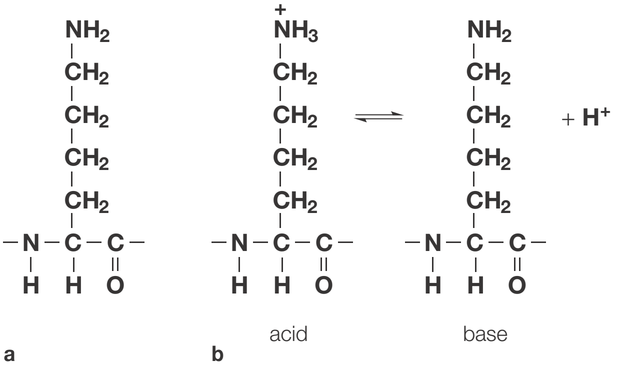
■ **Figure 23.159** Titration curve for alanine

At pH 6.0 there is a point of inflection between the two separate halves of the titration curve. There is no net or overall electrical charge on the molecule at this pH, and the amino acid will not move in an electric field. The pH is called the **isoelectric point (pI)**, and the amino acid has the least buffering effect at this point:

$$\text{isoelectric pH} = \tfrac{1}{2}(pK_{a1} + pK_{a2})$$

Consideration of the titration curve shows why amino acids can only act as acid–base buffers within a certain pH range, where both components of the conjugate acid–base pair are present in solution at sufficient concentrations. Maximum buffering capacity occurs where the concentrations of the two possible forms are equal – so at pK_{a1} and pK_{a2}. According to the Henderson–Hasselbalch equation (see Chapter 18 and the *IB Chemistry data booklet*), the ratio between the components of a conjugate pair increases or decreases by a factor of ten for a pH change of one unit. So an amino acid such as alanine or glycine can act as a buffer in the ranges pH = pK_{a1} ± 1 and pH = pK_{a2} ± 1. Outside these ranges the amino acid exists predominantly as a single ionic species and cannot act as a buffer.

The situation is further complicated by the presence of side-groups in the amino acid, some of which can also act as acids and bases. This is the reason that proteins can also act as buffers. Consider a protein that is rich in lysine (Figure 23.160a).



■ **Figure 23.160 a** The R group of lysine. **b** The acid–base equilibrium of the R group of lysine

At pH 10, some of the NH_2 side-groups will be protonated and others will not (Figure 23.160b). So there will be a mixture of acid and base present – and therefore a buffer.

Calculating the pH of a buffer solution

For a buffer solution made up of a mixture of HA (acid) and A^- (base), the pH of the buffer can be worked out using the Henderson–Hasselbalch equation:

$$pH = pK_a + \log_{10} [A^-(aq)]/[HA(aq)]$$

or

$$pH = pK_a + \log_{10} [\text{conjugate base}]/[\text{conjugate acid}]$$

The Henderson–Hasselbalch equation allows calculation of the pH of an amino acid solution with known acid–base composition or the concentration of conjugate acid and base in a solution of known pH.

For amino acids the pK_{a1} value characterizes the equilibrium between the cationic form of the amino acid (species 1 on the titration curve) and the zwitterion, so the equation becomes:

$$pH = pK_{a1} + \log_{10} [\text{zwitterion}]/[\text{cation}]$$

The pK_{a2} value characterizes the equilibrium between the zwitterion and the anionic form (species 3 on the titration curve), so the equation becomes:

$$pH = pK_{a2} + \log_{10} [\text{anion}]/[\text{zwitterion}]$$

Note that the same zwitterion is the conjugate base in the first acid–base equilibrium, but the conjugate acid in the second.

34 a The isoelectric point (pI value) of the amino acid serine (2-amino-hydroxypropanoic acid) is 5.7. Draw the main structural formula of serine, in the solid state and in an aqueous solution at pH values of 1, 14 and 5.7.

b Calculate the pH of an aqueous solution that contains 0.8 mol dm^{-3} zwitterionic and 0.2 mol dm^{-3} anionic forms of serine. (For serine $pK_{a1} = 2.2$ and $pK_{a2} = 9.1$.)

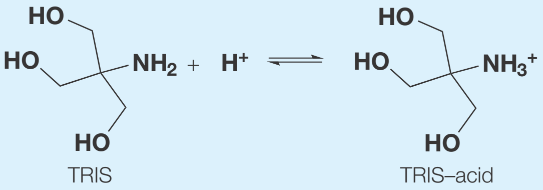
35 Identify the conjugate acid and conjugate base in a 0.500 mol dm^{-3} solution of glycine (pI = 6.0) at pH = 3.0. Calculate the concentration of both glycine species. (For glycine $pK_{a1} = 2.3$ and $pK_{a2} = 9.6$.)

36 The following examples illustrate the calculation of pH in some buffer systems used in working with biochemical systems:

a Calculate the pH of a solution containing 0.200 mol dm^{-3} ethanoic acid ($K_a = 1.74 \times 10^{-5}$ mol dm^{-3}) and 0.250 mol dm^{-3} sodium ethanoate.

b Phosphate buffers are routinely used in biochemical experiments, particularly for buffering around neutral pH. Calculate the pH of a buffer solution containing 0.055 mol dm^{-3} $H_2PO_4^-$ ($pK_a = 7.21$) and 0.045 mol dm^{-3} HPO_4^{2-}.

c TRIS is a buffer system used frequently in biochemistry. A buffer solution is prepared by adding hydrochloric acid to TRIS to form a mixture of TRIS and its protonated form (TRIS–acid). The equilibrium that exists in the buffer solution is:



Calculate the pH of a buffer solution containing 0.650 mol dm^{-3} TRIS–acid ($pK_a = 8.30$) and 0.750 mol dm^{-3} TRIS.

Note: When working out the pH of a buffer solution, you can check whether or not your answer is reasonable. If the solution contains a higher concentration of acid than base, the pH of the solution will be lower than the pK_a of the acid; if there is a higher concentration of base than acid, the pH will be higher than the pK_a.

Assay of proteins

A protein assay is a method of determining the concentration of a protein in solution. We introduced the idea behind this type of assay in Section 23.2 where we discussed a method using Biuret reagent. Protein assays usually involve ultraviolet–visible (UV–Vis) spectroscopy. There are two basic approaches:

- The absorption of radiation by the protein in the UV region of the spectrum (at a wavelength of 280 nm) can be measured
- A coloured dye is added that binds to the protein and the absorption in the visible region of the spectrum of the protein–dye complex can be measured.

In each case the concentration of the protein is determined by reference to a calibration curve which is constructed using known concentrations of a protein standard (often bovine serum albumin – BSA).

UV assay

Proteins absorb electromagnetic radiation in the UV region due to the presence of aromatic rings in the amino acid side-chains (R groups such as phenylalanine, tyrosine and tryptophan). The wavelength at which maximum absorbance occurs is 280 nm. The following steps are taken to construct a calibration curve and determine the concentration of an unknown protein solution:

- The UV–Vis spectrophotometer (Figure 23.161) is zeroed at 280 nm with a cuvette containing just the solvent. This is the blank sample. For UV spectroscopy the cuvette must be made of quartz or a plastic that allows the transmission of UV light.

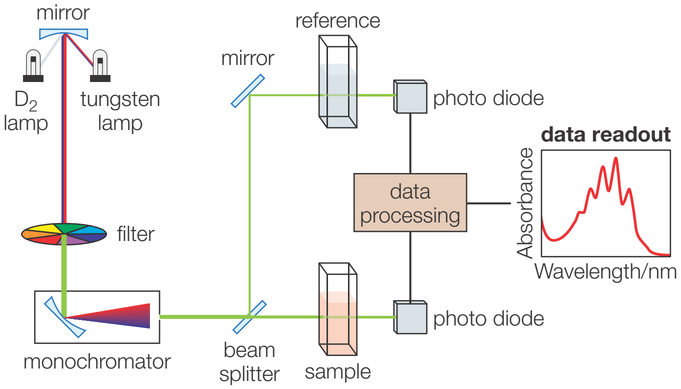


Figure 23.161 A schematic diagram of a UV–Vis spectrophotometer. The combination of the filter and monochromator ensures that only light of a particular wavelength is shone through the sample. A simple colorimeter just has the filter to select the light

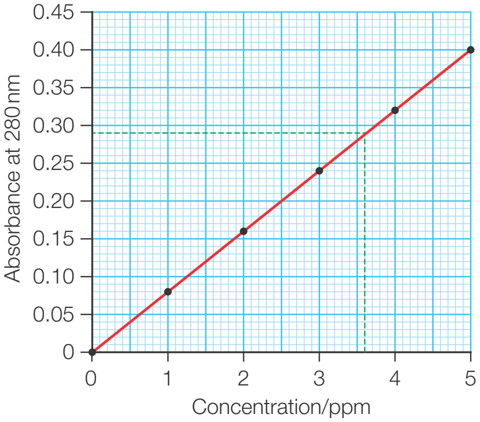


Figure 23.162 A calibration curve for the UV assay of proteins using absorbance at 280 nm (note that 1.00 mg dm^{-3} is the same concentration as 1 ppm)

- A series of protein solutions of known concentration are made up – for example, 1.00 mg dm^{-3}, 2.00 mg dm^{-3}, 3.00 mg dm^{-3} etc. This is usually done by serial dilution of a stock solution of accurately known concentration (note that 1.00 mg dm^{-3} is the same concentration as 1 ppm).
- The absorbance of each of these protein solutions is measured at 280 nm.
- A calibration curve of absorbance against concentration is plotted.
- The absorbance of the protein solution(s) of unknown concentration is then measured.
- The concentration of the unknown solution is read off the calibration curve.

If the absorbance of the unknown protein solution was 0.29, then its concentration, read from the calibration curve in Figure 23.162, is 3.6 ppm (or mg dm^{-3}).

Visible assay

There are various other methods for determining the concentration of protein using visible spectroscopy that rely on chemical reaction with the peptide bonds in the protein or dye binding to the amino acids in the protein chains. The basic method is the same as described above, with a calibration curve first being constructed using standard solutions of known concentration. For visible spectroscopy a glass or clear plastic cuvette can be used.

The visible assay of protein concentration using Biuret reagent has already been described in Section 23.2 (see Figure 23.30). This is based on a specific reaction between copper(II) ions (Cu^{2+}) in alkaline solution and the peptide bonds as found in proteins. A violet-coloured complex is formed. This test is very insensitive and will not detect proteins in solution with a content of less than 1 mg cm^{-3}. The Lowry method has been used quite extensively and is a combination of the Biuret method and the oxidation of tyrosine and tryptophan residues with Folin and Ciocalteu's reagent to give a blue–purple colour. This method is very sensitive but is easily affected by non-protein substances.

One such dye-binding method uses Coomassie Brilliant Blue G-250 dye (sometimes referred to as the Bradford assay). The bound and unbound forms of the dye have different colours and the concentration of the protein can be determined by measuring the absorbance at 595 nm, which is in the visible region of the spectrum.

The dye only binds to certain specific amino acids – the basic amino acids (arginine, histidine, lysine) and aromatic amino acid residues (tyrosine, tryptophan, phenylalanine). All proteins contain these amino acids but the total amounts of these amino acids vary between proteins and this can produce variation in the colour response with different proteins. It is important therefore to choose an appropriate standard. By testing with BSA and bovine plasma gamma globulin (IgG) it can be shown that the absorbance produced is dependent on the protein used – with BSA giving a higher absorbance than IgG.

It is worth noting that it is feasible to carry out this assay using a simple colorimeter which uses colour filters to produce light of a given wavelength. The absorption can be measured at 600 nm in this case.

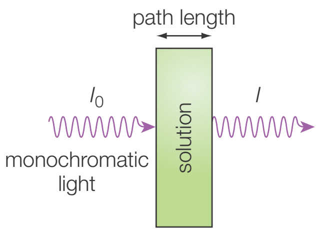


Figure 23.163 The Beer–Lambert law relates the absorbance of light by a solution to the path length and concentration

The Beer–Lambert law

This law relates the amount of light absorbed by a solution to its concentration and the path length (Figure 23.163).

The Beer–Lambert law is:

$$\log_{10}(I_0/I) = A = \varepsilon cl$$

where:

- I_0 is the intensity of the light before it passes through the sample
- I is the intensity of the light after it has passed through the sample
- A is the absorbance = $\log_{10}(I_0/I)$
- ε is the molar absorptivity or molar extinction coefficient (with units cm^{-1} mol^{-1} dm^{3}) – this is the absorbance of a 1.00 mol dm^{-3} solution in a 1.00 cm cell at the specified wavelength
- c is the concentration of the solution in mol dm^{-3}
- l is the path length – the thickness of the sample (usually in cm – a path length of 1 cm is used in most instruments).

The Beer–Lambert law is usually used in the form:

$$A = \varepsilon cl$$

It tells us that more radiation is absorbed by a more concentrated solution or if the radiation has to pass through a thicker sample. If the light encounters twice as many particles as it passes through the sample, then twice as much will be absorbed.

The concentration of protein in a sample can be determined from absorbance data and application of the Beer–Lambert law. First of all the value of the molar extinction coefficient,

ε, must be worked out – this can be done either by measuring the absorbance of a solution of known concentration or from the calibration curve. If the data in Figure 23.162 were obtained using a constant path length of 1 cm, we have:

$$A = \varepsilon c$$

Thus ε is the gradient of the graph of absorbance against concentration.

$$\varepsilon = 0.32/4.00 = 0.0800\,\text{cm}^{-1}\,\text{ppm}^{-1}$$

37 The absorbance observed for the protein solution in the experiment above was 0.29 using a 1 cm path length cuvette. Confirm using Beer–Lambert's law that the concentration of this solution was 3.6 ppm.

38 Using Beer–Lambert's law, calculate the concentration of each of the following protein solutions. All absorbances were measured at 280 nm in a cuvette of path length 1.0 cm.

a The molar absorptivity at 280 nm for a particular protein solution is 500 $\text{cm}^{-1}\,\text{mol}^{-1}\,\text{dm}^3$ and the absorbance = 0.31.

b The molar absorptivity at 280 nm for a particular protein solution is 63.5 $\text{cm}^{-1}\,\text{mol}^{-1}\,\text{dm}^3$ and the absorbance = 0.23.
(You will note that absorbance has no units.)

Nature of Science

The limitations and development of methodology

Protein analyses are used routinely in analytical chemistry, but it is important for scientists to have an understanding of the sensitivity, accuracy and precision of their data. They must be aware of possible systematic errors in their procedures which, although they might give reproducible results, are not accurate. Different protocols have been developed from the very simple methods described here to give reliable and accurate readings for protein concentrations. In introducing the methods we have noted the importance of the following:

- the possible interference of other components of the complex mixtures often encountered in biological samples
- the sensitivity of the assay used and the nature of the interaction upon which it is based
- the need to use an appropriate standard to construct the calibration curve.

Publication and collaboration in science results in the refinement of experimental protocols and the extension of methods into new applications.

23.8 Nucleic acids (AHL) *– DNA is the genetic material that expresses itself by controlling the synthesis of proteins by the cell*

DNA – the source of heredity

Deoxyribonucleic acid (DNA) was discovered in 1869, 10 years after the publication of Darwin's *The Origin of Species*. The Swiss biochemist Friedrich Meischer isolated a sample of DNA from white blood cells in pus sticking to discarded bandages. At this time there was no suspicion of the immense significance of the molecule as the 'vehicle' of heredity and evolution. Not until 1944 did Oswald Avery demonstrate that DNA was the material that transferred genetic information from one cell to another.

A one-page letter published on 25 April 1953 in the scientific journal *Nature* started the recent rapid increase in information about the origins of life, evolutionary development and the transfer of genetic information:

'We wish to suggest a structure for the salt of deoxyribonucleic acid (DNA). This structure has features which are of considerable scientific interest…It has not escaped our notice that the specific base pairing [inherent in the proposed structure] suggests a possible copying mechanism for the genetic material.'

The structure of DNA proposed thus by James Watson and Francis Crick led to the advent of molecular biology and genetic engineering. It was arguably the most important scientific development of the 20th century. In making their discovery, Watson and Crick used a model-building approach composition of DNA and, most importantly, data from X-ray crystallography (Figure 23.164).

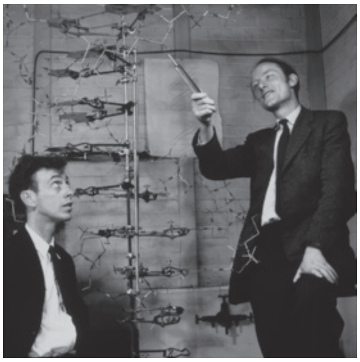


Figure 23.164 The discoverers of the structure of DNA: James Watson (on left) and Francis Crick, with their model of part of a DNA molecule in 1953

The development of science often builds on previous results. Elucidating the structure of DNA would have been impossible without the discovery of X-rays in 1895. Then in 1925, von Laue showed that the diffraction of X-rays could be used to find the arrangement of atoms in crystals. The method was successfully applied to determine the structure of proteins, including myoglobin and insulin, for example. Then, in 1952, Rosalind Franklin, working with Maurice Wilkins, shone X-rays on to crystalline forms of DNA and produced diffraction patterns that were both beautiful and complex (Figure 23.165). Watson, Crick and Wilkins were awarded the Nobel Prize for Physiology and Medicine in 1968. Tragically, Franklin died of cancer in 1958 and the Nobel Prize cannot be awarded posthumously.

The ordered X-ray patterns produced reflected the regularity of a double helical structure. Two DNA strands, running in opposite directions, are linked together in a ladder-like molecule – but a twisted ladder or a right-handed helix (Figure 23.166).

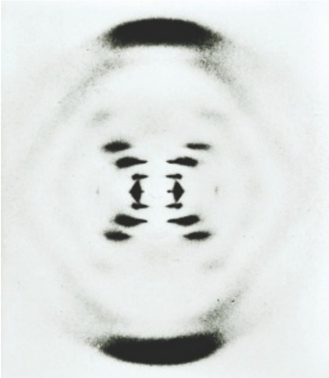


Figure 23.165 X-ray diffraction photograph of DNA. This image of the β-form of DNA was obtained by Rosalind Franklin in 1953. The cross of bands indicates the helical nature of DNA

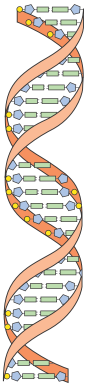


Figure 23.166 The double helix formed from two DNA strands. The sugar–phosphate backbone is on the outside and the bases are in the centre. London dispersion forces maintain the helical structure. Without such forces DNA would be a 'ladder' rather than a double helix

Each DNA strand is a condensation polymer of sugar molecules and phosphate groups. Attached to this sugar–phosphate backbone is a sequence of organic bases constructed from a choice of just four, often referred to simply by the first letter of their names: A, C, G and T. Heredity information is stored as the sequence of these bases along the chain. The genetic message is written in a language of only four letters.

Nucleic acids

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are essential components of all living cells. They are involved in the transmission of heredity information and in the production of the vast range of proteins made by cells. Both molecules are synthesized in cells by the condensation polymerization of **nucleotides** – they are polynucleotides. DNA controls heredity on a molecular level:

- it is a self-replicating molecule capable of passing genetic information from one generation to the next
- it contains in its base sequence the genetic code used to synthesize proteins.

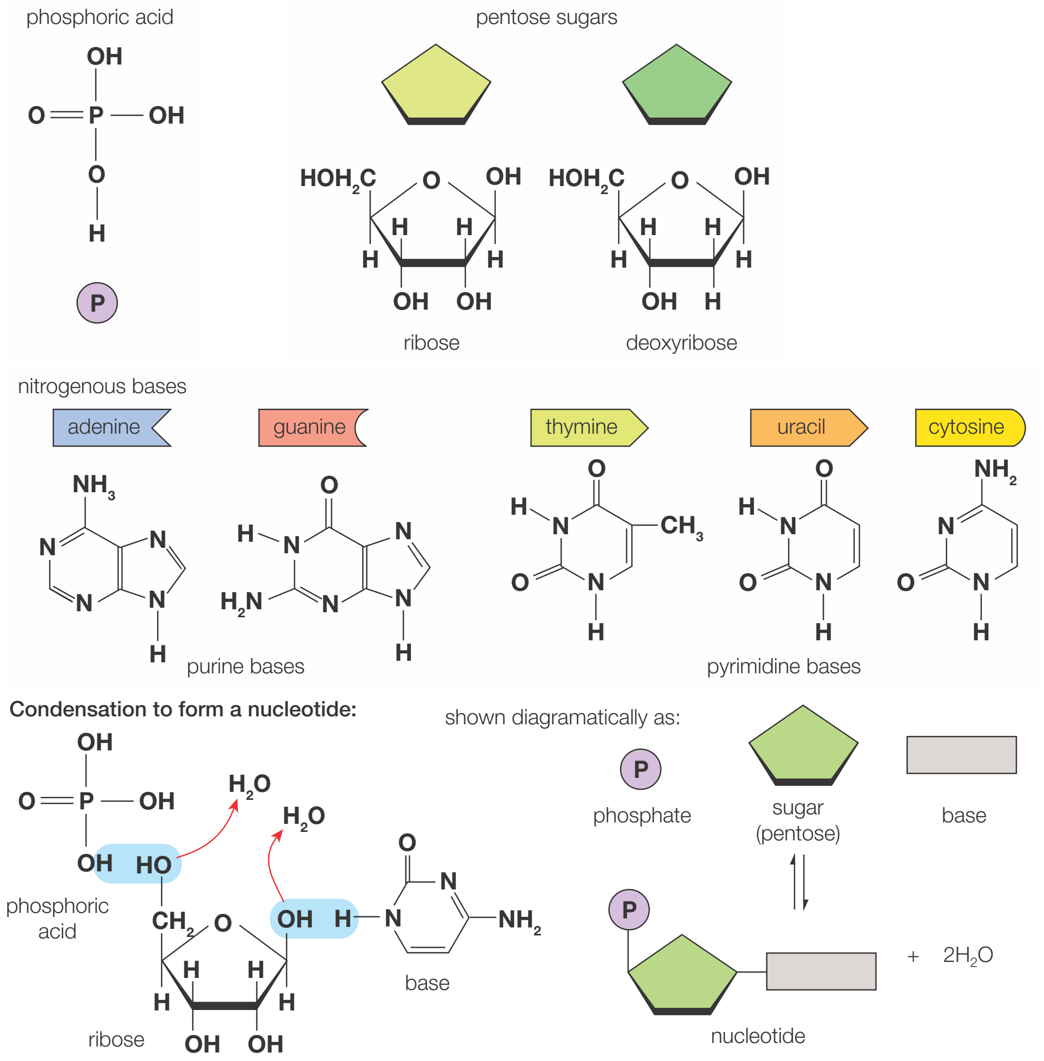
Various forms of RNA are involved in the processes of 'gene expression' that result in the production of proteins

Nucleotides are themselves made from three smaller types of molecule (phosphate, pentose sugar and base) covalently bonded together under enzyme control. The phosphate group is a chemically reactive functional group that allows new molecules to be added via a condensation reaction. Hence, nucleotides can form long chains (linear polymers). The phosphate groups are also ionized and are partly responsible for the solubility in water of nucleic acids.

The second component of a nucleotide is a pentose sugar (five-carbon monosaccharide) – deoxyribose in DNA and ribose in RNA. These sugars are chemically reactive and are involved in bonding different nucleotides together. This occurs via condensation reactions (under enzyme control) involving the hydroxyl groups located at carbon atoms 1 and 5.

The third component of each nucleotide is known as a **base**. It is covalently bonded to the pentose sugar via the carbon atom in position 1 of the ring. Four different bases are found in DNA: adenine (A), thymine (T), guanine (G) and cytosine (C). A fifth base (uracil) is found in RNA (see page 119). Cells continually synthesize nucleotides (Figure 23.167) and these form a 'pool' in the cytoplasm from which nucleotides can be used by the cell for synthesizing DNA.

Figure 23.167 Components of nucleotides



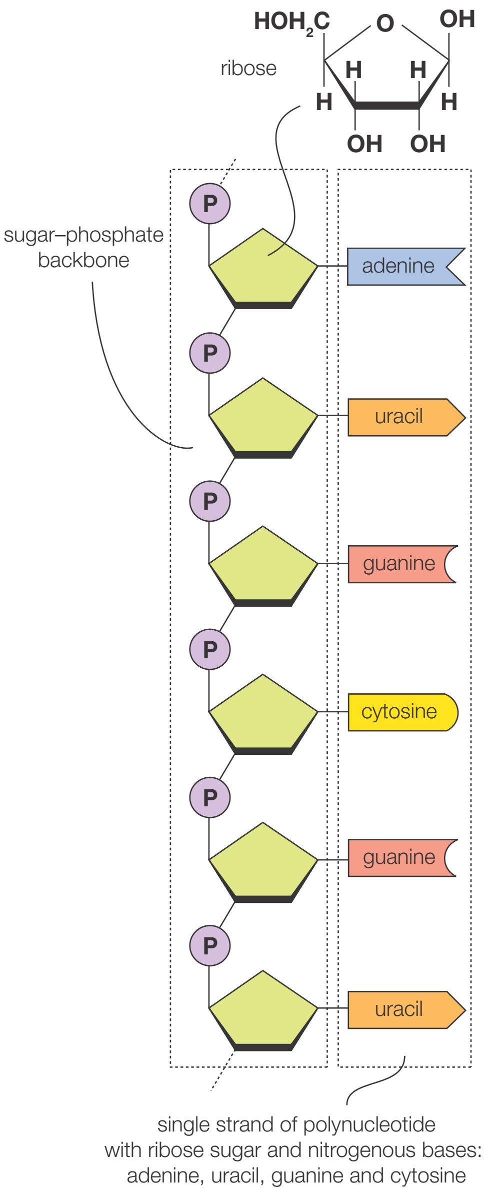

HOH$_2$C
O
OH
ribose
H
H
H
H
OH
OH
sugar–phosphate backbone
P
adenine
uracil
guanine
cytosine
guanine
uracil
single strand of polynucleotide with ribose sugar and nitrogenous bases: adenine, uracil, guanine and cytosine

Figure 23.168 Structure of RNA

Differences between DNA and RNA

Both RNA and DNA molecules are polynucleotides, but RNA molecules are considerably shorter than those of DNA. In RNA the nucleotides contain ribose and the bases are cytosine, guanine, adenine and uracil (Figure 23.168). In living cells there are three main functional types of RNA, known as messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). (Thymine only occurs as one of a number of minor bases in tRNA.) All three forms of RNA are directly involved in the synthesis of proteins.

DNA molecules occur in the chromosomes and form very long strands, containing several million nucleotides. In all DNA molecules the nucleotides contain deoxyribose. The bases in DNA are cytosine, guanine, adenine and thymine, but never uracil.

The DNA molecule consists of two polynucleotide strands held together by inter-molecular hydrogen bonding. The two strands take the shape of a double helix. Transfer RNA and ribosomal RNA contain both double-stranded (with an approximate helical shape) and single-stranded regions. In all cases RNA and DNA molecules are held together in the double-stranded form by complementary base pairing. Table 23.22 summarizes the differences between RNA and DNA.

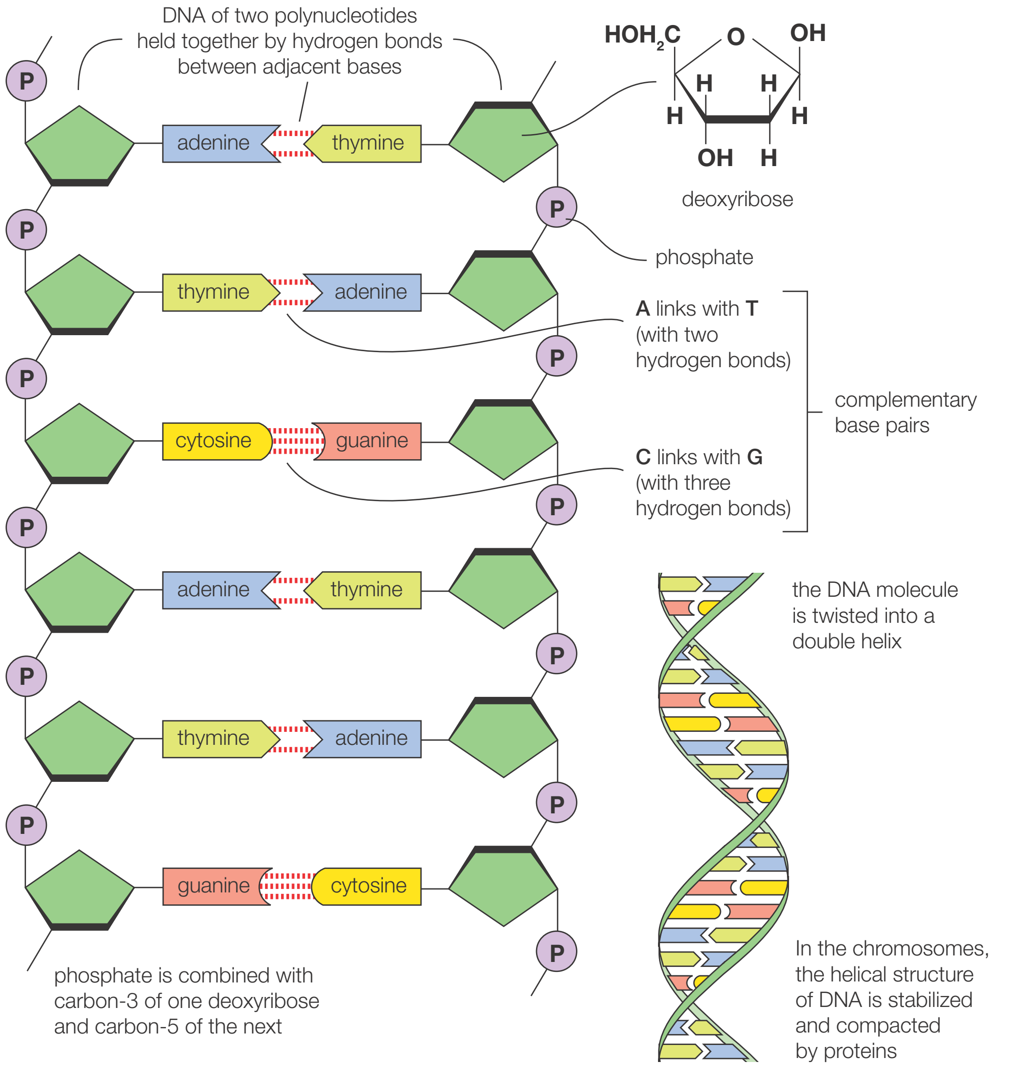
DNA	RNA
Generally very long strands, several millions of nucleotides long	Relatively short strands, 100–1000 nucleotides long
Contains deoxyribose	Contains ribose
Consists of two antiparallel polynucleotide strands of complementary base pairs: cytosine (C) with guanine (G), adenine (A) with thymine (T); the strands are held together by hydrogen bonds in the form of a double helix	Messenger RNA is single-stranded; transfer and ribosomal RNA have both single- and double-stranded sections; contains uracil (U) rather than thymine (T)
Relatively stable towards chemicals (especially alkalis) and enzymes	Less stable towards chemicals and enzymes

Table 23.22 Summary of the differences between RNA and DNA

Structure of DNA

DNA consists of two linear **polynucleotide** strands which are wound together in the form of a **double helix**. The double helix is composed of two right-handed helical polynucleotide chains coiled around the same central axis. The bases are inside the helix and the sugar–phosphate backbone is on the outside. The two chains of the double helix are held together by hydrogen bonds between the bases on the two polynucleotide chains (Figure 23.169).

Figure 23.169
The structure of DNA


DNA of two polynucleotides held together by hydrogen bonds between adjacent bases
P
adenine
thymine
thymine
adenine
cytosine
guanine
adenine
thymine
thymine
adenine
guanine
cytosine
HOH_2C
O
OH
H
H
H
H
OH
H
deoxyribose
phosphate
A links with **T** (with two hydrogen bonds)
C links with **G** (with three hydrogen bonds)
complementary base pairs
the DNA molecule is twisted into a double helix
In the chromosomes, the helical structure of DNA is stabilized and compacted by proteins
phosphate is combined with carbon-3 of one deoxyribose and carbon-5 of the next

On heating, the two strands of DNA separate from each other. This is known as the melting of DNA. The temperature at which the two strands of DNA separate completely is known as the melting temperature of DNA. A DNA molecule which is rich in GC pairs has a higher melting temperature than a DNA molecule (of the same length) rich in AT pairs. This is because GC pairs are held together by three hydrogen bonds, but the AT base pair is only held together by two hydrogen bonds. The two strands of the double helix are **antiparallel** – that is, they run in opposite directions. The 5'–3' **phosphodiester linkages** run in opposite directions. The **5'** (pronounced 'five prime') **end** designates the end of the DNA strand that has the fifth carbon in the sugar ring of the deoxyribose at its terminus. The **3'** (pronounced 'three prime') **end** of a strand ends at the hydroxyl (–OH) group of the third carbon in the deoxyribose sugar ring. The fact that the DNA chains have direction is significant in the context of the fact that they carry a 'coded message' that must be read in a particular direction.

The hydrogen bonding between the bases of the two strands is highly specific. Thymine (T) pairs with adenine (A), via the formation of two hydrogen bonds, and cytosine (C) pairs with guanine (G), via the formation of three hydrogen bonds. These base pairs are favoured energetically and are known as **complementary base pairs**. A further set of interactions that is important to the formation and stability of the double helix are the London forces between the base pairs as they stack on top of each other in the structure – this is referred to as 'base stacking'. Overall the stability of the helix is achieved by the fact that it maximizes hydrophobic interactions between the non-polar stacked bases in the sequestered environment in the middle of the molecule, while allowing polar and charged groups in the sugar–phosphate backbone to interact with the aqueous solution.

Consequently, the two strands of DNA are complementary to each other and the sequence of bases in one strand determines the sequence of bases in the other chain. Opposite each adenine on one chain there is always a thymine on the other chain, and opposite guanine there is cytosine. Complementary base pairing is the underlying basis for the processes of replication, transcription and translation.

Nature of Science

The build-up of ideas

The background to the realization of the importance of base pairing illustrates how findings from different experimental techniques come together to generate an overall picture. In 1950 Erwin Chargaff analysed the base content of DNA from different species of organism using paper chromatography (Table 23.23). His results showed the intriguing finding that, although the proportions of the bases varied between species, the number of purines equalled the number of pyrimidines in any sample of DNA. Chargaff went on to show that the DNA from any tissue of a particular species has equal numbers of adenine and thymine residues and equal numbers of guanine and cytosine residues. Watson and Crick realized that this evidence was crucial and used it alongside the X-ray crystallography data to build their model.

■ **Table 23.23** The molar proportions of A, T, C and G in samples of DNA

Source of DNA	Adenine, A/ molar %	Thymine, T/ molar %	Cytosine, C/ molar %	Guanine, G/ molar %
Bacteria	15.1	14.6	35.4	34.9
Wheat	27.3	27.1	22.8	22.7
Salmon	29.7	29.1	20.4	20.8
Human	30.9	29.4	19.8	19.9

The nitrogenous bases found in DNA are derivatives of **purine** or **pyrimidine**. Cytosine and thymine are pyrimidines and consist of one heterocyclic ring. Adenine and guanine are purines and consist of two heterocyclic rings. Heterocyclic rings contain atoms other than carbon. Only these particular base pairings will hydrogen bond together strongly and fit inside the double helix. Cytosine and thymine are too large to fit into the helix, and adenine and guanine are too far apart to form stable hydrogen bonds.

ToK Link

Information and knowledge

DNA stores information in the form of a linear chain of four bases: adenine, thymine, guanine and cytosine. The sequence of these bases along the coding strand is ultimately read as triplets and translated into a linear chain of amino acids (a protein). Each amino acid has one or more unique triplets of bases that code for it during translation in the ribosome. This is the genetic code. DNA holds information, but it generally does not actively apply that information. DNA does not make proteins directly. To extract the information and get it to the ribosomes the DNA sequence is transcribed into a corresponding sequence in a 'carrier' molecule called ribonucleic acid, or RNA. The portions of DNA that are transcribed into RNA are genes. The gene code is not a code for an organism in the same way as a blueprint does not build a building without the builders to build it and the technical know-how of the construction team. The information in the genome is at many different levels, in sequence, in structure, between the genes in the junk DNA and in the chemistry. It is not 0 and 1 – it is an intimate layering of information.

UniProt is a central resource for storing and interconnecting biological information from large and disparate (very different) sources. It is a comprehensive catalogue of protein sequence (primary structure of amino acids) and functional annotation, which gives information about the function of the protein. UniProt is built upon the bioinformatics infrastructure and scientific expertise at the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). UniProt has three different protein databases optimized for different uses, and is updated and distributed every month. It can be accessed online for searches or download at www.uniprot.org. Given the development of databanks such as this, and the stress placed on access to information in modern science, it is worth looking at the relationship between the terms.

Data represents raw numbers (binary 1 and 0 for computer data) or assertions. Data comprises facts, observations or perceptions. Information is data with context and relevance. In contrast, data, for example computer data, can include millions of useless garbage bits, which are nothing more than uninterpretable binary zeros and ones. Information involves manipulation of raw data. Often, information can be used to obtain a more meaningful indication of trends or patterns. Knowledge is information with decision-making and action-directed utility and purpose. The Greek philosopher Plato suggested knowledge is defined as a justified true belief. Knowledge is information that helps to produce information from data, or produce more valuable information from less valuable information (Figure 23.170).

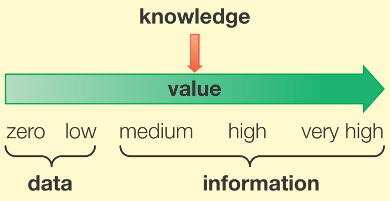


Figure 23.170 The interaction between knowledge and data that increases the value of the acquired information

We live in the information age. But every era of history has had its own information revolution: the invention of writing, the composition of dictionaries, the creation of the charts that made navigation possible, the discovery of the electronic signal, the cracking of the genetic code.

An interesting interpretation of the history that has led up to our 'information age' is given in the book *The Information: A History, a Theory, a Flood* by James Gleick. He aims to tell the story of how human beings use, transmit and keep what they know. From African talking drums to Wikipedia, from Morse code to the 'bit', it is one account of the modern age's defining idea and an exploration of how information has revolutionized our lives.

39 a Representing molecules of purine or pyrimidine bases by **B**, sugars by **S** and phosphoric acid or phosphate groups by **P**, and using no other symbols, draw a diagram to show how these are linked in a short length of double-stranded DNA. Use full lines (–) for normal covalent bonds and dotted lines (···) for hydrogen bonds.

b Your sketch makes the two strands look identical. Ignoring the difference between the bases:

i Explain how the two strands differ.

ii Give the technical term which describes this difference.

iii State how it is indicated on diagrams of DNA.

c When a DNA molecule is gently heated in solution, the two chains gradually separate. The temperature at which 50 per cent of the helical structure is lost is called the melting temperature. Explain why the melting temperature of a particular DNA sequence is dependent upon the percentage of GC base pairs in the DNA.

40 a What role do hydrogen bonds play in the accurate replication of DNA?

b DNA is replicated semi-conservatively. What is meant by this term?

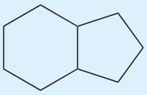
c i What type of interaction takes place between the bases in the two DNA strands?

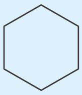
ii Condensation reactions are involved in the formation of DNA. What is the name given to the links which form the backbone of a DNA strand?

41 a State three ways in which the structure of DNA differs from that of RNA.

b i Name two nitrogenous bases present in DNA, one a purine and the other a pyrimidine.

ii Suggest the most appropriate name for each of the following outline structures: purine or pyrimidine.

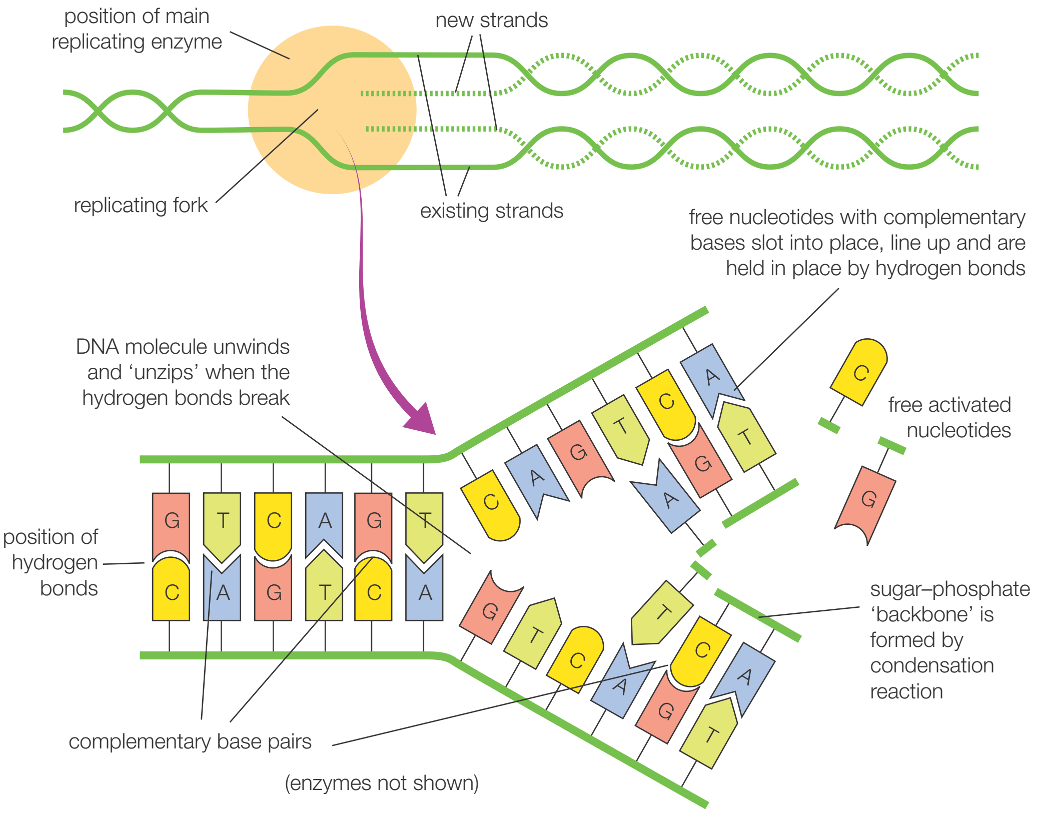

I

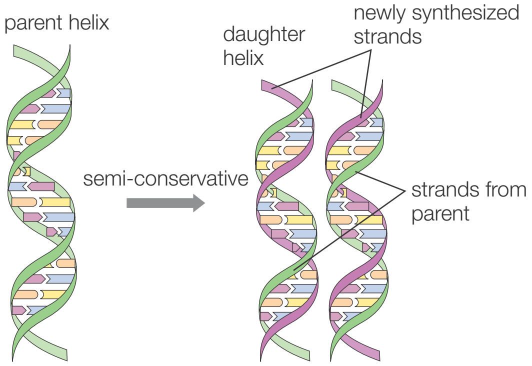

II

Replication

DNA has the unique property among biomolecules of duplicating itself in the presence of appropriate enzymes. The genetic information inside a cell is coded into the sequence of bases in its DNA molecules. During cell division, DNA molecules replicate and produce exact copies of themselves. Each daughter cell has DNA molecules identical to that of the parent cell. DNA replication is a very complex process but the underlying feature is that the two strands of DNA are unwound (under enzyme control) and each strand serves as a template pattern for the synthesis of a new complementary DNA strand (Figure 23.171). The specificity of complementary base pairing ensures the exact duplication of the sequence of bases in the new daughter strand of DNA.

■ **Figure 23.171** Simplified summary of the replication of DNA


position of main replicating enzyme
new strands
replicating fork
existing strands
free nucleotides with complementary bases slot into place, line up and are held in place by hydrogen bonds
DNA molecule unwinds and 'unzips' when the hydrogen bonds break
free activated nucleotides
position of hydrogen bonds
G T C A G T
C A G T C A
sugar–phosphate 'backbone' is formed by condensation reaction
complementary base pairs
(enzymes not shown)



■ **Figure 23.172** The outcome of semi-conservative replication

During replication, the hydrogen bonds and instantaneous dipole-induced dipole forces between the base pairs in the double helix are broken. The original strands act as templates for the synthesis of two new strands. Each new strand contains a sequence of bases complementary to the bases of the original strand.

Hydrogen bonds and London (dispersion) forces form between the original and new strands, creating a stable helical structure. Thus, two daughter molecules are formed from the parent double helix (Figure 23.172). This form of replication is known as **semi-conservative replication**, because each daughter molecule contains one newly produced strand and one strand from the original DNA molecule.

Nature of Science

Competition and collaboration in science

The story of the discovery of the structure of DNA shows how different approaches to solving the same problem can lead to a consistent conclusion. James Watson and Francis Crick, working in Cambridge, England, tackled the question largely by building molecular models, while Maurice Wilkins and Rosalind Franklin, working in London, approached it through X-ray crystallography. It is, however, also a story interwoven with tales of personal ambition and human conflict that reflects some failure of the different teams of scientists to collaborate and communicate effectively. Mingled in the events taking place in the UK is the intervention of the Nobel laureate, Linus Pauling, who, shortly before Crick and Watson's letter was published, put forward a triple-helical structure for DNA. His reasoning was flawed, probably in part because the data he was working was incorrect. One of the drawbacks of his structure was that he had placed the phosphate groups internal to the structure, where they would essentially repel each other.

The inter-relationships between the different groups involved in the 'race' to elucidate the structure serve to illustrate that the practice of science is not impersonal and abstract but involves careers, personal and group ambition, and competitiveness. There have been other more modern examples that also illustrate this – from research into AIDS, the sequencing of the human genome, and discussion on the nature of the evolutionary process, for instance – but the DNA story remains well known and well documented. It is partly coloured by the tragic story of Rosalind Franklin missing out on her acknowledgement in the form of a Nobel Prize. Her story is detailed in Brenda Maddox's biography *Rosalind Franklin: the dark lady of DNA*.

Having noted the above, it is worth acknowledging that interdisciplinary collaborative research is widely recognized for its importance in advancing scientific knowledge. Many of the great discoveries and achievements in the 20th and 21st centuries were the results of collaborative efforts between scientists and engineers from diverse fields. The discovery of the Higgs boson, the Intergovernmental Panel on Climate Change assessments, the mapping of the ocean floor, combatting bacteria via vaccines and antibiotics, and the discovery of DNA are just a few examples that exemplify the power of interdisciplinary collaborations. Great scientific work often occurs when two groups, given the same problem, ask different questions, notice different details, use different approaches to describe the problem, and come into the situation with different perspectives. This is the strength of interdisciplinary collaborative research. The mixing of different frameworks of thinking is a great way to stimulate the development of new approaches to a problem that a single group, or scientific discipline, would not be able to do.

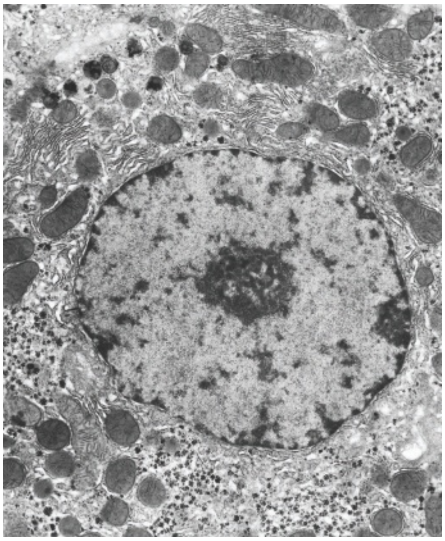


Figure 23.173 Transmission electron micrograph (TEM) of a section through a rat liver cell. At the centre is the nucleus, which contains the cell's genetic information

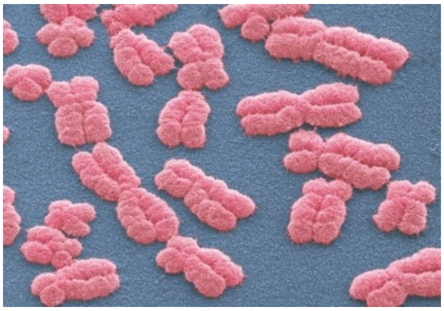


Figure 23.174 Coloured scanning electron micrograph (SEM) of pairs of human chromosomes

DNA and chromosomes

DNA molecules are enormous macromolecules. They need to be, as the DNA contains the essential genetic information that defines the organism concerned. The molecular sizes increase with the complexity of the organism:

- The DNA of a bacterial virus (bacteriophage A) is 166 000 base pairs (166 kilobases) in length.
- The DNA of the bacterium *E. coli* contains about 4 700 000 base pairs (4700 kilobases).
- In animal cells the DNA is compartmentalized in the nucleus (Figure 23.173), where it is packaged into chromosomes, which are each thought to consist of a single molecule of DNA with associated proteins; the length of DNA in each of the human chromosomes varies from 48 000 000 to 240 000 000 base pairs (48 000 to 240 000 kilobases) (Figure 23.174).

The folding of such lengths of DNA into the compact chromosomes that become visible under a light microscope at cell division represents an amazing feat of packaging. A point to note is that only a small proportion of all of this DNA codes for proteins. The remainder is a mixture of 'junk' and sections that regulate the decoding itself. The functional regions of DNA are known as genes.

When analysed by gel electrophoresis (see later) it is noted that all DNA fragments are observed to migrate towards the positive electrode in electrophoresis, indicating that they are negatively charged. The double-helix structure of DNA shows that the origin of this charge is on the phosphate groups that link the sugars together in the backbone of the molecule. The negative charge causes DNA to associate with proteins known as histones that have a high proportion of basic amino acids and so carry positive charges at cell pH. The combination of DNA and histones, known collectively as chromatin, helps to stabilize the DNA within the chromosomes in the nucleus (Figure 23.175).

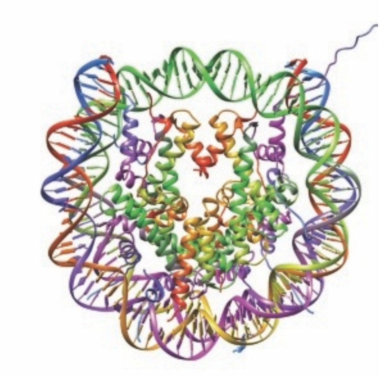


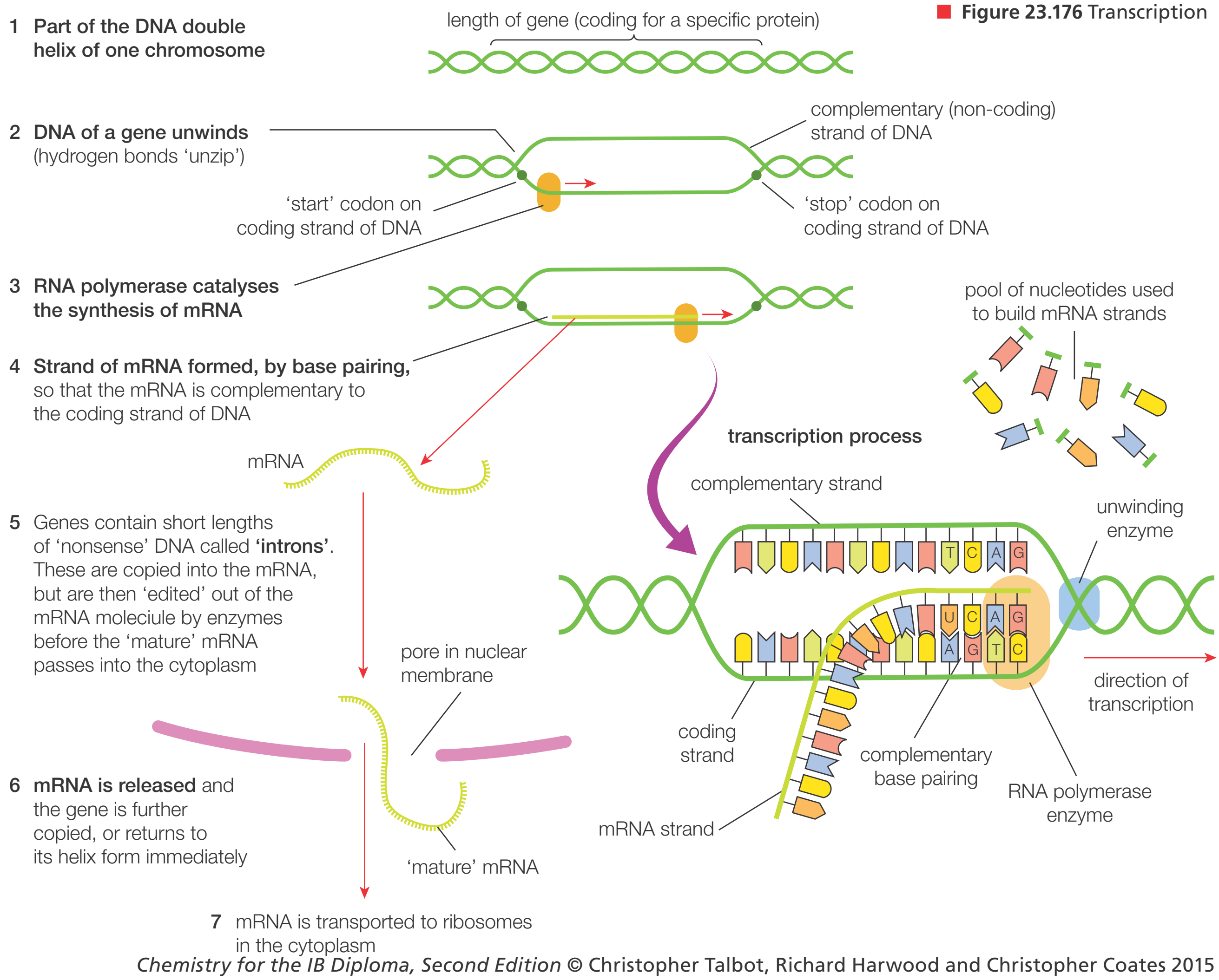
Figure 23.175 A molecular model of a nucleosome, the fundamental repeating unit used to package DNA inside cell nuclei. The DNA is coiled round a core of histone proteins (the multicoloured ribbons). Each set of two DNA loops around a histone core and is known as a nucleosome. Further compacting and packaging (not seen here) form the denser forms of chromatin, and eventually the cell's chromosomes

Protein synthesis

The DNA molecules in the nucleus of an animal cell hold the genetic code for protein synthesis. Each gene is responsible for the production of a single protein. The genetic information is coded in DNA in the form of a specific sequence of bases within a gene. The synthesis of a protein involves two steps: transcription and translation.

Messenger RNA

RNA (ribose nucleic acid) is a single-stranded molecule that is formed by **transcription** from DNA (Figure 23.176). The DNA molecule separates into two strands (under enzyme control) to reveal its bases, as in replication. However, in transcription, it is free ribonucleotides (and not deoxyribonucleotides) that base pair to it and form an RNA molecule. The RNA molecule, known as **messenger RNA** (mRNA), is transported out of the nucleus of the cell and attaches to a cell organelle known as a **ribosome**. Ribosomes are formed from protein and RNA, and are the sites at which proteins are synthesized from amino acids, during a process called **translation**. Messenger RNA is responsible for converting the genetic code of DNA into protein.



1 Part of the DNA double helix of one chromosome
length of gene (coding for a specific protein)
2 DNA of a gene unwinds (hydrogen bonds 'unzip')
complementary (non-coding) strand of DNA
'start' codon on coding strand of DNA
'stop' codon on coding strand of DNA
3 RNA polymerase catalyses the synthesis of mRNA
pool of nucleotides used to build mRNA strands
4 Strand of mRNA formed, by base pairing, so that the mRNA is complementary to the coding strand of DNA
mRNA
transcription process
complementary strand
unwinding enzyme
T C A G
U C A G
A G T C
5 Genes contain short lengths of 'nonsense' DNA called 'introns'. These are copied into the mRNA, but are then 'edited' out of the mRNA moleciule by enzymes before the 'mature' mRNA passes into the cytoplasm
pore in nuclear membrane
direction of transcription
coding strand
complementary base pairing
RNA polymerase enzyme
mRNA strand
6 mRNA is released and the gene is further copied, or returns to its helix form immediately
'mature' mRNA
7 mRNA is transported to ribosomes in the cytoplasm

Figure 23.176 Transcription

Additional Perspectives

p53

Protein 53 (p53) is a transcription factor that in humans is encoded for by the *TP53* gene. A transcription factor is a protein that binds to specific sequences of DNA and controls the transcription of genes. p53 regulates cell division and functions as a tumour suppressor gene (Figure 23.177). For this reason, p53 has been described as the 'guardian of the genome'. It has a number of anti-cancer mechanisms; for example, it can activate DNA repair proteins when DNA is damaged and needs to be repaired, and it can initiate programmed cell death if the damaged DNA cannot be repaired. More than half of human tumours contain a mutated *TP53* gene. The p53 molecules from mutated *TP53* genes either are misfolded or lack essential functional residues and do not bind to DNA.

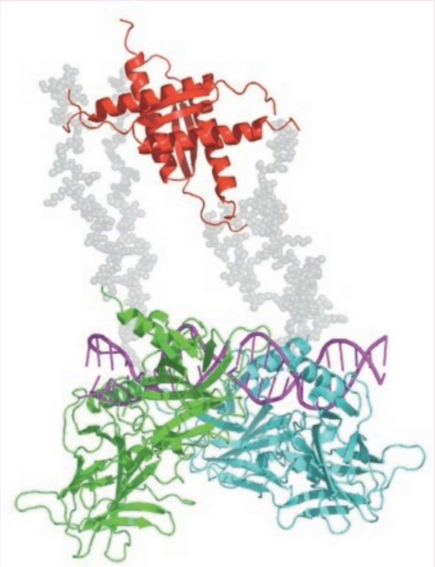
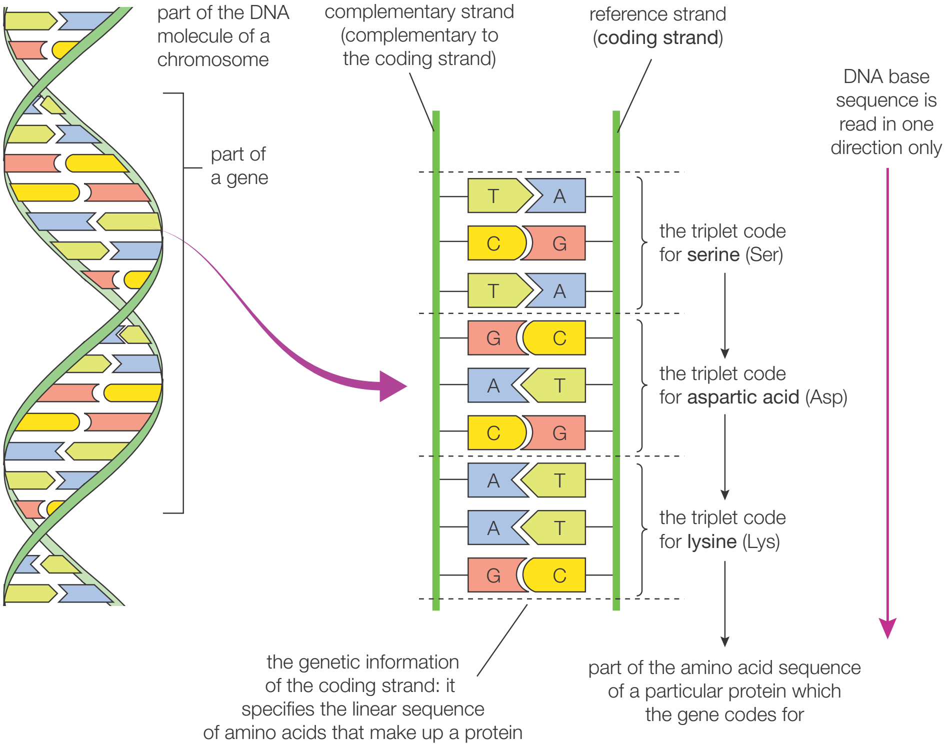


Figure 23.177 Tumour suppressor p53 bound to DNA

The triplet code

The primary structure of a protein consists of a chain of amino acids connected by peptide links. There are about 20 naturally occurring amino acids. The structure of DNA includes the four nitrogenous bases adenine (A), guanine (G), cytosine (C) and thymine (T). The code for each amino acid (called a **codon**) is a sequence of three bases. There are 64 (4^3) different triplets (sequences of three bases) that can be made up by four bases. As a result, some amino acids are encoded by more than one codon. The codons for some amino acids are given in Figure 23.178. Of the 64 codons, 61 code for amino acids and three act as 'stop' signals that terminate protein synthesis when the end of the polypeptide chain has been reached.

Figure 23.178 Part of a gene and how its DNA codes for amino acids



part of the DNA molecule of a chromosome
part of a gene
complementary strand (complementary to the coding strand)
reference strand (coding strand)
DNA base sequence is read in one direction only
T A
C G
T A
G C
A T
C G
A T
A T
G C
the triplet code for serine (Ser)
the triplet code for aspartic acid (Asp)
the triplet code for lysine (Lys)
the genetic information of the coding strand: it specifies the linear sequence of amino acids that make up a protein
part of the amino acid sequence of a particular protein which the gene codes for

Universality and redundancy in the genetic code

Many codons are redundant, which means that two or more codons can code for the same amino acid. Degenerate codons may differ in their third positions; for example, both GAA and GAG code for the amino acid glutamic acid. The **degeneracy** of the genetic code is what accounts for the existence of silent mutations. These are DNA mutations that do not result in a change to the amino acid sequence of a protein.

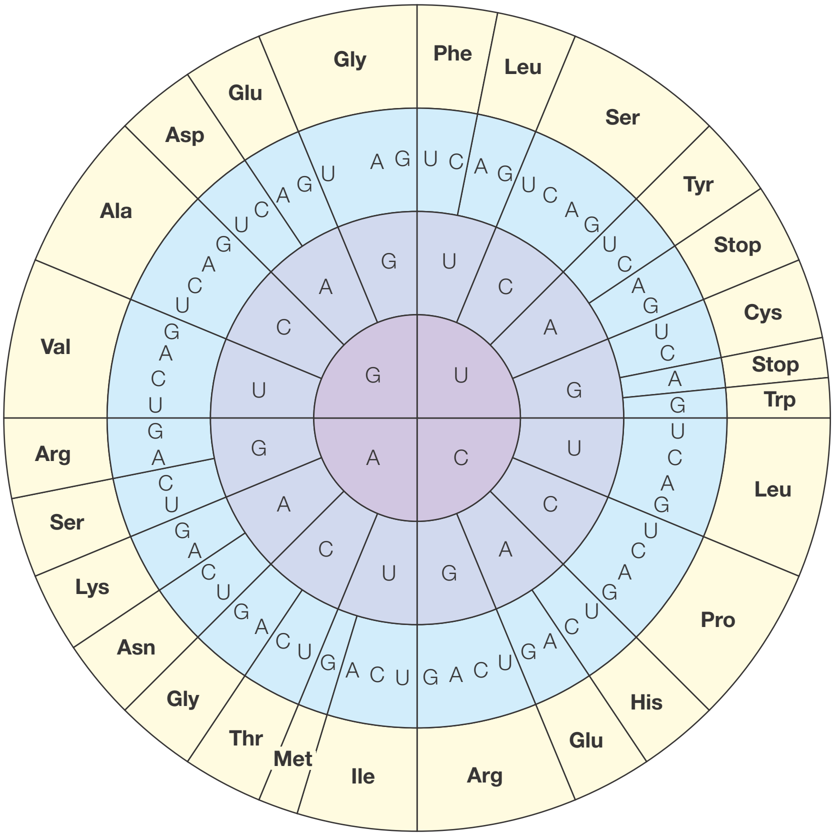
Degeneracy occurs because there are only 22 different codons required – one for each of the 20 amino acids and a stop and start codon. However, there are four bases arranged in triplet codons, which can produce 64 different codons ($4^3 = 64$). (Note that with four bases, if there were two bases per codon the number of possible codons would only be 16, as $4^2 = 16$.)

The genetic code in Figure 23.179 has also been called the universal genetic code. It is described as universal because it is used by all known organisms as a code for DNA, messenger RNA and transfer RNA. The universality of the genetic code encompasses animals (including humans), plants, fungi, bacteria and viruses.

■ **Figure 23.179** The genetic code

The genetic code in circular form
The codons are those of the DNA complementary strand and of messenger RNA (where uracil, U, replaces thymine, T).

Read the code from the centre of the circle outwards along a radius. For example, serine is coded by UCU, UCC, UCA or UCG, or by AGU or AGC.



In addition, some codons stand for **stop**, signalling the end of a polypeptide/protein chain.

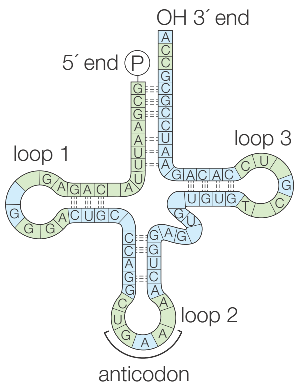
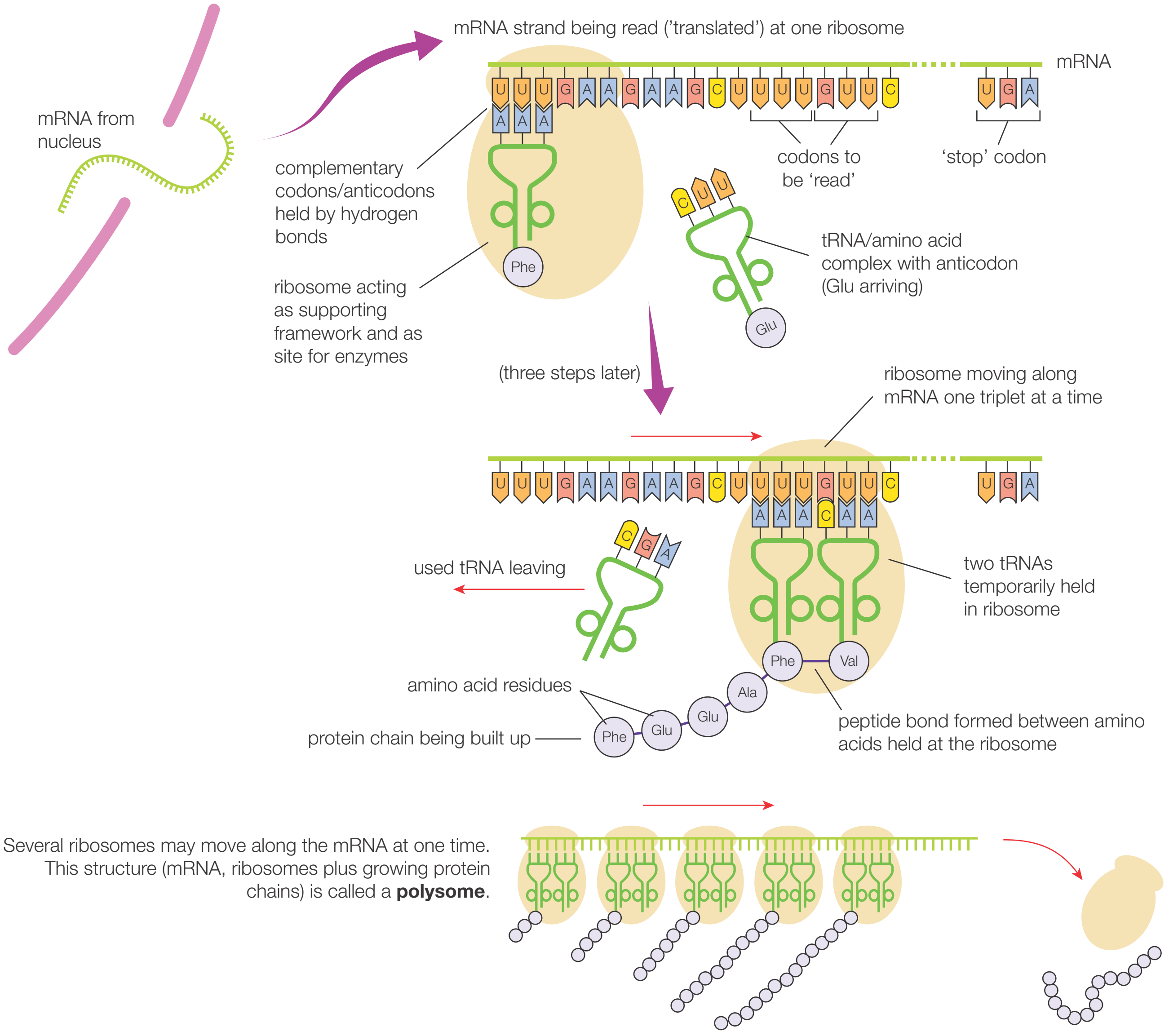


Figure 23.180 RNA, although single-stranded, can form helical loops by bending back on itself. Here, tRNA has a 'cloverleaf'-type structure with four sections containing base-paired regions

Role of ribosomes in protein synthesis

Protein synthesis takes place in ribosomes located in the cytoplasm. One end of an mRNA molecule binds to a ribosome, which moves along the mRNA strand three bases at a time. Molecules of another type of RNA, called **transfer RNA** (tRNA; Figure 23.180), bind to free amino acids in the cytoplasm. The tRNA molecules each carry a specific amino acid. They also each have their own base triplet, known as an **anticodon**, which binds via hydrogen bonding to the complementary codon triplet on the mRNA. In this way the mRNA determines the order of amino acids. Peptide links form between adjacent amino acids, and the protein chain steadily grows. Figure 23.181 summarizes the process of protein synthesis.


mRNA strand being read ('translated') at one ribosome
mRNA
mRNA from nucleus
U U U G A A G A A G C U U U U G U U C
U G A
A A A
codons to be 'read'
'stop' codon
complementary codons/anticodons held by hydrogen bonds
C U U
tRNA/amino acid complex with anticodon (Glu arriving)
Phe
Glu
ribosome acting as supporting framework and as site for enzymes
(three steps later)
ribosome moving along mRNA one triplet at a time
A A A C A A
C G A
used tRNA leaving
two tRNAs temporarily held in ribosome
Phe
Val
Ala
Glu
Glu
Phe
amino acid residues
protein chain being built up
peptide bond formed between amino acids held at the ribosome

Several ribosomes may move along the mRNA at one time. This structure (mRNA, ribosomes plus growing protein chains) is called a **polysome**.

Figure 22.181 Translation: protein synthesis

Ribosomes

All cells contain ribosomes (Figure 23.182) as they are the site of protein synthesis. They are complex assemblies of structural ribosomal RNA molecules and a substantial number of proteins. They are a further example of the process of self-assembly as isolated ribosomes can be dis-assembled into their various components and then reassembled into functional ribosomes in the test tube.

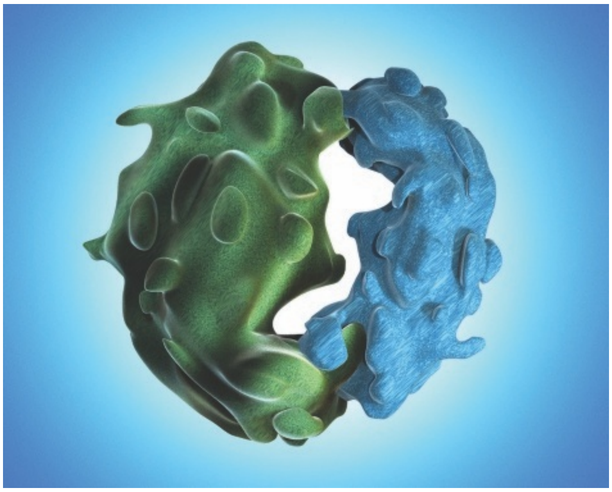


Figure 23.182 A computer-generated side-on image of a ribosome from a bacterial cell. Each ribosome consists of a large and a small subunit. Each subunit contains a different form of ribosomal RNA in association with a large number of ribosomal proteins

Nature of Science

Expressing and delivering the message

The language used in describing the stages involved in the complexity of genetic transfer and expression is understandably borrowed from normal linguistic expression and code-breaking. The genetic message encoded in the DNA of cells is used to form protein molecules by the processes of transcription and translation. These two processes together are termed gene expression. Thus DNA, by the processes of transcription and translation, is ultimately responsible for the structure of all the proteins synthesized by cells.

There is a one-way flow of coded information from the nucleus to the cytoplasm in animal cells. This was termed the 'central dogma' by Francis Crick (Figure 23.183). However, a group of viruses known as retroviruses (including HIV) contain RNA, which is reverse transcribed into DNA by an enzyme known as reverse transcriptase.

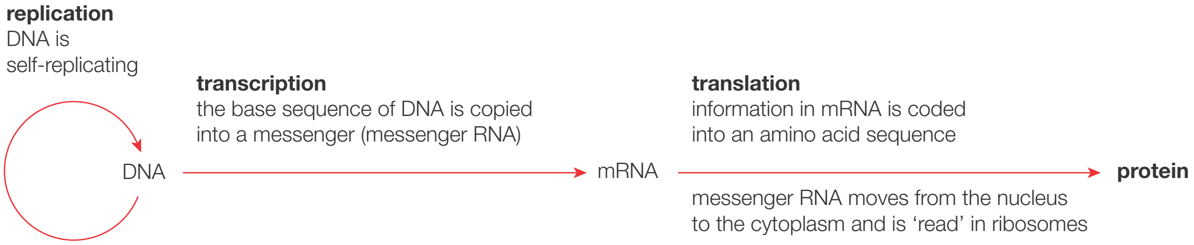

replication
DNA is self-replicating
transcription
the base sequence of DNA is copied into a messenger (messenger RNA)
translation
information in mRNA is coded into an amino acid sequence
DNA
mRNA
protein
messenger RNA moves from the nucleus to the cytoplasm and is 'read' in ribosomes

Figure 23.183 The central dogma

Additional Perspectives

Sickle cell anemia

Sickle cell anemia (Figure 23.184) is a genetic disorder in which red blood cells are distorted and have a reduced ability to transport oxygen. In normal hemoglobin, the amino acid at position 6 in the β-polypeptide chain is glutamic acid. In sickle cell anemia, this amino acid is valine. This change in one amino acid has come about because of a faulty sequence of nucleotides in the section of DNA that codes for hemoglobin synthesis. This disease is inherited: it can be passed from parents to their children. A selective advantage of this faulty sequence of nucleotides is that it protects carriers of sickle cell anemia from the worst effects of malaria, which is constantly present in the areas of Africa where sickle cell anemia is most common.

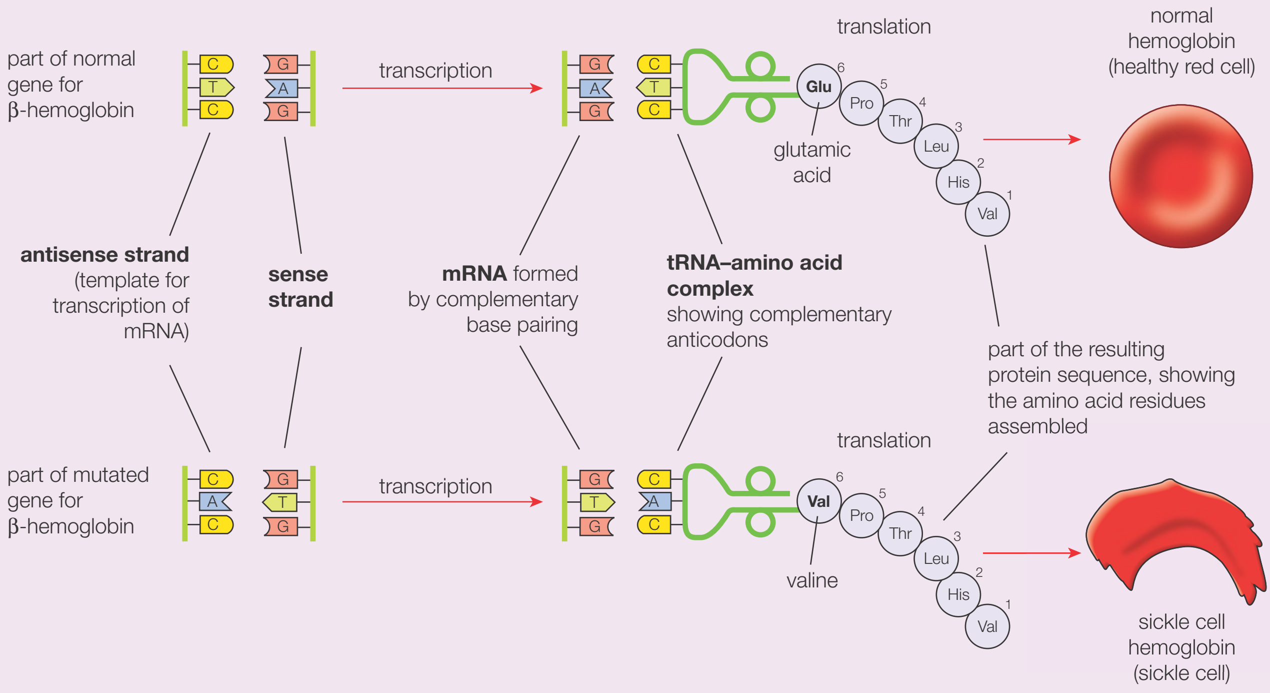

part of normal gene for β-hemoglobin
C
T
C
G
A
G
transcription
G
A
G
C
T
C
translation
6
Glu
5
Pro
4
Thr
3
Leu
2
His
1
Val
glutamic acid
normal hemoglobin (healthy red cell)
antisense strand (template for transcription of mRNA)
sense strand
mRNA formed by complementary base pairing
tRNA–amino acid complex showing complementary anticodons
part of the resulting protein sequence, showing the amino acid residues assembled
part of mutated gene for β-hemoglobin
C
A
C
G
T
G
transcription
G
T
G
C
A
C
translation
6
Val
5
Pro
4
Thr
3
Leu
2
His
1
Val
valine
sickle cell hemoglobin (sickle cell)

Figure 23.184 Sickle cell anemia: an example of a single mutation

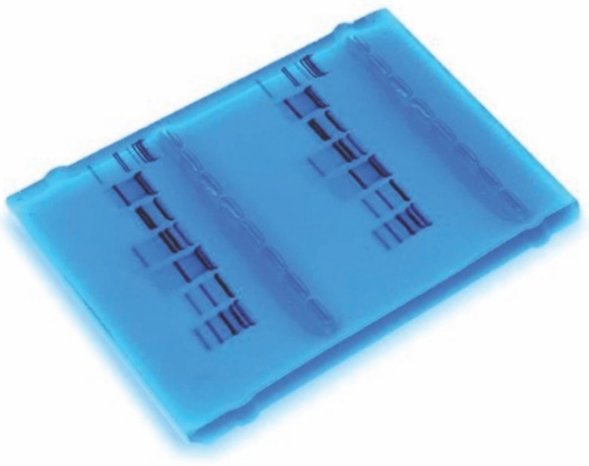


Figure 23.185 A simulation of DNA profiling in a school laboratory. Bacteriophage lambda DNA is treated with restriction enzymes and the fragments then electrolysed. The DNA fragments are visualized using Fast Blast DNA stain (Bio-Rad)

DNA profiling

DNA profiling (Figure 23.185) uses the techniques of genetic engineering to identify a person from a sample of their DNA. DNA profiling used to be referred to as 'DNA fingerprinting' but this was changed to avoid confusion with actual fingerprinting in forensic investigations. It is widely used to eliminate or charge suspects in crimes in which blood, tissue or body fluid samples, for example semen and blood, are available. It can also be used to establish evolutionary relationships between people and establish whether a person is the biological mother or father in a paternity case.

Large portions of a person's DNA are identical to every other person's DNA. In addition, large sections of DNA are not genes and do not code for proteins. However, small sections or fragments of our human DNA are unique to a particular individual. These non-coding fragments of DNA are termed **polymorphic** because they vary from person to person. DNA profiling is essentially the process of separating an individual's unique, polymorphic fragments from the common ones.

Figure 23.186 A PCR machine

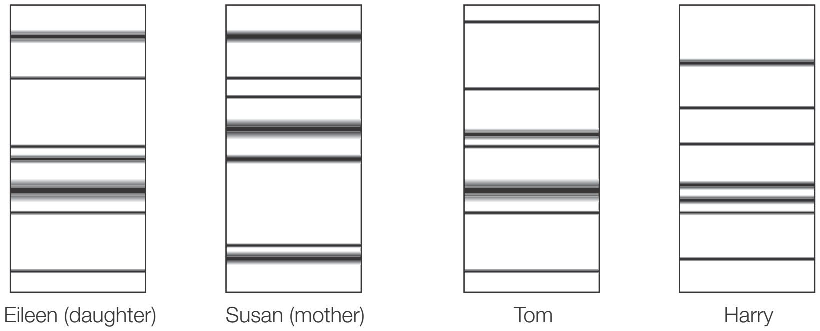
The process of DNA profiling is summarized in outline form:

- A sample of cells is obtained from blood, semen, hair root or body tissues, and the DNA is extracted from the disrupted cells.
- The DNA is copied and amplified by an automated process called **polymerase chain reaction (PCR)** (Figure 23.186). This technique separates the strands of the DNA (using high temperature) and then uses a heat-stable DNA polymerase to make thousands of exact copies of the original DNA. This process produces sufficient DNA to analyse.
- The DNA is then cut into small, double-stranded fragments using restriction enzymes. These enzymes recognize certain sequences within the non-coding DNA which often contains many highly repetitive regions of DNA sequences.
- The resulting DNA fragments of varying lengths are separated by gel electrophoresis into a large number of invisible bands.
- The gel is treated with alkali to split double-stranded DNA into single strands.
- A copy of the strands is transferred to a membrane and selected radioactively labelled DNA **probes** are added to the membrane to base pair with particular DNA sequences. The excess probes are washed away.
- The membrane is overlaid with X-ray film which becomes selectively 'fogged' by emission of ionizing radiation from the based-paired radiolabels.
- The X-ray film is developed, showing the positions of the bands (fragments) to which probes have base paired.

DNA profiling can be used to identify the parents of a child since the child will inherit half of its DNA from the father and half from the mother. Figure 23.187 shows the DNA profiles of a mother and a daughter together those of two men, one of whom is the father.

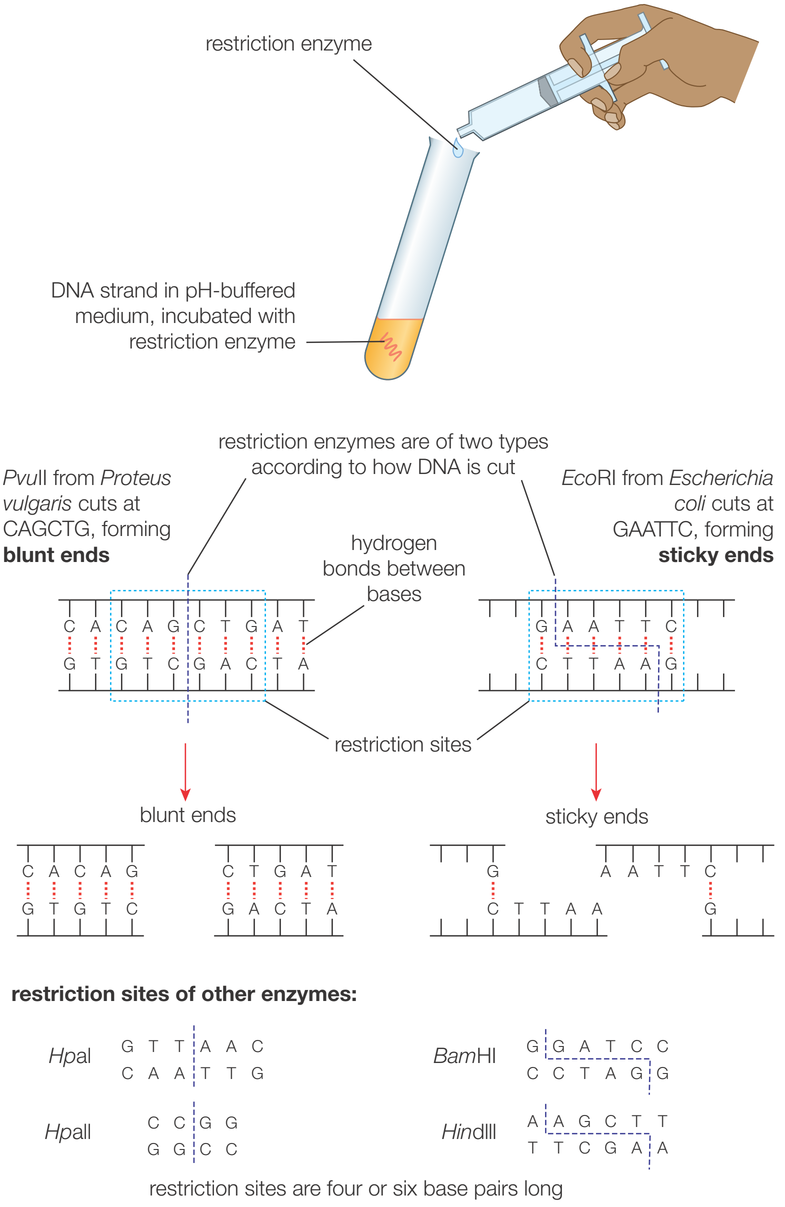
Ignoring the three bands in Eileen's DNA profile which occur in the same position as her mother's, you will see that all four of the remaining bands correspond with those of Tom, while only one matches with those from Harry. It is therefore unlikely that Harry is Eileen's father.

Figure 23.187 An example of DNA profiling



An important group of enzymes is known as restriction endonucleases. They are usually just known as **restriction enzymes** and occur naturally in bacteria. They protect bacteria against viral DNA by cutting it into small pieces, thereby inactivating it. Many restriction enzymes have been isolated and purified from bacteria. Restriction enzymes 'cut' DNA at certain specific target sequences called restriction sites. Their action results in either blunt ends or sticky ends (Figure 23.188). Restriction enzymes can be used to 'cut out' specific genes from an organism and then, using other enzymes, the genes can be introduced into a different organism.

■ **Figure 23.188** The role of restriction endonucleases (restriction enzymes)


restriction enzyme
DNA strand in pH-buffered medium, incubated with restriction enzyme
restriction enzymes are of two types according to how DNA is cut
*Pvu*II from *Proteus vulgaris* cuts at CAGCTG, forming **blunt ends**
*Eco*RI from *Escherichia coli* cuts at GAATTC, forming **sticky ends**
hydrogen bonds between bases
C A C A G C T G A T
G T G T C G A C T A
G A A T T C
C T T A A G
restriction sites
blunt ends
C A C A G
G T G T C
C T G A T
G A C T A
sticky ends
G
C T T A A
A A T T C
G
restriction sites of other enzymes:
*Hpa*I
G T T A A C
C A A T T G
*Bam*HI
G G A T C C
C C T A G G
*Hpa*II
C C G G
G G C C
*Hin*dIII
A A G C T T
T T C G A A
restriction sites are four or six base pairs long

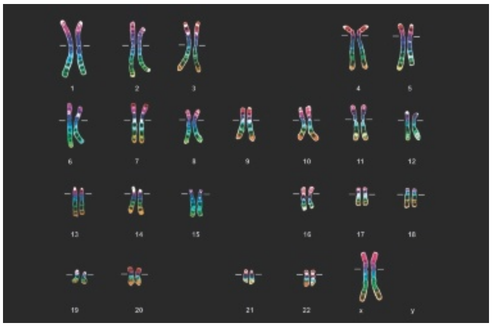
Restriction enzymes are named after the microorganisms they are found in. Roman numerals are added to distinguish different enzymes from the same microorganism.

Nature of Science

Owning the knowledge

The term genome refers to the all DNA sequences present in the chromosomes of a cell. The genome includes the genes and all of the non-coding sequences. In 1990 the Human Genome Project set out to identify all of the genes in human chromosomes (Figure 23.189) (around 30 000 of them) and to sequence the 3 billion base pairs which make up human DNA. As a result of advances in the technology used to sequence DNA, the task of producing the complete human genome was completed 2 years ahead of schedule in 2003. The Human Genome Project has been an example of successful international cooperation, with scientists in 18 countries all working on sequencing the DNA.

The success of the project has opened a range of possibilities, with the genomes of other species, from yeasts to mosquitoes, being targeted. In the UK, a government-backed project, the 100 000 Genomes Project, has been set up to sequence the whole genomes of participants by 2017. The bank of information generated by such a scheme offers clinical opportunities and has a range of medical and ethical implications.



■ **Figure 23.189** The set of human chromosomes

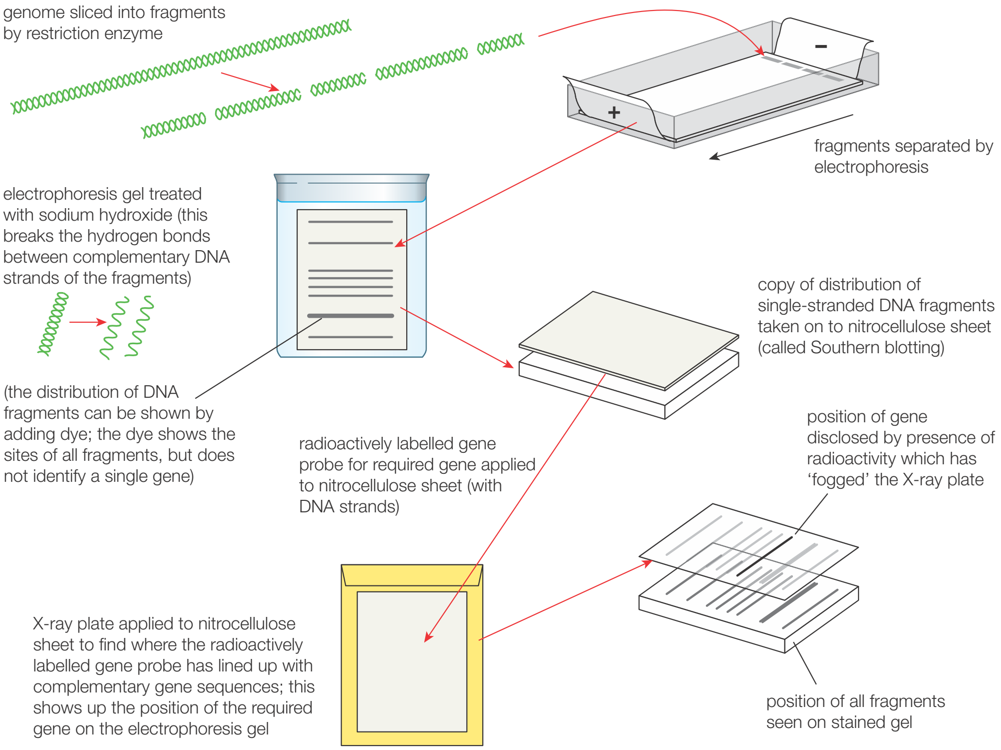
During the pursuit of the Human Genome Project tensions grew between business-funded research and the open public project. That issue has raised itself several times in the short history of the molecular biology 'story'. When asked if he and Crick would patent their discovery of DNA, Watson is reported to have laughed saying that 'there was no use for it'. Twenty years later when the process for making recombinant DNA was developed at Stanford University, USA, by Boyer and Cohen, the technique was patented – leading to a revenue of $255 million in the biotech industry before the patent expired in 1997.

For several years patents on genetically modified seeds and animals have been granted worldwide. There have been documented damaging impacts on famers, who are deprived of their right to save the seeds, and on breeders who can no longer freely use the patented seeds or further breeding. In North America, for example, the multinational seed company Monsanto has sued many farmers for alleged patent infringements. The possibility of patenting seeds has resulted in a highly concentrated market structure with only ten multinational companies controlling half of the international seeds. A number of farmers' organizations and non-governmental organizations are opposing these patents. Currently, as genetically modified organisms are still not grown or raised in most countries, the negative issues associated with these patents are not experienced in all countries. There is a trend for patents to be aimed not only at GM crops but also at conventional plants. For example, patent claims have been made for soy beans with a better oil quality covering part of the plant genome. Some of the most controversial examples are patent claims which cover large parts of the rice genome and its use in the breeding of any food crops that have similar genomic information to rice, such as maize and wheat.

This has opened many controversies, including the question of right of ownership of biological knowledge. What limits do you think should apply to patenting, which implies sole rights to use information?

Tests using gene probes (Figure 23.190) are being developed to diagnose inherited diseases. Some of the new treatments developed using knowledge from the Human Genome Project target precise cells in the body, such as cancer cells. Others work most effectively in people with a particular genetic make-up. In future, medicines may even switch genes on or off to control disease. In gene therapy the idea is to overcome genetic diseases by modifying the mutated DNA which causes disorders such as thalassemia and cystic fibrosis. Results from the Human Genome Project should help scientists make progress in this exciting new area of treatments.

Figure 23.190 Gene probe technology


genome sliced into fragments by restriction enzyme
–
+
fragments separated by electrophoresis
electrophoresis gel treated with sodium hydroxide (this breaks the hydrogen bonds between complementary DNA strands of the fragments)
copy of distribution of single-stranded DNA fragments taken on to nitrocellulose sheet (called Southern blotting)
(the distribution of DNA fragments can be shown by adding dye; the dye shows the sites of all fragments, but does not identify a single gene)
radioactively labelled gene probe for required gene applied to nitrocellulose sheet (with DNA strands)
position of gene disclosed by presence of radioactivity which has 'fogged' the X-ray plate
X-ray plate applied to nitrocellulose sheet to find where the radioactively labelled gene probe has lined up with complementary gene sequences; this shows up the position of the required gene on the electrophoresis gel
position of all fragments seen on stained gel

23.9 Biological pigments (AHL) *– biological pigments include a variety of chemical structures with diverse functions which absorb specific wavelengths of light*

One of the most pleasing natural sights, particularly in certain areas of the world, are the colours of woodland trees in the autumn (fall) (Figure 23.191). This colouration is in part due to the breakdown of green chlorophyll, which unmasks the already present orange and yellow pigments (carotenoids and xanthophylls) in the leaves. The red colours of the anthocyanins also appear at this time as their synthesis is initiated upon the breakdown of the chlorophyll.

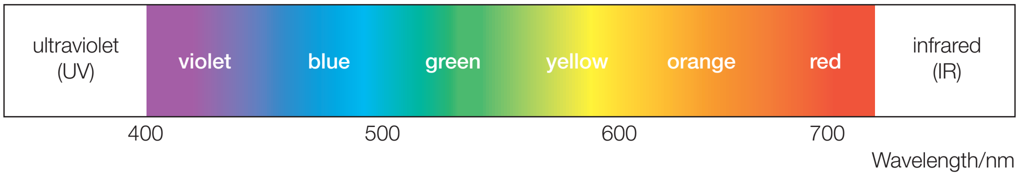
Figure 23.191 Green and red leaves on trees

Biological pigments are coloured compounds which are produced by metabolism. Many organic compounds are colourless, whereas others have very distinctive colours. Examples of naturally occurring coloured organic compounds include the anthocyanins, carotenoids (e.g. β-carotene) and the porphyrins (e.g. chlorophyll, hemoglobin and myoglobin). Their colour results from the absorption of certain wavelengths of visible light. They include the bright colours in the wings of insects and the feathers of birds, the wide variety of colours of flowers and seaweeds, and the chemicals that give colour to human skin, hair, eyes and blood. What does this diverse group of molecules have in common?

All pigment molecules have intense absorption bands in the visible region of the spectrum. The colour that we see is the light that is not absorbed but instead is reflected. For example, chlorophyll appears green because it absorbs red and blue light but reflects green.

Colour

Figure 23.192 The visible spectrum



The visible region of the electromagnetic spectrum lies between 400 nm and 750 nm. White light is a mixture of visible wavelengths or colours as shown in Figure 23.192. The colours gradually merge into each other. Any substance that reflects or transmits all of these wavelengths therefore appears white. An object that absorbs all visible light and transmits none looks black, for example graphite. Coloured substances are those that absorb only certain wavelengths of the visible spectrum (Figure 23.193).

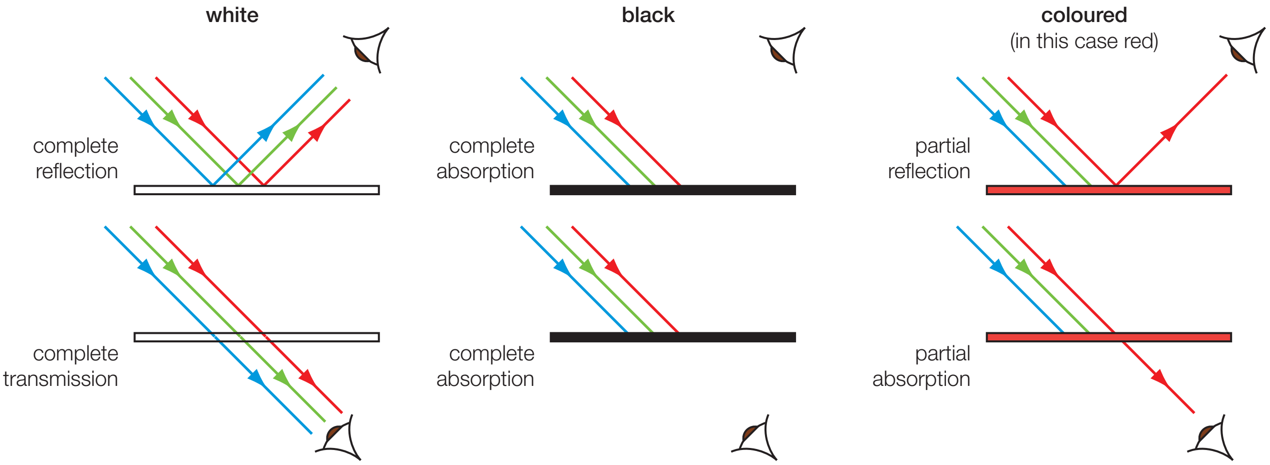


Figure 23.193 Colour and absorption of visible light wavelengths

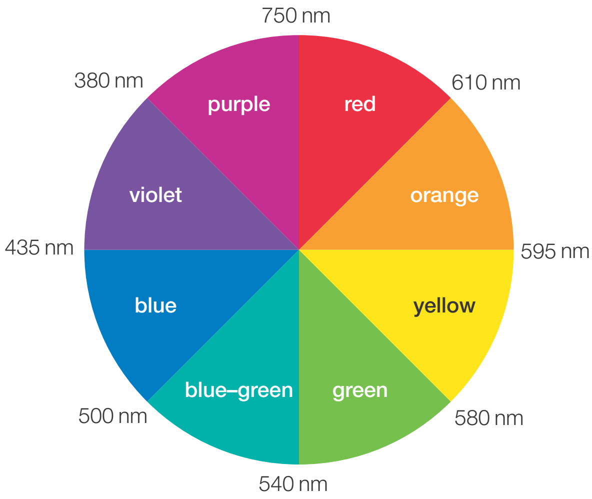
The colour perceived by the eye of the coloured compound is determined by which wavelengths of the visible light are reflected by the substance in question into the eye (Figure 23.194).

White light can be described as a mixture of red, green and blue light. These are known as the primary colours and when mixed together (in equal intensities) produce white light. All the colours can be generated from the three primary colours.

The colour that we see is white light minus the colour being absorbed. The colour we see is called the complementary colour to the colour being absorbed. A colour wheel (Figure 23.195) illustrates the approximate complementary relationship between the wavelengths of light absorbed and the wavelengths transmitted or reflected. For example, a blue substance will strongly absorb the complementary colour of light, orange. In this case, the absorption spectrum of a blue solution would have a maximum absorbance at a wavelength corresponding to orange light.

Colour of compound	Wavelength absorbed /nm	Colour of light absorbed
greenish yellow	400–430	violet
yellow to orange	430–490	blue
red	490–510	blue–green
purple	510–530	green
violet	530–560	yellow–green
blue	560–590	yellow
greenish blue	590–610	orange
blue–green to green	610–700	red

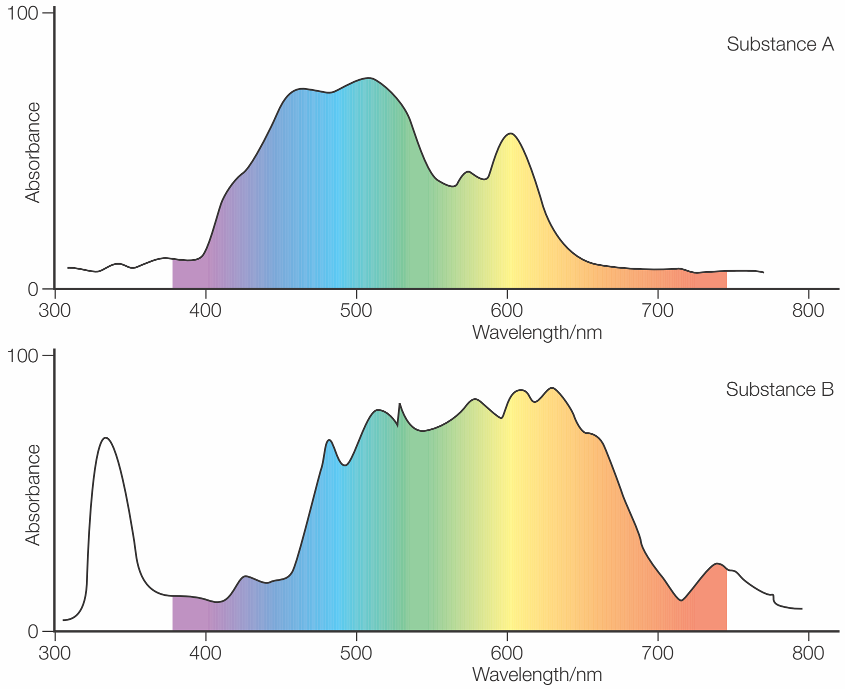
■ **Figure 23.194** A table of complementary colours. These are in the left- and right-hand columns – the colour of a compound is the colour complementary to the colour of the light it absorbs



■ **Figure 23.195** A simple colour wheel: complementary colours are opposite one another

Worked example

Absorption spectra for two coloured substances, A and B, are shown in Figure 23.196. What colour will these substances have? Substance A absorbs most light except red light and so appears red. Substance B absorbs most light except violet–blue and hence appears blue.



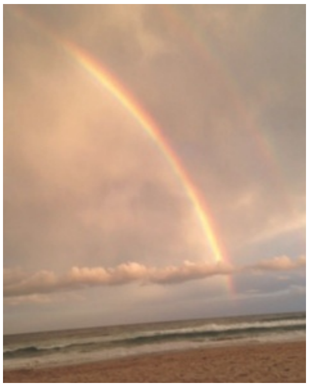
■ **Figure 23.196** Absorption spectra for substances A and B

Nature of Science

What is colour?

Colours play an important role in people’s lives. Our eyes detect the colours of objects and send nerve impulses to the brain providing us with a constant stream of sense perception information. The colour of a food determines how appetizing it appears and may indicate how fresh it is. Advertisers are aware of how we are physiologically influenced by colour. So what is colour?

We know from Chapter 2 that visible light is electromagnetic radiation with wavelengths between approximately 400 nm and 700 nm. Light can be viewed as an electromagnetic wave or a stream of ’packets’ of energy known as quanta. Isaac Newton was the first to realize that white light from the Sun is in fact made up of seven colours: red, orange, yellow, green, blue, indigo and violet. These colours are arranged in order of increasing frequency and energy, but decreasing wavelength.



■ **Figure 23.197** The colours of the spectrum shown in a rainbow

But it should be recognized that a rainbow (Figure 23.197) has an infinite number of different colours and that this classification into seven colours is somewhat arbitrary and subjective and based on Newton’s religious belief that seven was a special number.

Newton passed sunlight through a prism to obtain the visible spectrum. He also demonstrated that when this spectrum of colours was passed through a second prism it produced white light once more. One hundred and fifty years later, Thomas Young and George Palmer independently suggested that receptors in the eyes were sensitive to blue, green or red light and that different stimulation of these three colour receptors (cones) allows us to perceive all the different colours.

The three colours – blue, green and red – are called additive primary colours because they cannot be produced by the combination of other coloured lights. Traditional colour televisions make use of the three additive primary colours. The inside of the screen is coated with a large number of phosphor dots which glow either blue, red or green when they are struck by an electron beam. The dots combine to form coloured images.

The reason the world appears colourful is that many chemical compounds around us absorb and reflect different wavelengths from light that falls on to them. For example, chlorophyll in a leaf appears green because it absorbs red light and reflects light of other wavelengths. A substance appears coloured if it absorbs some but not all of the electromagnetic radiation from white light. When a compound absorbs wavelengths of one particular colour, a complementary colour appears. Pairs of complementary colours are represented on the colour wheel in Figure 23.195. The coloured wavelengths that are absorbed lie opposite their complementary colours.

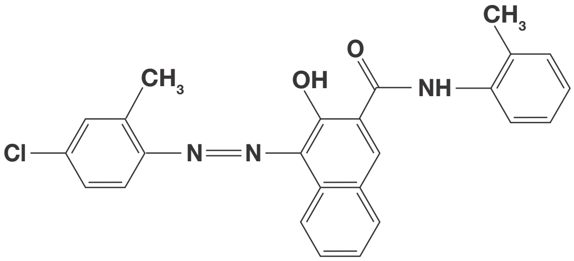
The colorimeter is a simple form of visible spectrometer. A fixed wavelength of visible light is selected using a coloured filter. This is then passed through the solution under test. The light that has not been absorbed by the solution is transmitted to a photocell. The light generates a small electric current which is measured by a meter. The more light that is transmitted by the solution, the greater the electric current produced. However, most colorimeters are calibrated so that they record the light absorbed by the solution rather than the light transmitted. The absorbance of a solution is proportional to the concentration of the coloured compound in the solution. Rates of reaction can be determined if one of the species in the reaction is highly coloured.

■ UV–Vis absorption

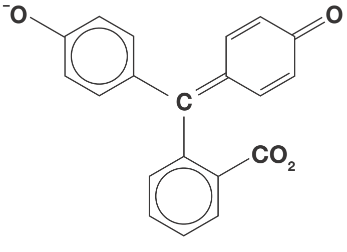
In order to absorb electromagnetic radiation in the ultraviolet–visual (UV–Vis) region of the spectrum, molecules must generally contain a double bond in the form of C=C, C=O or a benzene ring. These groups, which give rise to absorptions in the UV–Vis region, are called **chromophores**. Electromagnetic radiation in the UV–Vis region of the spectrum is absorbed to promote electrons from a low energy level (molecular orbital) in molecules to a higher energy level (molecular orbital). Molecules containing one or more double bonds will absorb ultraviolet radiation. The bonds may be carbon–carbon double bonds in alkenes and arenes, for example benzene, or carbon–oxygen double bonds in aldehydes and ketones. Molecules with conjugated systems absorb ultraviolet radiation and, often, visible light. Conjugated systems are molecules that have an alternating arrangement of carbon–carbon single and carbon–carbon double bonds. They are also termed delocalized systems.

■ **Figure 23.198** A rose – the colour is due to the presence of anthocyanins based on cyanidin 3,5-diglucoside

A number of biological molecules are highly conjugated, for example chlorophyll, anthocyanins (Figure 23.198), β-carotene, hemoglobin, retinol and myoglobin. The conjugation may be linear, for example β-carotene, or cyclic, for example chlorophyll. Artificial dyes, such as azo dyes (Figure 23.199), and acid–base indicators, such as phenolphthalein (Figure 23.200), are also conjugated systems.



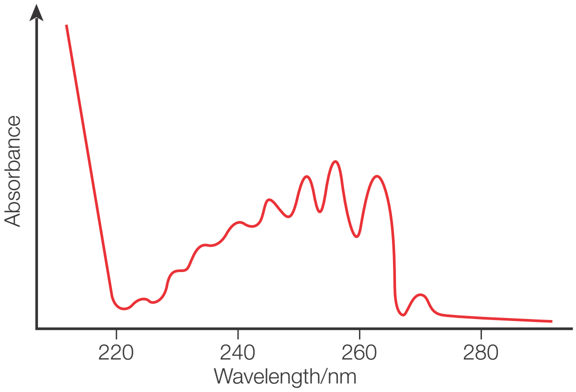
■ **Figure 23.199** Structure of a red azo dye



■ **Figure 23.200** Structure of the alkaline form of phenolphthalein

Ultraviolet spectra are usually recorded in solution. Water, ethanol and methanol are commonly used as solvents. Some solvents such as benzene cannot be used because they absorb ultraviolet radiation. Ultraviolet spectra can also be recorded for transparent solids and films and for gas phase molecules (Figure 23.201). The spectra (Figure 23.202) generally appear as broad bands with one or two humps. Chemists record the wavelength of the highest part of each peak and the relative degree of absorption.

■ **Figure 23.201** An ultraviolet–visible spectrometer



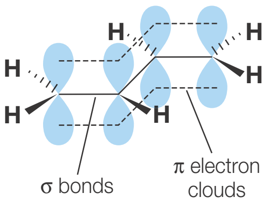
■ **Figure 23.202** Ultraviolet spectrum of benzene

The height of the absorption peak at a particular wavelength varies with the concentration of the absorbing substance. This means that ultraviolet spectroscopy can be used to estimate amounts of substance in solution for colourless substances (as can visible spectroscopy for coloured substances – see Section 23.8).

The effect of conjugation on the absorption of light by organic molecules

Table 23.24 summarizes the absorption of three related hydrocarbon molecules. Ethene contains a simple isolated carbon–carbon double bond, but the other two have conjugated carbon–carbon double bonds. In these cases, there is delocalization of the pi bonding orbitals over the whole molecule.

A 'conjugated' molecule is one that possesses alternating double and single bonds. Many conjugated molecules have alternating carbon–carbon double and single bonds. A consequence of conjugation is that the π electrons in the double bonds are 'spread' in a π molecular orbital that extends above and below over all the carbon atoms. Buta-1,3-diene is



■ **Figure 23.203** The conjugation of two alkene, >C=C<, chromophores in buta-1,3-diene

an example of a simple conjugated molecule with alternating carbon–carbon single and double bonds. The π molecular orbital is formed by the sideways overlap of unhybridized p_z orbitals and extends evenly over four carbon atoms (Figure 23.203).

The conjugation of additional π electrons within a molecule will shift the wavelength of the maximum absorption band (λ_{max}) from the ultraviolet region of the electromagnetic spectrum to the visible region – that is, towards longer wavelengths.

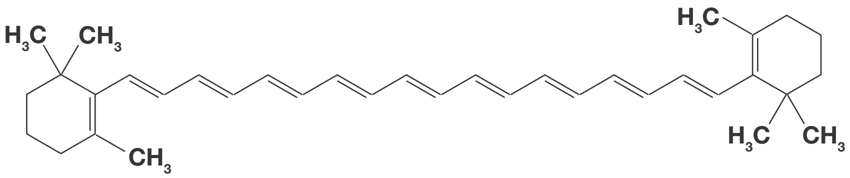
Name of molecule	Structure of molecule	Wavelength of maximum absorption, λ_{max}/nm
Ethene	$CH_2{=}CH_2$	171
Buta-1,3-diene	$CH_2{=}CH{-}CH{=}CH_2$	217
Hexa-1,3,5-triene	$CH_2{=}CH{-}CH{=}CH{-}CH{=}CH_2$	258

■ **Table 23.24** Ultraviolet absorptions of ethene, buta-1,3-diene and hexa-1,3,5-triene

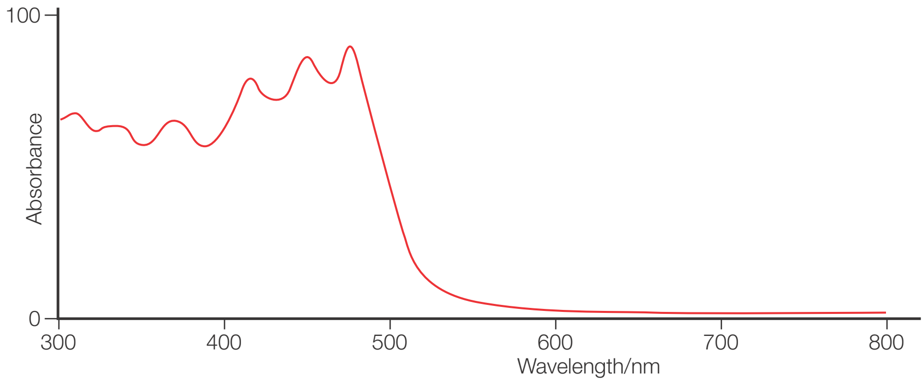
All three molecules in Table 23.24 give similar UV–Vis absorption spectra except that the absorption band in the ultraviolet region moves to a longer wavelength as the amount of delocalization in the molecule increases. Therefore the maximum absorption is moving to shorter frequencies as the amount of delocalization increases. Hence, absorption maxima move progressively to lower energy as the amount of delocalization increases.

If the conjugated system is very extensive then a number of the absorption bands may extend to the visible region of the electromagnetic spectrum and the compound will appear coloured. For example, β-carotene (Figure 23.204), the pigment responsible for the orange colour of carrots, has a conjugated chain containing 11 carbon–carbon double bonds. Figure 23.205 shows the visible spectrum of beta-carotene.

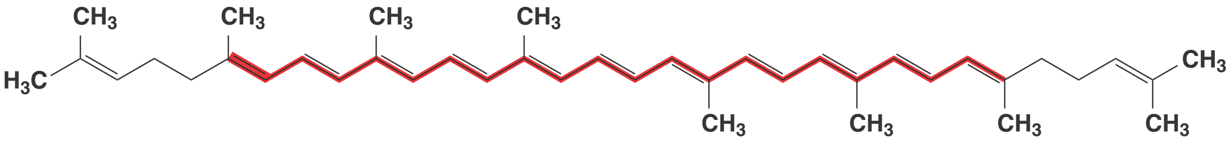
■ **Figure 23.204** The structure of β-carotene



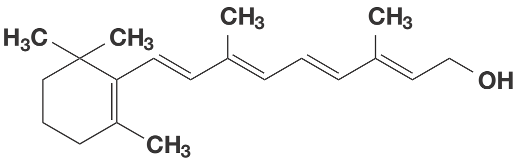
■ **Figure 23.205** The visible spectrum of β-carotene



Lycopene, which has a system of 11 conjugated double bonds (Figure 23.206), absorbs light in the blue–green part of the visible spectrum and therefore appears red. It is the red pigment in tomatoes.



■ **Figure 23.206** Lycopene, the red pigment in tomatoes, has 11 conjugated double bonds (highlighted). Note that not all the C=C bonds in lycopene are part of the conjugated system

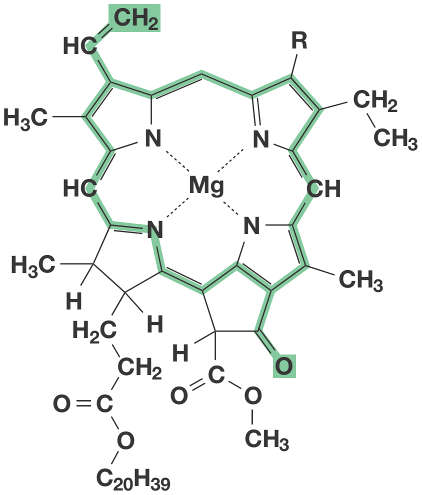


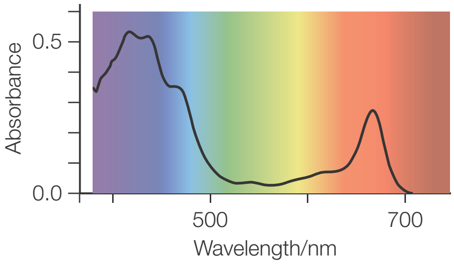
■ **Figure 23.207** Retinol has only five conjugated double bonds and absorbs only in the UV region of the electromagnetic spectrum

Retinol (Figure 23.207), however, only has a system of five conjugated double bonds and therefore does not absorb visible light (it only absorbs ultraviolet radiation) and is colourless.

Chlorophyll a and b have long conjugated systems (highlighted in Figure 23.208). They absorb light in the 400–500 nm region and in the 600–700 nm region (Figure 23.209). The green light in the middle of the spectrum is not absorbed, and so these molecules look green in natural light.

■ **Figure 23.208** The basic structures of chlorophyll a (R is CH_3) and chlorophyll b (R is CHO), showing the conjugated system. Chlorophyll b has an extra double bond that is part of the conjugated system





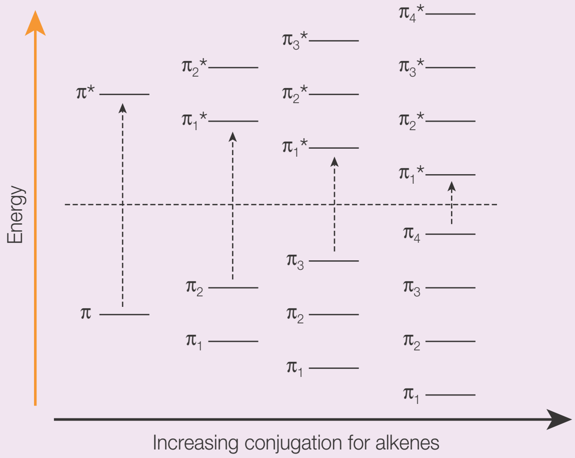
■ **Figure 23.209** The visible spectrum of chlorophyll

Additional Perspectives

Molecular orbital theory and conjugation

The increase in the length of conjugation in a carbon chain reduces the difference in energy between its pi bonding (π) and excited pi antibonding (π*) energy levels as shown in Figure 23.210. The energy required to produce a π → π* transition decreases and the wavelength of light increases. (This is known as a bathochromic shift.) Note that with increased conjugation, there is an increased number of π → π* transitions; hence multiple absorption bands will be observed for the ultraviolet spectrum of a highly conjugated system such as beta-carotene.

■ **Figure 23.210** Effect of increasing conjugation on the energy gap between π and π* orbitals



Nature of Science

The perception of colour – importance of quantitative data

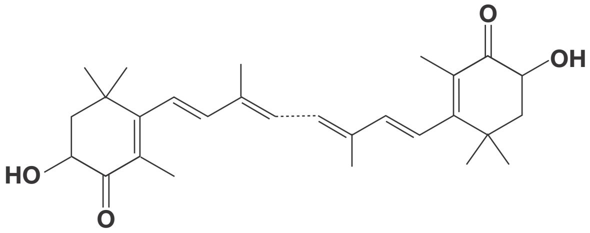
Quantitative data is very important in science and collection of data concerning the wavelength absorbed and the absorbance of a solution provides a lot more information than verbal descriptions just based on colour. Indeed, analysis of the wavelengths of light absorbed to produce a given colour goes towards answering the question as to whether we are all perceiving the 'same thing'. It provides an objectivity to the observations and a firmer basis for reliability than simple verbal descriptions.

As we have developed an understanding of how we see, and indeed of the various forms of colour-blindness, we have established a firmer reality for our descriptions of colour in whatever context.

Carotenoids

Carotenoids are the most widespread pigments in nature and are found in algae, where they provide protection against damage by light. They also act as accessory pigments and help to absorb blue light for photosynthesis. In humans, some carotenoids can act as a precursor for vitamin A synthesis (see Section 23.5). The dark greenish purple colour of the crab shell is caused by a complex of protein with the carotenoid astaxanthin (Figure 23.211).

Figure 23.211 The structure of astaxanthin



When the crab is boiled, the protein is denatured and the colour changes to the typical red colour of a carotenoid (Figure 23.212). Astaxanthin is also responsible for the pink colour of fresh salmon. The pink colour of flamingos is caused by carotenes absorbed from the algae in their diet (Figure 23.213)

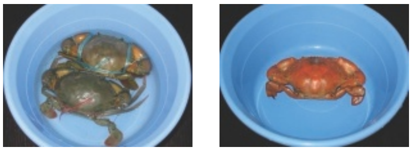


Figure 23.212 Fresh and boiled crabs

Figure 23.213 The pink colour of flamingos is caused by their dietary intake of algae

The carotenoids are found in many different plant groups including leafy green vegetables. Mixtures of carotenoids are often found, but in some fruits and vegetables, such as the mango and the carrot, β-carotene predominates. In the tomato the major carotenoid is lycopene, and in the red pepper it is capsanthin. The structures, classification and uses of carotenoids have been discussed previously.

Most carotenoids are derived from a 40-carbon polyene chain (multiple C=C double bonds). The ends of the chain may terminate in a cyclic (ring) group which may or may not have oxygen-containing functional groups attached. The carotenoids are divided into two groups:

- the carotenes, which are hydrocarbons
- the xanthophylls, which contain oxygen.

β-carotene is a carotene (see Figure 23.204), while astaxanthin is a xanthophyll (see Figure 23.211).

Carotenoids are highly unsaturated molecules and the presence of carbon–carbon double bonds makes them susceptible to chemical attack. Degradation pathways include isomerization, oxidation and decomposition of the carotenoid molecules. Heat, light and acids promote isomerization of the *trans* form of carotenoids to the *cis* form. Light, enzymes and reaction with hydroperoxides (from oxidation of unsaturated lipids) cause oxidation. This results in the bleaching of colour, the production of unpleasant odours and, for some carotenoids, loss of vitamin A activity.

Carotenoids are involved in light-harvesting in plants during photosynthesis – a photon of light is absorbed by a carotenoid molecule to promote an electron to an excited state. The energy it has absorbed is then transferred to chlorophyll.

Nature of Science

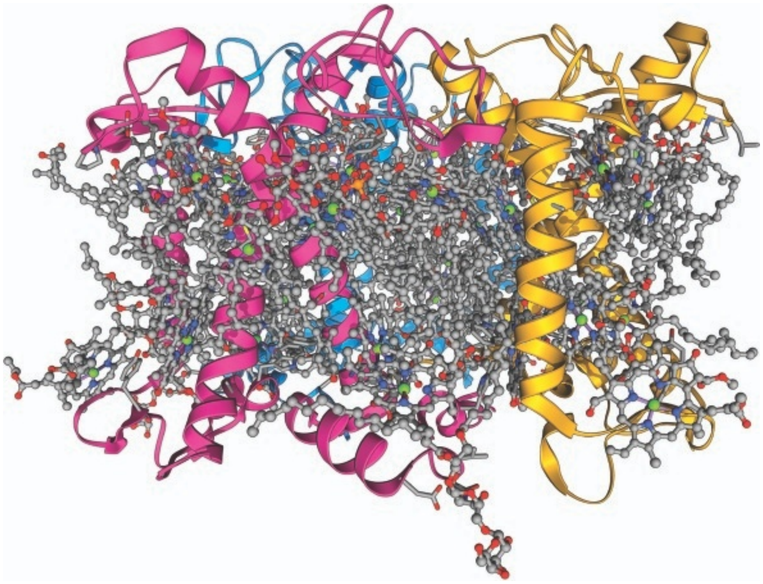
An evolutionary event

The appearance of photosynthesizing cyanobacteria on the early Earth is arguably one of the most significant events in the evolution of life on the planet (see Section 23.1). The success of these organisms took life in a new direction and changed the composition of the atmosphere. Inherent in that evolutionary step is the emergence of a viable system for harvesting light to provide energy for the cells.

Photosystem II (or water-plastoquinone oxidoreductase) is the first protein complex in the light-dependent reactions of oxygenic photosynthesis. It is located in the thylakoid membrane of plants, algae and cyanobacteria. Within the photosystem, enzymes capture photons of light to energize electrons that are then transferred through a variety of co-enzymes and co-factors to reduce plastoquinone to plastoquinol. The energized electrons are replaced by oxidizing water to form hydrogen ions and molecular oxygen. By replenishing lost electrons with electrons from the splitting of water, photosystem II provides the electrons for all of photosynthesis to occur.

Photosystem II (of cyanobacteria and green plants) is composed of around 20 subunits (depending on the organism) as well as other accessory light-harvesting proteins. Included in the complex are molecules of both chlorophyll a and β-carotene (Figure 23.214).

Figure 23.214 Structure of photosystem II



Carotenoids are generally poorly soluble in water, but freely soluble in non-polar organic solvents such as hexane. Carotenoids are essentially hydrocarbon in nature, despite the presence of polar functional groups. The properties of the polar functional groups, such as −OH, are outweighed by the much larger polyene backbone, which is hydrophobic. The dissolving of a carotenoid in water would be an energetically unfavoured process – strong hydrogen bonds between water molecules would be replaced by weaker interactions involving London dispersion forces. The only carotenoids that are soluble in water are those containing a carboxylic acid functional group, such as crocetin present in the spice saffron. It can form water-soluble salts in alkaline conditions.

Anthocyanins

The anthocyanins are a group of pigments responsible for the colours of flowers (Figure 23.215). The blue colour of cornflowers and the red colour of roses are both caused by the presence of anthocyanins. These pigments are also responsible for the colours in vegetables and fruits including cranberries, blueberries, strawberries, cherries, red apples, bilberries and raspberries (Figure 23.216).

Figure 23.215 Grape hyacinth, *muscari atlanticum*

Figure 23.216 Fresh raspberries

The anthocyanins are polyphenols and occur naturally as glycosides – the molecules are covalently bonded to a sugar. The aglycones, or the molecules without the attached sugar, are known as anthocyanidins. All anthocyanidins are flavanoids – their structure is based on the flavan structure shown in Figure 23.217. It has a C_6–C_3–C_6 skeleton with two conjugated benzene rings isolated by an oxygen-containing pyran ring.

Six different anthocyanidins occur in nature (due to the different atoms and functional groups represented by R^1 and R^2 in Figure 23.218) (see Table 23.25), but because these molecules will contain one or more sugar residues there are a huge number of different anthocyanins. Anthocyanidins are fully conjugated benzopyrylium (or flavylium) salts – note the positive charge on the oxygen atom.

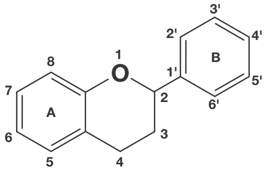


Figure 23.217 Structure of the flavan nucleus

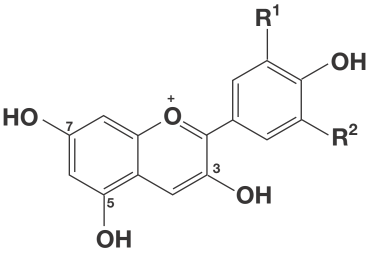


Figure 23.218 Generalized structure of the anthocyanidins

Name	R^1 functional group	R^2 functional group
Pelargonidin	–H	–H
Cyanidin	–OH	–H
Peonidin	$–O–CH_3$	–H
Delphindin	–OH	–OH
Petunidin	$–O–CH_3$	–OH
Malvidin	$–O–CH_3$	$–O–CH_3$

Table 23.25 The groups that contribute to the structures of the six different anthocyanidins

A number of different monosaccharides can be attached to the anthocyanins at a number of different positions. The sugar residue may also undergo a reaction with an organic acid to form an ester, so the range of different anthocyanins is quite large. Any single plant species will also contain a significant number of different anthocyanins.

Anthocyanins always have a sugar residue at position 3 (Figure 23.218), and glucose (or another monosaccharide, for example galactose) often occurs additionally at position 5 and occasionally at positions 7, 3' and 4'. The colours of anthocyanidins and anthocyanins are sensitive to factors such as pH, temperature and presence of metal ions. Quercetin (Figure 23.219) is one of the most bioactive of the flavonoids, and many medicinal plants owe much of their activity to their high quercetin content.

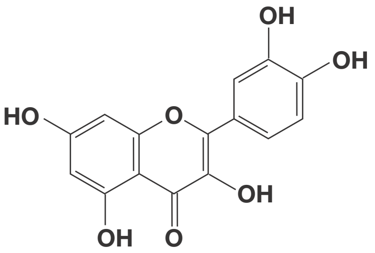


Figure 23.219 Structure of quercetin

Quercetin has demonstrated significant anti-inflammatory and antioxidant activity. Foods rich in quercetin include apples, honey, black and green tea, onions, raspberries, red wine, red grapes, citrus fruits, broccoli and cherries.

Anthocyanins are freely soluble in water, but poorly soluble in non-polar organic solvents. The presence of one or more sugar residues helps the anthocyanin maintain its solubility in water. The sugar residue contains one or more hydroxyl groups which can hydrogen bond with adjacent water molecules. If the sugar is hydrolysed, or lost, the solubility decreases.

The origin of the colour of anthocyanins

The absorbance spectra of anthocyanins are similar to those of cyanidin. This is the parent compound of anthocyanins but lacks the glucose residues. Under acidic conditions (low pH) thecyanidin is red. The absorbance spectrum shown in Figure 23.220 shows that cyanidin absorbs light at 375 nm and 530 nm, and compares the absorbance of the anthrocyanin to that of green chlorophyll a and b.

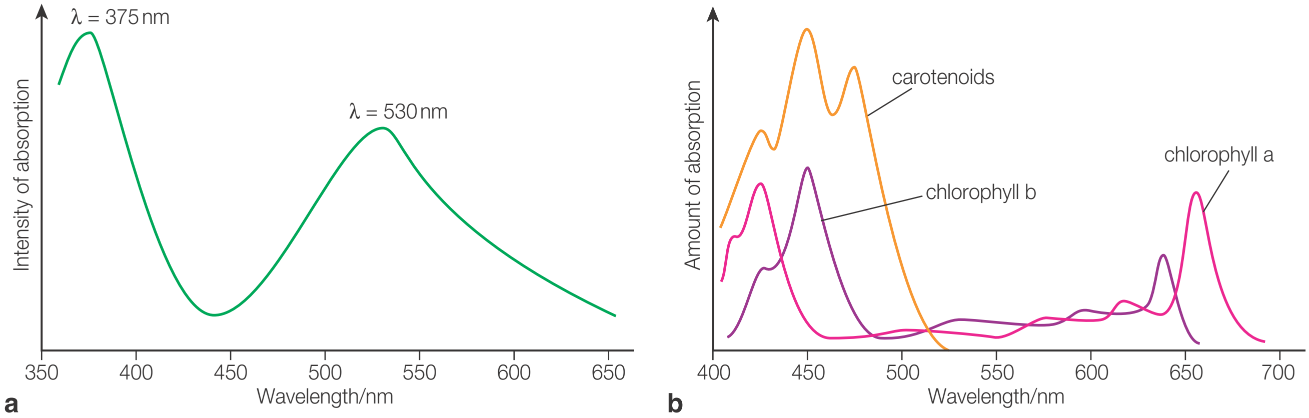


Figure 23.220 **a** Ultraviolet–visible spectrum of cyanidin compared to that of **b** chlorophyll a and b

The stability of anthocyanins

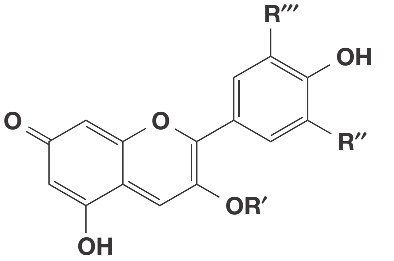
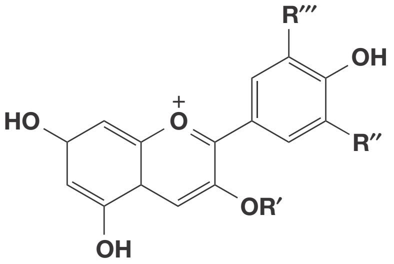
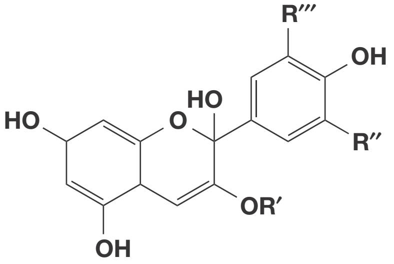
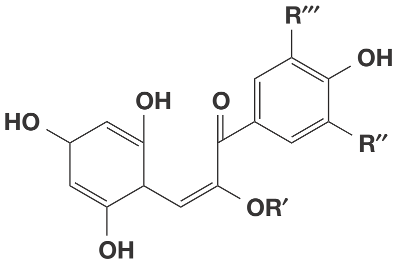
The range of colours exhibited by flowers petals and fruit is due, in part, to the presence of different mixtures of anthocyanins. However, the pH and presence of other substances are also strong influences on the range of colours.

In aqueous solution, anthocyanins exist in a complex equilibrium between four different structural forms (Table 23.26):

$$A \leftrightharpoons AH^+ \leftrightharpoons B \leftrightharpoons C$$

Which of these species predominates, and hence where the position of the equilibrium lies, depends on the pH and the temperature of the solution. At low pH the red flavylium cation form predominates. This is converted to the carbinol base form as the pH is increased. This has a shorter conjugated system than the flavylium cation and so only absorbs in the UV region of the spectrum and is therefore colourless. This species dominates at pH 4–5 and the colour of the mixture in this pH range will be quite pale. As the pH is increased the carbinol base form is converted to the yellow chalcone form. If the pH is increased still further, the purple quinoidal base form is also formed which is in equilibrium with its intensely blue-coloured anion. The complex equilibrium with its series of colour shifts is shown in Figure 23.221.

Table 23.26 The different forms of anthocyanin involved in the changes with pH

Anthocyanin	Structural form	Colour	Structure
A	Quinoidal base	Purple–blue	
AH^+	Flavylium cation	Red	
B	Carbinol base	Colourless	
C	Chalcone	Yellow	

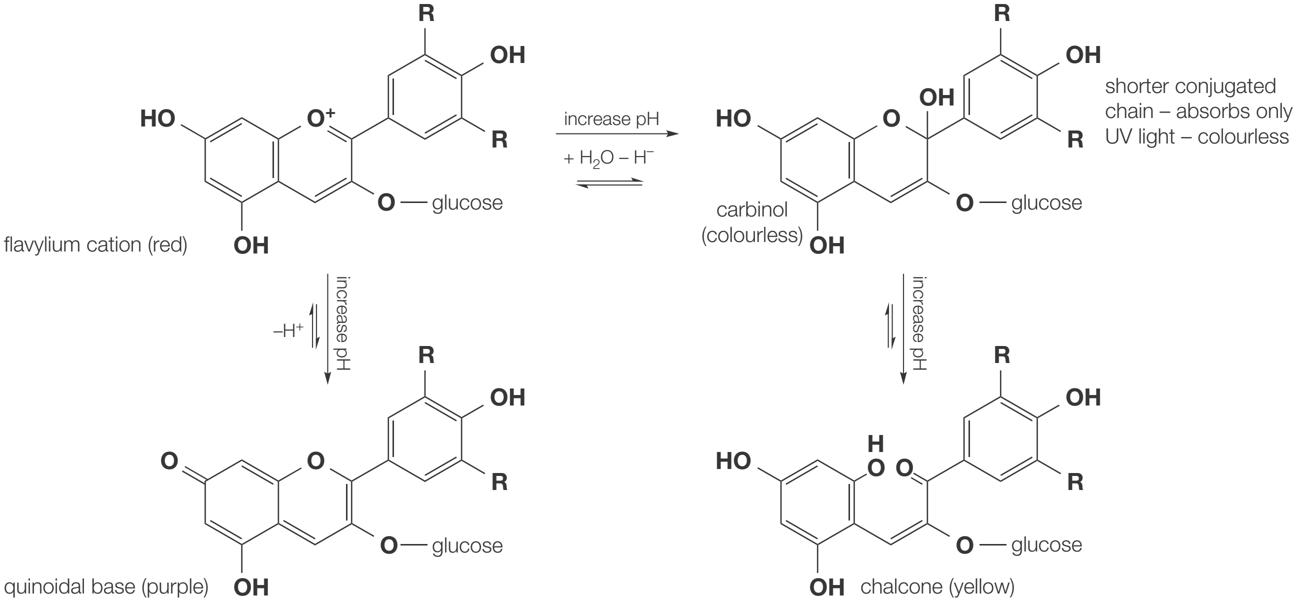


Figure 23.221 A complex equilibrium, which is very sensitive to pH, exists in an anthocyanin solution

Because the colour changes with pH, anthocyanins can be used as indicators. Many chemistry courses include the extraction of an indicator solution from red cabbage – the indicator extracted contains anthocyanins and demonstrates the colour changes just described. Figure 23.222 shows the range of colours that can be seen as the pH is increased (from left to right).

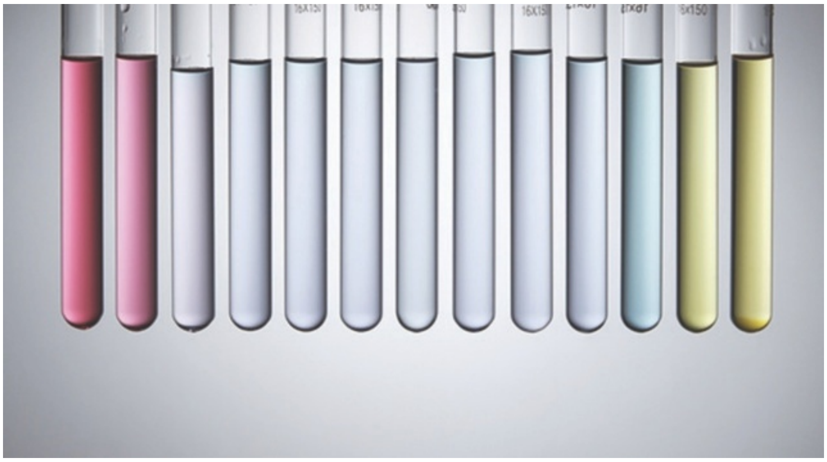


Figure 23.222 The colour changes observed with an anthocyanin indicator extracted from red cabbage as the pH is increased from low pH (left) to high pH (right)

Figure 23.223 The colour of the flowers on a hydrangea bush change colour in response to soil acidity

Anthocyanins also form complexes with metal ions – especially strongly polarizing ions such as $Fe^{3+}(aq)$ and $Al^{3+}(aq)$ – and these are responsible for the unusual colour developments observed in canned fruit. Interactions between the anthrocyanins in the petals of the hydrangea and the uptake of Al^{3+} ions from the soil is thought to cause the colour change that can be seen in the flowers in response to the acidity of the soil (Figure 23.223).

Temperature is another factor that affects the colour of anthocyanins. As the temperature is increased, the anthocyanin structure is increasingly destabilized and destroyed. There is a loss of initial colour and a brown colour develops.

Porphyrin rings in heme and chlorophyll

The principal structure in both heme and chlorophyll is based on a complex, planar macrocyclic unit called a porphin ring which contains a system of conjugated C=C double bonds. The term **macrocycle** can be used in many ways – here it means a large ring with multiple donor atoms (nitrogen atoms, each with a lone pair of electrons) that can bond to metal ions by coordinate bonding.

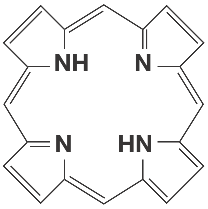


Figure 23.224 Structure of the porphin ring – a macrocycle

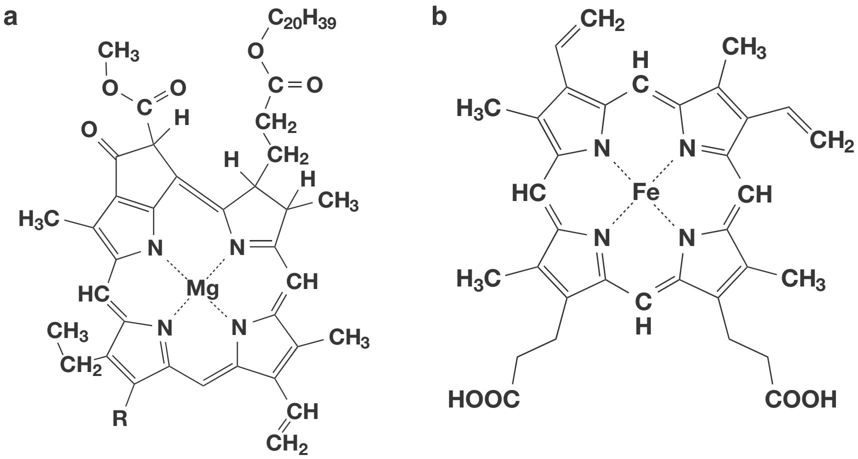
Porphin (Figure 23.224) is the parent compound for porphyrins, which are present in myogblobin, hemoglobin, vitamin B_{12} (see Section 23.5) and the chlorophylls.

Porphin is an organic compound which consists of four pyrrole rings joined by four =C– groups. Around the edge of the porphin cyclic system there is a conjugated system with extensive π electron delocalization. The ability of the porphin ring to form stable complexes with a variety of metal ions is often crucial to the biological properties of porphyrins.

Porphyrins are based on the porphin ring structure, but have a ring system which is more oxidized (by removal of hydrogen atoms) and has a number of functional groups attached. They still retain extensive conjugation and hence π delocalization, which makes them coloured.

Chlorophylls contain a porphyrin unit complexed to a central Mg^{2+} ion (Figure 23.225a), while heme is a complex between a porphyrin unit and an Fe^{2+} ion (Figure 23.225b).

Figure 23.225 The structure of the porphyrin unit in **a** chlorophyll and **b** heme



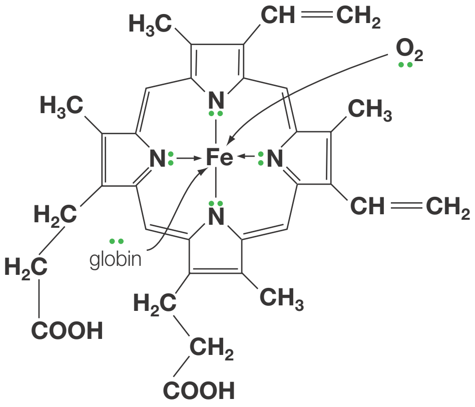


Figure 23.226 The structure of the heme group in hemoglobin

Heme in myoglobin and hemoglobin

Heme is present in the respiratory pigments hemoglobin and myoglobin. It acts as a prosthetic group in both myoglobin (the pigment in muscles) and hemoglobin (the pigment in red blood cells). In both the heme group is associated with a polypeptide chain in a 1 : 1 ratio. A hemoglobin molecule contains four protein chains and a myoglobin molecule contains one protein chain. Hemoglobin contains four heme subunits, while myoglobin contains one heme ring. Each heme ring contains one iron(II) ion (Figure 23.226).

Heme groups are responsible for the red and purple colours of oxidized hemoglobin and myoglobin. In the body, the iron in the heme group is coordinated to the four nitrogen atoms of the porphyrin ring, and also to a nitrogen atom from a histidine residue of the hemoglobin protein, known as globin. The sixth position (coordination site) around the iron of the heme is occupied by oxygen when the hemoglobin protein is oxygenated.

The binding of oxygen to hemoglobin

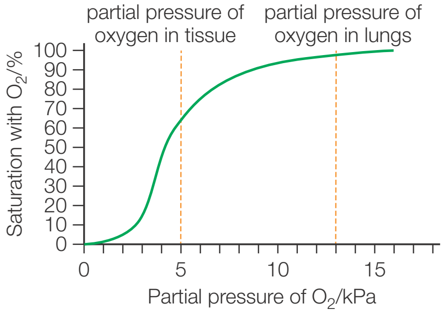
Hemoglobin transports oxygen from the lungs through the bloodstream and releases it to the cells of the tissues to carry out respiration.

Hemoglobin consists of four polypeptide subunits, each of which contains a heme prosthetic group with the iron at the centre of the heme having oxidation number +2. Each heme can carry one molecule of oxygen, so each hemoglobin unit can transport four molecules of oxygen simultaneously.

The iron in the heme can bond to six ligands. In the unbound state, the Fe^{2+} is bonded to five ligands – four are the nitrogen atoms of the porphyrin and the other is an amino acid (His) that attaches it to the protein. When molecular oxygen binds, this becomes the sixth ligand, and hemoglobin is said to be oxygenated (it is this oxygenated form that gives blood its red colour, because of the red colour of the heme prosthetic group). Binding of the oxygen molecules results in Fe^{2+} being oxidized to Fe^{3+}, though this is reversed when the oxygen is released. In hemoglobin, the oxygen binds reversibly, allowing its release to tissue cells to be used in cellular respiration.

The graph in Figure 23.227 shows how the affinity of hemoglobin for oxygen changes as the partial pressure of oxygen changes. The scale on the *y*-axis represents the fraction of iron ions bound to oxygen molecules. This is called an oxygen binding curve or oxygen dissociation curve.

Figure 23.227 The oxygen binding or dissociation curve for hemoglobin. The partial pressure of oxygen is the pressure of the oxygen in a mixture of gases



The type of curve in Figure 22.227 is described as sigmoidal. When the partial pressure of oxygen is low, hemoglobin has a low affinity for oxygen, but the affinity increases markedly as the partial pressure of oxygen increases – the gradient of the curve increases. This suggests that it becomes easier for oxygen to bind to hemoglobin when some oxygen molecules have already bound to the iron and that the binding of oxygen is cooperative. This cooperative effect is due to subtle changes that occur in the quaternary structure as oxygen binds. A conformational shift caused by the binding of oxygen at one heme group makes the other heme groups more receptive to oxygen. It is an allosteric effect. From the graph we can deduce the following about how this affects hemoglobin's ability to bind O_2.

- At low concentrations of O_2, hemoglobin has a low affinity for O_2.
- At high concentrations of O_2, hemoglobin has a high affinity for O_2.

In other words, the equilibrium shifts to the right in the lungs, causing oxygen uptake where oxygen concentration is high, and shifts to the left in respiring cells, releasing oxygen where oxygen concentration is low:

$$Hb(aq) + 4O_2(g) \leftrightharpoons Hb(O_2)_4(aq)$$

Simplified form: $Hb(aq) + O_2\ (g) \leftrightharpoons HbO_2(aq)$

This type of curve is important for the functioning of hemoglobin as an oxygen carrier – the affinity for oxygen is high when the blood passes through the lungs (high partial pressure of oxygen) so the hemoglobin binds lots of oxygen, but the affinity is much lower in the tissues and so the hemoglobin gives up the oxygen to the tissue (because it is bound to more oxygen than it can be at that partial pressure).

The effect of pH, carbon dioxide and temperature on the binding of oxygen by hemoglobin

Hemoglobin also transports H^+ ions and CO_2 molecules around the body. Both pH and concentration (partial pressure) of carbon dioxide affect the ability of hemoglobin to bind oxygen. As pH decreases ([H^+] increases) the affinity of hemoglobin for oxygen decreases. This can be represented by the equilibrium:

$$\underset{\text{oxygenated hemoglobin}}{HbO_2} \quad + \quad H^+ \quad \leftrightharpoons \quad \underset{\text{protonated form of deoxyhemoglobin}}{HbH^+} \quad + \quad O_2$$

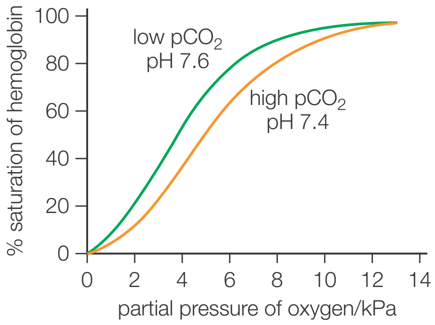


Figure 23.228 The effect of pH and carbon dioxide on the affinity of hemoglobin for oxygen

As the H^+ concentration increases, the position of this equilibrium shifts to the right and O_2 is released from the hemoglobin. The H^+ does not bind to the same site as the O_2 but rather to an amino acid side-chain; it is acting in a similar way to a non-competitive inhibitor of an enzyme. Binding H^+ changes the shape (conformation) of the protein slightly to reduce the affinity for oxygen (Figure 23.228).

The curve shows that decreasing pH reduces the affinity of hemoglobin for O_2, because the curve has shifted to the right as the acidity increases. Note that increasing the concentration of CO_2 has the same effect. Carbon dioxide produced in respiration in cells diffuses into red blood cells, where it dissolves to form an acidic solution (carbonic acid, H_2CO_3). This lowers the pH so that more H^+ binds to hemoglobin and causes a release of oxygen.

Hemoglobin also binds carbon dioxide, but not at the same site as O_2. This carbon dioxide reacts with the $-NH_2$ group on the terminal amino acid of each polypeptide chain that makes up hemoglobin (Figure 23.229). This process has two effects – it releases H^+ and also changes the shape of the protein; both of these reduce the affinity of hemoglobin for oxygen.

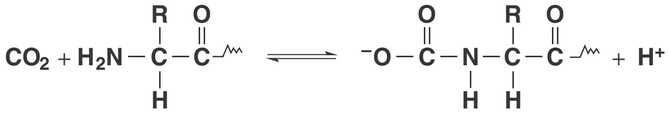
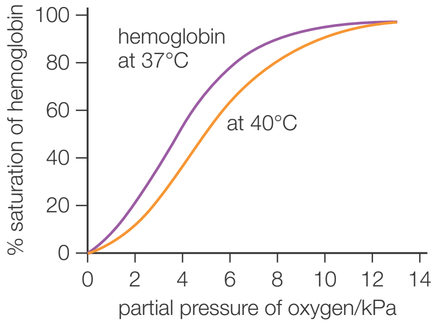


Figure 23.229 The reaction of carbon dioxide with a terminal amino acid of a polypeptide chain



■ **Figure 23.230** The effect of temperature on the affinity of hemoglobin for oxygen

The ability of hemoglobin to bind oxygen decreases as the temperature increases, as illustrated by the curves in Figure 23.230.

This suggests that the oxygenation process is exothermic, and that the deoxygenation process is endothermic:

$$\underset{\text{oxygenated hemoglobin}}{HbO_2} \leftrightharpoons \underset{\text{deoxygenated hemoglobin}}{Hb} + O_2 \quad \Delta H = +\text{ve}$$

Increasing the temperature causes the position of equilibrium to shift to the right – towards the deoxygenated form, causing the release of oxygen. This means that oxyhemoglobin more readily releases its oxygen in conditions of higher temperature, for example in cells during high metabolic activity such as exercise.

The equation below summarizes the factors that influence the equilibrium position in the oxygenation of hemoglobin:

$$Hb + 4O_2 \underset{\substack{\text{in respiring cells, low } pO_2 \\ \text{higher temperature} \\ \text{low pH/high } pCO_2}}{\overset{\substack{\text{in lungs, high } pO_2 \\ \text{lower temperature} \\ \text{high pH/low } pCO_2}}{\rightleftharpoons}} Hb\text{–}(O_2)_4$$

Factors that increase the affinity of hemoglobin for O_2 displace the oxygen dissociation curve to the left; factors that decrease hemoglobin's affinity for O_2 displace the oxygen dissociation curve to the right.

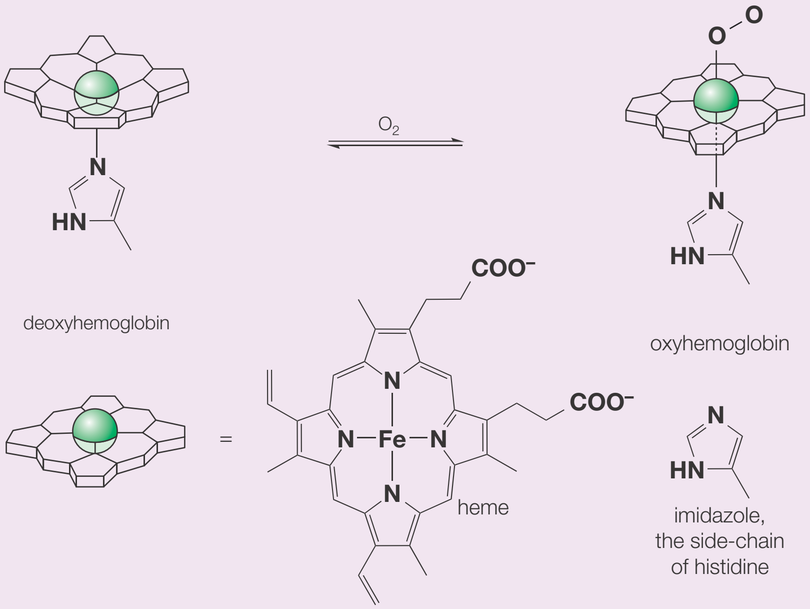
Additional Perspectives

The mechanism of oxygen transport

The mechanism of oxygen transport has been extensively studied. 'Naked' heme, in which the iron(II)–porphyrin ring is free of globin protein, is easily and irreversibly oxidized by molecular oxygen (O_2) to iron(III), forming a dimer in which the two Fe^{3+} ions are linked by a peroxide group, –O–O–. One of the functions of the globin protein chain is to prevent this oxidation to the dimer.

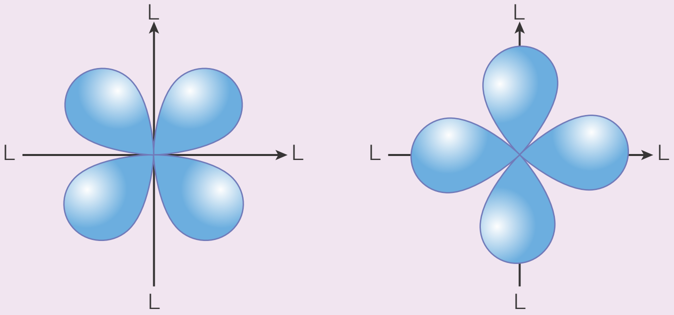
The oxygen-free form of hemoglobin contains a five-coordinate iron(II) ion in which four of the coordination positions are occupied by the nitrogen atoms of the porphyrin ring and the fifth by the nitrogen atom of a histidine in the globin protein (Figure 23.231).

■ **Figure 23.231** The binding of oxygen to deoxyhemoglobin



Crystal field theory (see Chapter 13) is required in order to understand what happens when hemoglobin transports oxygen. This simple bonding theory assumes that when a cation interacts with ligands (lone pair donors) electrostatically, it attracts an anion or the negative end of a molecule of a polar ligand.

Ligands in the horizontal *x*–*y* plane (containing the four nitrogen atoms of the porphyrin ring) will repel the electrons in the $d_{x^2-y^2}$ orbital more strongly because the lobes of that orbital lie along the *x* and *y* axes (Figure 23.232). Similarly, the ligands along the *z* axis (the nitrogen of the histidine amino acid) repel electrons in the d_{z^2} orbital, but there is less repulsion because there is only one ligand. The other d orbitals (d_{yz}, d_{xz} and d_{xy}) have lobes that lie between the axes, so they interact less strongly with the ligands. As a result, the d sub-shell is 'split' into three sets of d orbitals.



Ligands on the x and y axes do not interact very strongly with the d_{xy} orbital

but the $d_{x^2-y^2}$ orbital is aligned along the axes, so it interacts more strongly with the ligands

Figure 23.232 The interaction between ligands and the d orbitals of a metal ion

The iron(II) ion in hemoglobin has the condensed electron configuration [Ar] $3d^6$ and the six d electrons arrange themselves into a high-spin arrangement. The diameter of a high-spin Fe^{2+} ion is a little larger than the 'central' hole of the porphyrin ring, so it sits just above it.

When the sixth coordination position is occupied by an oxygen molecule an octahedral complex is formed containing a Fe^{3+} ion, and the splitting of the d orbitals is changed and increased (Figure 23.233).

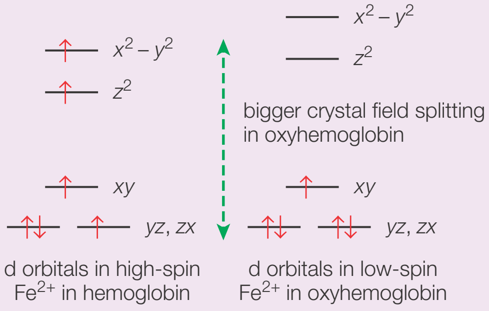

x² – y²
z²
xy
yz, zx
x² – y²
z²
bigger crystal field splitting in oxyhemoglobin
xy
yz, zx
d orbitals in high-spin Fe²⁺ in hemoglobin
d orbitals in low-spin Fe²⁺ in oxyhemoglobin

Figure 23.233 Crystal field splitting in oxyhemoglobin

The Fe^{3+} ion has a d^5 configuration and because the splitting of the d orbitals has increased, it becomes energetically favourable for the d electrons to rearrange themselves, with all of them paired up in the lower orbitals (the low spin arrangement).

This low-spin Fe^{3+} ion is much smaller than the high-spin Fe^{3+} ion and it can now fit into the hole in the porphyrin ring. As the Fe^{3+} ion drops into the hole of the porphyrin ring, it pulls the nitrogen of the histidine ring with it, altering the shape of the globin protein chain, which increases its affinity for oxygen.

The effects of altitude

In the human body, many chemical equilibria are maintained to ensure the health and functioning of the human body. If environmental conditions change, the body must adapt to keep functioning. The consequences of a sudden and large change in altitude demonstrate this fact. This may happen during a flight or mountain climbing and can cause headache, nausea, extreme tiredness and other discomforting symptoms. These conditions are all symptoms of hypoxia, a deficiency in the amount of oxygen reaching the cells of body tissues. However, a person living at a high altitude for several months gradually recovers from altitude sickness and adjusts to the lower oxygen concentration in the atmosphere.

The reversible combination of oxygen with hemoglobin can be represented by the simplified equation:

$$Hb(aq) + O_2(g) \rightleftharpoons HbO_2(aq)$$

where HbO_2 is oxyhemoglobin and the equilibrium expression constant is

$$K_c = [HbO_2(aq)]/[Hb(aq)]\ [O_2(g)]$$

According to Le Chatelier's principle, a decrease in oxygen concentration will shift the equation shown in the equation above from right to left. This change decreases the supply of hemoglobin causing hypoxia. Given sufficient time, the body deals with this problem by synthesizing more hemoglobin molecules. The equilibrium will then gradually shift back towards the formation of oxyhemoglobin. It takes several weeks for the increase in hemoglobin production to meet the body's basic metabolic needs adequately. The number of red blood cells in the blood increases during acclimatization to high altitudes.

People who live permanently at high altitude (Figure 23.234), such as in the Andes in South America or the Himalayas, show a number of adaptations to their low-oxygen environment. They are not genetically different from people who live at low altitudes, but their exposure to a low-oxygen environment from birth stimulates the development of these adaptations from an early age. They often have especially broad chests, providing larger than normal lung capacities. The heart is often larger than in a person who lives at low altitude, especially the right side, which pumps blood to the lungs. They also have more red blood cells and hemoglobin to increase the efficiency of oxygen transport.

■ **Figure 23.234** Living in a mountain environment

The effect of carbon monoxide

Carbon monoxide, CO, is a toxic gas as it is a better ligand than oxygen, binding to the iron ions in hemoglobin more strongly than oxygen does, so carbon monoxide is a competitive inhibitor for the binding of oxygen to hemoglobin. Its affinity for hemoglobin is about 200 times that of oxygen, and so it effectively makes the hemoglobin unavailable to carry oxygen to respiring cells. The lone pair on the carbon atom in carbon monoxide binds to the iron ion.

This can be compared with the effect on hemoglobin of CO_2 and H^+ which is non-competitive because they do not bind at the same site as oxygen. Binding of carbon monoxide to hemoglobin forms carboxyhemoglobin, which does not readily dissociate.

Carbon monoxide poisoning can occur from the burning of fossil fuels with insufficient ventilation and from smoking tobacco.

Fetal hemoglobin

Fetuses (embryos) in the womb have a different type of hemoglobin (hemoglobin F) to adult humans (hemoglobin A). Fetal hemoglobin has a higher affinity for oxygen under the same conditions. This is important because it allows the transfer of oxygen from the mother's hemoglobin to that of her fetus.

The four polypeptide chains in adult hemoglobin are two α-chains and two β-chains ($\alpha_2\beta_2$). Before birth hemoglobin in the fetus has a different structure with two α-chains and two γ-chains

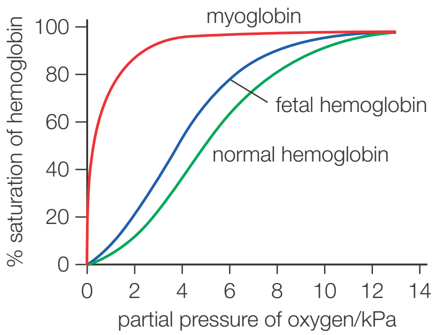


Figure 23.235 Dissociation curves for fetal hemoglobin and myoglobin

($\alpha_2\gamma_2$). Fetal hemoglobin has a higher affinity for oxygen under the same conditions (Figure 23.235). The fact that its oxygen dissociation curve lies to the left of adult blood indicates that it is able to extract oxygen from the maternal blood. Following birth the fetal hemoglobin levels decline and after 6 months adult hemoglobin becomes the predominant form.

Myoglobin also has an oxygen dissociation curve to the left of that of hemoglobin, which means it has greater affinity for oxygen and can pick it up from hemoglobin for storage. The dissociation curve for myoglobin is not sigmoidal in shape as there can be no cooperative binding within its one heme structure.

The oxygen bound to the iron(II) in myoglobin is responsible for the red colour of raw meat. When meat is cooked this oxygen is replaced by water and the meat loses its original colour and turns brown.

Chlorophyll

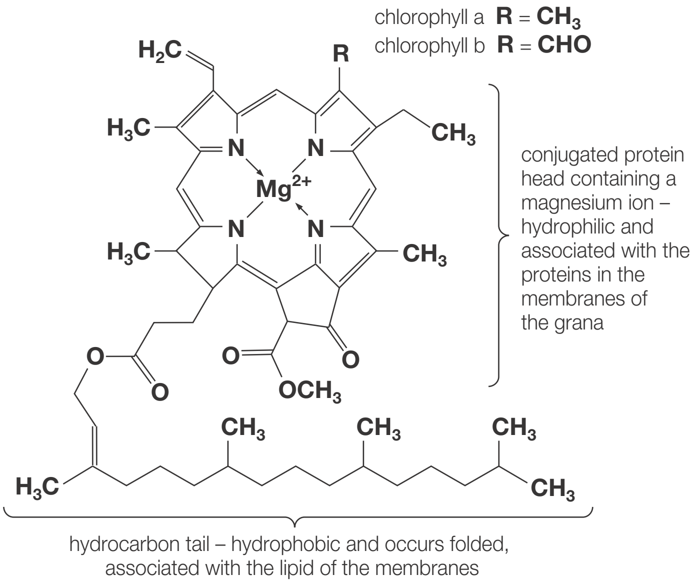
The chlorophylls are the green pigments of leafy vegetables – they also give the green colour to the skin of apples and other fruits, especially when they are unripe. Chlorophylls are involved in absorbing the light needed in the process of photosynthesis.

Chlorophyll (Figure 23.236) occurs in plants in two forms:

- chlorophyll a – blue–green
- chlorophyll b – yellow–green.

Chlorophyll b differs from chlorophyll a by having an aldehyde (–CHO) group in place of a methyl group (–CH_3). It is a porphyrin pigment, composed of four pyrrole rings linked to form a tetrapyrrole, with a magnesium ion complexed in the centre of the rings.

Figure 23.236 The structure of chlorophyll


chlorophyll a R = CH3
chlorophyll b R = CHO
H2C
R
H3C
CH3
N
N
Mg2+
N
N
H3C
CH3
O
O
O
O
OCH3
CH3
CH3
CH3
H3C
CH3
conjugated protein head containing a magnesium ion – hydrophilic and associated with the proteins in the membranes of the grana
hydrocarbon tail – hydrophobic and occurs folded, associated with the lipid of the membranes

Chlorophyll absorbs light strongly in the blue part of the spectrum and to a lesser extent in the red (see Figure 23.209). Other photosynthetic pigments, known as accessory pigments, harvest light in different parts of the spectrum and pass their energy to chlorophyll. As a result, chlorophyll undergoes a redox change, passing electrons to a series of electron transport carriers. Ultimately, chlorophyll is reduced back to its original state by gaining electrons from water, the water being oxidized to oxygen. This oxygen is released to the atmosphere. The process stores 'reducing power' which is able to reduce carbon dioxide to carbohydrate in reactions that do not depend on light energy (Figure 23.237).

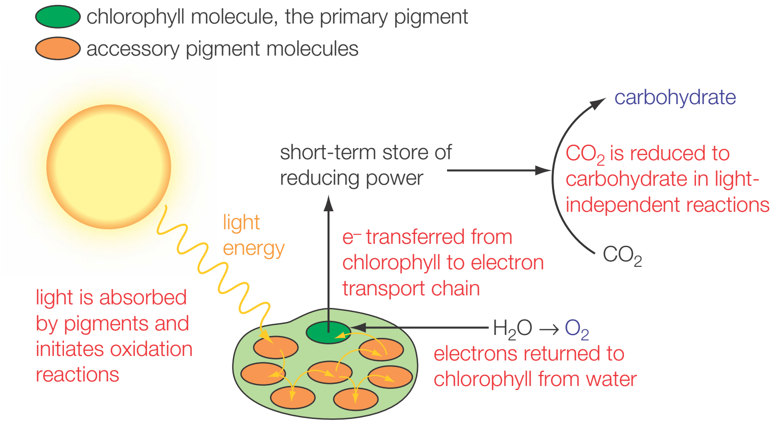

chlorophyll molecule, the primary pigment
accessory pigment molecules
carbohydrate
short-term store of reducing power
CO_2 is reduced to carbohydrate in light-independent reactions
light energy
e⁻ transferred from chlorophyll to electron transport chain
CO_2
light is absorbed by pigments and initiates oxidation reactions
$H_2O \rightarrow O_2$
electrons returned to chlorophyll from water

Figure 23.237 The absorption of light by photosynthetic pigments initiates redox reactions that ultimately lead to the storage of energy as carbohydrate

The thermal stability of chlorophyll is pH dependent. In acidic solution magnesium is lost from the porphyrin ring and replaced by two H^+ ions. This causes a colour change from green to olive-brown as the chromophore is altered. Cooking food often breaks cell membranes, releasing acids which decrease the pH and bring about this colour change. Chlorophyll is more stable in alkaline conditions, which is why sodium hydrogencarbonate is sometimes added to water during cooking. The bright green colour of chlorophyll is often used as an indication of the freshness of food.

Cytochromes: electron transport carriers

Molecular oxygen, which is supplied to tissues of the body by hemoglobin, is reduced to water during the final step of aerobic respiration (see Section 23.1). This takes place in the mitochondria and involves a group of enzymes known collectively as cytochromes.

Cytochromes are a varied group of protein molecules that also contain the heme prosthetic group. They are found embedded in mitochondrial membranes and are responsible for electron transport during the redox reactions of aerobic respiration and photosynthesis. During the reactions they become successively reduced and then re-oxidized as they in turn accept and then pass on electrons. They are organized in sequence, corresponding to their electrode potentials, so that the electrons effectively flow down an electrochemical gradient.

In cytochromes the iron of the heme group interconverts its oxidation state between +2 and +3 as the cytochrome undergoes redox change. The final cytochrome involved in aerobic respiration passes its electrons to the terminal acceptor oxygen with the formation of water:

$$4Fe^{2+}{}_{(\text{cytochrome c})} + 4H^+ + O_2 \rightarrow 4Fe^{3+}{}_{(\text{cytochrome c})} + 2H_2O$$

This is also the site of inhibition of the poison cyanide. By blocking the electron transport chain it prevents aerobic respiration from occurring, which is why it is such a potent poison.

Historical use of dyes and pigments

Dyes and pigments are coloured substances that are different in one important way: dyes are soluble in the liquid in which they are applied and pigments are insoluble. Dyes and pigments have been used by humans since early times. Some cave paintings discovered in southern France and northern Spain are up to 30 000 years old. The artists used mineral pigments to colour them: iron(III) oxide, Fe_2O_3, provided the red colour, iron(II) carbonate, $FeCO_3$, provided the yellow colour and carbon (charcoal) or manganese(IV) oxide, MnO_2 provided the black colour. Pigments are spread as a surface layer in paints or in coloured plastic articles. Artists' paints used to contain hazardous lead and chromium compounds. They have been replaced by safer red and yellow organic pigments. In contrast to pigments, dyes are soluble. Dye molecules attach themselves to the molecules of the substance they are colouring, sometimes by ionic or covalent bonds but often by London dispersion forces. The dyeing of fabrics has a long history and until the 19th century dyes came from animals or plants. One highly prized dye from the Roman

Empire was Tyrian purple obtained from a mollusc. Only members of the Emperor's family could wear clothes dyed with Tyrian purple, so it became known as royal purple.

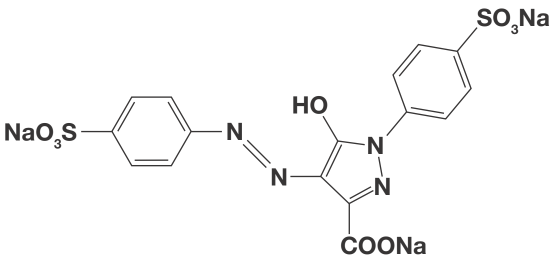


Figure 23.238 The structure of tartrazine (E102)

Investigation of pigments by chromatography

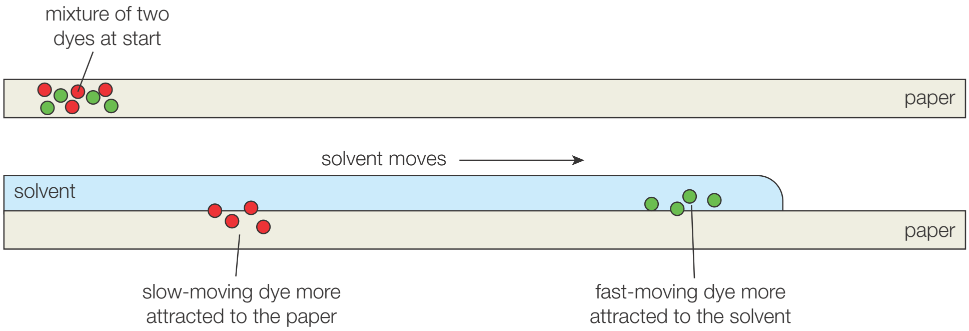
Dyes and pigments are highly appropriate candidates for separation by chromatographic techniques. Such methods can be applied readily to the biological pigments described in this section and to the separation of synthetic food dyes.

A food dye is a food-grade synthetic water-soluble colour – many food dyes are used as food additives. Natural food dyes include caramel, chlorella, saffron and paprika. Artificial food dyes include tartrazine (Figure 23.238), sunset yellow FCF and fast green FCF.

Paper chromatography

Paper chromatography is the simplest of the chromatographic techniques, and is often used to separate dyes in a mixture. In this technique the stationary phase is made up of water molecules that are trapped in the cellulose fibres of paper. The mobile phase is the aqueous or organic solvent that moves up the paper by capillary action. This capillary action is caused by the forces between the cellulose fibres of the paper and the solvent. Dyes that are more soluble in the solvent than they are in the water molecules of the stationary phase move rapidly up the paper, while those that are more soluble in the water are not carried as far up the paper (Figure 23.239).

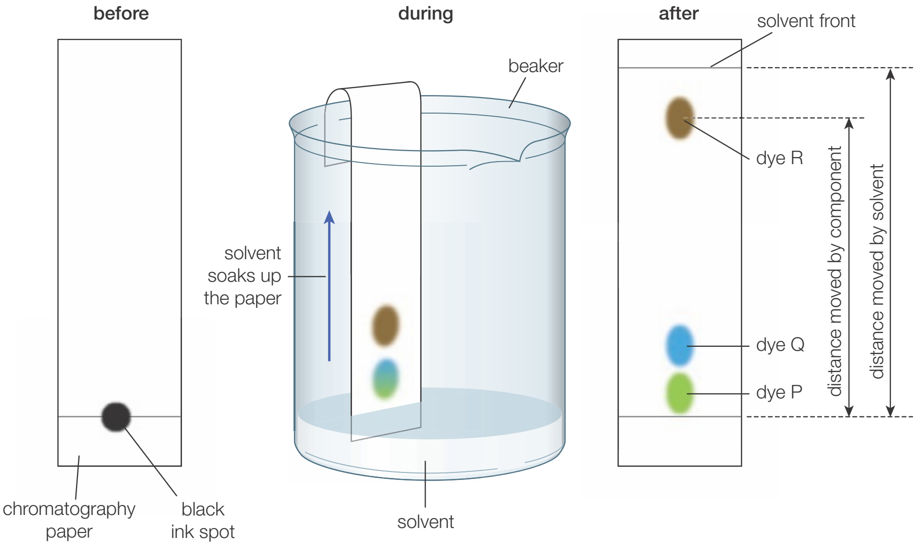
Figure 23.239 The principles of paper chromatography


mixture of two dyes at start
paper
solvent moves
solvent
paper
slow-moving dye more attracted to the paper
fast-moving dye more attracted to the solvent

When the solvent front has almost reached the top, the paper is removed and left to dry in a fume cupboard (Figure 23.240). For each of the different components, a retention factor (or R_f value) can be calculated:

$$R_f = \frac{\text{distance moved by component}}{\text{distance moved by solvent}}$$

Figure 23.240 Ascending paper chromatography



Standard retention factors are tabulated for a wide variety of substances, using particular solvents under standard conditions. These allow the identification of unknown components by comparison with these standard retention factor values. Note that the retention factor is not dependent on the distance travelled by the solvent. It is dependent on the nature of the mobile and stationary phases as well as on the components of the substance being separated.

It is sometimes necessary in these cases to carry out a variation of simple chromatography. After the initial separation process the paper is allowed to dry and then rotated 90 degrees. A different solvent is used to repeat the process after the paper has been rotated. This technique is known as two-dimensional chromatography and allows the separation of complex mixtures (see also Section 23.2).

The substances to be separated do not have to be coloured. Colourless substances can be made visible by spraying or treating the chromatogram with a locating agent. For example, ninhydrin (Figure 23.241) can be used as a locating agent for amino acids produced by the hydrolysis of proteins. The locating agent reacts with the colourless substance, either to form a coloured product or to allow the spots to be located by, for example, using ultraviolet light.

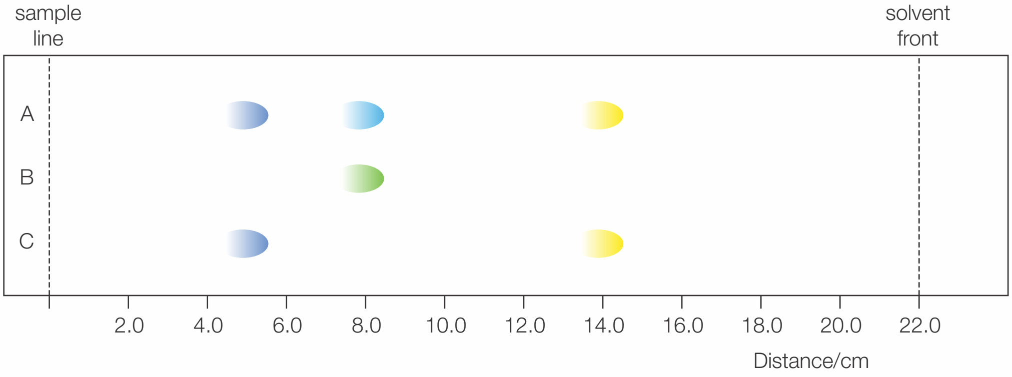
■ **Figure 23.241** Ninhydrin and spray bottle

Sugars can be detected by allowing a paper chromatogram to dry and then pulling it through a solution of silver nitrate in aqueous propanone. The solvent is allowed to evaporate and the paper sprayed with sodium hydroxide in aqueous ethanol. Reducing sugars, for example glucose and maltose, produce black spots of silver. Unreacted silver ions are removed by immersing the paper in ammonia solution.

If required, the individual components can be extracted from the paper by cutting out the individual spots using scissors and then extracting the components using a suitable solvent.

Worked example

Paper chromatography was performed on three green food colourings, A, B and C. The chromatogram (Figure 23.242) was developed at a temperature of 25 °C using a solvent consisting of 60% by volume propanone in water.



■ **Figure 23.242** Chromatograms for three green food colourings

1 Which of the food colourings are mixtures and which could be a pure compound?

2 Calculate the retention factor for food colouring B.

3 The food colouring known as sunset yellow has an R_f of 0.64 (under the stated conditions, on the same paper strip). Could any of these samples contain this compound?

4 Explain whether A could be a mixture of B and C.

1 Food colouring B may be a pure compound (one spot), but the other food colourings, A and C, are mixtures (three and two spots).

2 The retention factor of B = $\dfrac{\text{distance travelled by B}}{\text{distance travelled by solvent front}}$

$$= \frac{8.0}{22.0} = 0.36$$

3 $R_f \text{ value} = \dfrac{\text{distance travelled by component}}{\text{distance travelled by solvent front}}$

$0.64 = \dfrac{\text{distance travelled by component}}{22\,\text{cm}}$

Hence, sunset yellow will travel 14 cm. Food colourings A and C both contain this compound.

4 A could contain C as it has spots with identical colour and position. However, it cannot contain B since although the spot is in the correct position, it has a different colour and so cannot be the same dye.

Thin-layer chromatography

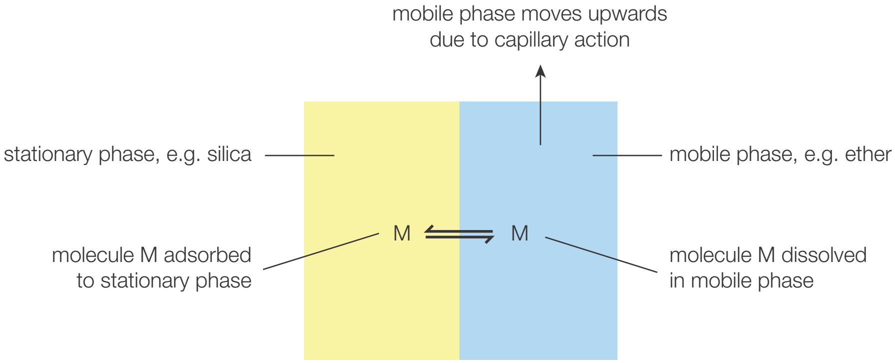
To prepare thin-layer chromatography (TLC) plates, the silica or alumina is first heated to a high temperature so that all the water is removed from it. The compounds then act as polar solids and the solutes are transferred from the liquid mobile phase by adsorption on the surface. However, both these stationary phases attract water molecules and the surfaces become hydrated: $SiO_2.xH_2O$ (silica gel) and $Al_2O_3.xH_2O$ (hydrated alumina). The water present then forms the stationary phase and the solutes are separated by partition (Figure 23.243). A thin layer of cellulose can also be used as the stationary phase, but, since it retains water, the separation is by partition.

The basis of TLC is similar to that of paper chromatography. The paper stationary phase is replaced by a TLC plate, which is a thin layer of a substance such as silica (SiO_2) or alumina (Al_2O_3) coated on a glass, aluminium foil or plastic plate.

A spot of the sample solution is placed near the bottom of the plate. The plate is placed in a closed vessel containing solvent (the mobile phase) so that the liquid level is below the spot. The solvent ascends the plate by capillary action, the liquid filling the spaces between the solid particles. The technique is usually carried out in a closed vessel to ensure the atmosphere is saturated with solvent vapour and that evaporation from the plate is minimized.

The components may be recovered by scraping the areas containing the spots into a suitable solvent. Alternatively, TLC plates and paper chromatograms can be scanned by a device known as a densitometer to give quantitative information about the components of the mixture.

■ **Figure 23.243** The principles of TLC


mobile phase moves upwards due to capillary action
stationary phase, e.g. silica
mobile phase, e.g. ether
M
M
molecule M adsorbed to stationary phase
molecule M dissolved in mobile phase

Thin-layer chromatography is faster than paper chromatography and will work with smaller samples. Since the thin layer can be made from different solids, a wide range of mixtures can be separated. A further advantage is that it is easier to retrieve the spots after separation simply by scraping them off the supporting backing of the plates. Thin-layer chromatography is mainly used for the separation of organic compounds. It is a simple, reliable and low-cost technique that is often used to select the conditions for larger-scale separations.

Once a thin-layer chromatogram has been developed, it is often necessary to use some method to make the separated components visible, because most organic compounds are colourless. This is often done by illuminating the plate with a short- or long-wave ultraviolet lamp. The solutes are identified in the same way as for paper chromatography, using R_f values and pure compounds as references.

There are a number of semi-permanent methods for visualization which not only allow you to see these compounds but also provide a method for determining what functional groups are contained within the molecule. This method is referred to as staining the TLC plate.

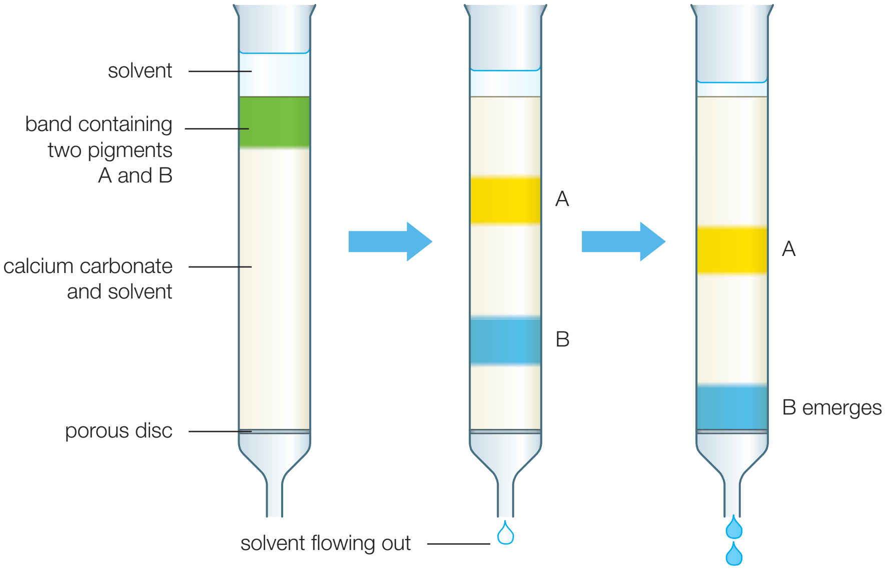
The staining of a TLC plate with iodine vapour is among the oldest methods for the visualization of organic compounds. It is based upon the observation that iodine has a high affinity for both unsaturated compounds and those containing a benzene ring.

Column chromatography

Column chromatography is a convenient variation of the chromatography technique that allows for the large-scale separation of mixtures. Traditionally, a glass column is packed with an inert substance such as silica or alumina, which acts as an inert stationary phase. The column must be tightly packed as trapped air bubbles will hamper the separation of the components. The mixture to be separated is introduced at the top of the column and then the solvent is allowed to run through under the force of gravity, more being added as needed. As the separation takes place the different components will arrive at the bottom of the column at different times and may be collected in different flasks for further use (Figure 23.244). The components may be obtained by evaporating the solvent.

Column chromatography can be used to separate and identify the constituents in a mixture of dyes and the constituents of vegetable extracts. In pharmacy it is used for the separation of vitamin A from fish liver oil and in the detection of adulterants in foods and wines. Adulterants are chemical substances that should not be present in foods or drinks.

Figure 23.244 Separation of the components in plant pigments during column chromatography


solvent
band containing two pigments A and B
calcium carbonate and solvent
porous disc
solvent flowing out
A
B
A
B emerges

23.10 Stereochemistry in biomolecules (AHL) – *most biochemical processes are stereospecific and involve only molecules with certain configurations of chiral carbon atoms*

Handedness in humans has a long and culturally troubled history, with unhelpful connotations and stigma being attached at times to individuals depending on whether they are left-handed or right-handed. There are many aspects to this phenomenon and there are extensive studies being carried out into its origins, advantages and social implications. What concerns us here is the notion of 'handedness', or chirality, in biochemistry and how it relates to the way molecules 'see' each other, interact, assemble into highly complex structures, and control activity in the many metabolic pathways that control the functioning of organisms.

Shape – three-dimensional configuration – is a key concept in biochemistry. Molecules that have the same sequence of atoms and chemical bonds but different arrangements of atoms in space are known as stereoisomers (see Chapter 20). Stereoisomers that cannot be transformed into one another without breaking chemical bonds are known as **configurational isomers** and include two classes: **optical isomers (enantiomerism)** and ***cis-trans* isomers**.

The discussion in this section covers the stereochemistry of some biologically important compound,s namely the proteogenic 2-amino acids, monosaccharides and carbohydrates, lipids and retinoids.

Stereochemical notations

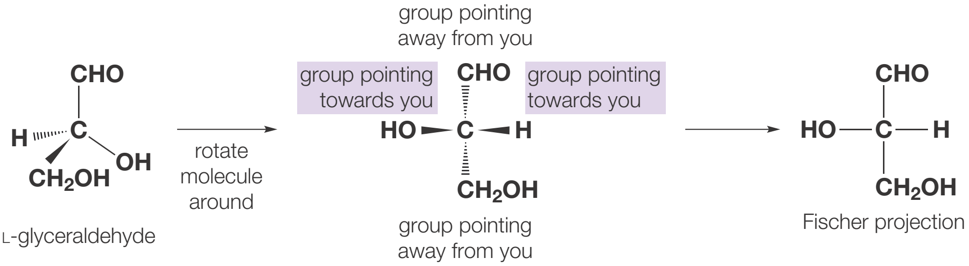
The D/L notation

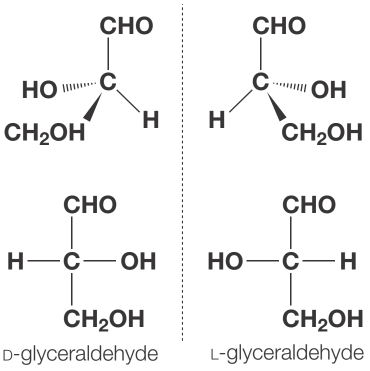
The D and L convention for naming enantiomeric forms of a molecule tends to be used for carbohydrates and amino acids. The system has a long history having been developed before X-ray crystallography allowed the determination of the absolute configuration of molecules. When the D and L system is used, everything is compared to glyceraldehyde (2,3-dihydroxypropanal).

Carbohydrates

Representing in two dimensions the three-dimensional structures of open-chain monosaccharides and their various isomers requires a systematic approach. The system suggested by Emil Fischer in the 1890s is still the agreed method for drawing these structures. In a Fischer projection, a tetrahedral carbon atom is represented by two crossed lines. By convention, the horizontal lines represent bonds coming out of the page. The vertical lines represent bonds going into the page. Using molecular models is the easiest way to understand how the two-dimensional Fischer projection represents a three-dimensional structure (Figure 23.245). In fact, Emil Fischer initially developed his projection using molecular models made of toothpicks and bread rolls!

■ **Figure 23.245** Deriving a Fischer projection for glyceraldehyde, starting from a tetrahedral model and moving to a line drawing


CHO
H
C
OH
CH₂OH
L-glyceraldehyde
rotate molecule around
group pointing away from you
group pointing towards you
CHO
group pointing towards you
HO
C
H
CH₂OH
group pointing away from you
CHO
HO
C
H
CH₂OH
Fischer projection



■ **Figure 23.246** The structures of D- and L-glyceraldehyde, showing the Fischer projections

The enantiomers of glyceraldehyde are shown in Figure 23.246 and their Fischer projections are shown beneath them.

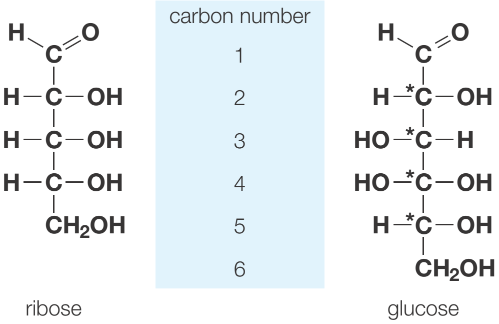
Note that, by definition, when shown in Fischer projection, if the OH group on the chiral carbon is on the right, the molecule is assigned the label D; if the OH group is on the left, it is given the label L.

These projections are a method of representing the structure of any straight-chain forms of a sugar by projection on to a plane. In a Fischer projection, the sugar molecule is shown with the carbon numbered 1 at the top – according to the normal naming rules the aldehyde/ketone group will get the lowest possible number. Groups that point away from you are drawn vertically and those that point towards you are drawn horizontally (see Figures 23.245 and 23.246).

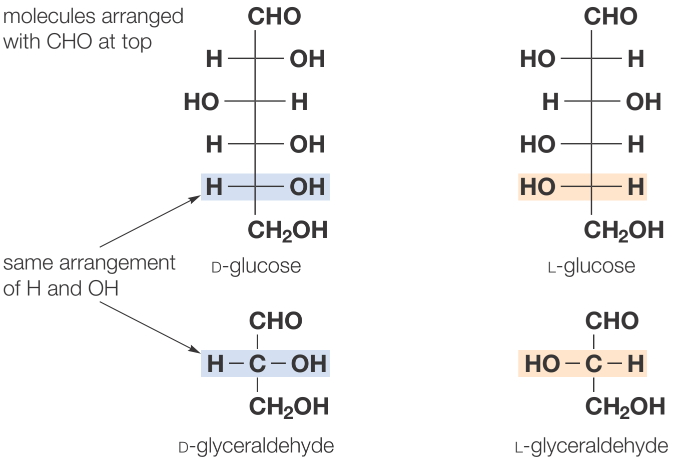
In larger structures, such as ribose or glucose (see Figure 23.247), the tetrahedral centres are simply stacked on top of each other, with the aldehyde group at the top.

The open-chain structure of glucose has four chiral centres (marked * in Figure 23.247). Looking at two forms of glucose in the Fischer projection, we can compare them to glyceraldehyde to determine which is the D enantiomer and which is the L form (Figure 23.248). In D-glucose the OH group on the chiral centre furthest from the carbonyl group is on the right and in L-glucose it is on the left.

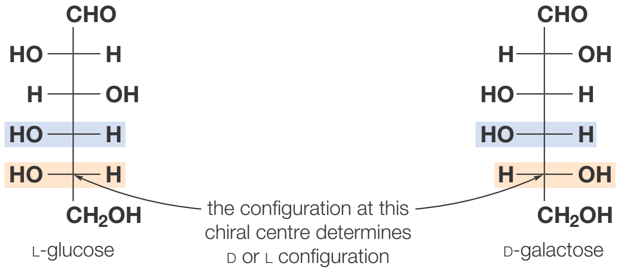
There are four chiral centres in glucose and in D- and L-glucose all chiral centres have the opposite configuration, so that the two molecules are mirror images. If only some of the chiral centres have the opposite configurations, so that the molecules are not mirror images (Figure 23.249), then the two molecules are diastereomers and are given different names, the configurations at the other chiral centres determining the name of the substance.



■ **Figure 23.247** Open-chain structures of ribose and glucose. These structures are drawn as Fischer projections, which enables stereoisomers to be compared easily



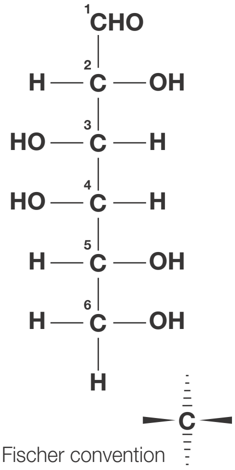
■ **Figure 23.248** Fischer projections of the two stereoisomers of glucose and their relationship to glyceraldehyde, showing how the structure of glucose at carbon-5 is related to the structure of glyceraldehyde


CHO
HO H
H OH
HO H
HO H
CH₂OH
L-glucose
the configuration at this chiral centre determines D or L configuration
CHO
H OH
HO H
HO H
H OH
CH₂OH
D-galactose

■ **Figure 23.249** The two molecules here are not mirror images – L-glucose and D-galactose are diastereomers. Note the configuration at the chiral centre highlighted in yellow is the same for both

The D/L notation is especially useful for naming sugars (Figure 23.250). For example, the name D-galactose defines the relative stereochemistry of all the chiral centres in the molecule. 'D' or 'L' determines the absolute stereochemistry of the highest-position chiral carbon, and hence the absolute stereochemistry of all centres can be defined in a very convenient manner. D-Galactose has four chiral centres – the number of isomers is 2*n*, where *n* represents the number of chiral carbon atoms. Hence, this molecule has 16 stereoisomers – only one of these is D-galactose. All of the D-aldohexoses have the same arrangement at C-5 and it is this carbon atom which is configurationally related to D-glyceraldehyde (see Figure 23.248), the Fischer representation of which has the –OH group on the right-hand side of the projection.

The D sugars are the most abundant form in nature.



■ **Figure 23.250** Structure of D-galactose

The ring forms of sugars

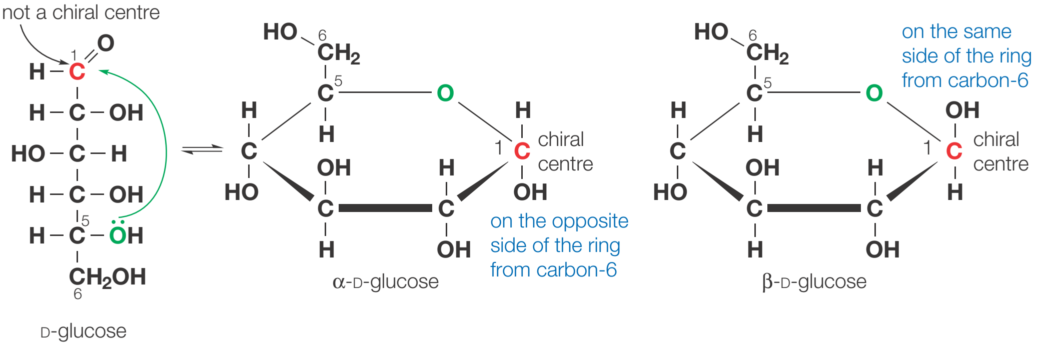
The conversion of sugars in the straight-chain form to the ring form, described in Section 23.4, creates an additional type of isomer, known as α- and β- forms (Figure 23.251). These are distinguished by the relative position of the groups attached to the carbon atoms that close the ring by forming an ether link with oxygen. The α- and β-forms are isomers known as **anomers**.

When a sugar cyclizes from the straight-chain form, an extra chiral centre is formed (Figure 23.251). The oxygen on carbon 5 can attack the C=O group from either above the plane of the group or below and thus two possible cyclic molecules can be formed. These are called α- and β-forms. If the OH on the new chiral centre (anomeric carbon) is on the same side of the ring as carbon-6 then the isomer is β, and if it is on the opposite side then it is α.

An additional consideration to this is that if the ring is drawn as a Haworth projection with carbon-6 above the ring, then:

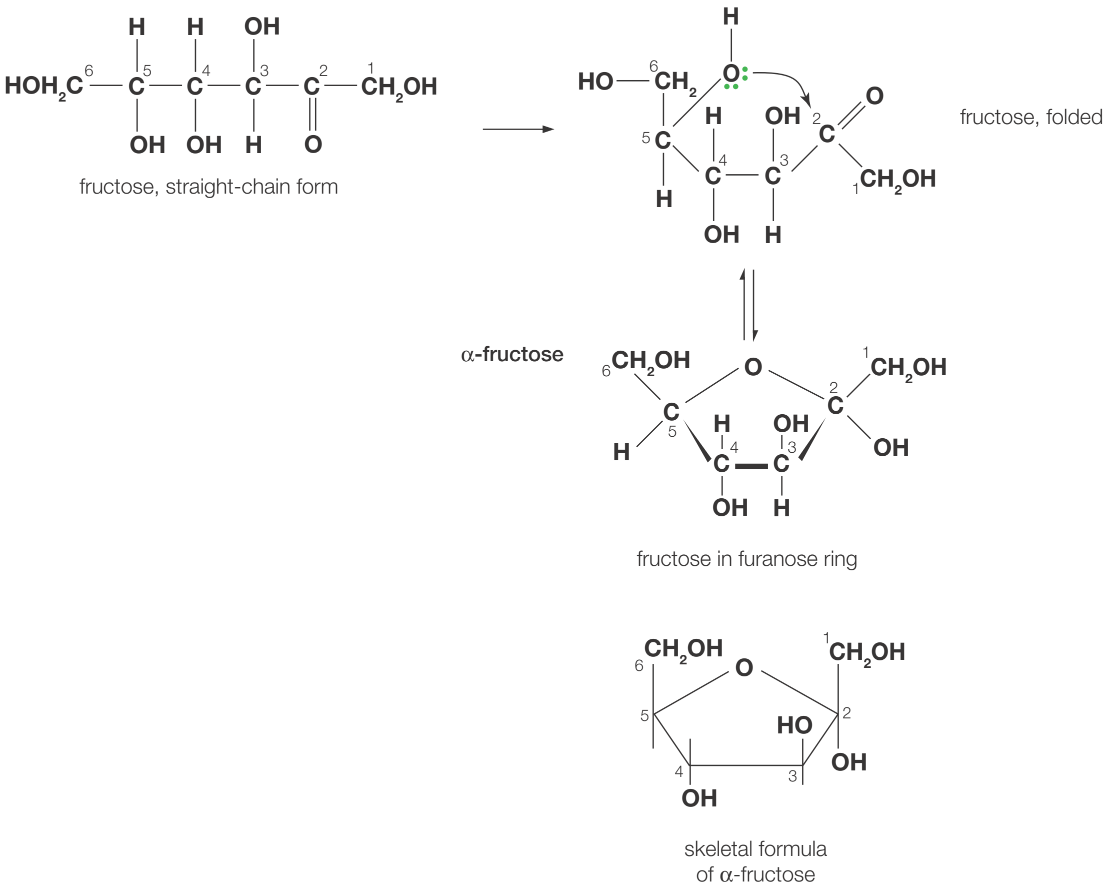
- if the OH on the anomeric carbon is below the ring, it is the α-form
- if it is above the plane of the ring it is the β-form.

■ **Figure 23.251** The structures of glucose in aqueous solution; the formation of the ring structures


not a chiral centre
D-glucose
α-D-glucose
chiral centre
on the opposite side of the ring from carbon-6
β-D-glucose
on the same side of the ring from carbon-6

Whereas glucose is an aldohexose, fructose is a ketohexose and also forms cyclic anomers. Figure 23.252 shows the formation of α-fructose.

Figure 23.252 The formation of α-fructose


fructose, straight-chain form
fructose, folded
α-fructose
fructose in furanose ring
skeletal formula of α-fructose

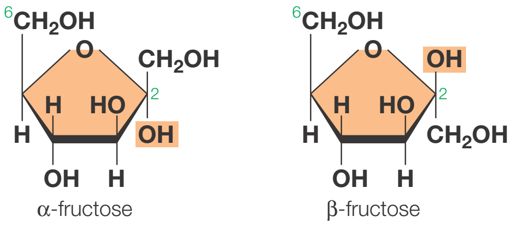


Figure 23.253 The anomeric forms of fructose

The α- and β-forms of fructose are shown in Figure 23.253. Note that in fructose it is carbon-2 that is the new chiral centre, and it is the orientation of the –OH group on this carbon anomeric carbon that distinguishes the two forms. Glucose and fructose both exist in aqueous solution in an equilibrium between the acyclic (straight-chain) form and the two cyclic (ring) α- and β-forms. The two cyclic forms are heavily favoured in the equilibrium.

A further piece of terminology that you may encounter is that sugars, such as glucose, with six atoms in the ring (including the oxygen atom) are in the pyranose form, and structures with five atoms in a ring, such as fructose, are in the furanose form.

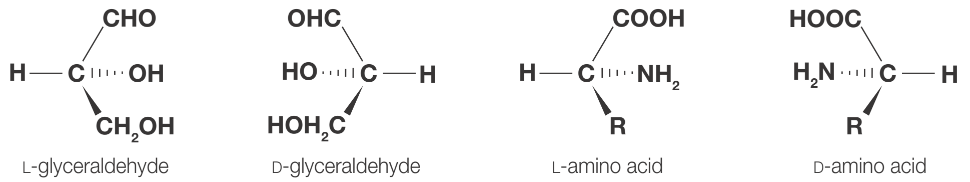
Additional Perspectives

Mutarotation

The two anomers of glucose have very slightly different effects on plane-polarized light (see Chapter 20). For example, α-D-glucose has an optical rotation of +112° and its anomer, β-D-glucose, has an optical rotation of +19°. If the pure anomer is placed in water the optical rotation will change until an equilibrium mixture is formed. This process is known as mutarotation.

Amino acids

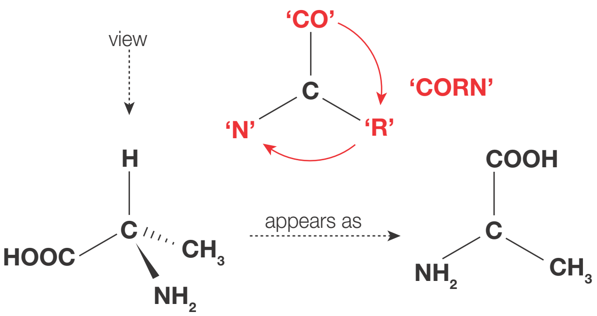
The structures of naturally occurring amino acids can also be related to L-glyceraldehyde (Figure 23.254). It is assumed that the amine (–NH_2) and carboxylic acid (–COOH) functional groups of an amino acid are analogous to the hydroxyl (–OH) and aldehyde (–CHO) groups of glyceraldehyde. The variable side-chain (R) is analogous to the hydroxymethyl group (–CH_2–OH).



■ **Figure 23.254** Diagrams showing the absolute configurations of L- and D-glyceraldehyde and L- and D-amino acids (as a generalized molecular structure)

For amino acids, the CORN rule is applied to determine whether the amino acid is the D or the L form. Arrange the substituents COOH, R, and NH_2 around the chiral carbon so that the hydrogen atom is pointing towards you, in front of the plane of the paper. If the CORN groups are arranged clockwise, then it is the L enantiomer (Figure 23.255); if they are arranged anticlockwise it is the D enantiomer.

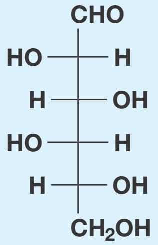
All amino acids are chiral, except glycine, which has only three different groups around the central carbon. The L and D forms of amino acids have identical physical properties and chemical reactivities apart from the direction in which they rotate plane-polarized light and their reactions with reagents that are chiral. This last point is crucial in biochemistry. As enzymes, themselves made of proteins, are chiral molecules, they distinguish completely between the L and D forms of amino acids. Biological systems have evolved to use only the L forms of amino acids. All amino acids found in naturally occurring proteins are L-amino acids.



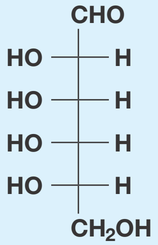
■ **Figure 23.255** The CORN rule illustrated for L-alanine – the view shown is for looking down the C–H bond of alanine

42 Work out whether each of the sugars shown below is the D form or the L form:

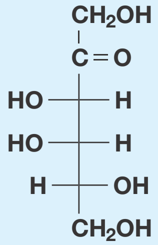
A



B

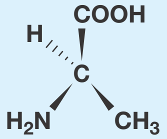


C

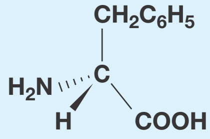


43 Classify the following amino acids below as either the D or the L form, and name them by referring to the *IB Chemistry data booklet*.

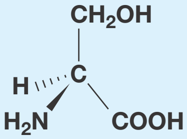
A



B

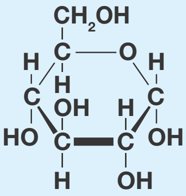


C

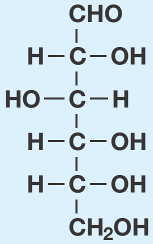


44 Glucose is an example of an aldohexose sugar which dissolves in water. In aqueous solution, glucose exists in three different structures which are in equilibrium with each other. The three structures are shown below:

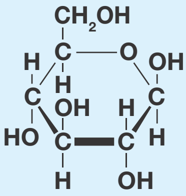
A



B



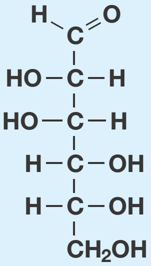
C



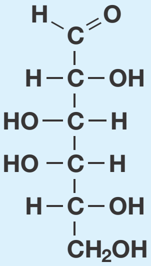
a What chemical term is used to describe the different forms shown above?
b Explain why glucose is very soluble in water,
c Which of the two cyclic forms of glucose can be described as α-glucose? Give a reason for your answer.

45 The diagrams below show the structures of three hexoses: mannose, galactose and glucose.

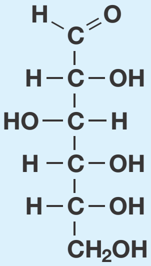
A



B



C



a Explain why are all three are reducing sugars.
b Which form are they all in – D or L form? Explain why you reached that conclusion.
c By suitably numbering the carbon atoms, explain which chiral centres are different in the three molecules. What is the name given to such molecules?

The *R*/*S* notation

It is now relatively easy to establish the absolute configurations of chiral centres within molecules using X-ray crystallography (see Chapter 4) or two-dimensional NMR (see Chapter 21). There is no need for reference molecules such as glyceraldehyde.

The chemists Cahn, Ingold and Prelog jointly developed a set of rules for assigning the absolute configuration of any chiral centre. The term 'absolute configuration' refers to the actual three-dimensional structure of a chiral molecule. The Cahn–Ingold–Prelog rules involve assigning 'priorities' to the functional groups or atoms attached to the chiral centre, and then relating these priorities to a description of the chiral centre. The two possibilities are known as *R* and *S* configurations.

The letters used for the two forms are derived from Latin: *R* is from the Latin word *rectus* meaning right, and *S* is from the Latin word *sinister* meaning left.

The Cahn–Ingold–Prelog rules are summarized below.

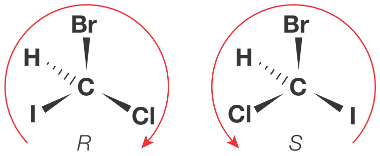


Figure 23.256 Describing bromochloroiodomethanes as *R* and *S* isomers

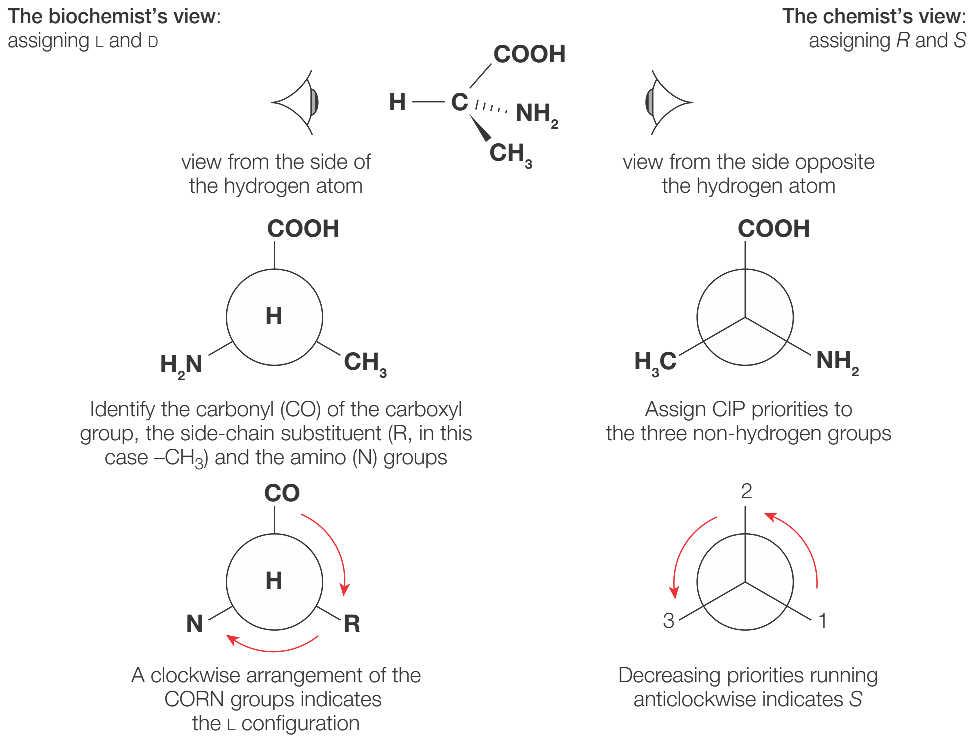
Decreasing priorities – numbered 1, 2, 3 and 4 – are assigned to the atoms attached to the carbon atom that is the chiral centre. Higher priority is given to atoms of higher atomic number. In the case of all chiral amino acids found in natural proteins, this leads to the priorities nitrogen (N) > carbon (carboxylic acid) and C (variable side-chain) > hydrogen (H). So the amine (amino) group is numbered 1 and the hydrogen atom 4.

The chiral centre is viewed from the side opposite the lowest priority atom. If the direction of decreasing priority (1, 2, 3) of the other three atoms is clockwise then the configuration is designated *R*; an anticlockwise arrangement is designated *S*. Figure 23.256 shows how this applies to the bromochloroiodomethanes.

These rules are also illustrated in Figure 23.257 for the amino acid alanine. In the case of alanine, the priorities of the atoms attached to the 2-carbon atom are nitrogen (A_r = 14) > carbon (A_r = 12) > hydrogen (A_r = 1). Hence, the amine group is numbered 1 and the hydrogen is numbered 4. The carbon atom of the carboxyl group is bonded to oxygen and hence has higher priority than the carbon in the methyl group, which is bonded to hydrogen. The overall priority order is therefore nitrogen > carbon (carboxylic acid functional group) > carbon (methyl side-chain) > hydrogen. When the 2-carbon is viewed from the side opposite the hydrogen (priority 4), the decreasing priority of the other groups follows an anticlockwise pattern – and therefore the absolute configuration is designated as *S*.

The *R*/*S* system has no fixed relationship to the D/L system described earlier. Nor indeed does it have any fixed relationship with the final type of notation we consider next – the (+)/(−) system based on the rotation of the plane of polarized light by the enantiomers. An *R* isomer can be either dextrorotatory or levorotatory, depending on its exact groups.

Figure 23.257 Identifying the absolute configuration of amino acids, using alanine as an example



The biochemist's view: assigning L and D

The chemist's view: assigning *R* and *S*

view from the side of the hydrogen atom

view from the side opposite the hydrogen atom

Identify the carbonyl (CO) of the carboxyl group, the side-chain substituent (R, in this case $-CH_3$) and the amino (N) groups

Assign CIP priorities to the three non-hydrogen groups

A clockwise arrangement of the CORN groups indicates the L configuration

Decreasing priorities running anticlockwise indicates *S*

The (+)/(–) notation: optical activity

An enantiomer can also be named by the direction in which it rotates the plane of plane-polarized light. If it rotates plane-polarized light clockwise (as seen by a viewer towards whom the light is travelling), that enantiomer is labelled (+) and its mirror image is labelled (–). The (+) and (–) isomers have also been termed d and l, respectively – for dextrorotatory (Latin for right-handed) and levorotatory (for left-handed), but this terminology is no longer favoured.

The D/L labelling system is unrelated to (+)/(–) system: it does not indicate which enantiomer is dextrorotatory and which is levorotatory.

Nature of Science

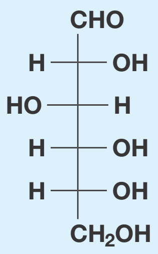
The origin of chirality

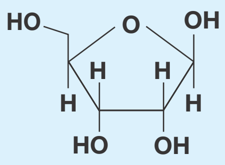
There are a number of reasons why living organisms have a strictly controlled chiral environment. The existence of regular structures in nucleic acids (the double helix) and proteins (the alpha helix and beta sheet) requires pure enantiomers. A mixture of two enantiomers, for example, sugars and amino acids, would not lead to regular secondary structures. And if the active site of an enzyme is to bind a substrate in a very stereospecific arrangement, then a precise arrangement of binding groups is required, and this will only operate for one stereoisomer and not the other. If living systems have L-and D-amino acids they would need twice as many enzymes, which would be very wasteful metabolically.

46 Research using the internet, journals and the library about the evolutionary origin of chirality in amino acids.

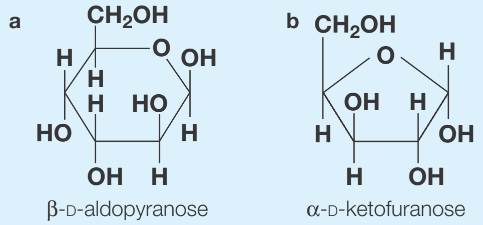
47 The open-chain form of D-glucose has the structure shown here (on the left).
- a Draw the Haworth projection of beta-D-glucose.
- b Explain why D-glucose is water soluble.

48 The structure of ribofuranose is shown here (right). It is a monomer involved in the formation of RNA.
- a State which anomer is shown.
- b State whether the sugar is a reducing or non-reducing sugar.
- c Draw the Fischer projection of D-ribose.





49 Consider the following two monosaccharides:



Draw Fischer projections of the open-chain forms of **a** and **b**.

The stereochemistry of carbohydrates

The structural differences between α- and β-glucose have a large effect on the properties of the disaccharides that can be formed from these monomers, and also on the polymers that can be built when these are extended.

Disaccharides

Two important disaccharides, maltose and cellobiose, are both dimers of glucose. The differences between them arise from how the two glucose molecules are linked.

- In maltose, a glycosidic link is formed between carbon-1 on the first glucose molecule and carbon-4 on the second (Figure 23.258a). The first glucose molecule (drawn on the left) is in the α-form and so the bond is an α-glycosidic link: the link in maltose is written as α-1,4. Maltose is an α-glycoside since the left-hand glucose ring is locked in the α configuration. Because carbon-1 on the right-hand ring is still free, further glucose molecules can be added and the chain can be extended to form the polysaccharide starch.
- Cellobiose, like maltose, is also built from two glucose molecules. However, in this case, the left-hand glucose molecule is in the β-form. A β-1,4 glycosidic link is formed between the two glucose units (Figure 23.258b). Cellobiose is a β-glycoside since the left-hand ring is fixed in the β-form. Again, further glucose molecules can be added, to form cellulose.

Three other disaccharides involving glucose linked to other sugars are lactose, sucrose and trehalose. These are involved in energy transport in various situations and were discussed in Section 23.4.

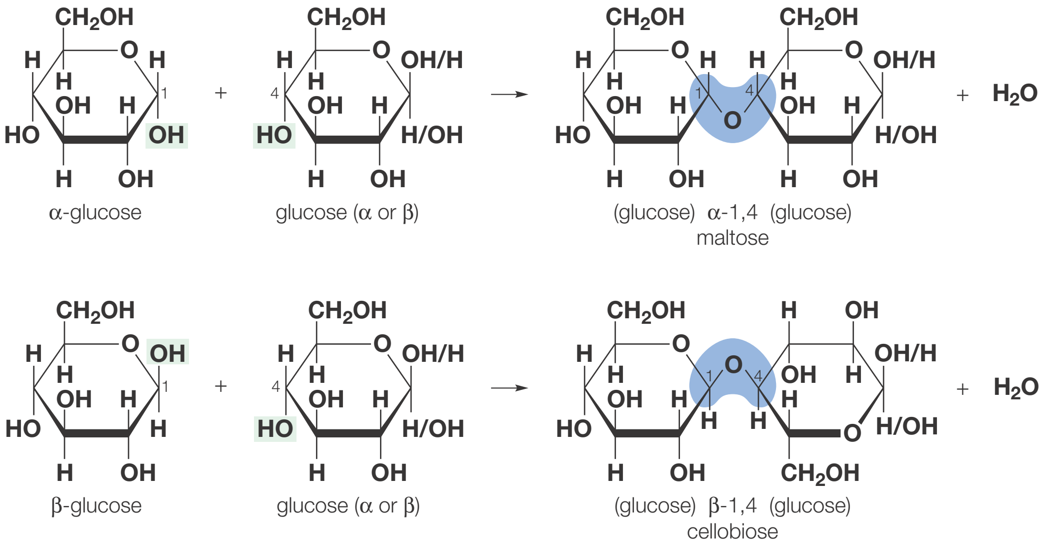


Figure 23.258 **a** The formation of maltose, with an α–1,4 glycosidic link. **b** The formation of cellobiose, with a β–1,4 glycosidic link. Note that the right-hand glucose ring in cellobiose is inverted relative to the first ring

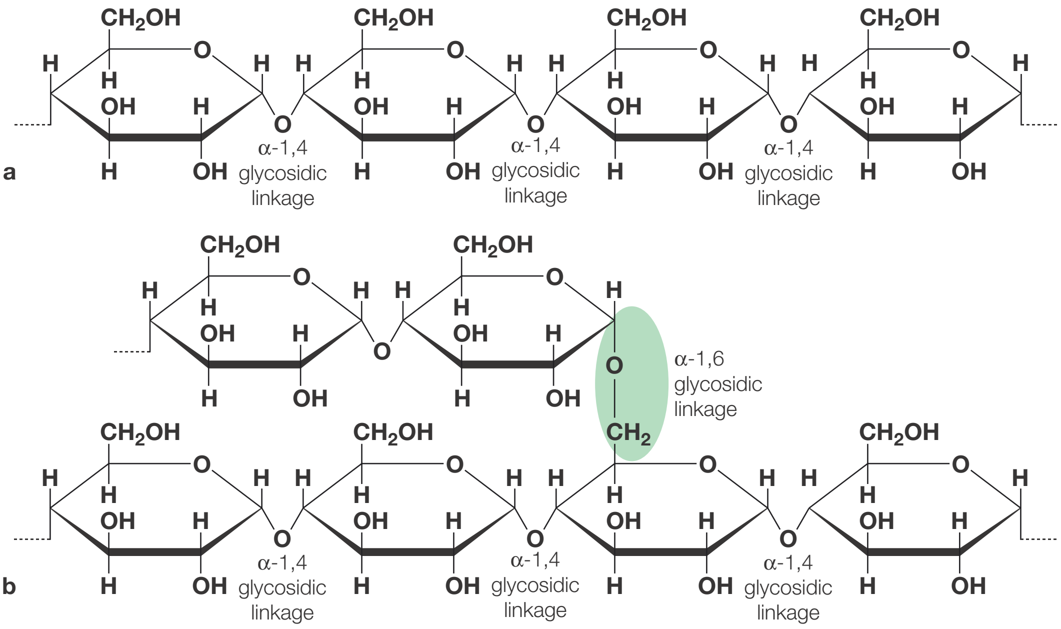
Polysaccharides

The structural differences between α- and β-glucose have a large effect on the properties of their polymers. The differences in structure between maltose and cellobiose are carried forward in their extended polymers, starch and cellulose respectively.

- Starch and glycogen are polymers of α-glucose. Starch forms a relatively compact spiral structure and is stored as starch grains in plant cells. Starch consists of a mixture of two types of glucose polymers, called α-amylose and amylopectin. α-Amylose consists of thousands of α-glucose units linked together to form linear, unbranched chains. These glucose units are linked by α-1,4-glycosidic linkages (Figure 23.259a). An α-1,4-glycosidic linkage is formed between the C-1 of an α-sugar molecule and the C-4 of another sugar molecule.

  Amylopectin is a branched polymer of α-glucose units linked by α-1,4-glycosidic linkages and α-1,6-glycosidic linkages at the branch points (Figure 23.259b).

Figure 23.259 How the glucose units are joined in **a** amylose and **b** amylopectin



- Cellulose is a polymer of β-glucose. It is a linear polymer with β-1,4-glycosidic links. These position the sugars at a different angle from the α-glycosidic links found in starch so the cellulose chain forms an uncoiled linear structure with alternate glucose monomers 'upside down' with respect to each other (Figure 23.260). This enables the hydroxyl groups of one molecule to form hydrogen bonds with the hydroxyls of other cellulose molecules lying parallel to it.

  Consequently, cellulose forms cables, known as microfibrils, of parallel chains that give it a rigid structure. Cellulose is found in all plant cell walls and is one of the main sources of support in plant cells. This is why wood, which is rich in cellulose, is such a useful building material. Aspects of the structure of starch, glycogen and cellulose were discussed in Section 23.4.

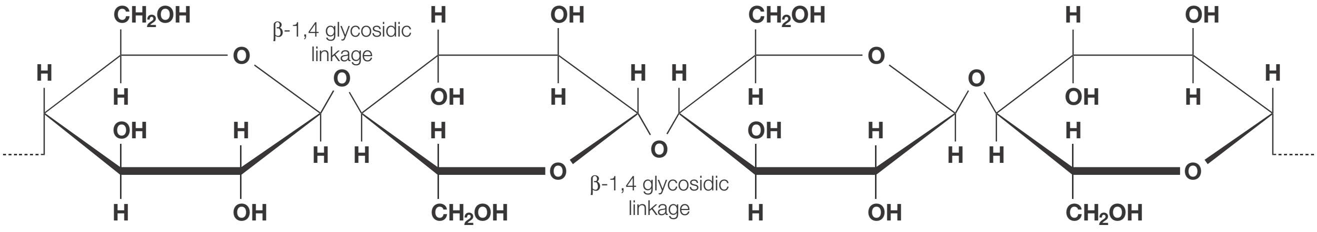


Figure 23.260 The linear structure of a cellulose chain

Digestion and dietary fibre

Starch and glycogen can be relatively easily hydrolysed into glucose by the action of digestive enzymes (α-glycosidases such as α-amylase), but the human body does not produce an enzyme to hydrolyse the β-glycosidic links in cellulose. Most cellulose therefore passes through the human gut largely chemically intact, contributing to the bulk of the faeces. As such it is now commonly referred to as dietary fibre and is seen as an important component of the human diet. There are two types of dietary fibre:

- Insoluble fibre includes cellulose, hemicellulose and lignin found in plant cell walls.
- Soluble fibre includes pectin found in plant cells; sources include oats, oat bran and beans.

20 CAPSULES
míntec
IBS RELIEF
ABDOMINAL PAIN • CRAMPS • BLOATING • WIND
Soothing relief anywhere, anytime
TARGETED RELEASE
CLINICALLY TESTED PEPPERMINT OIL

Figure 23.261 Medication for relief from irritable bowel syndrome

Medical studies have indicated that this dietary fibre is of benefit to the health of the large intestine. The cellulose fibrils abrade the wall of the digestive tract and stimulate the lining to produce mucus which helps in the passage of undigested food through the gut. Fibre in the diet helps to reduce conditions such as constipation, hemorrhoids and, possibly, colorectal cancer. In addition dietary fibre helps regulate the absorption of sugars and bile acids, reducing the risk of diabetes and cholesterol-related heart diseases. Generally, foods derived from plants with little or no processing are likely to be a good source of fibre.

In contrast to humans, many microorganisms produce cellulase and other β-glycosidases which allows them to digest cellulose and use it as a principal source of food. Ruminants such as cattle and sheep, horses and insects such as termites can extract energy from plant tissue and wood using cellulase-producing bacteria in their digestive systems.

In fact, a certain amount of dietary fibre can be fermented by cellulase-producing bacteria in the human large intestine. These bacteria produce short-chain fatty acids, together with other metabolites, which help to prevent the development of various adverse health conditions such as hemorrhoids, Crohn's disease, irritable bowel syndrome (Figure 23.261) and bowel cancer. In many countries dietary fibre is now a recommended macronutrient contributing to a healthy diet.

Understanding health issues

Many governments are stressing concern over health issues and promoting programmes to reinforce advice on healthy eating. This involves social concern for the public and also a growing awareness of the economic costs of lost working time and health service demands.

The World Health Organization cites a low fruit and vegetable intake as a key risk factor in chronic diseases such as diabetes, obesity and some cancers. Our growing understanding of the significance of fibre in the diet has led to an increase in the marketing of whole grain foods and plant foods such as vegetables and salads.

Fruit and vegetable intake varies considerably from country to country, largely reflecting the prevailing economic, cultural and agricultural environments. In developed countries fresh fruit and vegetable intake has decreased with increasing dependence on fast foods that are highly processed. It is estimated that globally 2.7 million deaths per year can be traced to low fruit and vegetable intake.

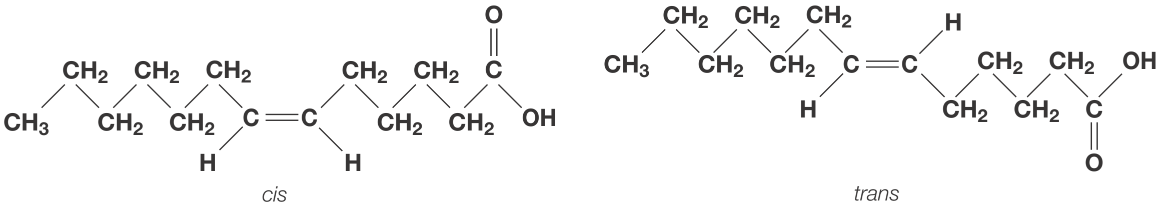
There is also growing awareness of the link between *trans* fats and cardiovascular and other diseases, but again that has generated different responses around the world. Canada was the first country to introduce mandatory labelling of *trans* fats on food products, while Denmark, Iceland and Switzerland have banned their use. The World Health Organization has called for the elimination of *trans*-fatty acids from the global food supply. The cost of this, however, will be considerable as it will involve very major changes in food formulations and eating habits. The next section gives some background on the chemistry of *cis* on and *trans* fats and the issues surrounding their part in our diet.

Stereochemistry of lipids

Fats and oils that are saturated have no carbon–carbon double bonds between the carbon atoms in their fatty acids. The unsaturated fatty acids in fats and oils contain carbon–carbon double bonds. These can be termed mono- or poly-unsaturated fats depending on whether they have one or more carbon–carbon double bonds in their fatty acids.

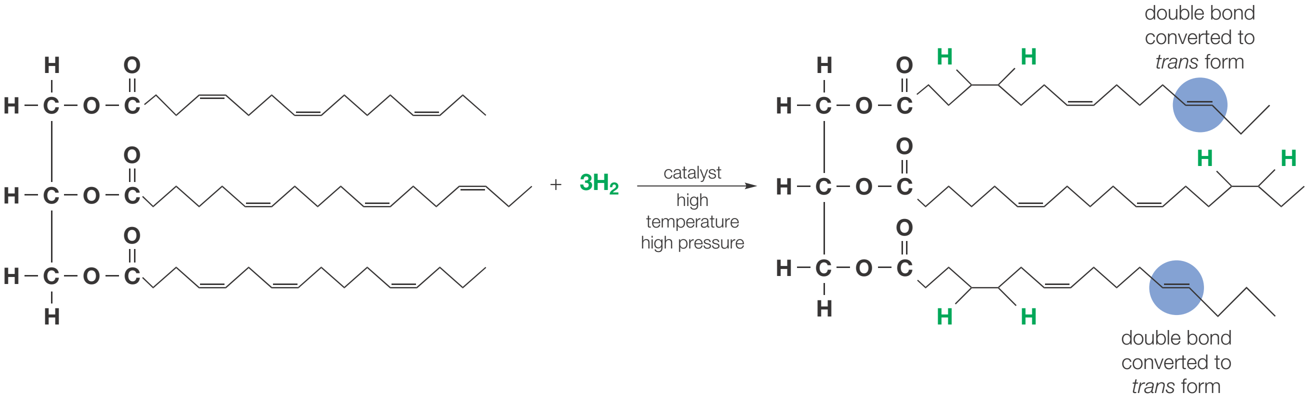
Unsaturated fats exist in two forms, known as *cis*–*trans* isomers, which arise due to the restriction on rotation around the double bond. This type of isomerism was introduced in Chapter 20, while examples of unsaturated fatty acids and their role in lipids were discussed in Section 23.3.

- The *cis* form of a long-chain fatty acid occurs when the hydrogens attached to the carbons involved have the same orientation relative to the double bond.
- The *trans* form of a long-chain fatty acid occurs when the hydrogens attached to the carbons involved have the opposite orientation across the double bond.



With the exception of some dairy products, most naturally occurring unsaturated fats are in the *cis* form. As discussed earlier, *cis* C=C double bonds introduce kinks into the fatty acid chains, thus molecules of the *cis* isomer cannot easily arrange themselves side by side to solidify, so they tend to have lower melting points than the corresponding *trans* isomer.

In order to make products derived from vegetable oils more presentable and yet able to be spread from the refrigerator, the food industry introduced the process of hydrogenation to reduce the level of unsaturation. **Hydrogenation** of vegetable oils takes place in the food industry when hydrogen is added across the carbon–carbon double bonds using a finely divided metal catalyst such as nickel (Figure 23.262).


H-C-O-C
O
H
+ 3H2
catalyst
high temperature
high pressure
double bond converted to trans form
double bond converted to trans form

■ **Figure 23.262** Partial hydrogenation of a poly-unsaturated vegetable oil – showing only the hydrogens that have been added in the process

The product is a fat that, as it is more saturated, has a higher melting point and is therefore a more convenient form for packing and storage as a solid or semi-solid. Fats made in this way also break down less easily under conditions of high-temperature frying and usually have a longer shelf life than liquid oils. Most margarines and shortening come into this category.

However, there is a problem with the process. In partial hydrogenation, only some of the carbon–carbon double bonds in the oil are broken, and those that remain often get chemically modified (isomerized) from the *cis* position to the *trans* position. The resulting fatty acids are therefore known as *trans* fats and are particularly prevalent in processed foods. These have been implicated in a number of cardiovascular diseases. The body finds it difficult to metabolize *trans* fatty acids (lipase enzymes seem to recognize only *cis* forms) and so they tend to accumulate in the adipose tissue rather than be excreted. Medical evidence shows that consuming *trans* fats raises the level of the more harmful forms of circulating cholesterol – low-density lipoproteins (LDL cholesterol) – which is a risk factor for heart disease. A further associated effect is that increased levels of *trans* fats reduce the blood levels of HDL cholesterol, which has a protective function against heart disease. Finally, *trans* fats are a lower-quality source of energy compared to their *cis* counterparts.

Nature of Science

The benefit of hindsight

Sometimes the practice and developments of science can have unintended consequences. For instance, the development of hydrogenation reactions to convert liquid oils into solid fats that are less prone to going rancid would, at first sight, seems to be advantageous progress. As knowledge develops about both the hydrogenation reaction and biochemical processes, fresh insights about unexpected consequences come to light. Nowadays, hydrogenated fats are generally regarded as things to be avoided at all costs.

We learn from the experience and should aim to build in as much rigorous pre-testing as possible, but there will always be aspects of the complex interactions in biochemistry which will catch us unawares. Theories can be used to explain natural phenomena. The understanding that many biological molecules are chiral is essential to an understanding of the biochemical processes that occur in cells.

Stereochemistry in vitamins – retinal and vision chemistry

Vitamin A is a collective name referring to a group of poly-unsaturated compounds with a range of biological functions. One of these compounds is retinal, a long-chain aldehyde involved in what is termed the visual cycle – the photochemical changes associated with our ability to detect light.

The retina of the eye contains millions of two types of light-sensitive cells, known as rods and cones. The rods are stimulated by light of lower intensity and do not provide colour vision. Rod cells are concentrated at the outer edges of the retina and are used in peripheral vision and night vision. The major photoreceptor pigment ('visual purple') on the surface of rod cells is a large conjugated protein molecule called rhodopsin (Figure 23.263). This consists of a protein, opsin, tightly bound to 11-*cis*-retinal, which is derived from vitamin A (see Figure 23.264).

Figure 23.263 A computer model of rhodopsin, the complex of opsin and 11-*cis*-retinal. Note the extensive triple-helical regions of opsin which enable the molecule to sit across the membrane of the cell

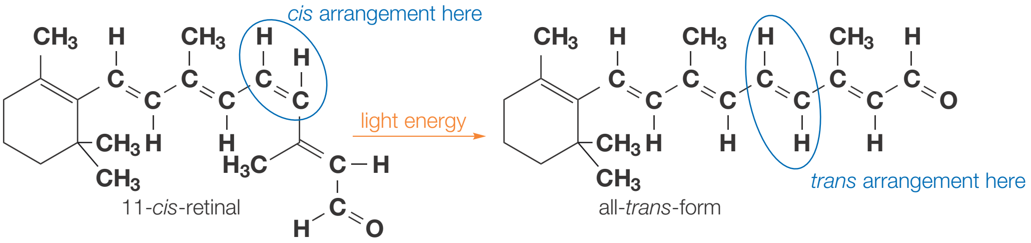


Figure 23.264 The light-induced transformation of 11-*cis*-retinal to all-*trans*-retinal

In *cis*-retinal, one of the C=C double bonds in the chain is in the *cis* configuration, and all the others are *trans*. The molecule is named 11-*cis*-retinal referring to the position of the double bond. In the light-sensitive rhodopsin complex the aldehyde group of 11-*cis*-retinal is reversibly bound to a lysine residue in the protein opsin. When this light-sensitive pigment (rhodopsin) is exposed to light, a transformation of 11-*cis*-retinal occurs, changing it to all-*trans*-retinal (Figure 23.264).

This causes the all-*trans*-isomer to dissociate from the opsin, which triggers a nerve impulse to the brain. Rhodopsin is regenerated from opsin and 11-*cis*-retinal after the all-*trans* form has isomerized back to the 11-*cis* form in a series of steps catalysed by enzymes. A summary of the visual cycle is given in Figure 23.265.

In the human body retinal can only be synthesized from other vitamin A group compounds such as retinol or β-carotene. Insufficient production of retinal resulting from vitamin A deficiency can cause night blindness, a common medical condition in developing countries.

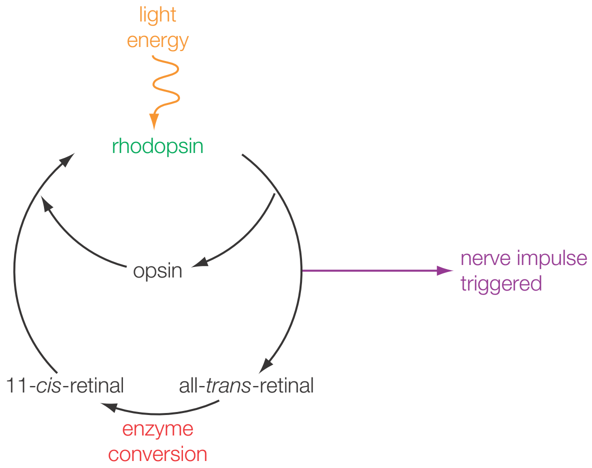
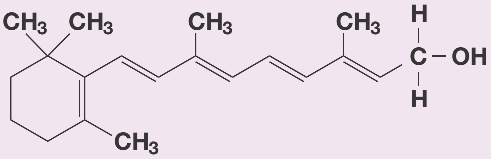


Figure 23.265 The visual cycle

Additional Perspectives

Vision chemistry

In the body, β-carotene is converted into retinol (vitamin A) (Figure 23.266), which is essential for vision. The retina at the back of the eye contains two types of receptors: rods (black and white vision) and cones (colour vision).

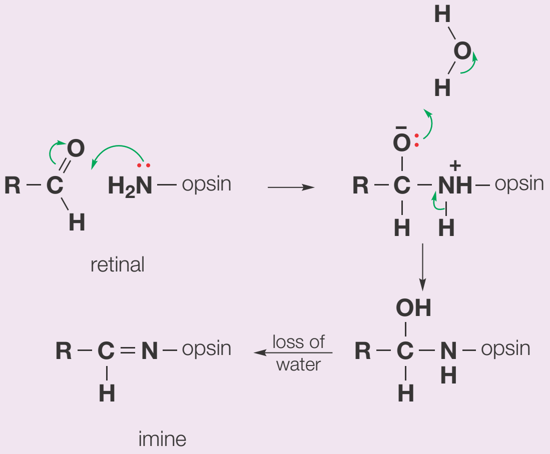


■ **Figure 23.266** The structure of all-*trans*-retinol (one of the molecules classed as vitamin A)

The rods contain retinol which is converted to retinal in an enzyme-controlled reaction. The *trans* isomer of both molecules is the more stable isomer and an enzyme in the rods converts *trans*-retinal to 11-*cis*-retinal. This immediately bonds to the protein opsin to form the light-detecting molecule known as rhodopsin.

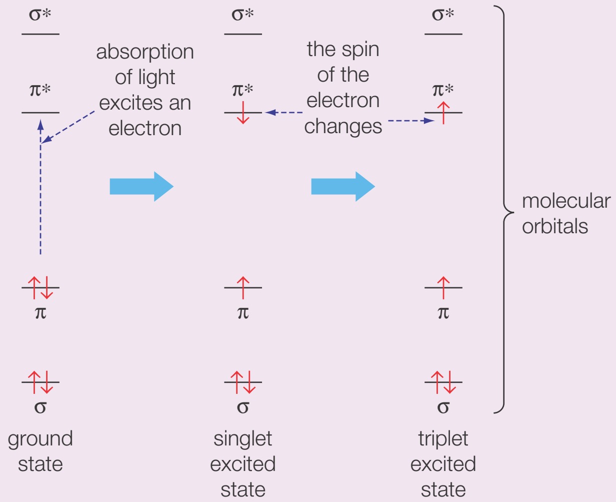
Rhodopsin is formed by the nucleophilic attack of the primary amine group, $–NH_2$, of the protein on the $–CH=O$ group of the retinal molecule (Figure 23.267). The 11-*cis*-retinal is covalently bonded to the protein and also interacts with the protein through London (dispersion) forces. The complex of retinal and opsin shifts the wavelength of maximum light absorption from close to 400 nm to 493 nm. The absorbed light triggers the nerve impulse that is sent to the brain.

When a rhodopsin molecule absorbs light the carbon–carbon double bond at position 11 causes an electron in the pi orbital of the π bond to undergo a transition to the pi antibonding orbital (π*).



■ **Figure 23.267** Formation of an imine from opsin and retinal

This π–π* transition can lead to a singlet excited state (with the electron spins of the ground state and excited electrons still opposed), but this can be followed by transfer to the triplet state, when the spin of electron in the π* orbital reverses. This triplet excited state has a relatively long life time. The pi bond has been weakened by the removal of one electron (Figure 23.268).



■ **Figure 23.268** π–π* excitation of an alkene by light (of the appropriate energy)

The 11-*cis*-retinal now converts to the more stable *trans* form and it is this molecular event that allows vision to occur. The all-*trans*-retinal is a very different shape from 11-*cis*-retinal and will not bond and combine with opsin.

Hence the absorption of light energy causes the conversion of 11-*cis*-retinal to the all-*trans* form causing the dissociation of rhodopsin into opsin and retinal. This isomerization causes an impulse in the optic nerve by altering the permeability of the rod cell membrane to sodium ions.

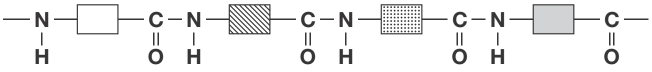
Examination questions – a selection

Q1 Glucose is the major source of energy for the human brain. It is completely oxidized to carbon dioxide and water in aerobic respiration in brain cells. Calculate the mass of carbon dioxide produced in the brain per day if its daily consumption of glucose is 135 g. [3]

Q2 State the difference between anabolic and catabolic pathways in cell metabolism. [2]

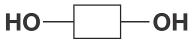
Q3 Explain the difference between hydrolysis and condensation reactions. [2]

Q4 **a** The diagram below represents the structure of part of a protein chain.



Draw the structures of two of the amino acids that went to make up this protein (using the form of representation used here in the structure of the polymer). [2]

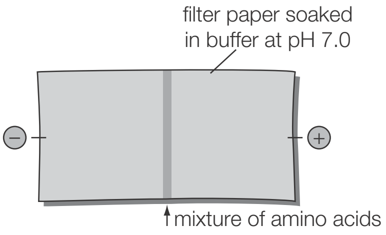
b Complex carbohydrates such as starch are another group of condensation polymers. Enzymes such as amylase, a carbohydrase, can hydrolyse complex carbohydrates to simple sugars which can be represented as:



Draw the structure of the complex carbohydrate chain (showing at least three monomer units). [2]

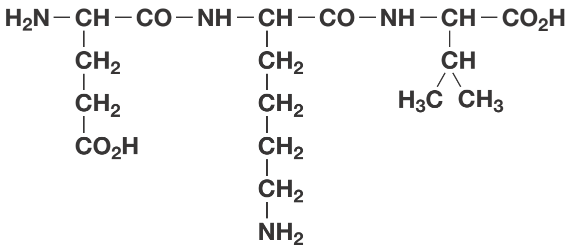
Q5 Aspartame is an artificial sweetener that is 160 times sweeter than sucrose. It is a dipeptide of the amino acids aspartic acid and phenylalanine, with the methyl ester of phenylalanine forming the C-terminal end. By referring to Figure 23.24, draw the structure of aspartame. [3]

Q6 **a** Electrophoresis is a technique that separates molecules according to their electrical charge. The diagram below shows, in outline, a technique that could be used for separating the amino acids that are obtained when a protein is hydrolysed.



After electrophoresis of the amino acid mixture using the technique shown above, the filter paper is stained with ninhydrin, which gives a purple colour in the presence of amino acids.

A small peptide consisting of three amino acid units, shown below, is hydrolysed and the mixture is separated by electrophoresis (at pH 7.0).

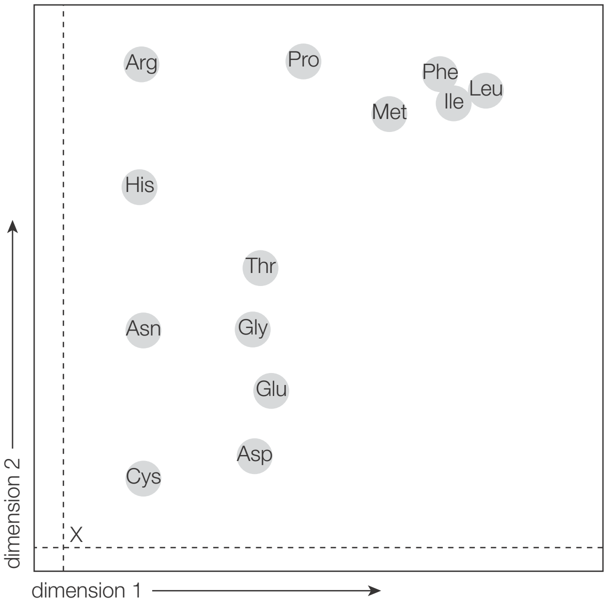


i Sketch the pattern of separation of amino acids that will be obtained on the filter paper after staining. Explain how you came to your conclusions. [3]

ii Explain how the result will be different if the electrophoresis is performed at pH 12.0. [2]

b When a protein chain is folded into its three dimensional structure, it is normally found that amino acids such as valine and isoleucine are found buried in the interior of the protein structure, whereas amino acids such as glutamic acid and lysine are found on the protein surface. By referring to the structures of these amino acids in Section 3 of the *IB Chemistry data booklet*, explain this observation. [3]

Q7 The diagram below shows a two-dimensional chromatogram of an amino acid mixture. The amino acid mixture was initially placed at point X and separated with an aqueous mixture of butan-1-ol and ethanoic acid (in dimension 1). The chromatogram was allowed to dry and then the paper turned around by 90° and a second solvent mixture of phenol and ammonia was applied.



a Define the term retention factor, R_f. [1]

b Calculate the R_f values of glutamic acid and isoleucine (in aqueous butan-1-ol and ethanoic acid) (to one decimal place). [1]

c Calculate the R_f values of asparagine, threonine and methionine (in aqueous phenol and ammonia) (to one decimal place). [1]

d State which amino acids were completely separated by developing the chromatogram in dimension 1 and dimension 2. [2]

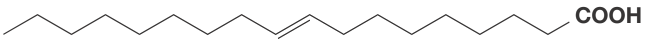
e Name another technique that can be used to separate a mixture of amino acids. [1]

Q8 **a** By means of simple labelled diagrams indicate the structural differences between a triglyceride (fat) molecule and a phospholipid. [3]

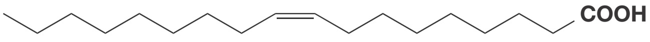
b **i** What is the difference between a saturated and an unsaturated fat? [1]

ii Olive oil is said to be 'rich in mono-unsaturated fats'. What do you understand by the term 'mono-unsaturated'? [1]

iii Elaidic acid (octadec-trans-9-enoic acid) and oleic acid can be represented by the structures



elaidic acid



oleic acid

What type of isomerism is shown by these two fatty acids? What is the systematic name for oleic acid? [2]

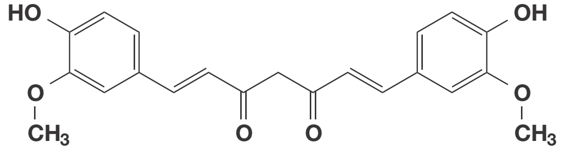
Q9 Some foods contain natural antioxidants which help to prolong their shelf life. The shelf life of oily fish decreases upon exposure to light.

a Identify the chemical feature in the oil in fish that is susceptible to photo-oxidation. [1]

b State the specific term given to food that is unsuitable for eating as a result of photo-oxidation. [1]

c Suggest how light initiates this process. [1]

d Some foods contain a yellow spice called turmeric. The active ingredient in turmeric is curcumin, shown below.

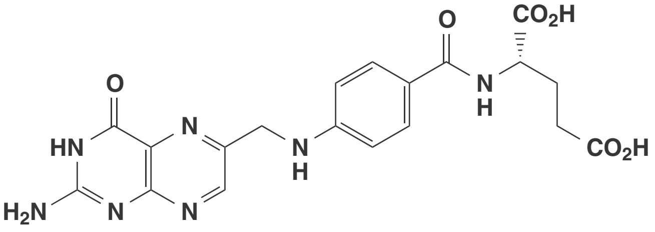


Suggest which structural feature of curcumin is responsible for extending the shelf life of such a food. [1]

Paper 3, May 2012, QF3

Q10 Vitamins are organic compounds needed in small amounts for normal metabolism in the body. Vitamins can be classified as water-soluble or fat-soluble.

a Vitamin B_9 is water-soluble and is important in the repair of DNA. The structure of vitamin B_9 is given below. Suggest why this vitamin is water-soluble. [1]



b Vitamin A (retinol) is important for maintaining healthy skin. The structure of vitamin A (retinol) is given in Table 35 of the *IB Chemistry Data Booklet*.

i State one disease caused by a deficiency of vitamin A in the body. [1]

ii The livers of polar bears and seals contain a very large amount of vitamin A. Some early explorers in the Arctic died from consuming too many livers. Suggest an explanation for this even though males require at least 0.9 mg of the vitamin per day (females require at least 0.7 mg per day). [1]

Q11 The organic compound 1-phenylethanone ($C_6H_5COCH_3$) is being investigated as a potentially useful enzyme inhibitor.

There are two possible ways of synthesizing 1-phenylethanone from 1-phenylethanol under consideration.

Method 1:

$$3C_6H_5CH(OH)CH_3 + 2CrO_3 + 3H_2SO_4 \rightarrow C_6H_5COCH_3 + Cr_2(SO_4)_3 + 6H_2O$$

Method 2 (using a catalyst):

$$C_6H_5CH(OH)CH_3 + \tfrac{1}{2}O_2 \rightarrow C_6H_5COCH_3 + H_2O$$

a Calculate the atom efficiency of both of these methods. [3]

b Identify two other considerations which would need to be borne in mind, besides the atom efficiencies, when deciding which method is 'greener' and therefore the preferred option. [2]

Q12 An amino acid buffer has been prepared by mixing 0.60 dm^3 of 0.20 mol dm^{-3} HCl and 0.40 dm^3 of 0.50 mol dm^{-3} glycine solutions. Calculate the following:

a The pH of the original buffer solution [3]

b The pH of the solution after the addition of 1 cm^3 of 1.0 mol dm^{-3} HCl solution. [2]

c The pH of the solution after the addition of 0.40 g of solid NaOH. [1]

Q13 DNA is the molecule that carries genetic information in nearly all cells. Two months before Watson and Crick published their paper describing the double helical nature of DNA in 1953, Linus Pauling published a suggested structure for DNA based on a triple helix. Pauling's model soon proved to be incorrect; it had the phosphate groups facing into the core of the helix and the nitrogenous bases facing out.

a Suggest why Pauling's model would not have been a stable structure for DNA. [2]

b DNA has the unusual property of being able to replicate. State the type and position of the bonds that break at the start of the replication process. [1]

Q14 Thymine is one of the four nitrogen-containing bases present in DNA.

a Explain how thymine forms part of a nucleotide in DNA. [2]

b The four nitrogen-containing bases are responsible for the double helix structure of DNA. Using the structure of thymine and the structure of one of the other bases in the *IB Chemistry data booklet*, draw a diagram to explain how thymine is able to play a role in forming a double helix. Identify the type of interactions between the two bases. [3]

c Describe how the order in which the four bases occur in DNA provides the information necessary to synthesize proteins in the cell. [2]

d It is now possible to purchase a work of art made from your own DNA profile. Outline the role that restriction enzymes play in making a DNA profile. [2]

Paper 3, May 2010, QF4

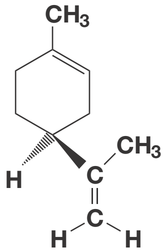
Q15 The pigment in blueberries is an anthocyanin.

a With reference to the colour wheel (see *IB Chemistry data booklet*) explain how the pigment in blueberries causes them to be blue. [2]

b State the combination of pH and temperature that produces the strongest colour in anthocyanins. [1]

Paper 3, November 2010, QF5

Q16 Different conventions are used to classify enantiomers. Orange and lemon peel contain different enantiomers of the compound limonene. One of the enantiomers is represented below.



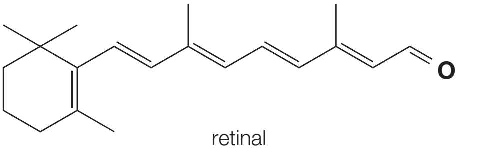
a Identify the chiral centre in this enantiomer with an asterisk (*). [1]

b The (+)-enantiomer has the characteristic smell of oranges and the (–)-enantiomer has the characteristic smell of lemons. Explain the meaning of the (+) and (–) symbols used in this notation. [1]

c The *R/S* notation is also used. The (+)-enantiomer is often described as

R-limonene and the (–)-enantiomer as *S*-limonene. Explain what is meant by the *R*/*S* notation and state whether the structure shown is *R* or *S*. [2]

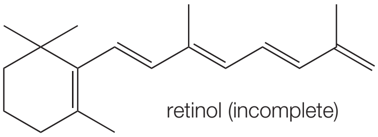
Q17 a The rod cells in the retina at the back of the eye contain a primary alcohol called retinol which is responsible for their sensitivity to visible light. Retinol is oxidized by an enzyme-catalysed reaction to the aldehyde retinal.



retinal

i Deduce the molecular formula of retinal from its skeletal formula shown above. [1]

ii Suggest the structure of the alcohol retinol by completing the skeletal formula below.[1]



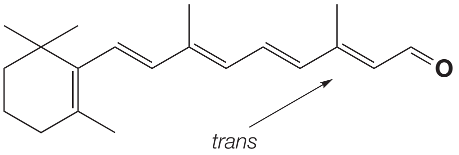
retinol (incomplete)

iii Identify a functional group which is present in both the retinol and retinal molecules. [1]

iv Outline the reagents and conditions that could be used to convert an alcohol to an aldehyde in a laboratory. [4]

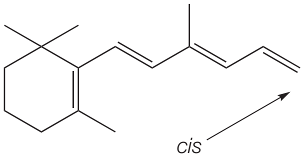
v Deduce how many moles of hydrogen molecules you would expect to react with one mole of retinol (in the presence of a heated nickel catalyst). [1]

b When light enters the rod cells an enzyme-catalysed reaction occurs. This changes the arrangement around the double bond at position 11 from *trans* to *cis*, as indicated in the structure below.



retinal

i Suggest the structure of *cis*-retinal by completing the skeletal formula below. [1]

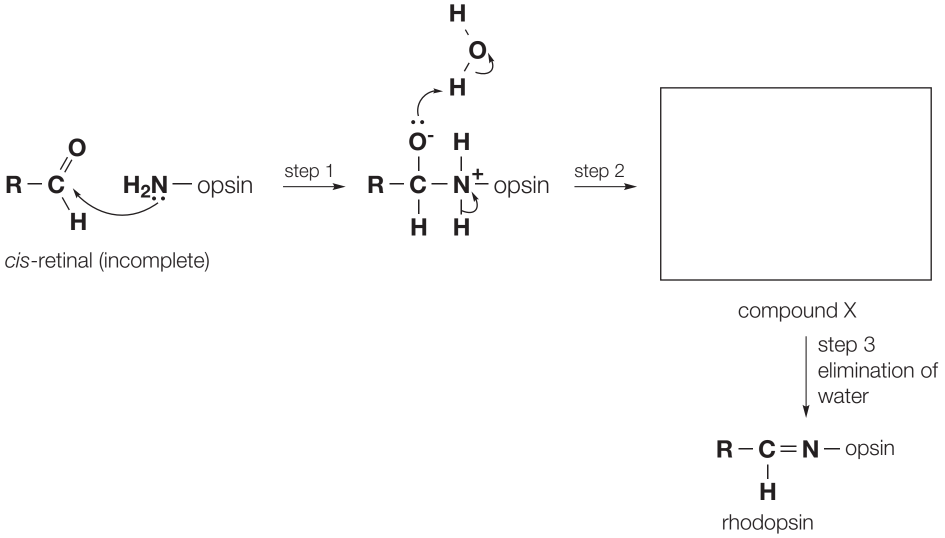


cis-retinal (incomplete)

ii The *cis*-retinal covalently to the protein opsin to form rhodopsin. Part of the mechanism of this reaction is shown below. State the name of the functional group on opsin which is reacting with the aldehyde group on the *cis*-retinal molecule. [1]

iii State the type of reaction mechanism which starts in step 1 and is completed in step 2. [1]

iv Deduce a structure for compound X that would be drawn in the box below. [1]



24 Option C Energy

ESSENTIAL IDEAS

- Societies around the world are completely dependent on energy resources.
- The quantity of energy is conserved in any conversion but the quality is degraded.
- The energy of fossil fuels originates from solar energy which has been stored by chemical processes over time. These abundant resources are non-renewable but provide large amounts of energy due to the nature of chemical bonds in hydrocarbons.
- The fusion of hydrogen nuclei in the Sun is the source of much of the energy needed for life on Earth. There are many technological challenges in replicating this process on Earth but it would offer a rich source of energy. Fission involves the splitting of a large unstable nucleus into smaller stable nuclei.
- Visible light can be absorbed by molecules that have a conjugated structure with an extended system of alternating single and multiple bonds. Solar energy can be converted to chemical energy in photosynthesis.
- Gases in the atmosphere that are produced by human activities are changing the climate as they are upsetting the balance between radiation entering and leaving the atmosphere.

Additional higher level (AHL)

- Chemical energy from redox reactions can be used as a portable source of electrical energy.
- Large quantities of energy can be obtained from small quantities of matter.
- When solar energy is converted to electrical energy the light must be absorbed and charges must be separated. In a photovoltaic cell both of these processes occur in the silicon semiconductor, whereas these processes occur in separate locations in a dye-sensitized solar cell (DSSC).

24.1 Energy sources *– societies are completely dependent on energy resources. The quantity of energy is conserved in any conversion but the quality is degraded*

Nature of Science

The quality of energy

In this chapter we will discover that the kinetic and potential energy changes at the molecular level that we have studied in chemistry have profound implications for our world.

The first law of thermodynamics (see Chapter 5) states that *energy cannot be created or destroyed – it can only change form.* This implies then that all the energy we need already exists in nature. So why are we concerned about energy crises or energy shortages? Why can we not just extract the energy we need from our surroundings?

The key concept here is one of 'quality'. Although the quantity of energy is constant, the quality of that energy varies. The idea of the 'quality' of energy was first analysed by William Thomson (later Lord Kelvin) in 1851. He proposed the concept of the 'availability' of energy: that is, energy that we can actually utilize to do work. This idea was highly significant at the time, as the Industrial Revolution was in progress and steam engines were causing major changes to society – what we would today call a disruptive technology.

During this topic we will explore different types of 'high-quality' or 'available' energy. These sources, whether they are plants, fossil fuels or atomic nuclei, all have potential energy 'locked up' in them – they have the potential to provide us with useful work.

What is an energy source?

You know from your studies of enthalpy in Chapter 5 that all chemical substances have energy associated with them – potential energy is stored in chemical bonds and intermolecular forces, and during chemical reactions energy may be released to the surroundings (exothermic reactions) or absorbed by the substances (endothermic reactions).

However, not all exothermic reactions are used as a source of energy for heating, lighting, electricity generation, transport and so on. Historically, substances have been adopted as energy sources for a variety of reasons.

What are these factors that lead to certain chemical substances and their reactions being used as sources of energy? Four factors will be discussed here: availability, energy density, pollution and renewability.

Availability

In many parts of the world, availability of the fuel is the most important factor. About half the world's population burns solid fuels that are easily obtainable locally: wood, coal or dried dung. In richer countries, natural gas (methane), liquid hydrocarbons and electricity are more prevalent, but these energy sources require more infrastructures to bring them to the population. Liquefied natural gas (e.g. propane or butane) can be piped into homes or provided in pressurized bottles. Liquid hydrocarbons are provided at filling stations by the roadside for easy replenishment of motor vehicle fuel tanks. Electricity is provided by a power grid, which typically extends across the entire country via overhead or underground cables. However, the electricity is not a source of energy in itself – it must be generated.

The question of availability is also connected to cost. Some sources of fossil fuels have been extractable using known technology but to do so has not been economically worthwhile, as other sources have been plentiful and relatively cheap to extract. However, as fossil fuels start to decrease, other extraction methods have started to be employed. For example, hydraulic fracturing ('fracking') is a controversial method used in parts of the USA.

Energy density or specific energy

In Chapter 5 the combustion of hydrocarbons in terms of the enthalpy change involved in burning 1 mole of the fuel was discussed. In real-life applications it is more useful to consider the energy that can be provided by the combustion of a certain mass or volume of the fuel, as this will influence the ease of transport of the fuel.

There are two ways of conveniently expressing the quantity of energy in a fuel: energy density and specific energy.

- *Energy density:* the energy released from the fuel, per unit volume of fuel consumed. Usually expressed in megajoules (MJ) per cubic decimetre (dm^3) or litre, $MJ\,dm^{-3}$.
- *Specific energy:* the energy released from the fuel, per unit of mass of fuel consumed. Usually expressed in megajoules (MJ) per kilogram (kg), $MJ\,kg^{-1}$.

Some approximate energy densities and specific energies are summarized in Table 24.1.

Table 24.1 Energy densities and specific energies of some common energy sources

Energy source	Energy density/$MJ\,dm^{-3}$	Specific energy/$MJ\,kg^{-1}$
Liquefied petroleum gas (LPG) – propane	26	46
Gasoline (petrol)	36	46
Kerosene	34	43
Coal		24
Wood		16
Lithium-ion battery	0.9–2.2	0.7
Uranium-235 (nuclear fission)		80 000 000
Deuterium-tritium (nuclear fusion)		330 000 000 (theoretical)

Nuclear fuels clearly have the highest energy densities, but they have enormous infrastructure requirements (discussed in Section 24.3). Fossil fuels have lots of energy stored within them, which can be easily and conveniently released on an industrial scale in power stations, or on a smaller scale in homes or vehicles. Compare these to lithium-ion batteries, as used in laptop computers and mobile phones and a small number of automobiles. These represent the cutting-edge of battery technology, but their energy densities are nowhere near those of fossil fuels – it can be seen that carrying a relatively small mass of fuel will give a motor vehicle or aeroplane a very long range, far in excess of that achievable using a battery for energy storage.

How can energy densities and specific energies be calculated?

To calculate the energy density or specific energy we can use the enthalpy of combustion, the molar mass of the fuel and the density of the fuel. If the fuel is a mixture of compounds (e.g. kerosene, which is a mixture of hydrocarbons with chain lengths around 6–16 carbon atoms) then the composition must also be known.

Worked examples

Example 1: specific energy

Octane is a major component of gasoline. Calculate the specific energy of octane.

Necessary data: $\Delta H^{\ominus}_{c}$: $-5530\,\text{kJ mol}^{-1}$; molar mass: $114.23\,\text{g mol}^{-1}$

1 mole of octane releases 5530 kJ when burnt, and this is 114.23 grams, so 1.00 g would release:

$$\frac{5530\,\text{kJ}}{114.23\,\text{g}} = 48.4\,\text{kJ g}^{-1}$$

Converting kJ to MJ involves dividing by 1000, but dividing by kg instead of g would involve multiplying by 1000, so the answer remains $48.4\,\text{MJ kg}^{-1}$.

Example 2: energy density

Ethanol can be obtained from plant matter by fermentation and is increasingly used as a liquid fuel in applications similar to gasoline (petrol). Calculate the energy density of ethanol.

Necessary data: $\Delta H^{\ominus}_{c}$: $-1371\,\text{kJ mol}^{-1}$; density: $0.785\,\text{g cm}^{-3}$; molar mass: $46.07\,\text{g mol}^{-1}$

1 mole of ethanol releases 1371 kJ when burnt, and this is 46.07 g, so 1.00 g would release:

$$\frac{1371\,\text{kJ mol}^{-1}}{46.07\,\text{g mol}^{-1}\,\text{kJ/g}}$$

1.00 g of ethanol is 1.00/0.785 cm^3 ethanol, so the energy released by 1.00 cm^3 is:

$$\frac{1371\,\text{kJ mol}^{-1} \times 0.785\,\text{g cm}^{-3}}{46.07\,\text{g mol}^{-1}} = 23.4\,\text{kJ cm}^{-3}$$

Converting kJ to MJ involves dividing by 1000, but dividing by m^3 instead of cm^3 would involve multiplying by 1000, so the answer remains $23.4\,\text{MJ dm}^{-3}$.

1 Calculate the specific energy and the energy density of liquefied propane (LPG).
Necessary data: $\Delta H^{\ominus}_{c}$: $-2202\,\text{kJ mol}^{-1}$; density: $0.493\,\text{g cm}^{-3}$; molar mass: $44.10\,\text{g mol}^{-1}$

The importance of energy density or specific energy in fuel choice

Fossil fuels such as gasoline and kerosene have a high specific energy, meaning that a comparatively small mass of fuel is needed for a given energy content. Fossil fuels remain the fuel of choice for most transport applications as they offer very long ranges. A modern car with a high-technology gasoline or diesel engine can travel around 800 miles (1290 km) on a tank of fuel, which far exceeds the range of the few electric cars on the market. In addition, the infrastructure for refuelling already exists.

Fossil fuels are used for air travel because airliners must carry their own fuel, so the maximum energy for the minimum mass is needed. (A Boeing 747's maximum take-off weight is about 330 tonnes, of which over a third is fuel!)

The energy density of the fuel itself does not tell the whole story. Different fuels require different amounts of energy to produce them. This may involve harvesting (in the case of biofuels like ethanol), drilling, refining and cleaning up pollution. When these factors are taken into account we arrive at a figure called the **embodied energy**, which must be subtracted from the energy content of the fuel if we are to have an accurate idea of the amount of energy obtained from that fuel.

Pollution

For the large proportion of people around the world who rely on solid fuels, the biggest pollution concern is that of particulates. When solid fuels are burnt, frequently there is an inadequate supply of oxygen, meaning that some of the carbon in the fuel is released as tiny particles of soot, or 'black carbon'. The US Environmental Protection Agency states that clean air should contain fewer than 15 micrograms of particles per cubic metre. An open fire can produce 300 times as much. Among the 3 billion people worldwide who rely on such fires, respiratory illnesses are a major cause of death.

In richer countries, great efforts were made to decrease localized pollution during the 20th century. Electronic management of vehicle engines, and catalytic converters in the exhaust, mean that motor vehicles release a fraction of the black carbon and carbon monoxide they released 30 years ago, which means that smog is far less common in cities in the USA and Europe. However, in rapidly growing economies such as China and India, pollution from motor vehicles remains a huge problem, with respiratory illnesses alarmingly prevalent in large traffic-congested cities.

The massive reduction in carbon, carbon monoxide and other sources of exhaust pollution, such as nitrogen oxides, achieved by mandating the use of catalytic converters, led many to believe that the problem of vehicle pollution was largely solved. However, it is impossible to use hydrocarbon fuels without releasing carbon dioxide, which is a greenhouse gas.

Current thinking on motor vehicle pollution is that, as well as cleaning the exhaust emissions by chemical means, it is increasingly important to improve the fuel economy of engines, to minimize the emission of carbon dioxide.

Energy efficiency

Proponents of electric vehicles suggest that they offer a greener alternative to gasoline or diesel, because the vehicles themselves produce no pollution. The vehicle is plugged into the power grid overnight and then can be used for commuting the next day. The average round-trip commute in the US is 32 miles (51 km) which is comfortably within the range of the few electric vehicles currently available.

However, the energy to charge the batteries must be obtained from somewhere. A proportion of the electricity supply may have come from renewable sources such as hydroelectric or wind power, but most is likely to have come from fossil fuel sources. Critics of electric vehicles argue that they just move the pollution somewhere else, at the cost of the convenience of a conventional hydrocarbon-fuelled vehicle.

Manufacturers of electric vehicles claim that the energy from the fossil fuels is more efficiently transferred to the vehicle by burning the fuel in a large, highly optimized industrial power station, and providing this energy to a vehicle via the power grid rather than burning the fuel in the vehicle itself.

A series of simple efficiency calculations can be used for a first comparison:

$$\text{efficiency of an energy transfer} = \frac{\text{useful output energy}}{\text{useful input energy}} \times 100\%$$

In the case of a typical petrol engine, the efficiency of the engine is approximately 18–20 per cent, meaning that only 20 per cent of the chemical potential energy available in the fuel is converted to kinetic energy moving the vehicle. Most of the rest of the energy is released as thermal energy, because of mechanical friction between the moving parts of the vehicle and heat losses from the engine block.

The electric motors used to move electric vehicles have an efficiency of around 95 per cent.

The lithium-ion batteries used in modern electric vehicles have a charge-discharge efficiency of around 85 per cent, meaning that 85 per cent of the energy transferred to the battery is usefully recovered and can be used to drive the motors.

What is the efficiency of the fossil fuel power station used to generate the electricity? For coal and oil-fired stations, the figure varies between 30 per cent and 49 per cent. We will use a middle value of 40 per cent for our calculation.

We can imagine we begin with 100 J of energy from the fuel and determine how much useful energy remains when it reaches the wheels of the vehicle.

Overall efficiency = efficiency of power station × efficiency of battery × efficiency of motor

$$\frac{40}{100} \times \frac{85}{100} \times \frac{95}{100} \times 100\% = 32\%$$

This overall efficiency exceeds that of a typical gasoline engine, meaning that, in theory, the electric car is a more efficient use of resources than the gasoline car.

The *theoretical* efficiency of a gasoline engine is higher than 18–20 per cent – it is about 37 per cent. This is determined by thermodynamic principles and cannot be bettered. In theory, would a highly efficient gasoline engine be preferable? Actually no, because the gasoline used for vehicle engine has a very large 'embodied energy'. The extraction and refining of the gasoline uses energy equivalent to 1.4 times the actual energy content of the gasoline itself. Coal and oil used in fossil fuel power stations require far less refining.

Worked example

Energy efficiency – a sample calculation:

A car carries 50 litres of gasoline. How much energy is stored in this volume of gasoline? Assuming the car's engine is 18 per cent efficient, how much energy is actually available to propel the car?

Energy density of gasoline: $36\,\text{MJ}\,\text{dm}^{-3}$

Total energy in fuel = energy density × volume of fuel

$$36\,\text{MJ}\,\text{dm}^{-3} \times 50\,\text{dm}^{3} = 1800\,\text{MJ}$$

Efficiency of engine = 18%

So energy available for car = 18/100 × 1800 = 320 MJ (to 2 significant figures)

2 A coal-fired power station consumes 500 tonnes of coal per hour. How much energy is stored in this mass of coal? Assuming the power station is 40 per cent efficient, what is the output of the station per hour? What is the output in watts (joules per second)? Specific energy of coal: $24\,\text{MJ}\,\text{kg}^{-1}$.

Renewability

Renewable energy sources are those that are naturally replenished on a time scale useful to humans.

Non-renewable energy sources are finite, so once they have been used up they will no longer be available to us.

Non-renewables

Non-renewable energy is basically synonymous with fossil fuels. Coal, oil and natural gas will eventually run out, and so alternative energy sources are being sought. In addition, the carbon dioxide, and resultant global warming, produced by fossil fuels has led to great interest in alternative sources of energy.

The realization that fossil fuels will run out led to the concept of 'peak oil'. Peak oil refers to the point at which production of liquid hydrocarbons is at a maximum, before going into decline owing to decreasing availability of such fuels. Peak oil was first predicted to occur in the late 1960s, but discoveries of new oil deposits keep pushing it back. The International Energy Agency says that production of 'conventional' crude oil peaked in 2006, but 'unconventional' sources of liquids have kept production at high levels. Unconventional sources refers to processes such as coal-to-liquid conversion, gas-to-liquid conversion and the extraction of oil from 'tar sands' – loose sand saturated with very heavy hydrocarbons, which require heating and dilution with lighter hydrocarbons to extract them. If unconventional sources are included, then peak oil may still be in the future, but not too far ahead. Some estimates place it in about 2020.

Renewables

Renewable energy sources are often referred to as 'alternative' energy or 'green' energy.

The International Energy Agency defines renewable sources as follows: '*Renewable energy is derived from natural processes that are replenished constantly. In its various forms, it derives directly from the Sun, or from heat generated deep within the Earth. Included in the definition is electricity and heat generated from solar, wind, ocean, hydropower, biomass, geothermal resources, and biofuels and hydrogen derived from renewable resources.*'

In contrast to non-renewable sources, renewables are highly dispersed over the Earth, meaning they should be equally available to everyone, in varying proportions. Non-renewables, in the form of fossil fuels, tend to be concentrated in certain countries or regions, meaning that some countries have become very wealthy on the basis of an accident of geography.

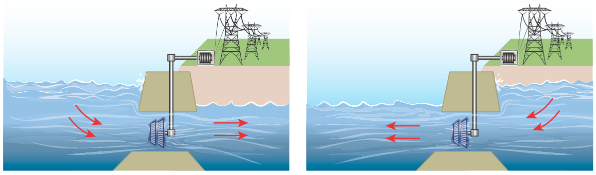
The advantages and disadvantages of energy sources

Table 24.2 summarizes the advantages and disadvantages of energy sources.

Source	Where does it come from?	Renewable or non-renewable?	Advantages	Disadvantages
Coal	Fossilized plants found in seams between layers of rock in the Earth Mostly comprised of carbon with some inorganic impurities such as sulfur Mined by drilling shafts into the Earth	Non-renewable	Cheap to extract High specific energy Energy is easily released – simply burn it Large infrastructure already exists for transport and burning Coal reserves will last much longer than oil reserves – still several hundred years remaining	Releases carbon dioxide when burnt – a greenhouse gas Sulfur impurities release sulfur dioxide when burnt, which leads to acid rain – factory emissions must be scrubbed to remove this Coal mining is a dirty and dangerous activity
Oil	Fossilized animal remains in liquid form, found soaked into layers of porous rock in the Earth; a layer of non-permeable rock above the oil reservoir prevents it from escaping Pipes are sunk into the Earth to release the oil and it is pumped to the surface	Non-renewable	Cheap to extract High specific energy Energy is easily released – simply burn it Large infrastructure already exists for transport and burning Incredibly versatile – different crude oil fractions have various physical properties making them suitable for many different fuel applications Also a chemical feedstock for plastics industry and pharmaceutical industry	Releases carbon dioxide when burnt – a greenhouse gas Will run out quite soon – conventional reserves may have already peaked

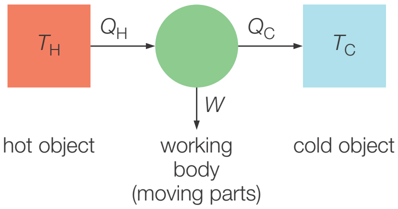
■ **Table 24.2** Advantages and disadvantages of energy sources

Natural gas	Usually associated with crude oil. Natural gas consists of lighter molecules (mostly methane, CH_4) which are gaseous at surface pressure and temperature Pipes are sunk into the Earth to release the oil and it is pumped to the surface	Non-renewable	Cheap to extract High specific energy Energy is easily released – simply burn it Large infrastructure already exists for transport and burning Can be piped directly into homes	Releases carbon dioxide when burnt – a greenhouse gas. Will run out quite soon
Nuclear	Uranium is mined and refined Nuclear 'chain reaction' involving fission of uranium nuclei generates large amounts of heat Heat used to generate steam which can then generate electricity via turbines	Non-renewable	Uranium is cheap to mine Enormous specific energy Fission process itself generates no carbon dioxide or other greenhouse gases	Uranium reserves are finite, although some reactor types can regenerate nuclear fuel Nuclear waste is highly toxic, long-lasting and difficult to dispose of safely Danger of weapons proliferation
Wind	Wind turns turbines that generate electricity	Renewable	In order to produce a useful amount of electricity large numbers of turbines are placed together in a 'wind farm'	Some people object to wind farms on the grounds that the noise and appearance spoils the landscape The wind does not always blow – wind farms are more effective in some regions than others
Solar	Sunlight falls on a solar panel which captures its energy and converts it into electricity	Renewable	In theory, solar power could offer infinite supplies of energy, especially in very hot, sunny countries	Currently, solar panels are inefficient and too costly to be adopted widely, but technology is improving
Tidal/wave	Energy from water moving back and forth is captured by underwater turbines and used to generate electricity (Figure 24.1)	Renewable	Countries with long coastlines could potentially benefit from tidal or wave power	Only suitable for countries with long coastlines Infrastructure is costly and technology is undeveloped Tidal or wave generators need to be very robust and reliable to resist conditions at sea
Hydroelectric power (HEP)	Flow of water from high ground towards oceans is used to generate electricity. A river is dammed, and the flow of water through the dam is used to turn turbines (Figure 24.2)	Renewable	In suitable geographical locations, HEP offers a clean, cheap electricity supply	Limited to mountainous areas Building dams often requires flooding of surrounding areas which may displace populations and lead to loss of biodiversity Dammed rivers also lead to build-up of silt and affects the river ecosystem
Biofuels	Plant or animal material is used to make a liquid or gaseous fuel Biodiesel is an oil extracted from plants such as palm trees Ethanol can be made from corn (maize), sugar cane or the inedible cellulose from grasses or plant waste Biogas is made by fermenting plant or animal waste	Renewable	In theory, biofuels are 'carbon neutral' – the carbon that is released when they are burnt is the same carbon that they absorbed during photosynthesis Liquid biofuels allow existing fuel infrastructure and technology to be retained as they are a direct replacement for fossil fuels	Only renewable if plants are replanted In some countries (e.g. Malaysia), tropical rainforest has been replaced with palm oil plantations, leading to loss of biodiversity Some methods of extraction of biofuels are very energy inefficient Economic demand for biofuels leads to countries displacing food crops. This could lead to lack of food for population
Wood/ charcoal	Trees can be felled and burnt for energy Slowly roasting wood in the absence of oxygen drives off water and organic matter and turns wood into charcoal, which has a higher energy density	Renewable	Wood is widely available to most of the Earth's population and is a convenient source of energy for cooking and heating	Wood is only renewable if trees are replanted Often, wood is burnt in highly inefficient open fires and stoves, with enormous amounts of dangerous particulate emissions
Geothermal	Water is pumped into the Earth where the natural heat turns it into steam. The steam can be used directly for heating or used to generate electricity	Renewable	Geothermal energy is most practical in regions where there is a lot of volcanic activity such as Iceland and New Zealand. In these countries it has the potential to provide unlimited clean energy	Infrastructure can be difficult to engineer and is costly



■ **Figure 24.1** Tidal power captures the energy of tidal water moving backwards and forwards. This tidal electricity generation works as the tide comes in and again when it goes out. The turbines are driven by the power of the sea in both directions

■ **Figure 24.2** The Hoover Dam, on the Colorado River, USA, generates about 2000 MW from its hydroelectric turbines



■ **Figure 24.3** Carnot's heat engine. The thermal energy flows (Q_H, Q_C) from the hot body (T_H) to the cold body (T_C) via the working body (the circle). This working body does useful work (W) on the surroundings

■ The quality of energy and available work

During his analysis of energy 'quality' in the 1850s, William Thomson studied the steam engine. He recognized that the kinetic energy within the hot steam led to its expanding against a piston, driving the engine.

The steam engine, in very basic terms, is a heat engine. The flow of heat from one region to another produces useful work. Nicolas Leonard Sadi Carnot proposed the heat engine concept in 1824 (Figure 24.3).

After use, however, all the components of the engine are at thermal equilibrium and there is no longer any useful work to be extracted. However, the total energy has been conserved.

So, to use Thomson's terminology, the availability of the energy has been lost. During the workings of the steam energy, the quality of the energy has been degraded. The driving force of the engine is the transfer of heat from a region of highly concentrated, highly useful energy (the fuel) to the surroundings.

International cooperation in energy provision

Two international agencies exist to monitor and coordinate the provision of energy for different nations:

- The International Energy Agency (IEA) is an autonomous organization based in Paris, whose goal is to ensure reliable, affordable and clean energy for its 28 member countries and beyond.
- The International Renewable Energy Agency (IRENA), based in Abu Dhabi, seeks to promote increased adoption and sustainable use of renewable energy sources.

The IEA was established in the wake of the 1970s oil crisis, in an attempt to ensure that supplies of oil worldwide remained steady and were not subject to periodic 'shocks' that could cripple the world economy. The IEA maintains a stock of oil that can be released at times of crisis to prevent oil supply disruptions. The IEA is an autonomous organization, not under the control of any particular government, so the intention is that countries are less able to use oil supply as a 'weapon' against other countries. However, the IEA is criticized for two main reasons. Firstly, several very large and powerful countries are not members, including India, China and Russia. Secondly, critics claim that the IEA is too heavily influenced by the more powerful members, such as the USA, which has affected its responses to big issues such as 'peak oil' and climate change.

Although one of the stated aims of the IEA is to consider environmental factors surrounding energy, IRENA was founded in 2009 with the goal of becoming the main driver for increased adoption of renewables worldwide. In particular, it aims to help countries share expertise around the provision of infrastructure for renewables and setting up regulatory frameworks. IRENA also has its critics: when it was announced that it would be based in Abu Dhabi, critics pointed to the enormous carbon footprint of the United Arab Emirates and suggested that the country was a poor role model for renewables adoption. Despite this, over 100 states have joined IRENA, and China has announced its intention to join in the near future.

ToK Link

Factors affecting research priorities

'*I have no doubt that we will be successful in harnessing the Sun's energy. If sunbeams were weapons of war we would have had solar energy centuries ago*' (Lord George Porter). In what ways might social, political, cultural and religious factors affect the types of research that are financed and undertaken, or rejected?

Lord Porter's quote suggests that sociopolitical factors (war being an extreme example) can be catalysts for technological developments. Money is spent on acquiring knowledge for many reasons – social, political, cultural or ideological. This was certainly the case in the development of atomic energy in the 20th century. According to the Brookings Institute, the USA invested about 20 million dollars developing atomic bombs, the expense being justified in terms of beating the Germans to it. After the war, the USA wanted to stay ahead of the Soviet Union, so continued to invest heavily in atomic research. Images of the tragedies at Hiroshima and Nagasaki caused a backlash against this enormous and dangerous source of power, and so the USA tried to improve the image of nuclear power by promoting it as a clean source of energy – 'Atoms for Peace'. For a while it was thought that civilian nuclear power generation would result in electricity becoming 'too cheap to meter' and a new golden age of practically unlimited energy would commence. It never happened – plentiful fossil fuels and the difficulties of waste disposal and heavily publicized accidents meant that the nuclear industry soon became moribund. This led the *Economist* to describe nuclear power as 'the dream that failed' in 2012.

The Cold War in the latter half of the 20th century led to other politically driven projects. The Apollo space programme in the USA cost 25 billion dollars. Although it is sometimes claimed that many useful inventions arose from the space programme, many believe that the ultimate reason for the rapid development of the moon rocket, and the enormous costs associated with it, was competition with the Soviet Union. This might be seen as an attempt to demonstrate the superiority of one political ideology over another.

Sometimes economic and social factors are important. In the early 20th century, when the automobile industry was starting out, electric cars were more prevalent than gasoline cars. What factors led to gasoline winning the edge, leading to the colossal growth in oil industry infrastructure and power through the century? Economic factors were important: the discovery of crude oil reserves all over the world brought the price of gasoline down. As countries developed better road networks, the limited range of electrics became a problem. This led to a cycle of infrastructure development – more roads led to more cars (and mostly gasoline cars, to take advantage of the range available), which led to more roads to accommodate the cars.

There were also social factors at work. In the 20th century the automobile became synonymous with freedom, progress and opportunity as the automobile makers successfully made having a car seem necessary to a fulfilling life. The process is now repeating itself in China and India as they develop.

ToK Link

Ethics of energy generation

There are many ethical issues raised by energy generation and its consequent contributions to pollution and climate change. What is the influence of political pressure on different areas of knowledge?

The European Group on Ethics in Science and New Technologies (EGE) is a committee of the European Commission in Brussels, which attempts to explore the ethical issues surrounding energy production. They propose that international regulatory policies should take into account three main elements: societal impact, economic impact and environmental impact. The committee tries to apply the principle of justice to ensure that all EU citizens have access to energy, that the EU maintains a secure energy supply, and that the policies implemented are sustainable. At times these different considerations come into conflict, so it is the job of politicians to find a balance between them. Different political parties in different countries may disagree with the balance proposed; for example, some feel that sustained economic growth should override all other concerns. Others feel that we have a moral obligation to future generations to protect our environment.

At present, renewable energy sources have yet to match fossil fuels in terms of low cost and ease of production so governments offer funds to promote investment. Governments have to justify this expenditure and the funding is subject to political pressure. Some countries, such as Germany, have very active 'green' movements with a great deal of political influence. In other countries large energy companies are able to exercise their influence to ensure that energy policies continue to favour fossil fuels.

24.2 Fossil fuels *– the energy of fossil fuels originates from solar energy which has been stored by chemical processes over time. These abundant resources are non-renewable but provide large amounts of energy due to the nature of chemical bonds in hydrocarbons*

Nature of Science

The importance of fossil fuels in science

Large-scale fossil fuel use is a relatively recent development. Underground coal mining was developed in the late 18th century, and oil drilling began in earnest in the 19th century.

The industrialization of the world was driven by fossil fuels. Coal led to steam engines, which led to steam ships and mass manufacturing in factories. Factories drove urbanization and steam ships drove migration. As oil gradually overtook coal, motor vehicles and aircraft were developed, which led to further migration, mass travel and globalization, and cities grew outwards, with suburban zones for those who could afford cars.

This progress was driven by science and technology, as scientists around the world invented and developed more efficient ways to utilize fossil fuels for transport and power generation.

However, the existence of fossil fuels has also provided the scientific community with an enormous range of chemicals to draw upon in other applications. Before the mid-19th century, those drugs that existed were usually extracted from plants. In 1869, the first synthetic drug, chloral hydrate (a sedative) was introduced, leading to the multi-billion-dollar pharmaceutical industry today. Petroleum by-products are also used in manufacture of dyes and plastics.

In short, fossil fuels have provided both a 'means' to scientific progress, as they have provided raw materials for many scientific fields, as well as an 'end' as scientists have reacted to demand for fossil fuels to develop more and more efficient ways to use them.

■ What are fossil fuels?

Fossil fuels are fuels that were formed by the anaerobic decay of biological material from prehistoric times that was buried and acted upon by bacteria over millions of years.

Fossil fuels are rich stores of energy, because they are stores of solar energy, which drove the photosynthetic processes in prehistoric plants. Photosynthesis takes carbon dioxide from the atmosphere and converts it into more complex molecules, starting with glucose. The plants then use some of the glucose to form structural molecules (cellulose). Plant matter therefore contains carbon along with hydrogen and oxygen. In addition, protein molecules found in biological systems contain nitrogen and sulfur in significant quantities.

Photosynthesis is an endothermic process – it requires an energy input to drive the reaction. This energy input from the Sun becomes 'locked up' within the biological molecules present in the plants. Animals eat the plants and the energy passes into them (where it is used to provide energy for the life of the organism). When the plant or animal dies, much of the energy may stay 'locked up' in the remains.

Over geological time, prehistoric plant and animal matter is compressed and becomes highly concentrated. This means that fossil fuels have energy highly concentrated within them, which can be released by combustion.

The three major categories of fossil fuels are coal, crude oil (petroleum) and natural gas.

Coal

Coal is currently the most plentiful of the fossil fuels. It originated as prehistoric forests, which were flooded, buried and then gradually compressed by layer upon layer of soil.

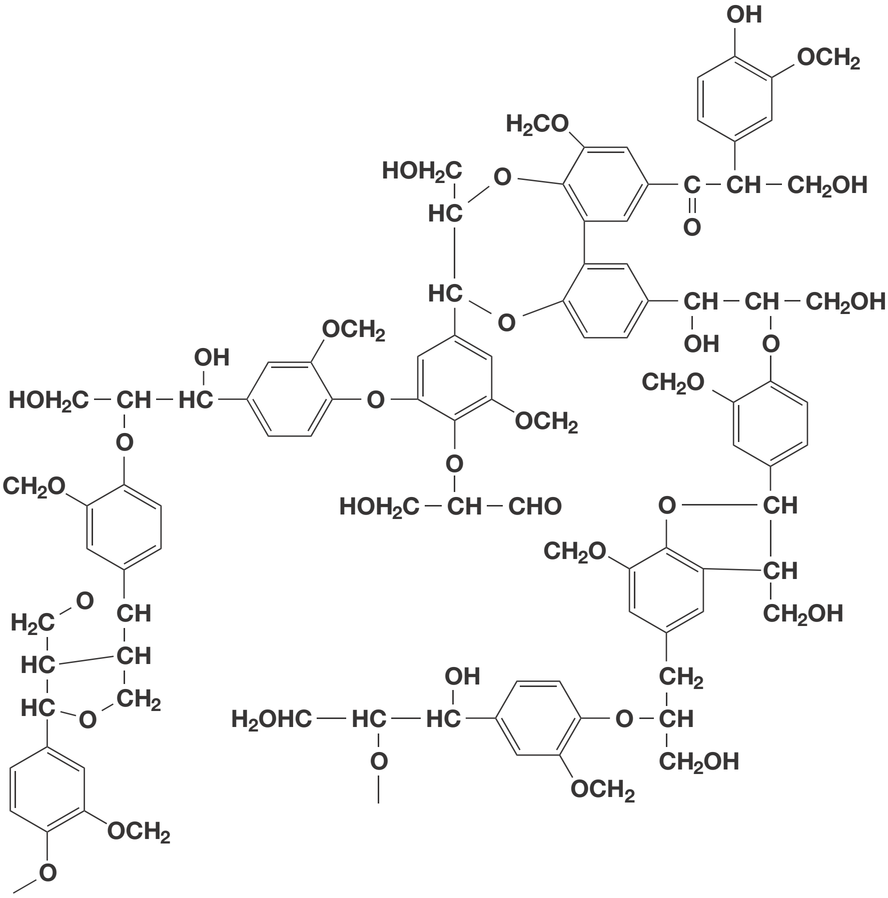
Most dead plant material decomposes where it falls. However, when large quantities of plant matter are buried and isolated from oxygen, it is acted upon by anaerobic bacteria which gradually reduce the biological molecules until a substance consisting mostly of carbon results.

Not all coal is the same, however; different types of coal reflect different stages in the decomposition process:

Peat still contains significant amounts of other chemical elements – the decay process is incomplete. In particular, peat contains quite a lot of water, and the carbon is still found combined with other chemical elements such as hydrogen, oxygen and sulfur.

Lignite (sometimes called 'brown coal') is only about 25–35 per cent carbon and contains a high percentage of water – up to 60 per cent. However, the biological molecules have started to polymerize forming a substance called lignin, characterized by large numbers of benzene rings (Figure 24.4).

Figure 24.4 Lignin



Bituminous coal ('black coal') is a sedimentary rock containing 60–80 per cent carbon but with hydrogen, sulfur and oxygen still present. In particular, the rock is saturated with tar-like hydrocarbon molecules called 'bitumen'. When heated, bituminous coal releases hydrocarbon vapours.

Anthracite coal is the hardest form of coal and the closest to pure carbon. When bituminous coal is subjected to intense heat and pressure beneath the Earth, most of the hydrogen, oxygen and sulfur are driven off, leaving a hard, glossy rock containing 92–98 per cent carbon. Anthracite does not release hydrocarbons when heated.

As the carbon content of the coal increases through the formation processes, the energy density of the coal also increases. Although anthracite coal is the most desirable form in terms of the energy density, other forms of coal are often used, mostly because they are plentiful and cheap in local situations. For example, peat is an economically important fuel in Ireland.

Crude oil and natural gas

The term *petroleum* comes from the Latin *petra,* meaning rock, and *oleum*, meaning oil. Some definitions of petroleum include both crude (unprocessed) oil as it is retrieved from the rock, and the products of refining. The 'petroleum industry' is usually taken to include both the oil and gas industries.

Crude oil remains one of the most important raw materials in the world today. It is a complex mixture of hydrocarbons and supplies us with fuels for a range of transport types and for electricity generation. In addition it is an important chemical feedstock for the production of important organic polymers, pharmaceuticals, dyes and solvents.

Crude oil was formed over geological periods of time (that is, millions of years) from the remains of marine animals and plants. These creatures died and were sedimented at the bottom of the oceans, where they became trapped under layers of rock. Under these conditions of high temperature and high pressure, this matter decayed in the presence of bacteria and the absence of oxygen to form crude oil (petroleum) and natural gas. It is a limited resource and eventually reserves will be so depleted that chemists will need to consider other sources of carbon, both as a fuel and as a chemical feedstock. Indeed, the balance of these two uses is an issue if we are to conserve this non-renewable resource for as long as possible.

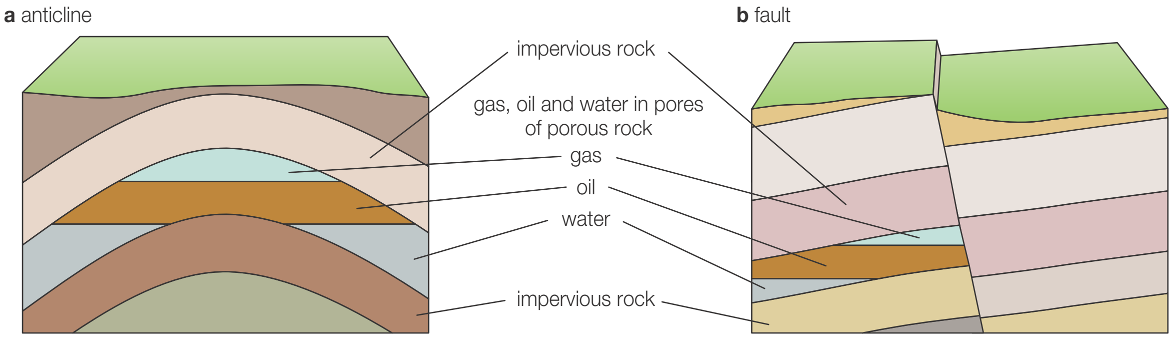


Figure 24.5 Crude oil is found in underground reservoirs in certain geological situations. It permeates the rock in these reservoirs and is usually found in association with natural gas

Crude oil varies greatly in appearance depending on its composition. It is usually black or dark brown. In the underground reservoirs it is usually found in association with natural gas, which forms a gas cap over the crude oil (Figure 24.5). Crude oil may also be found in a semi-solid form mixed with sand and water, as in the Athabasca Oil Sands in Canada, where it is usually referred to as crude bitumen. Venezuela also has large amounts of crude oil in the Orinoco Oil Sands. These oil sands resources are called unconventional crude oil to distinguish them from oil which can be extracted using traditional oil well methods.

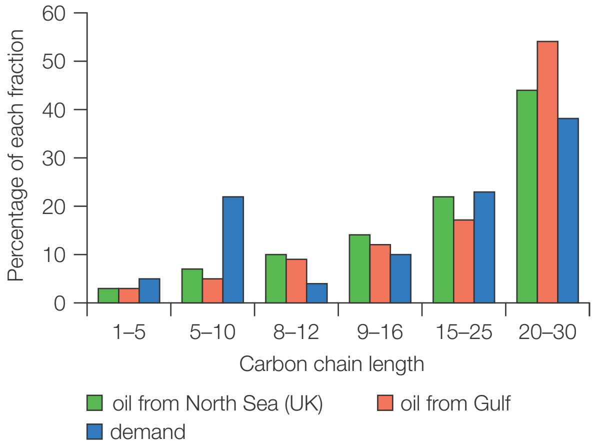


Figure 24.6 Supply and demand of crude oil fractions

The proportion of hydrocarbons in the crude oil (petroleum) mixture differs between oil fields, with samples varying in the balance between the lighter, more profitable fractions and the heavier oils and bitumen fractions. The distribution of the different fractions in the oil from a particular field may well not match the local or international demand and so post-distillation methods of chemically modifying the initial fractions are economically important.

Figure 24.6 shows the distribution of crude oil fractions from North Sea oil and oil from the Persian Gulf, compared to demand for fractions in the UK.

Hydrocarbon chains with five to ten carbon atoms are in high demand as these are used for motor vehicle fuels (gasoline). The supply of these fractions is much lower than the demand. For longer, heavier fractions, supply outweighs demand. However, these longer-chain molecules can be 'cracked' to form shorter chains.

Fractional distillation

Properties of hydrocarbons

The various hydrocarbons in crude oil have different properties depending on their molecular structures.

Short-chain molecules (those with small numbers of carbon atoms) have low boiling points, because they have small formula masses and relatively few electrons in each molecule. The instantaneous dipoles generated during molecular vibrations are quite weak, as are the induced diploes in neighbouring molecules, so the dispersion forces are also weak. Short-chain molecules (one to four carbons) are actually gases at room temperature.

Low-boiling-point hydrocarbons are more volatile, meaning that they release more vapour into the surrounding air. When the surrounding air is saturated with fuel vapour the fuel ignites easily.

As the carbon chains get longer, the formula masses increase, along with the number of electrons in the molecules. In addition, the contact area between adjacent molecules in the liquid will also increase. This leads to stronger dispersion forces and higher boiling points. When the boiling point is higher, the volatility is lower. There is less vapour in the air surrounding the fuel, and the fuel is more difficult to ignite.

Stronger dispersion forces also mean that longer hydrocarbon molecules are more viscous (less 'runny').

Separating the compounds in crude oil

Crude oil is a mixture of hundreds of different hydrocarbons, all with slightly different properties. Some are heavy, thick, viscous liquids with a high energy density but are difficult to ignite. Others pour easily and are easily ignited but have less energy stored within them. These different hydrocarbons with different properties find different uses.

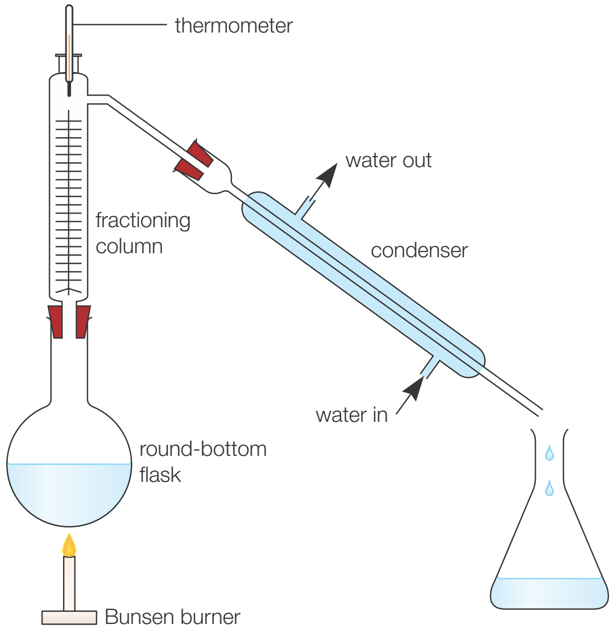


Figure 24.7 Apparatus for fractional distillation of crude oil in a school laboratory

The first step in oil refining involves separating the different hydrocarbons according to their boiling points. It would be impractical to separate every single compound, but it is possible to separate the crude oil into broad groups of similar compounds called fractions.

This process of separating out the oil by boiling point is called **fractional distillation**.

You may have seen a sample of crude oil heated to separate it into its components in the school laboratory (Figure 24.7). As the mixture of compounds increases in temperature, the components will reach their boiling points one by one. They become vapour and can then be condensed into separate containers.

Industrial fractional distillation turns this process on its head. Instead of gradually raising the temperature of the oil mixture and boiling off each fraction, the crude oil is heated to a high temperature so that almost all the compounds within it become vapour. This vapour is injected into the base of a vessel called the **fractionating column**.

The vapours rise up the column. As they get further from the heat source at the base of the column, the temperature falls. Gradually, each fraction condenses out of the vapour mixture. Heavy fractions with high boiling points condense first, close to the bottom of the column. As the column is ascended, successively lighter, shorter chain fractions condense. At the top of the column some very short-chain molecules are still in the vapour phase and are collected as refinery gas. Figure 24.8 summarizes the different fractions from crude oil and their uses.

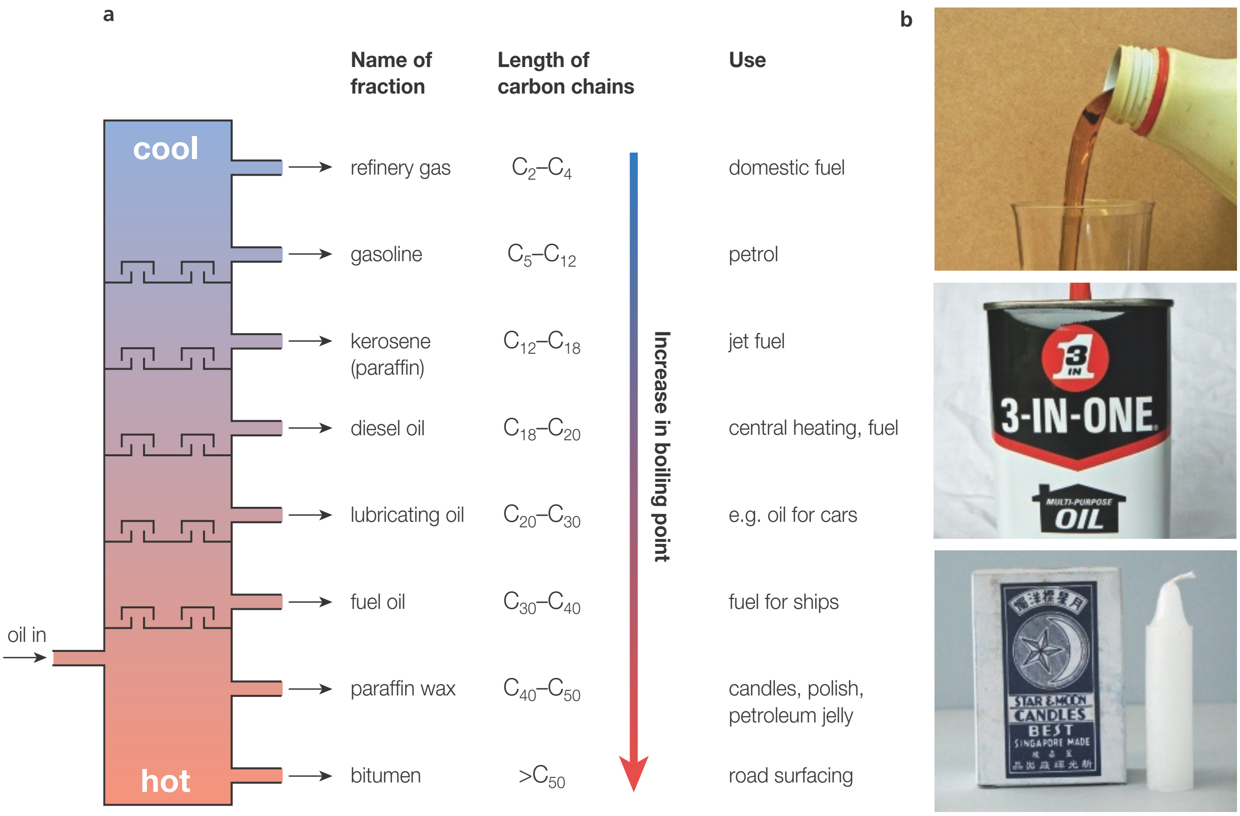


Figure 24.8 a The fractions obtained industrially from crude oil using a fractionating tower; **b** some examples of useful components

The importance of branched-chain alkanes in gasoline

In the internal combustion engine, hydrocarbon fuels are mixed with air, injected into a cylinder and ignited with a spark. The resultant explosion forces a piston to move. The movement of a series of pistons is converted to the rotation of a crankshaft, which in turn drives the wheels of the vehicle.

If we consider a molecule of octane (a component of petrol, or gasoline), we can see that for complete combustion to occur, with full oxidation of every atom of carbon and hydrogen, we require 12.5 moles of oxygen for each mole of octane:

$$2C_8H_{18}(l) + 25O_2(g) \rightarrow 16CO_2(g) + 18H_2O(l)$$

This equates to a mass ratio (mass of air : mass of fuel) of approximately 15 : 1. (Remember that only 20 per cent of the air is oxygen.)

If the air : fuel ratio is lower than this, the mixture is said to be 'rich' and the amount of oxygen is insufficient for complete combustion. This will result in the formation of some carbon monoxide. In this example only 10 moles of oxygen is available for each mole of octane:

$$C_8H_{18}(l) + 10O_2(g) \rightarrow 3CO_2(g) + 5CO(g) + 9H_2O(l)$$

A rich mixture may also lead to the emission of unburned hydrocarbon molecules (called 'volatile organic compounds' or VOCs) from the exhaust.

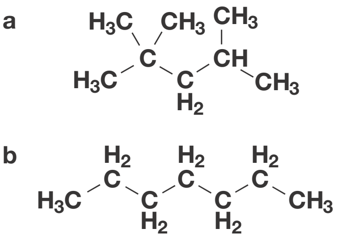
If the air : fuel ratio is increased, the mixture is said to be 'lean'. A lean mixture will not produce carbon monoxide. A disadvantage of a lean mixture is that the mixture may 'pre-ignite' while under compression. Pre-ignition means that the rapid compression of the fuel–air mixture in the cylinder leads to some of the fuel igniting before the spark, causing a sudden pressure increase in the cylinder. This causes a metallic 'knocking' sound. This inefficient combustion actually increases the proportion of unburnt hydrocarbons in the exhaust gases.

It is advantageous to blend the hydrocarbons in the fuel to minimize the chances of knocking. This means that the fuel–air mixture can be as lean as possible, reducing fuel consumption and decreasing the chances of VOCs, carbon monoxide and soot being emitted by the vehicle. Also, high-performance engines in luxury or sports cars often have very high compression ratios, which makes pre-ignition likely with poor quality fuels.

It is known that branched-chain hydrocarbons resist pre-ignition better than straight-chain hydrocarbons. The measure of resistance to pre-ignition is called the **octane number**.

The effect of chain branching on the octane number

Octane number is the measure of a fuel's tendency to resist pre-ignition. Octane number is measured on a scale of 0–100, where 100 refers to 2,2,4-trimethylpentane (historically called *iso*-octane – a highly branched molecule), and 0 refers to heptane (a straight-chain molecule). Figure 24.9 shows these two compounds.



■ **Figure 24.9 a** 2,2,4-trimethylpentane: octane number 100; **b** heptane: octane number 0

The octane number of a fuel blend is measured by comparing it to these two benchmarks. This is done by measuring the compression that the fuel sample can withstand before igniting.

The octane number can be increased by adding branched-chain alkanes to the mixture. Arenes (aromatic compounds, those with benzene rings) also increase the octane number.

Gasoline direct from the fractionating column has a large proportion of straight-chain hydrocarbons in it, giving an octane number of around 70, which is not good enough to be sold as motor fuel. Further stages of refining must take place, and these stages cost energy and money.

■ Cracking and reforming

As we have seen, the supply of hydrocarbon fractions rarely matches demand for those fractions. Typically, the larger fractions are more plentiful, whereas the highest demand is for the gasoline fraction, with carbon chain lengths ranging from around five to ten carbon atoms.

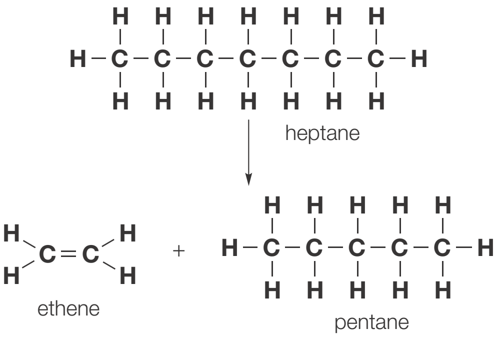
An obvious solution to this problem is to break down the larger molecules into smaller ones. This process is called catalytic cracking. The heavier fractions are heated and passed over a catalyst, which 'cracks' them into shorter chains.

Fractional distillation is a physical process, relying on boiling the crude oil and then condensing its component compounds. Intermolecular forces are overcome but no chemical bonds are broken. Catalytic cracking, by contrast, is a chemical reaction – a form of thermal decomposition. Longer-chain molecules have their covalent bonds broken to split the longer molecules into shorter ones.

The original alkane, having the general formula C_nH_{2n+2}, usually results in two products: a shorter alkane (also C_nH_{2n+2}, and an alkene, C_nH_{2n}). It is not possible to crack an alkane into two shorter alkanes – there are not enough hydrogen atoms present.

For example:

$$\underset{\text{heptane}}{C_7H_{16}} \rightarrow \underset{\text{pentane}}{C_5H_{12}} + \underset{\text{ethene}}{C_2H_4}$$



■ **Figure 24.10** Cracking a long saturated hydrocarbon molecule

The reaction shown in Figure 24.10 is a straightforward example starting from a relatively short alkane to illustrate the type of reaction being discussed. Note that the alkene splits off the end of the molecule. The process is not a highly specific one. Ethene is not always the alkene product; it is possible to obtain propene and butene as products that are useful in their own right. In certain circumstances hydrogen can also break off the end of the alkane molecule. Some more complex examples of possible reactions are:

$$\underset{\text{hexadecane}}{C_{16}H_{34}} \rightarrow \underset{\text{undecane}}{C_{11}H_{24}} + \underset{\text{ethene}}{C_2H_4} + \underset{\text{propene}}{C_3H_6}$$

$$\begin{array}{ccccc} C_{12}H_{26} & \rightarrow & C_8H_{18} & + & C_4H_8 \\ \text{dodecane} & & \text{octane} & & \text{butene} \end{array}$$

$$\begin{array}{ccccc} C_{10}H_{22} & \rightarrow & C_8H_{18} & + & C_2H_4 \\ \text{decane} & & \text{octane} & & \text{ethene} \end{array}$$

Different types of cracking processes have different applications. Some are more useful for producing large quantities of alkenes, which can then be used to make plastics. Other kinds are used predominantly in fuel production, as they produce large proportions of branched products. The different types of cracking are summarized below:

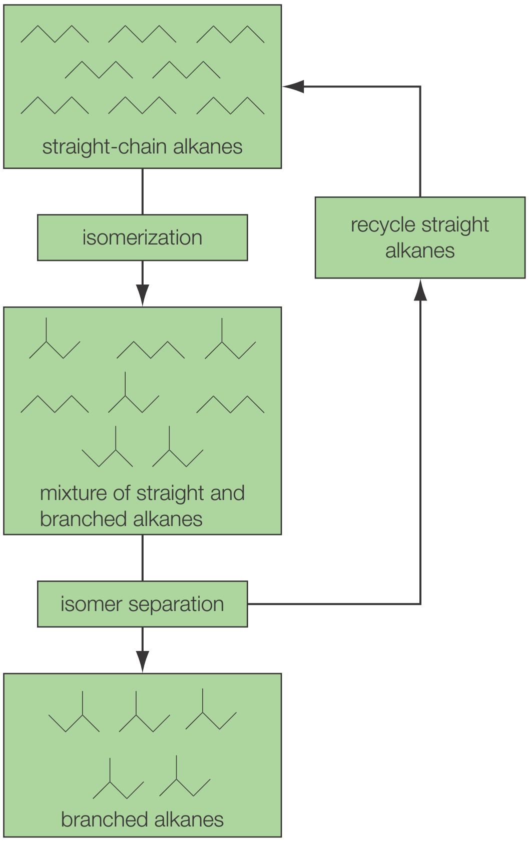
- *Thermal cracking* is carried out by heating the very long-chain alkanes found in the heavy fractions to temperatures of 800–850 °C at pressures of up to 70 atmospheres, and then cooling the mixture rapidly. Under such conditions, a free radical mechanism occurs and a mixture of products is produced that includes shorter-chain alkanes, alkenes and coke (an impure form of carbon). Ethene is a favoured product as it is the key starting material for the preparation of other chemicals such as poly(ethene) and ethanol.
- *Steam cracking* is a different form of thermal cracking. The initial reaction mixture of ethane, butane and alkanes up to eight carbon atoms long is preheated, vaporized and mixed with steam at 1250–1400 °C. The steam dilutes the feedstock and the reaction produces a higher yield of ethene and other short-chain alkenes. The addition of steam also reduces the amount of coke produced, which otherwise produces an unwanted lining in the reaction vessel.
- *Catalytic cracking* uses a catalyst in the cracking process to allow the reaction to occur at lower temperatures, around 500 °C, and gives a higher degree of specificity to the process by controlling the steps involved in the mechanism. The reactions are complicated and involve an ionic (carbocation) mechanism. The carbocations are produced and then undergo a rearrangement reaction on the catalyst surface. Large and intermediate-sized alkanes are passed over a mixed alumina (Al_2O_3) and silica (SiO_2) catalyst, which is in powdered form to increase its surface area. **Zeolites** (see below), naturally occurring minerals containing aluminium, silicon and oxygen, are also very good catalysts for this process as their crystal structures contain an extensive network which offers the hydrocarbons a large surface area for reaction. Catalytic cracking produces a mixture of alkanes, alkenes and molecules containing the benzene ring (arenes). A high proportion of branched alkanes is produced which can then be blended into petrol to increase the octane number. Some carbon is, however, formed during the process and this can coat the catalyst and stop it working. The catalyst must be cleaned or regenerated regularly by steam jets followed by heating, otherwise it becomes coated in carbon.
- *Hydrocracking* occurs when the feedstock is mixed with hydrogen at a pressure of about 80 atmospheres and cracked over a platinum or silica/alumina catalyst. This gives a high yield of branched alkanes, cyclic alkanes and aromatic compounds for use in unleaded petrol (gasoline). The presence of hydrogen ensures that no alkenes are produced in this type of cracking.

Petrol produced by cracking longer hydrocarbons will tend to have reasonably high octane numbers because the cracking method chosen (usually catalytic cracking or hydrocracking) will produce lots of branched molecules.

However, it is also necessary to utilize the so-called ‘straight-run’ gasoline produced in the fractionating column. The octane number of straight-run gasoline can be increased by subjecting it to a series of **reforming** processes that will increase the proportion of branched and aromatic compounds in the mixture.

Zeolites

The term zeolite is derived from the Greek words *zeo,* meaning to boil, and *lithos,* meaning stone. Zeolites release large amounts of steam when heated. A zeolite has a three-dimensional structure in which the silicon, aluminium and oxygen atoms form a framework of tunnels and cavities into



■ **Figure 24.11** Isomerization

which small molecules, such as water, can fit. Many occur naturally as minerals and are extensively mined in many parts of the world. Others are synthetic and are made commercially for specific uses. Because of their unique porous properties, zeolites are used in a variety of applications with a global market of several million tonnes per year. Major uses are in catalytic cracking, ion-exchange (water softening and purification) and the separation and removal of gases and solvents.

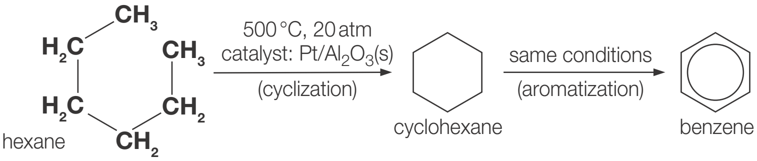
Reforming

In reforming processes, alkane molecules are rearranged to form branched or cyclic molecules.

Isomerization is a process in which straight-chain molecules, such as butane, are converted into their branched isomer(s), 2-methylpropane in this case. The process involves heating the alkanes in the presence of an aluminium chloride catalyst. The straight chains break apart and then re-join as branched chains. The resulting mixture consists of branched chains and straight chains in equilibrium. The equilibrium mixture can then be passed over another form of zeolite catalyst which acts like a molecular 'sieve'. Branched chains are separated from straight chains. The branched chains are used as fuel additives, and the straight chains can be returned to the reaction vessel (Figure 24.11).

Cyclization and **aromatization** are processes in which cycloalkanes and aromatic hydrocarbons are made from the straight-chain C_6–C_{10} alkanes in the naphtha fraction from distillation. The naphtha vapour is heated to 500°C and passed over a platinum catalyst on an aluminium oxide support. In a typical reaction under these conditions, hexane could be converted into cyclohexane initially, and then into benzene, for instance (Figure 24.12).

■ **Figure 24.12** The cyclization and aromatization of hexane through to benzene via cyclohexane



The reforming and cracking reactions of hydrocarbons and explanation of how these processes improve the octane number

Table 24.3 summarizes the products and primary advantages of the various cracking and reforming reactions.

■ **Table 24.3** Advantages and main products of different cracking and reforming processes

Reaction	Main products	Primary advantage(s) of this reaction
Thermal cracking	Shorter-chain alkanes Very small alkenes such as ethene	Supply of long-chain hydrocarbons outweighs demand – process produces more shorter-chain alkanes Ethene is important for plastics manufacture
Steam cracking	Mostly short-chain alkenes such as ethene Some hydrogen	High yield of short-chain alkenes for plastics manufacture Less coke produced, which clogs up the reaction vessel
Catalytic cracking	Branched alkanes, alkenes and arenes	Highly specific catalysts can be used which direct the reaction towards certain products, such as branched alkanes, which are used to increase the octane number of gasoline
Hydrocracking	Branched alkanes and arenes	Branched alkanes and arenes are used to increase the octane number of gasoline No alkenes produced
Isomerization	Branched alkanes	'Straight-run' gasoline contains relatively few branched alkanes and octane number is low. Isomerization increases the proportion of branched alkanes
Cyclization and aromatization	Cyclic hydrocarbons and arenes	Unbranched hydrocarbons with low octane number are converted to arenes with a much higher octane number

Worked example

Write balanced equations to illustrate an example of:

a thermal cracking

b steam cracking

c catalytic cracking.

a Thermal cracking

Thermal cracking starts with relatively long-chain hydrocarbons (up to C_{20}) and cracks them to shorter alkanes suitable for use in gasoline (such as C_8). Short-chain alkenes such as ethene and propene are also produced. A suitable equation might be:

$$\begin{array}{ccccccc} C_{18}H_{38} & \rightarrow & C_8H_{18} & + & 2C_3H_6 & + & 2C_2H_4 \\ \text{octadecane} & & \text{octane} & & \text{propene} & & \text{ethene} \end{array}$$

Ensure that the total number of carbons and hydrogens is conserved on each side of the equation. Note that at least one product is an alkane and the others are alkenes.

b Steam cracking

Steam cracking uses shorter-chain alkanes (up to C_8) and cracks them to form mostly short-chain alkenes such as ethene. Some hydrogen may be formed. A suitable equation might be:

$$\begin{array}{ccccc} C_8H_{18} & \rightarrow & 4C_2H_4 & + & H_2 \\ \text{octane} & & \text{ethene} & & \text{hydrogen} \end{array}$$

Again, the numbers of carbon and hydrogen are conserved. No alkanes are formed here but the presence of hydrogen allows the hydrogen number to be conserved.

c Catalytic cracking

Catalytic cracking in industry is highly specific, with catalysts chosen carefully to give the required mixture of branched alkanes. The feedstock for catalytic cracking is intermediate-length carbon chains from the naphtha fraction (around C_{15}).

$$\begin{array}{ccccccc} C_{16}H_{34} & \rightarrow & C_7H_{16} & + & C_7H_{14} & + & C_2H_4 \\ \text{Hexadecane} & & \text{heptane} & & \text{heptene} & & \text{ethene} \end{array}$$

Note that in this example, the products are likely to be branched isomers of heptane and heptene rather than straight chains

Write a structural equation showing how heptane could be cyclized and aromatized to form methyl benzene, and state the conditions necessary for this conversion.

Comparison with Figure 24.12 suggests that heptane, although a seven-carbon chain, might cyclize to a six-membered ring with a methyl branch. This cyclic hydrocarbon will then aromatize to a benzene ring with a methyl branch – methyl benzene.

3 Suggest equations for the following processes:

a Steam cracking of decane ($C_{10}H_{22}$) forming ethene and hydrogen.

b Thermal cracking of eicosane ($C_{20}H_{42}$) to octane plus short-chain alkenes.

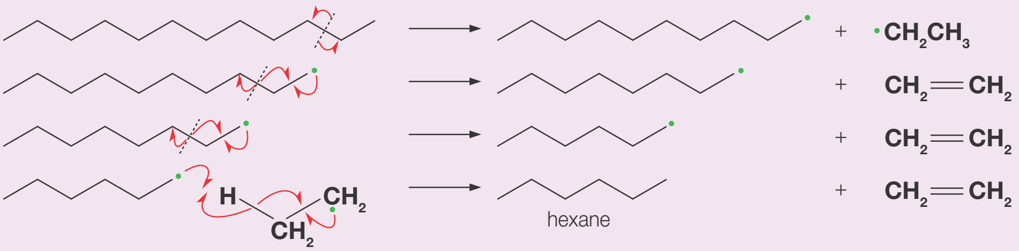
Additional Perspective

The mechanisms of cracking

Thermal cracking

Homolytic carbon–carbon bond fission produces alkyl radicals near the end of the chain. These alkyl radicals tend to split apart two carbon atoms along the chain, producing an ethene molecule and leaving a new alkyl radical with two fewer carbon atoms (Figure 24.13). This can then undergo further splitting, or a hydrogen atom can be transferred from another radical, if two radicals were to collide and react.

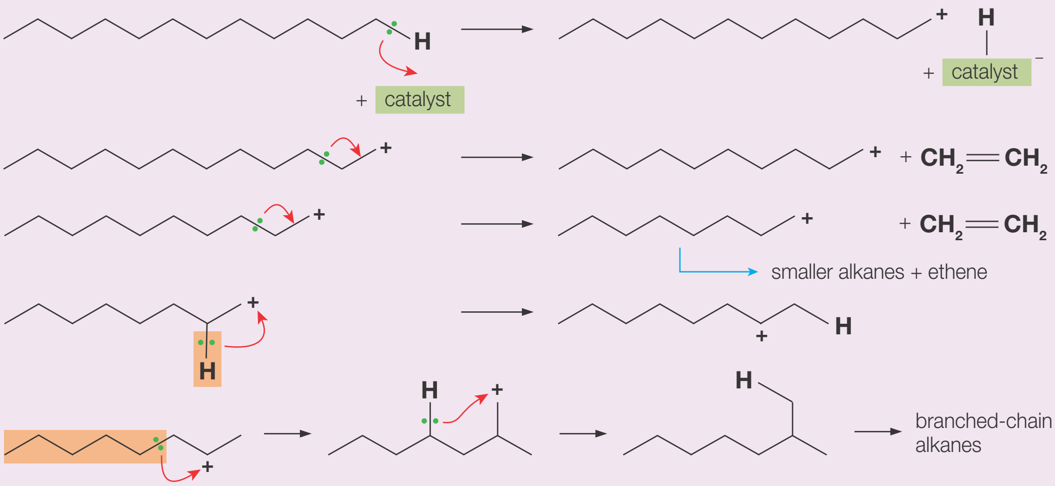
■ **Figure 24.13** The free radical mechanism of thermal cracking



Catalytic cracking

The transfer of a hydrogen atom to the surface of the catalyst produces a carbocation. In a similar way to the free radicals formed during thermal cracking, carbon–carbon bond fission occurs at the next-but-one bond to the electron deficient positive carbon, producing a molecule of ethene and an alkyl carbocation with two fewer carbon atoms (Figure 24.14). The shift of a hydrogen atom from a carbon atom in the centre of the chain to the carbon of the primary carbocation produces a more stable secondary carbocation. Further rearrangements produce branched-chain alkanes. Rearrangement reactions are another fundamental type of organic reaction, in addition to substitution, elimination and addition.

■ **Figure 24.14** The carbocation mechanism of catalytic cracking



Octane number ratings in different countries

When you visit a petrol station, you will often find a range of numbers quoted on the fuel pumps, telling you the octane ratings of different types of fuel. There are three main octane number standards worldwide:

- *RON or Research Octane Number:* An engine is run at 600 rpm and different compression levels, and the fuel is compared to 2,2,4-trimethylpentane and heptane. An octane number of 92 means that the fuel has the same tendency to pre-ignite as a mixture of 92 per cent 2,2,4-trimethylpentane and 8 per cent heptane.
- *MON or Motor Octane Number:* The fuel is tested at higher engine speeds, and with a variety of ignition (spark) timings, to give a better indication of how the fuel performs under driving conditions. Typically the MON is 8–10 points lower than the RON.
- *PON or Pump Octane Number (sometimes called AKI or Anti-Knock Index):* In the USA, Canada, Brazil, Mexico and some other countries, the number displayed on the fuel pump is the average of the RON and the MON (PON = (RON + MON)/2).

In most countries, oil companies market different grades of gasoline. These are often termed 'regular' and 'super' or 'premium' gasoline. The octane numbers of regular and super vary from country to country, depending on the quality of fuel available and the costs the market will bear.

In the UK, regular fuel is 95RON and super is 97RON. Some companies also market 99RON fuel at a higher price for high-performance cars.

In the USA, the PON/AKI standard is used. Regular is usually 85AKI (about 90RON). In some areas 94AKI (about 98RON) is available.

Countries whose populations drive mostly cheap, economical vehicles with low-compression engines typically offer predominantly low-octane fuels. For example, in India fuel ratings range from 89 to 91RON, and in Indonesia regular fuel is only 88RON. Conversely, in Hong Kong, where luxury vehicles are commonplace, consumers rejected 95RON fuel and now only 98RON is offered there.

In some countries, such as Brazil, gasoline fuel is mixed with ethanol (a biofuel). Originally this was for economic reasons – sugar, the raw material for ethanol manufacture, is cheap and plentiful in Brazil. In recent years other countries, such as the USA and Sweden, have adopted ethanol–gasoline mixtures for environmental and political reasons as they try to lower their dependence on imported oil. Biofuels are covered in detail in section 24.4.

■ Liquid fuels from coal

Coal is the most plentiful of the fossil fuels, but it has the disadvantage that the energy within it is most efficiently extracted in large-scale power plants.

Liquid fuels have the enormous advantage that they can be carried by vehicles and aircraft and used *in situ*. Gaseous fuels are easily piped to homes and used directly for heating and cooking.

The chemistry of converting coal into liquid fuels that can use the existing infrastructure designed for natural gas, gasoline, diesel, kerosene and so on has been widely investigated and implemented.

Coal gasification

The production of gaseous fuels from coal pre-dates the industrial production and domestic use of natural gas.

'Coal gas' is a mixture of carbon monoxide and hydrogen (both of which are combustible gases) formed by the reaction of coal with steam:

$$C(s) + H_2O(g) \rightarrow CO(g) + H_2(g)$$

Coal gas (sometimes called town gas) was widely used in the USA and the UK before the adoption of natural gas in the 1940s and 1950s.

Fossil fuel use in different countries

Hong Kong does not have natural gas resources, so it remains dependent on town gas manufactured from other fossil fuels. Some countries use coal gas (a mixture of carbon monoxide and hydrogen only, produced by combining coal with steam). Hong Kong uses a gas obtained from naphtha.

Until 1960, Hong Kong used coal gas. Then the change was made for a number of reasons. At that time, before the rise in the price of oil in 1973, oil was cheaper than coal. There was also the greater convenience of transporting and storing naphtha. Pollution from the combustion of naphtha is less than pollution from coal.

Naphtha consists of alkanes with 7–14 carbon atoms. Vaporized naphtha is passed with steam over a hot nickel catalyst. The products are hydrogen, carbon monoxide and other products.

naphtha + steam → carbon monoxide + hydrogen

Taking naphtha as $C_{10}H_{22}$, an example equation for the production of town gas is:

$$C_{10}H_{22}(l) + 10H_2O(g) \rightarrow 10CO(g) + 21H_2(g)$$

Substances added to town gas are:

- tetrahydrothiophene (THT) – to give town gas an unpleasant smell
- naphtha – to increase the energy density
- air – to increase the density.

It is essential that town gas has an unpleasant smell to allow leaks to be detected quickly. Leaks must be detected because an explosion might result, and also because carbon monoxide is a very poisonous gas.

The Hong Kong and China Gas Company, which supplies town gas to the territory, was founded in 1862 and is one of the oldest companies in Hong Kong.

This example illustrates how the balance of fossil fuel use can be highly specific to a country or territory, and will depend upon factors such as fuel availability, cost, and locally available technology and infrastructure.

Methods of coal liquefaction

Coal is made up mostly of carbon, but there is also hydrogen present. The proportion of hydrogen in the coal depends on the stage of fossilization reached by a particular sample of coal. Lignite contains more hydrogen than anthracite coal.

In any case, the proportion of hydrogen in liquid hydrocarbon fuels is much higher than in coal. Coal liquefaction relies on increasing the hydrogen : carbon ratio in the fuel. This is done in one of two ways:

1. Increase the amount of hydrogen in the coal.
2. Decrease the amount of carbon in the coal.

These processes lead to the formation of a mixture of hydrocarbons.

Direct hydrogenation of coal

Coal is crushed, mixed with a solvent (often heavy fuel oil) and heated in the presence of an iron-based catalyst. Hydrogen is added to the mixture. A mixture of hydrocarbons is formed, according to the following general equation:

$$nC + (n+1)H_2 \rightarrow C_nH_{2n+2}$$

The products are typically heavy hydrocarbons that require significant further refining (cracking, reforming) before they are of use as motor fuels.

Distillation of coal

Some coal types contain large amounts of hydrocarbons, so the liquefaction process involves separating the hydrocarbons from the solid carbon. As discussed above, bituminous coal is saturated with heavy hydrocarbon molecules. Such coals can be heated (a process called destructive distillation, or pyrolysis), which produces a mixture of tar (containing mostly cyclic or aromatic hydrocarbons), heavy oil and a solid residue of mostly carbon.

The tar and heavy oil can be further refined by cracking and reforming in order to make them useful as motor fuels.

Liquid fuels from coal gas

A third approach to coal liquefaction is to use the coal gas produced in the gasification process outlined previously to produce liquid hydrocarbons.

Recall that the products of coal gasification were carbon monoxide and hydrogen. These two products can be re-combined to make hydrocarbons:

$$(2n+1)H_2 + nCO \rightarrow C_nH_{2n+2} + nH_2O$$

The gas mixture is passed over an iron- or cobalt-based catalyst at 150–300 °C and pressures up to several dozen atmospheres.

Coal liquefaction and energy costs

Conversion of coal into liquid fuels by any of the processes outlined above requires a considerable input of energy, because high temperatures and pressures are necessary. Although the resulting liquid fuels are more convenient than coal for transport applications, they represent a very inefficient way of obtaining the maximum useful energy output from coal. Therefore, coal liquefaction is unlikely to present a long-term solution to our liquid fuel energy needs, although it does become economically viable when the cost of crude oil is particularly high or when supply is limited.

Worked example

Octane, C_8H_{18}, and its isomers are important components of gasoline. Deduce equations for the production of octane by:

a direct hydrogenation of coal

b liquefaction of coal gas, carbon monoxide, CO (g).

a The general equation for the production of an alkane containing *n* carbon atoms via hydrogenation is $nC + (n+1)H_2 \rightarrow C_nH_{2n+2}$ where C represents carbon within the coal.

For octane, $n = 8$, so the equation becomes:

$8C + 9H_2 \rightarrow C_8H_{18}$

b The general equation for the production of an alkane containing *n* carbon atoms via hydrogenation is:

$(2n+1)H_2 + nCO \rightarrow C_nH_{2n+2} + nH_2O$

For octane, $n = 8$, so the equation becomes:

$17H_2 + 8CO \rightarrow C_8H_{18} + 8H_2O$

Carbon footprints

Concern about the quantity of carbon dioxide entering the atmosphere due to human activities (see Section 24.5) has led to the adoption of a standard measure of the carbon dioxide emissions of particular activities. In simple terms it is a measure of the total carbon dioxide emissions resulting from a given activity, such as owning and using a car for a year, taking a transatlantic flight, or producing a certain output of a product such as computers, newspapers or foodstuffs.

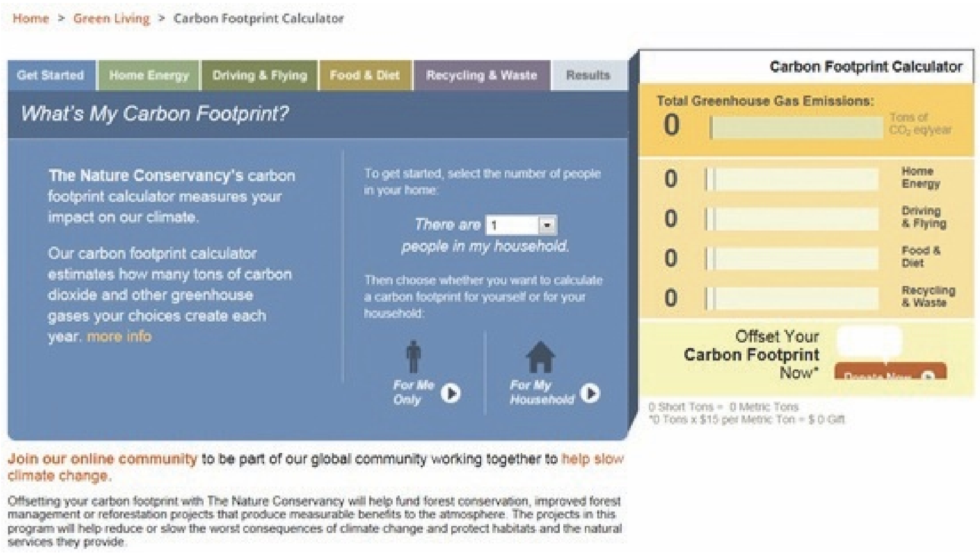
However, most of these activities involve the emission of numerous gases in addition to carbon dioxide, which might also contribute to global warming. For example, methane has a global warming potential approximately 25 times greater than carbon dioxide, for a given mass, so emissions of methane must be multiplied by this factor.

To simplify matters, instead of quoting the emissions of a whole series of greenhouse gases, the global warming potential of each, relevant to carbon dioxide, is factored in and the overall emissions are quoted as 'tonnes of carbon dioxide equivalent', which is abbreviated to CO_2-e. The carbon footprint expressed in this way offers a clear and simple measure of the climate change impact of the activity in question.

In order to calculate the carbon footprint, it is necessary to know the mass of each greenhouse gas emitted by the particular activity and the global warming potential of each. This can be a challenge, because the majority of emissions for many groups or activities actually take place far away from their actual location. For example, if we wished to calculate the carbon footprint of a typical household in the USA, we would need to know the amount and type of fuel consumed to generate their electricity, to fuel their cars, to produce all the goods they buy (food, clothing, cleaning materials and so on), to provide clean water, to remove their waste, and many more contributing factors. Most of these emissions are termed 'indirect' emissions

– the householders are not actually emitting the greenhouse gases themselves. Instead they are responsible for a small share of a much larger emission by a power station, a steelworks, a water treatment plant, and so on. The 'direct' emissions might only include using gas for cooking and heating and gasoline for driving, but even here there are indirect emissions arising from the extraction or refining of the fuels.

Despite the difficulty in estimating the thousands of carbon-emitting processes indirectly linked to everyday activities, carbon-footprint calculators have proliferated on the internet. By entering estimates of the size of your home, the type of car you own, the number of flights taken each year and various other factors, you can receive an estimate of the carbon footprint of your household. Corporations can undertake similar calculations to estimate their environmental impact. Often they will hire a consultancy firm to do the research and produce the final footprint figure. The danger here is that it is very easy to underestimate the carbon footprint by omitting indirect emissions. An example might be a magazine publisher that considers the heating and lighting of their offices and the cars driven by its staff, while ignoring the emissions of the printing firm that produces the final product; or the computer company that considers its own factories and workers, but omits the emissions due to mining the minerals needed to make the microchips and the oil needed to make the plastic casings. There is a marketing advantage in quoting a low carbon-footprint figure, but consumers need to examine the small print carefully to see what has been left out. Table 24.4 displays the typical carbon footprints of some different people, groups and activities. Many websites are available that enable the calculation of a personal carbon footprint (Figure 24.15).



What's My Carbon Footprint?

The Nature Conservancy's carbon footprint calculator measures your impact on our climate.

Our carbon footprint calculator estimates how many tons of carbon dioxide and other greenhouse gases your choices create each year. more info

To get started, select the number of people in your home:

There are 1 people in my household.

Then choose whether you want to calculate a carbon footprint for yourself or for your household:

0 Short Tons = 0 Metric Tons
*0 Tons x $15 per Metric Ton = $ 0 Gift

Join our online community to be part of our global community working together to help slow climate change.

Offsetting your carbon footprint with The Nature Conservancy will help fund forest conservation, improved forest management or reforestation projects that produce measurable benefits to the atmosphere. The projects in this program will help reduce or slow the worst consequences of climate change and protect habitats and the natural services they provide.

Figure 24.15 An online carbon-footprint calculator from the Nature Conservancy Council

Table 24.4 Some examples of carbon footprints

Person, group or activity	Carbon footprint (tonnes CO_2-e)
USA, average per person per year	19.8
UK, average per person per year	9.2
China, average per person per year	4.2
India, average per person per year	1.1
Google, all activities worldwide per year	1 500 000
Manufacture of one Ford Mondeo sedan	17
Manufacture and use of Apple iPhone over whole product lifecycle	0.075

Carbon calculations: CO_2 and motor vehicles

Most countries charge a road tax on motor vehicles. This is payable to the government and is intended to cover the cost of upkeep of the road network.

Many countries charge more road tax for larger and heavier vehicles, as these are presumed to cause more wear and tear to the roads. However, in 2005 the UK introduced new road tax charges based on carbon emissions. This was an incentive to purchase lower-emitting vehicles.

All new vehicles sold in the UK are given an official CO_2 emissions figure, quoted in grams per kilometre driven (g/km). This figure is obtained by measuring the CO_2 emissions under different driving conditions.

For some vehicles, usually hybrids (those with a supplementary electric motor alongside the gasoline engine) or those with very small engines, the CO_2 emission figure is below 100 g/km. Owners of these vehicles pay no road tax. As the CO_2 emission figure increases, the road tax increases, up to a maximum of £1100 (approximately US$1750) for the most powerful cars, which emit more than 255 g/km.

Worked example

Estimate the carbon footprint (tonnes CO_2-e) of driving a VW Polo BlueMotion with a CO_2 rating of 91 g/km, for 1 year, with a total distance covered of 20 000 km.

Total mass of CO_2 = 91 g/km × 20 000 km = 1 820 000 g

1 tonne CO_2 = 1000 kg = 1 000 000 g

So footprint tonnes $CO_2\text{-e} = \dfrac{1\,820\,000}{1\,000\,000} = 1.82$ tonnes CO_2-e

4 A car with an average fuel consumption of 7 litres/km covers 10 000 km in one year. Assume the fuel is octane, C_8H_{18}. Calculate the carbon footprint (tonnes CO_2-e). Necessary data: density of octane 0.703 g cm^{-3}, molar mass = 114.23 g mol^{-1}.

Oil spills and international cooperation

Oil spills at sea, whether from oil tanker leaks or drilling accidents, are major sources of ocean pollution. If such an accident occurs in international waters, who is responsible for cleaning it up? Such questions can be sources of enormous social political friction, so international committees exist to mediate such disputes.

In 1989 the International Maritime Organization prepared the International Convention on Oil Pollution Preparedness, Response and Co-operation (OPRC). Signatories to the OPRC are required to make preparations for dealing with pollution incidents and establish rules for ships and oil drilling operations under their jurisdiction. They are also required to provide assistance to other countries where necessary.

Conventions such as these are intended to ensure that there are rapid and efficient responses to pollution incidents without delays caused by political disagreements.

24.3 Nuclear fission and fusion *– the fusion of hydrogen nuclei in the Sun is the source of much of the energy needed for life on Earth. There are many technological challenges in replicating this process on Earth but it would offer a rich source of energy. Fission involves the splitting of a large unstable nucleus into smaller stable nuclei*

Nature of Science

The ethics of nuclear research

This section outlines the scientific concepts underpinning the production of useful energy by nuclear processes. The energy densities involved are enormous, meaning that nuclear energy is perhaps the only energy source that can hope to compete with fossil fuels in terms of meeting the world's energy needs.

The debate surrounding nuclear energy raises questions about the ethical nature of science. The main question is sometimes expressed as '*just because we* can *do something, does that mean we should?*'

In the 1940s, the goal of 'splitting the atom' was pursued for essentially militaristic reasons. After the atomic bombing of Japan, many (but by no means all) of the atomic scientists expressed regret that they had participated in the research while giving insufficient attention to the morality of what they were trying to achieve.

In the early part of the Cold War, despite the threat of annihilation, there was an optimism surrounding nuclear energy that was never fully realized. Atomic aircraft, ships and even cars were proposed, and atomic power was to become 'too cheap to meter', but the momentum of weapons development never translated into a peaceful, clean, atomic-powered utopia. Instead, a series of nuclear accidents, and the realities of the costs of nuclear infrastructure, led to disillusionment and nuclear power became less important to the worldwide energy mix than was predicted (currently nuclear power contributes about 12 per cent of worldwide electricity production).

Today, the threat of climate change means that even environmentalists disagree about the desirability of nuclear energy. Some feel that the utilitarian calculus (the cost/benefit analysis of the potential outcomes) leads to the conclusion that nuclear power is a lesser evil. That is, the potential dangers of nuclear energy are far outweighed by the potential benefits, in terms of lowering carbon emissions. Even nuclear risks are not rationally understood, with statistics on the harm caused by fossil fuels (for example, respiratory illness and coal-mining accidents, before even considering the climate-related dangers) exceeding those from the nuclear industry. However, others reject the utilitarian argument, suggesting that we have a duty to future generations not to build a nuclear 'mess' that will be their responsibility to clean up.

Many scientists believe that the capability to lower carbon emissions exists, and that there is an imperative to invest in and develop nuclear power. We *should* do it because we *can*. However, the 'should' involves a value judgement that is not straightforward. The connection between nuclear power and nuclear weapons cannot be ignored, and this, combined with high-profile accidents, means that many people do not trust science to solve this problem.

Nuclear power from fission

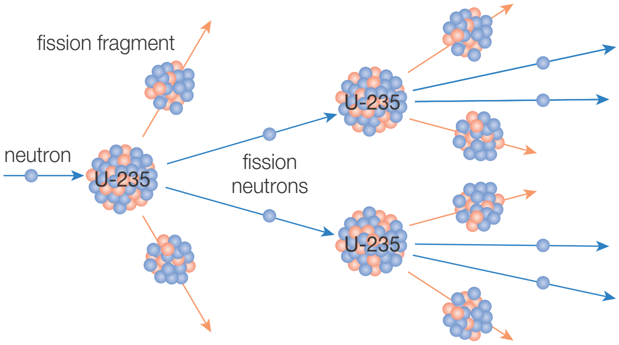
In a nuclear power station, radioactive nuclear fuel (uranium or sometimes plutonium) undergoes a process of nuclear fission. The process generates large amounts of thermal energy that can be used to boil water, producing steam. The pressurized steam can then be used to turn turbines for electricity generation. Naturally occurring uranium is mostly isotope ^{238}U, with a little ^{235}U. For fuel, it is enriched; that is, the proportion of ^{235}U is increased to allow the chain reaction to take place.

Nuclear power accounts for around 12 per cent of worldwide electricity generation, with the USA, France and Japan responsible for over half of this total. These figures looked likely to rise, as some governments seek to lower their dependence on fossil fuels, for environmental and economic reasons. However, the Fukushima Daiichi disaster of March 2011 led some governments, such as Germany's, to re-think their reliance on nuclear power. Some environmentalists see a shift from fossil fuels to more nuclear power as a useful way of cutting carbon emissions, while others see the production of larger quantities of dangerous radioactive waste, and the risk of accidents, as too great a price to pay.

Nuclear reactors

A nuclear reactor contains an array of ceramic uranium oxide rods separated by a so-called moderating material, or moderator, such as graphite. Nuclear fission typically involves the nucleus of a uranium-235 atom breaking apart and releasing slow-moving neutrons. These neutrons move through the uranium, colliding with other nuclei, which in turn causes them to split apart, releasing more neutrons (Figure 24.16). Therefore an ongoing chain reaction occurs. If this reaction is not controlled carefully, it proceeds rapidly, releasing massive amounts of energy. This is the basis of an atomic explosion. However, in a nuclear power station, this cannot happen, firstly because the uranium fuel contains only a small percentage of fissile uranium-235 (being mostly composed of uranium-238, which requires much faster neutrons and is therefore very unlikely to undergo fission), and secondly because the moderator absorbs and slows the neutrons moving through the material, preventing the chain reaction from running out of control.

Figure 24.16
A nuclear chain reaction



Deducing fission equations

Nuclear reactions must be balanced, just like chemical equations. Balancing is done by comparing the mass numbers and the atomic numbers of all the components in the equation. These numbers must be conserved.

$${}^{1}_{0}\text{n} + {}^{235}_{92}\text{U} \rightarrow {}^{236}_{92}\text{U}$$

In this first step, the U-235 nucleus (92 protons, 143 neutrons) absorbs an incoming neutron (1 neutron, no proton), becoming U-236 (92 protons, 144 neutrons)

$${}^{236}_{92}\text{U} \rightarrow {}^{92}_{36}\text{Kr} + {}^{141}_{56}\text{Ba} + 3{}^{1}_{0}\text{n}$$

Then the U-236 atom decays into two smaller nuclei, krypton-92 (36 protons, 56 neutrons) and barium-141 (56 protons, 85 neutrons). So far this adds up to 141 neutrons, so the remaining three neutrons from the U-236 atom are released. It is these neutrons that can go on to split more U-235 nuclei.

5 The first nuclear reaction ever observed was Ernest Rutherford’s transmutation of nitrogen-14 into oxygen using alpha particles in 1917. Protons were produced as a by-product. Write an equation for this reaction and deduce which isotope of oxygen was formed.

Explaining energy changes during fission using binding energies

During the nuclear chain reaction, enormous amounts of energy are released. This energy is released as electromagnetic radiation (mostly gamma rays) and kinetic energy of the fast-moving fission products (in the example above, the krypton and barium nuclei). This heats up the nuclear material.

This energy has to come from somewhere. It is accounted for by comparing the nuclear binding energy of the original nucleus and the fission products.

Nuclear binding energy holds the subatomic particles of the nucleus together. In order to break up a nucleus, energy must be provided. Upon formation of a nucleus, energy will be released. (Compare this with chemical bonds, in which bond breaking is also endothermic, while formation is exothermic.)

In the example cited, the total binding energy of the two lighter elements, krypton and barium, is greater than the binding energy of the uranium nucleus. The amount of energy released when krypton and barium are formed is greater than that needed to split the uranium nucleus.

Why do large nuclei split but small nuclei fuse?

The binding energy depends on the relationships between two forces. The nucleus is held together by the **strong nuclear force**. This force acts over very small distances, and is sufficient to overcome the **electrostatic repulsion** which acts between protons, which repel each other because they all have a positive charge. At the distances involved within the nucleus, the strong nuclear force is about 1000 times as strong as the electrostatic repulsion.

The graph in Figure 24.17 shows that for small nuclei, the binding energy increases with nucleon number. (This graph is also available in the *IB Chemistry data booklet*, Section 36.) This occurs because the strong nuclear force within the nucleus increases as the number of nucleons increases. The electrostatic repulsion increases too, but is less important when there are relatively few protons to repel each other. Overall, small nuclei experience an increase in binding energy as nucleon number increases.

The graph is then approximately horizontal for a while, as the balance between the strong nuclear force and the electrostatic repulsion gives similar binding energies for nuclei of intermediate size.

At very high nucleon numbers, the binding energy starts to fall with increasing numbers of nucleons. The electrostatic repulsion factor starts to outweigh the increase in the strong nuclear force. For each successive nucleon, more repulsion than attraction is experienced, so the binding energy falls.

This explains why there are a limited number of chemical elements. At higher nucleon numbers the nucleus becomes increasingly unstable. For elements higher than $Z = 83$ (83 protons; bismuth), the nuclei are prone to decay, so they are radioactive. Uranium atoms ($Z = 92$) are the largest to occur naturally.

Recall that a nuclear reaction will take place, and release energy, if the binding energy of the products is larger than the binding energy of the reactant particles.

For large nuclei, the binding energy is lower than that of a pair of smaller nuclei, so these large nuclei are prone to fission, and will release energy when they do so.

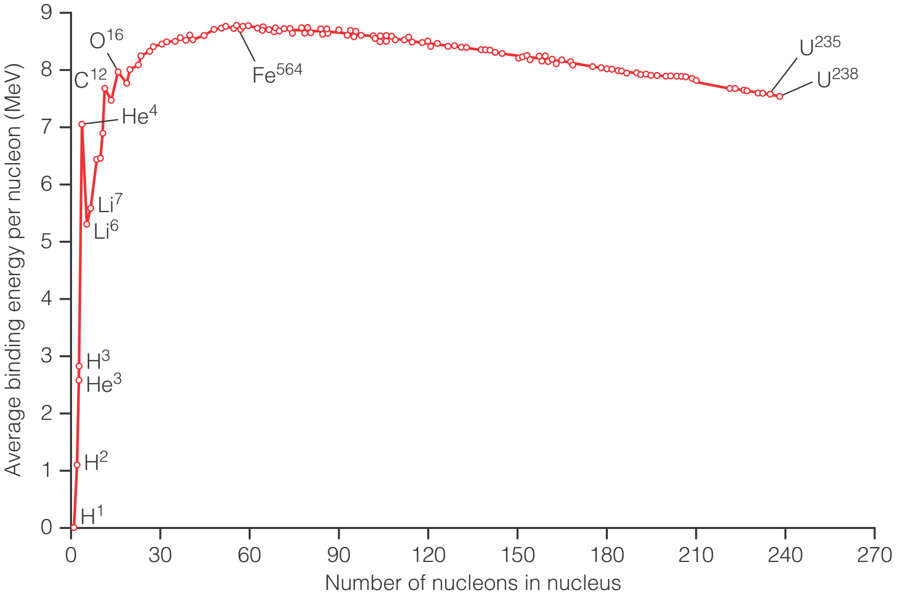
For small nuclei, the binding energy of a larger nucleus (the fusion product) is greater, so when small nuclei fuse together, energy output (binding energy of the product nucleus) is greater than the energy input (binding energy of the two smaller nuclei).

This explains the perhaps counter-intuitive observation that both nuclear fission and nuclear fusion are exothermic processes. Nuclear fission involves very large nuclei splitting into medium-sized products. Nuclear fusion involves very small nuclei combining into products that are still rather small.

The mega electronvolt (MeV) shown on the graph is a unit of energy often used when describing nuclear processes. For a full explanation see Section 24.7.

It should be noted that while most energy diagrams in chemistry have the lowest (most stable) energy state at the bottom of the diagram, this one is an exception. The most stable state is at the top.

Figure 24.17 The binding energy curve



Additional Perspective

The liquid drop model of the nucleus

We now understand that atomic nuclei are held together by the strong nuclear force. Atomic physicists used analogies to attempt to explain the behaviour of nucleons under this force.

George Gamow suggested that the nucleus be thought of as a droplet of incompressible fluid. In the model, each nucleon behaves like a molecule in the liquid, and the strong nuclear force is represented by the intermolecular forces within the liquid drop.

Niels Bohr and John Wheeler developed the model and found that it closely modelled many features of the nucleus (Table 24.5).

Table 24.5 Similarities between liquid drops and atomic nuclei

Liquid drop	Atomic nucleus
Liquid drops are spherical because there are no intermolecular forces acting on the outside of the outermost molecules in the drop. The spherical shape maximizes the intermolecular forces between the molecules and is the lowest energy configuration	Nuclei are spherical because the spherical shape is the lowest energy configuration, maximizing the number of strong nuclear force interactions
The liquid drop has the same density throughout, which is independent of the size of the drop	The density of the nucleus is independent of the size of the nucleus
Liquid molecules attract one another by dispersion forces but cannot get too close, otherwise repulsive forces (e.g. between the outer electrons of each molecule) start to become important	Nucleons attract one another by the strong nuclear force. However, there are also electrostatic repulsive forces between similarly charged particles (i.e. protons within the nucleus)
Liquid drops reach a maximum size before breaking apart into smaller drops	Nuclei reach a maximum size before breaking apart into smaller nuclei
When they split apart, liquid drops form a dumb-bell shape, which then splits apart at the 'neck' of the dumb-bell, forming two similarly sized fragments (Figure 24.18)	When they split apart, nuclei tend to form similarly sized fragments

Carl Friedrich von Weizsäcker quantified many of the energy terms outlined above and used his equations to make predictions of nuclear binding energy.

The liquid drop model is not a perfect one. For example, the number of molecules in a liquid drop is far larger than the number of nucleons in a nucleus, and the amount of movement available to a molecule within a drop is much larger than the size of the molecule itself, whereas the amount of movement available to a nucleon is similar to the size of the nucleon itself.

Nonetheless the liquid drop model proved useful when theorizing the contribution to the binding energy from different interactions within the nucleus.

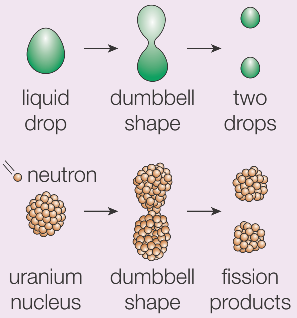


Figure 24.18 Liquid drops and atomic nuclei can be modelled in similar ways

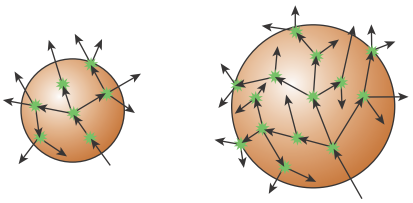


Figure 24.19 The critical mass of fuel is the minimum size necessary to sustain a chain reaction: **a** in a small sphere neutrons can easily escape so the chain reaction does not become self-sustaining; **b** for a sphere with a mass greater than the critical mass the chain reaction is self-sustaining

Sustaining the fission chain reaction

Naturally occurring uranium contains only 0.7 per cent uranium-235, which in most reactors is insufficient to sustain a chain reaction. Uranium for reactors is processed to increase the percentage of uranium-235 to around 4 per cent, a process called 'enrichment'.

In order for the chain reaction to proceed, there must be a sufficient density of neutrons moving through the reactor. If too small a quantity of nuclear fuel is used, too many neutrons 'escape' from the edges of the material. The minimum quantity of nuclear fuel needed to sustain a chain reaction is called the 'critical mass'. Figure 24.19 shows how the size of the mass of fuel is important in sustaining the chain reaction.

For a pure sphere of uranium-235 the critical mass is 52 kg, which represents a sphere of about 17 cm diameter. For much less pure uranium-235, as used in reactors, the critical mass is hundreds of kilograms.

Reactor design

Most nuclear reactors in current service were built from designs originating in the 1960s and 1970s. Most of these are light-water reactors, which use ordinary water as a moderator and coolant. Water pumped through pipes inserted into the nuclear fuel vessel acts as a moderator (to slow down the neutrons, making neutron capture by U-235 nuclei more likely) and a coolant. When it leaves the reactor the water is at a temperature of more than 300°C, but it remains liquid because the pressure in the cooling system is about 150 atmospheres. A heat exchanger is used to transfer this heat to another, unconnected water supply. This new water is boiled to form high-pressure steam that turns turbines.

Concern about global warming has put nuclear power back on the energy agenda in the UK, the USA and many other countries. Japanese companies are working on modular sealed unit reactors that can be quickly manufactured in factories and shipped to the required location, simplifying the setting up of reactors. In South Africa, pebble-bed reactors have been developed which use spheres of uranium oxide combined with graphite. They can be piled together and a chain reaction starts. The energy is transferred from the reactor using an inert gas such as helium. These are solutions to the massive investment required for earlier generations of reactors, most of which had to be subsidized by governments. These simple modular designs are easier to scale up or down according to need. They also address the problem of decommissioning reactors when they reach the end of their useful lives. Traditional reactors accrue enormous costs for dismantling and decontamination.

Breeder reactors

Uranium is not a renewable resource. Although quite plentiful, only a small fraction of naturally occurring uranium is fissile uranium-235, and reserves of uranium ore will eventually run out. When the nuclear fuel is used up, the resultant waste contains more than 90 per cent unreacted uranium-238. The waste is very dangerous, owing to the presence of radioactive actinoid (lower row f-block) elements such as plutonium-239 and neptunium-236. A breeder reactor aims to solve two problems:

1 Traditional light-water reactors use only a small percentage of the uranium in the nuclear fuel – the fissile uranium-235 portion. Can the left-over uranium-238 be utilized?
2 Highly radioactive waste fuel rods must be disposed of safely. Can they instead be employed to useful purpose?

A breeder reactor is designed differently from a commercial light-water reactor, to take advantage of another reaction that occurs in a nuclear reactor.

$$^{1}_{0}\mathrm{n} + ^{238}_{92}\mathrm{U} \rightarrow ^{239}_{92}\mathrm{U} \rightarrow ^{239}_{93}\mathrm{Np} + ^{0}_{-1}\mathrm{e} \rightarrow ^{239}_{94}\mathrm{Pu} + ^{0}_{-1}\mathrm{e}$$

In this scheme, a uranium-238 nucleus captures a neutron, and then undergoes two stages of nuclear decay, in which beta particles (electrons – see Dangers of nuclear power, pp. 34–5) are released, leading to conversion into neptunium-239 and then plutonium-239. Plutonium-239 is fissile, unlike uranium-238, so can contribute to the energy output of the reactor.

Although this reaction does occur in ordinary reactors, the plutonium fissions, meaning its abundance in the reactor decreases. However, in a breeder reactor, the plutonium-generating reaction occurs more quickly than the plutonium is used up by fission. This is achieved by using a coolant that is less effective at moderating (slowing down) the neutrons. Fission of uranium-235 is more efficient at low neutron speeds (as in water-cooled and moderated 'ordinary' reactors). Breeder reactors use alternative coolants such as helium gas or liquid sodium.

The breeder reactor, then, is so-called because it 'breeds' nuclear fuel: fissile plutonium-239. More fissile material comes out than went in.

Once the reaction is under way, spent uranium fuel rods, which contain mostly non-fissile uranium-238, along with the highly radioactive plutonium and neptunium products from traditional fission, can be introduced. In the breeder reactor these substances are recycled and fissioned. Hence radioactive waste can be put to use generating energy.

A slightly different reactor design leads to the consumption of the dangerous actinoid waste in fuel rods, so that the waste only contains fission products (such as barium and krypton). These have much shorter half-lives (less than 100 years) meaning the waste is much more easily disposed of. Such reactors are called **burner reactors**.

Problems with breeder reactors

The relatively low cost of uranium ore, combined with the greater cost of breeder reactors, means that they have not been widely adopted despite their advantages. However, as waste stockpiles increase and uranium reserves run low, they have been subject to renewed interest.

Safety considerations are even more important with breeder reactors because the coolants might be hazardous. For example, if hot liquid sodium coolant is exposed to air, it is likely to ignite, leading to a sodium fire.

The other problem with breeder reactors is that they produce plutonium, which is used in building nuclear weapons. A country that proposes to build breeder reactors is likely to come under scrutiny from the International Atomic Energy Agency, which seeks to limit the proliferation (spreading) of nuclear weapons.

■ Nuclear power from fusion

Nuclear fusion involves smaller nuclei joining together to make larger ones. It is a subject of intense interest and study because in theory fusion reactions can also produce larger amounts of energy per kilogram of fuel than fission reactions.

Solar fusion

The most well-known fusion reaction is that which takes place in the Sun. The Sun is primarily made of hydrogen. Under the intense gravitational pressures and enormous temperatures inside the star, hydrogen nuclei fuse together to make helium nuclei.

In order for this to occur the hydrogen nuclei must be moving fast enough to overcome the electrostatic repulsion between them. To fuse together, two nuclei must approach within 10^{-15} metres of each other, so that the strong nuclear force is able to bind them together.

The most abundant isotope of helium is ${}^{4}_{2}\mathrm{He}$ (helium-4) and the most common isotope of hydrogen is ${}^{1}_{1}\mathrm{H}$. Conservation of mass numbers and atomic numbers makes it clear that the formation of helium does not occur by simple fusion of two ${}^{1}_{1}\mathrm{H}$ nuclei, as this would result in a new nucleus with two protons and no neutrons.

Instead the formation of helium-4 occurs by a multi-step mechanism involving a less common isotope of hydrogen called **deuterium**. Deuterium differs from common hydrogen in that it has one neutron in addition to the usual single proton. The symbol for deuterium is ${}^{2}_{1}\mathrm{H}$ (sometimes written as ${}^{2}_{1}\mathrm{D}$).

Construction of nuclear equations for fusion processes

The following scheme is proposed for the formation of helium-4 in the Sun. The mechanism also involves formation of a positron (symbol ${}^{0}_{1}\mathrm{e}$): a particle with the same mass as an electron, but a positive charge instead of a negative charge. (Physicists refer to the positron as a particle of antimatter, or an antiparticle.)

$${}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{2}_{1}\mathrm{H} + {}^{0}_{1}\mathrm{e} + \nu$$

These products are a deuterium nucleus, a positron and a neutrino (symbol ν (nu), an electrically neutral, almost massless particle). Note that the mass numbers and atomic numbers are conserved in this equation (follow the subscript numbers and superscripts numbers on each side of the arrow).

The deuterium combines with another hydrogen nucleus, with the production of a helium-3 nucleus and gamma radiation:

$${}^{2}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{He} + \gamma$$

The two reactions above take place twice, forming two ${}^{3}_{2}\mathrm{He}$ nuclei. In total six ${}^{1}_{1}\mathrm{H}$ nuclei are involved. Then, the two ${}^{3}_{2}\mathrm{He}$ nuclei fuse giving a ${}^{4}_{2}\mathrm{He}$ nucleus and reforming two ${}^{1}_{1}\mathrm{H}$ nuclei.

$${}^{3}_{2}\mathrm{He} + {}^{3}_{2}\mathrm{He} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H}$$

The overall nuclear equation is:

$$4\,{}^{1}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + 2\,{}^{0}_{1}\mathrm{e} + 2\nu$$

Each of the fusion processes above releases lots of energy, because the binding energy of the fusion product is greater than that of the two colliding nuclei.

The process by which small nuclei 'build up' to form larger nuclei is called **nucleosynthesis** and is the source of all the chemical elements heavier than hydrogen in the universe.

6 An important nucleosynthetic process in stars is the formation of nitrogen from carbon. In this series of reactions, carbon-12 fuses with hydrogen (${}^{1}_{1}\mathrm{H}$) forming nitrogen-13. This then releases a positron to become a new isotope. This new isotope fuses with a further hydrogen nucleus, forming nitrogen-14. Write equations for this series of reactions and deduce the identity of the unknown isotope.

How do we know the composition of stars?

The elements present in the Sun and their abundance were determined by examining the absorption spectra of elements in the laboratory and comparing these spectra with solar radiation.

Absorption spectra

In Chapter 2 we discussed the emission spectrum of hydrogen and observed that electron transitions from higher energy levels to lower energy levels in hydrogen atoms led to the emission of photons with specific wavelengths. These appear as a line -spectrum – a series of bright lines of light against an otherwise dark background.

An absorption spectrum is the opposite of this. If white light is passed through a gaseous sample of atoms at low temperature and pressure, and the emerging light is analysed, the light beam will be found to be missing certain wavelengths of light. This is because the gaseous atoms have absorbed certain wavelengths. Electrons within the sample will have been promoted (or 'excited') to higher energy levels. Again, the wavelengths absorbed will be specific and will correspond to the difference in energy between the energy levels in the sample atoms. Each chemical element has its electron energy levels at slightly different energies, meaning that the wavelengths of light absorbed when electrons are excited will also be slightly different.

The relationship between emission and absorption spectra is shown in Figure 2.40 in Chapter 2.

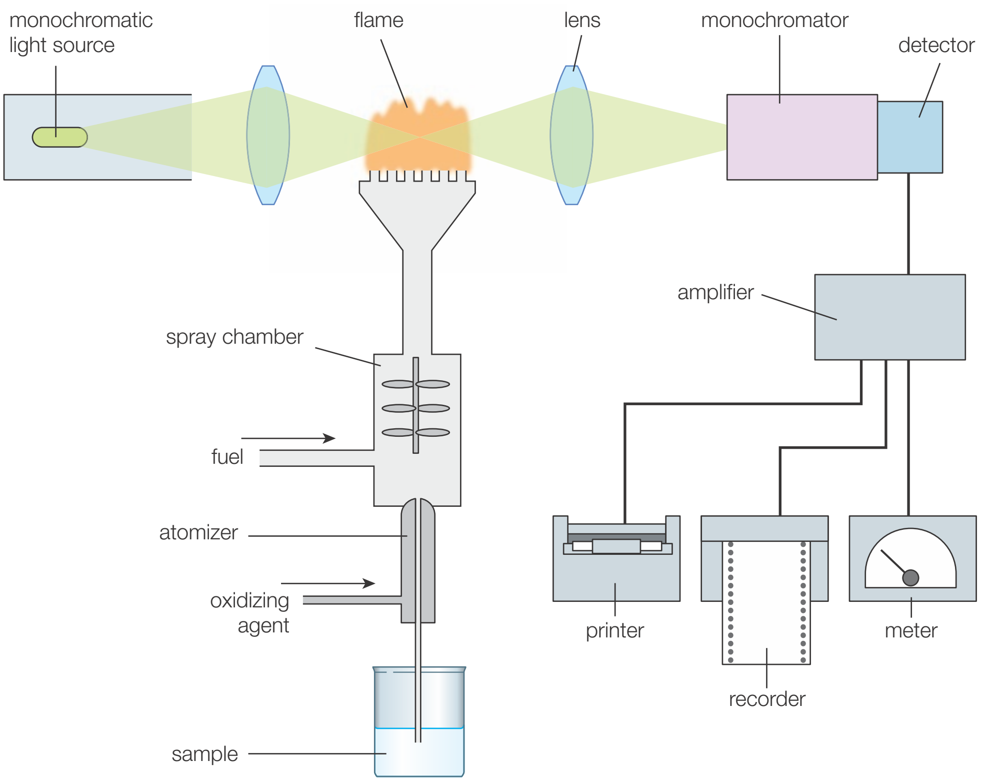
Examining the Sun using absorption spectra

In very simple terms, an atomic absorption spectrometer basically consists of a source of white light (light containing all the visible wavelengths), a means of vapourizing the sample and a detector. Light from the source passes through the sample and is then examined to determine which wavelengths are present and which have been absorbed. Figure 24.20 shows a schematic of a laboratory atomic absorption spectrometer. In the laboratory set-up, a flame is used to vapourize the sample, and a monochromator is used to filter out unwanted frequencies of radiation (those not likely to be absorbed by the sample).

The Sun can be thought of as a gigantic absorption spectrum experiment. The core of the Sun acts as the white light source, generating the full range of visible light wavelengths. As this light moves from the core outwards, it passes through the layers of gas in the outer part of the Sun. As it does so, atoms in this gaseous region will absorb some of the visible light wavelengths. Sunlight that reaches Earth therefore has narrow bands of missing wavelengths, corresponding to the absorption frequencies of the elements in the Sun.

The dark lines in the solar spectrum were discovered by the German optician Joseph von Fraunhofer (1787–1826) and named Fraunhofer lines. By comparing Fraunhofer lines with absorption spectra obtained in the laboratory, scientists were able to identify the presence of hydrogen, oxygen, sodium, magnesium, calcium and iron in the Sun. However, there was absorption at 587.5 nm, which was not present in any elements examined on Earth. Norman Lockyer, a British astronomer, proposed that the line was due to a new element, which he named helium, after Helios, the Greek Sun god. Lockyer is jointly credited with discovering the helium spectral line, along with Jules Janssen, a Frenchman, in 1868. Helium was not identified on Earth until 1895, when it was isolated from the air in small quantities.

Figure 24.20 Schematic diagram of a flame atomic absorption spectrometer



The solar spectrum, and the use of spectral lines to identify atoms present in the Sun, is the basis of our understanding of the fusion reactions occurring in the Sun and our knowledge of its composition (Figure 24.21).

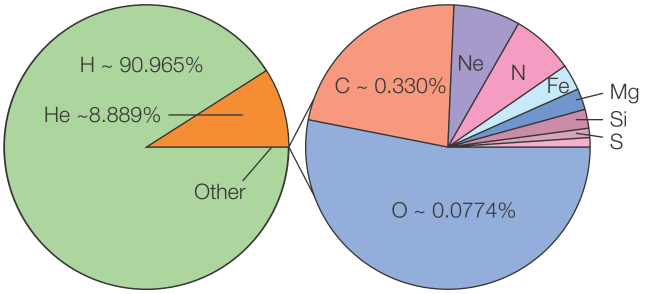


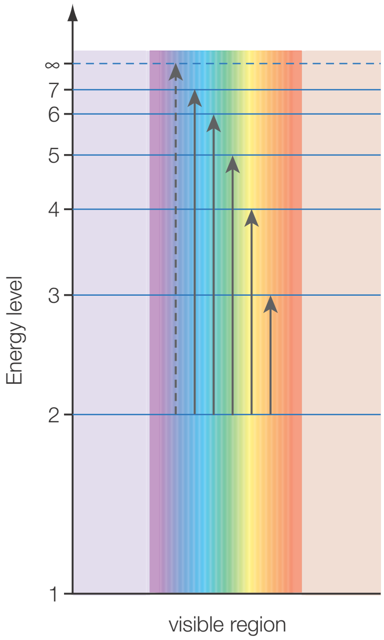
Figure 24.21 The composition of the Sun (data from NASA's planetary factsheet)

Explaining energy levels and absorption lines

As the Sun is mostly made up of hydrogen, the solar spectrum has strong absorptions in the lines corresponding to the hydrogen spectrum. Transitions from the first energy level to higher levels have energies greater than photons in the visible spectrum. This series of lines appears in the ultraviolet.

Transitions from the second energy level ($n = 2$) to the third, fourth, etc. correspond to wavelengths in the visible spectrum. This is called the Balmer series of the absorption spectrum (Figure 24.22). When an electron is excited from $n = 2$ to $n = 3$, light of wavelength 656 nm is absorbed from the white light spectrum, leading to a dark line on the spectrum. The solar absorption spectrum includes other dark lines at 486 nm, 434 nm and 410 nm, which correspond to the $n = 2 \rightarrow 4$, $n = 2 \rightarrow 5$ and $n = 2 \rightarrow 6$ electron transitions, respectively.

Figure 24.22 The Balmer series



None of these lines corresponds to the strong absorbance (a dark line) observed by Lockyer and Janssen in 1868. Lockyer initially thought that the line must be attributed to hydrogen under unique temperature or pressure conditions, but he eventually concluded that an entirely new element, with differently spaced energy levels, must be responsible. This was helium.

Man-made fusion

The hydrogen bomb

Soon after the invention of the atomic bomb in 1945, the USA began developing an even more powerful weapon: the hydrogen bomb. This weapon uses the fusion of hydrogen to produce massive amounts of energy very quickly, leading to an enormous explosion.

An atomic bomb uses uranium or plutonium and involves a fission reaction. However, the size of an atomic bomb is limited, because when the mass of plutonium or uranium exceeds a certain size, it starts to blow apart before the fission reaction is complete. The smaller pieces of fuel do not then have sufficient mass for the chain reaction. Eventually, a size is reached where adding more uranium or plutonium to the bomb does not lead to an increase in the power of the explosion. By 1952 the USA had developed a fission weapon with an explosive yield equal to 500000 tonnes (500kT) of TNT (compared with about 15kT for the Hiroshima weapon). This bomb used 60 kilograms of uranium-235.

It was theorized that a weapon using a fusion reaction could be much more powerful. In fact there is no theoretical upper limit to the size of the weapon. If more nuclear fuel is added, the fusion reaction will continue, just as the Sun consumes its hydrogen fuel continuously.

In order to generate the enormous temperatures required, a fusion weapon uses a fission bomb to ignite the fusion reaction. When the fission bomb explodes, the energy is channelled into a vessel containing deuterium fuel. A combination of the very high temperature and the compressive force of the intense radiation generated by the fission bomb is enough to start the deuterium nuclei fusing together. A possible fusion reaction is:

$$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{2}He + ^{1}_{0}n$$

This process is termed **thermonuclear fusion**, as thermal energy is used to ignite the fusion reaction. Hydrogen bombs are sometimes called thermonuclear weapons.

The most powerful hydrogen bomb tested by the United States ('Castle Bravo', 1954) had an explosive yield of 15000000 tonnes (15 megatonnes (Mt), or 1000 Hiroshimas), while the Soviet Union's '*Tsar Bomba*' ('King of Bombs', 1961) reached 50Mt.

Fusion reactors

Fusion weapons rely on igniting a fusion chain reaction that generates heat quickly, leading to more fusion. A commercial fusion reactor is more difficult to operate, because the fusion reaction must be carefully controlled, producing sufficient heat to maintain fusion, without 'running away' and exploding.

Most commercial attempts at producing a sustainable fusion reaction have used deuterium fuel, or a combination of deuterium and a third hydrogen isotope, called tritium, which is hydrogen with two neutrons: ${}^{3}_{1}H$.

Deuterium–deuterium fusion is written:

$${}^{2}_{1}H + {}^{2}_{1}H \rightarrow {}^{3}_{2}He + {}^{1}_{0}n$$

Deuterium–tritium fusion is written:

$${}^{2}_{1}H + {}^{3}_{1}H \rightarrow {}^{4}_{2}He + {}^{1}_{0}n$$

Unlike fission reactors, there are no dangerous actinoids produced, so there are no long-term fuel disposal problems. However the reactor vessel itself becomes highly radioactive due to its being bombarded by neutrons.

If they are to fuse, overcoming their mutual electrostatic repulsion, the nuclei must be heated to very high temperatures while containing them in a vessel without contaminating them. At such high temperatures, a vessel made of steel, for example, would release vaporized iron atoms, contaminating the nuclear fuel. Instead the nuclear fuel is suspended and isolated in a magnetic field – a process called **magnetic confinement**.

The confinement vessel is in the shape of a torus (simply described as a 'doughnut shape'). The fuel is accelerated around the 'doughnut' to very high speeds. The particles reach such high temperatures that their electrons are stripped away, resulting in the formation of a 'plasma' of highly charged nuclei. The nuclei have sufficient kinetic energy to collide and fuse, overcoming the electrostatic repulsion.

Although fusion reactions have been achieved in this way, the power consumption of the magnetic confinement and heating the fuel has so far been greater than the energy output from fusion. Further research and development is needed until the energies 'break even' and then an overall production of energy is achieved.

A further difficulty of building a fusion reactor is developing the materials required. A sustained fusion reaction would produce enormous numbers of neutrons. These neutrons would constantly bombard the walls of the vessel, so materials must be used which can maintain their properties under this bombardment.

If these problems can be overcome, nuclear fusion is an attractive prospect, as it produces much less dangerous waste than nuclear fission; and deuterium, although not the most common isotope of hydrogen, is plentiful enough in natural sources of hydrogen (for example, sea water) to offer cheap and abundant fuel almost indefinitely.

■ Dangers from nuclear power

The dangers of nuclear power arise from two main factors. Firstly, the by-products of nuclear fission are highly radioactive, meaning that they release ionizing radiation. Secondly, many of these substances have long half-lives, meaning that they remain dangerously radioactive for very long periods of time.

Types of radiation

Nuclei decay in order to stabilize their structure. In general, radioactive decay decreases the neutron/proton ratio of the nucleus in question. There are three types of radiation that are emitted from decaying atomic nuclei: alpha, beta and gamma radiation. The types of radiation differ in their ability to ionize atoms in materials with which they come into contact. If ionizing radiation comes into contact with biological systems it can damage genetic material such as DNA. This can lead to mutations, which in turn can lead to cancers or deformities in unborn children.

Alpha radiation

Alpha radiation is the most harmful type, as alpha particles are relatively heavy, meaning they are more likely to ionize nearby atoms. Alpha particles are composed of two protons and two neutrons – they are helium nuclei. The symbol for an alpha particle is ${}^{4}_{2}\mathrm{He}$ or α.

In the equation below, a uranium-238 nucleus decays by ejecting an alpha particle, forming a nucleus of thorium-234. Note that in doing so, the neutron/proton ratio has decreased from 146 : 92 (or 1.59 : 1) to 142 : 90 (1.58 : 1). In the equation the mass numbers and atomic numbers are conserved.

$${}^{238}_{92}\mathrm{U} \rightarrow {}^{234}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He}$$

Alpha radiation is highly ionizing, meaning that it is very likely to damage nearby atoms. However, since it interacts so strongly with material it enters, it is very quickly absorbed by atoms in the material. This makes poisoning by alpha radiation relatively unlikely, since clothing or even human skin is able to prevent alpha radiation penetrating the body. However, if alpha-producing material is ingested or inhaled, then serious radiation poisoning results, as the radiation is able to damage soft tissues inside the body.

Beta radiation

Beta radiation consists of high-energy electrons ejected from the nucleus. However, the nucleus does not normally contain electrons! In fact, a neutron transmutes into a proton and an electron, and the electron is ejected. In this way the neutron number falls by one, and the atomic number (proton number) increases by one. The neutron/proton ratio therefore decreases. The symbol for a beta particle is ${}^{0}_{-1}\mathrm{e}$ or β.

In the following nuclear equation, a caesium-137 nucleus decays into a barium-137 nucleus.

$${}^{137}_{55}\mathrm{Cs} \rightarrow {}^{137}_{56}\mathrm{Ba} + {}^{0}_{-1}\mathrm{e}$$

Beta radiation can penetrate clothing and body tissues. Exposure to beta radiation outside the body is potentially harmful, but inhalation and ingestion is more so.

Gamma radiation

Gamma radiation is a form of electromagnetic radiation. Gamma radiation (sometimes referred to as 'gamma rays') is usually released in addition to either alpha or beta decay, since these decay types often leave the nucleus in an energetically 'excited' state. Energy is emitted in the form of gamma radiation, allowing the nucleus to return to the ground state. The symbol for gamma radiation is γ. In the equation below, iodine-131 decays to xenon-131 with emission of beta and gamma radiation.

$${}^{131}_{53}\mathrm{I} \rightarrow {}^{131}_{54}\mathrm{Xe} + {}^{0}_{-1}\mathrm{e} + \gamma$$

Gamma radiation is weakly ionizing but highly penetrating. It is still able to cause damage to DNA but is more likely to pass straight through tissues without interacting with them.

The likelihood of harm from gamma radiation increases with prolonged exposure, and gamma sources are difficult to shield, since they are able to penetrate metal containers and brick walls. Most radioactive waste is placed in deep storage sites, to prevent radiation harming humans, plants or animals (see below). This is mainly because of the risk of prolonged exposure to gamma rays.

The properties of alpha and beta particles and gamma rays are summarized in Table 24.6.

Table 24.6 Summary of properties of alpha, beta and gamma radiation

Radiation	Relative charge	Relative mass	Nature	Penetration	Deflection by electric field
Alpha particles	+2	4	2 protons and 2 neutrons (He^{2+} ion)	Stopped by a few sheets of paper	Low
Beta particles	–1	$\frac{1}{1837}$	Electron	Stopped by a few millimetres of plastic or aluminium	High
Gamma rays	0	0	Electromagnetic radiation of very high frequency	Stopped by a few centimetres of lead	None

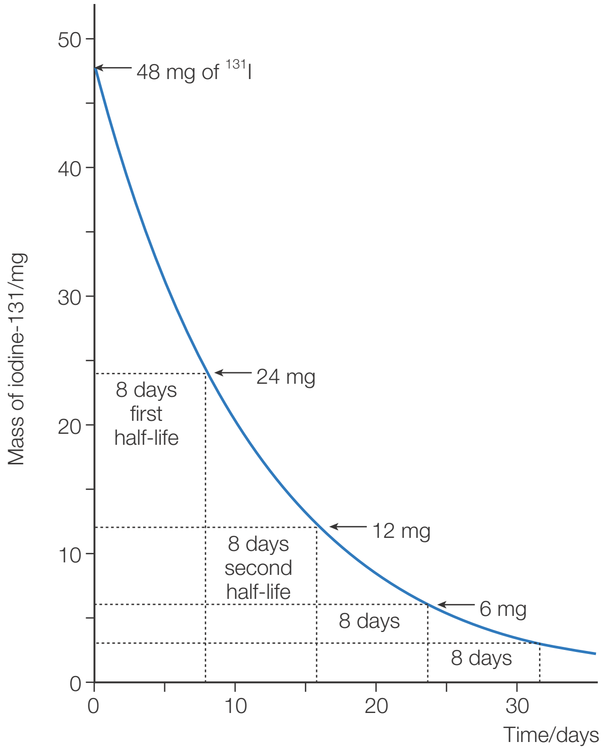


Figure 24.23 Half-life curve for iodine-131

Half-life

The rate at which nuclei undergo radioactive decay varies between chemical elements. Radioactive decay is an exponential process. The rates of radioactive decay are compared using the half-life, which is the time taken for half of the radioactive nuclei to undergo decay.

Each radioisotope has its own unique half-life which is unaffected by temperature or pressure. For example, iodine-131 has a half-life of 8 days. This means that every 8 days the number of radioactive atoms present halves (Figure 24.23).

Waste from nuclear fission contains a mixture of fission products, with nuclei of intermediate size, along with small amounts of actinoids – very heavy elements formed by neutron capture by uranium-238 nuclei. These actinoids, including plutonium and neptunium isotopes, have half-lives of thousands of years. Fission products, such as isotopes of krypton, barium, caesium and strontium, typically have half-lives of less than 100 years. As the radioactive elements in the waste fall into these two distinct categories, at first the radioactivity falls relatively quickly over the first few centuries, as the fission products decay, and then the radioactivity settles to a low but almost constant level for thousands of years. Typically, nuclei with high activity have short half-lives and those with low activity have longer half-lives. Table 24.7 summarizes the half-lives of some isotopes.

Table 24.7 Examples of half-lives of isotopes in nuclear waste

Isotope	Half-life
Fission products	
Iodine-131	8 days
Barium-140	12.7 days
Krypton-85	10.7 years
Strontium-90	28.8 years
Caesium-137	30 years
Actinoids	
Plutonium-239	24 000 years
Plutonium-242	376 000 years
Neptunium- 237	2.2 million years

The presence of actinoids in fission waste creates enormous difficulties for waste disposal. After 100 years the majority of radioactivity from fission products has disappeared, but the presence of actinoids means that the waste is still dangerously radioactive and will be almost indefinitely. Breeder reactors, which can 'burn up' the actinoids in nuclear waste, help solve this problem.

Solution of radioactive decay problems

Half-life calculations

As each half-life elapses, the number of radioactive nuclei in the sample will drop to half of its original value. The amount of radioactive material in the sample after a given time can therefore be calculated.

The equation $N = N_0e^{-\lambda t}$ is given in the *IB Chemistry data booklet*.

N = amount of nuclear material remaining

N_0 = amount of nuclear material present at the beginning

λ = decay constant for the material

t = time elapsed

The equation $t_{\frac{1}{2}} = \dfrac{\ln 2}{\lambda}$ is also given in the *IB Chemistry data booklet.*
$t_{\frac{1}{2}}$ = half-life
Rearranging the first equation for λ and substituting in the second equation gives:

$$t_{\frac{1}{2}} = t\,\frac{\ln 2}{\ln\left(\frac{N_0}{N}\right)}$$

Which can also be expressed as $N = N_0\left(\frac{1}{2}\right)^{\textit{no. of half lives}}$

Worked examples

A sample contains 10 g of iodine-131, with a half-life of 8 days. Calculate the mass of iodine-131 remaining in the sample after 24 days.

No of half-lives: $\frac{24}{8} = 3$ half-lives

Mass of ^{131}I in sample after 24 days, $= N = N_0\left(\frac{1}{2}\right)^{\textit{no. of half lives}} = 10 \times (0.5)^3 = 1.25\,\text{g}$

The mass of radioactive carbon-14 in a fossilized mammal bone is 1.16 g. The mass of the same isotope in a recently killed mammal is 18.56 g. The half-life of carbon-14 is 5700 years. How old is the fossilized bone?

Rearranging gives $t = t_{\frac{1}{2}}\dfrac{\ln\left(\frac{N_0}{N}\right)}{\ln 2} = 5700 \times \dfrac{\ln\left(\frac{18.56}{1.16}\right)}{\ln 2} = 5700 \times 4.00 = 22\,800$ years

7 Neptunium-237 is an actinoid produced in nuclear reactors. It has a half-life of 2.14 million years. How long will it take for its radioactivity to fall to 10 per cent of its original value?

Discussion of the storage and disposal of nuclear waste

Hazards from radioactive waste

Nuclear accidents or accidental exposure to high-level nuclear waste can result in acute radiation poisoning which results in death within a matter of days. However, the risks of such occurrences are minimized by careful reactor design and regulation of waste disposal. Nonetheless, many people are worried about possible exposure to low levels of radiation over prolonged periods. Such exposure can lead to cancers and birth defects. Concern about the dangers of such exposure leads to local opposition to the siting of nuclear power stations and waste storage facilities.

High-level waste

When the uranium-235 fuel in a reactor is used up, the fuel rods are extremely radioactive. They contain actinoids, such as uranium-234 and plutonium-241 (the f-block on the periodic table consists of actinoids and lanthanoids), which emit alpha particles, and other fission products such as strontium-90 and caesium-137, which emit beta particles and gamma rays, respectively. Much waste also contains plutonium-239, which can be used to make nuclear weapons.

The long half-lives of these materials, combined with their very dangerous levels of radioactivity means that they must be disposed of very carefully. Alpha and beta sources present the most risk when ingested, so it is vital that the materials are well contained and not allowed to enter groundwater.

Some high-level waste is treated by a process called vitrification. The waste is heated in a furnace with broken glass. The resulting liquid glass is poured into steel containers. The glass solidifies into cylinders with the radioactive materials fused into the glass structure, which is fully waterproof. It is thus impossible for the radioactive materials to be washed out if exposed to water.

However, the vitrified waste is still radioactive. It must be stored in places where people are unlikely to accidentally encounter it, and a safe distance from plant and animal life. Much vitrified waste is actually held in temporary storage facilities under guard, but countries such as the USA and Australia are exploring options for burying the waste in very deep bunkers, in isolated areas. The thick surrounding rock will absorb any gamma rays from the waste, minimizing the chances of radiation harming any plant and animal life.

High-level waste such as fuel rods contains radioactive material that, with further processing, could be used as nuclear fuel. Temporary burial allows for the possibility that the waste could be reclaimed in the future when technology makes such reprocessing economically viable.

Low-level waste

The removal of items from nuclear power station sites is carefully controlled. Even materials that have not been near the reactor itself are designated as radioactive waste, and their disposal is monitored. These articles include protective clothing, paper filters and plastic bags. The radioactivity is low level and short lived. Although these materials are classified as radioactive waste as a precautionary measure, they are usually barely more radioactive than normal domestic waste. Such waste is treated in a similar way to municipal solid waste – it is incinerated to reduce its volume and then buried in dedicated landfills.

Medical waste

Radioactive isotopes are used in numerous medical applications. Examples include:

- Use of short half-life substances as medicines or diagnostic tools: iodine-131 is used to treat thyroid cancers, as the thyroid gland absorbs iodine. The radiation is therefore carried to the source of the tumour, killing the cancer cells. The short half-life means that the radiation level falls rapidly, so the rest of the body is not subject to dangerous levels of radiation.
- Use of long half-life substances to generate radiation in machines used for radiotherapy: caesium-137 is used to generate beta radiation, which can be directed in a beam at tumours.
- Long half-life substances such as cobalt-60 and iridium-192 are also used to generate X-rays for radiography.
- Technetium-99m is used as a radioactive tracer. This is a gamma-emitting isotope with a short half-life of only about 1 day. It can be injected into the body and the gamma emission detected by a gamma camera. It is carried by the blood into the organs (for example, the brain, the kidneys or the liver) and can be used to determine blood flow in these regions. Abnormal blood flow might indicate the presence of tumours.

Most radioactive materials used in hospitals have a short half-life which means they are rendered safe by leaving them to decay for some time. Materials used by medical professionals such as clothing and gloves are classified as low-level waste and are disposed of as described above.

Disposal of isotopes from old X-ray or radiotherapy machines is carefully regulated. These materials are buried, as with high-level waste from reactors.

ToK Link

Risk assessment in the nuclear industry

The assessment of risk in the field of nuclear energy presents challenges that illustrate the difficulty of applying a consequentialist (utilitarian) ethical framework.

Under a consequentialist framework, the ethical decision is reached via a 'calculation' of the costs and benefits of an action. When comparing nuclear power with fossil fuels, for example, the difficulty arises when deciding which factors to include in the calculations.

Some commentators suggest that the risks of nuclear power are overemphasized. Even allowing for the victims of prominent nuclear accidents, the number of victims of the fossil fuel industry (for example, coal mining in China or the gas industry in Africa) is far larger. There are also larger-scale externalities to these industries. For nuclear power, these are the risk of accident and the disposal of waste. For fossil fuels, the major issue is climate change. The utilitarian calculation differs depending on what we include: do we just consider the direct impact on human life right now, or do we consider wider implications or possible risks? Should the lives that might be lost in both industries be treated as statistics to be inserted into a formula?

Others resist the utilitarian calculus altogether, and suggest that we should reject nuclear power on principle, as we have a duty not to inflict the waste problem on later generations. Similar arguments could be applied to fossil fuels and climate change, of course.

So how do we decide on the best course of action? There is no simple answer. Various international organizations exist to monitor and police nuclear energy and carbon emissions, but countries are always likely to consider their own interests before those of the wider world. Less economically developed countries want to utilize their fossil fuel resources in order to develop their economies. Other countries might see nuclear energy as a means to ensure their energy security as fossil fuels supplies start to dwindle.

International cooperation in the nuclear industry

The International Atomic Energy Agency (IAEA)

The International Atomic Energy Agency was founded in 1957 with the twin goals of promoting international cooperation in the development of peaceful applications of nuclear energy and policing the proliferation of atomic weapons.

It is essential that international standards exist in nuclear safety, because the harmful effects of a nuclear accident do not respect international borders. After the 1986 accident in Chernobyl, Ukraine, a cloud of radiation led to elevated levels of radioactivity in soil throughout Europe. Subsequently the IAEA performed checks on other countries' nuclear safety standards. A similar process occurred after the Fukushima nuclear accident in 2011.

In the anti-proliferation role, the IAEA is authorized to establish safeguards designed to limit the spread of both nuclear material and information relevant to weapons manufacture.

In both its roles, the IAEA cannot actually prevent sovereign states from contravening international agreements. Instead it aims to work with states to ensure that their own laws regarding nuclear safety meet international standards and that these laws are properly enforced.

International cooperation in particle research

Research into particle physics involves enormously expensive facilities called particle accelerators. The most famous of these is probably the Large Hadron Collider at CERN in Switzerland, which cost around US$ 9 billion, funded by 20 European nations. There are other facilities at Fermilab (near Chicago, USA), SLAC (California, USA) and DESY (Germany).

These facilities offer perhaps the ultimate example of international collaboration in the sciences. Although in one sense they work in competition with one another, their research findings are published in international peer-reviewed journals, and scientists will move from one facility to another to share expertise and gain experience.

Figure 24.24 Architect's rendition of the International Thermonuclear Experimental Reactor in France

The International Thermonuclear Experimental Reactor (ITER)

ITER is an internationally funded project aiming to build a nuclear fusion reactor. Participating countries include India, Japan, China, Russia, South Korea, the USA and as well as the European Union, which hosts the project and is the major funding body.

The reactor is being constructed in the south of France, and will be the largest magnetic confinement fusion reactor in the world (Figure 24.24). The purpose of the reactor is to demonstrate the feasibility of getting more energy out of the fusion reaction than is put in, and to develop containment materials capable of withstanding the intense neutron bombardment of a sustained fusion reaction. It is hoped that the success of ITER will lead to the first commercial fusion reactor.

24.4 Solar energy – *visible light is absorbed by molecules that have a conjugated structure with an extended system of alternating single and multiple bonds, when the electrons within those bonds absorb light energy and are excited to higher energy levels. Solar energy can be converted to chemical energy in photosynthesis*

Nature of Science

Science, technology and public understanding

The field of solar energy can be used to exemplify two important aspects of scientific endeavour.

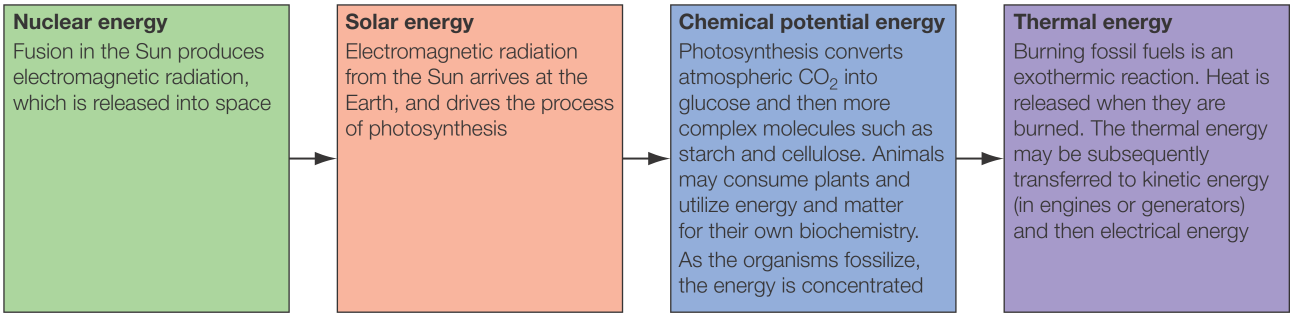
The first is the relationship between science and technology. Frontier research on solar cells is reliant on highly specialized experimental design that proceeds incrementally as the efficiency of such solar cells is measured and evaluated and successful designs carried forwards. In contrast, the technological side is more concerned with how the science will actually be used, and ensuring that it can be applied in cost-effective ways. Real-world solar panels must be as reliable as possible, as cheap as possible and as aesthetically pleasing as possible.

The second is the public understanding of science, and how that understanding is reached. It is widely understood that biofuels are desirable, as they are potentially carbon neutral and therefore offer a solution to climate change. This is a reassuring prospect. This view of biofuels has been heavily promoted by industry and by governments keen to demonstrate action against climate change. This is in contrast to the actual situation, in which the advantages of biofuels must be balanced against the impact on food production and biodiversity. Such a view of solar energy and biofuels cannot be conveyed in a sound bite and is a much more difficult concept to present to the public.

Solar energy

The phrase 'solar energy' probably prompts thoughts of solar panels capturing sunlight as part of a 'green' electricity-generation project. Indeed, research is progressing in this area, and the technology used to convert sunlight into electricity will be studied in Section 24.8.

However, energy that originated in the Sun is already widely used. Fossil fuels can be thought of as concentrated solar energy, as the photosynthetic process that drove the capture of atmospheric carbon into living organisms was powered by sunlight. As the living things decayed, the energy density of the dead organisms increased, until the fossil fuels we use today were produced (Figure 24.25).



Nuclear energy
Fusion in the Sun produces electromagnetic radiation, which is released into space

Solar energy
Electromagnetic radiation from the Sun arrives at the Earth, and drives the process of photosynthesis

Chemical potential energy
Photosynthesis converts atmospheric CO_2 into glucose and then more complex molecules such as starch and cellulose. Animals may consume plants and utilize energy and matter for their own biochemistry.
As the organisms fossilize, the energy is concentrated

Thermal energy
Burning fossil fuels is an exothermic reaction. Heat is released when they are burned. The thermal energy may be subsequently transferred to kinetic energy (in engines or generators) and then electrical energy

Figure 24.25 The historical energy transfers involved in the burning of fossil fuels

Photosynthesis

Fossil fuels are non-renewable, meaning that they will not be replenished within human time scales, as fossilization takes many millennia.

However, plants are still photosynthesizing and converting solar energy into chemical potential energy. Can the chemical energy in plants be efficiently converted into useful fuels?

The biochemical processes in photosynthesis:

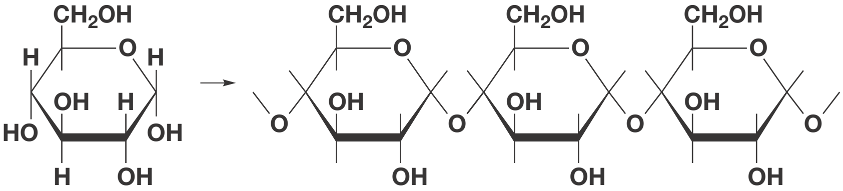
Photosynthesis converts atmospheric carbon dioxide and water vapour into glucose and oxygen:

$$6CO_2 + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2$$

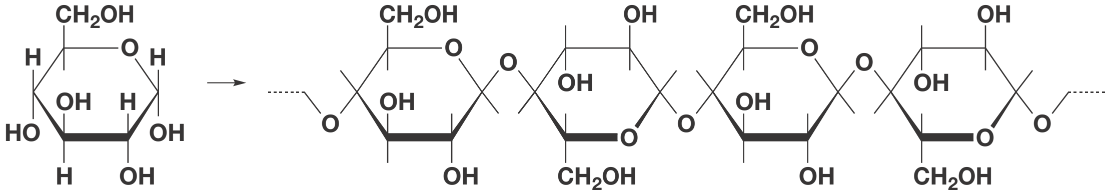
Glucose is a carbohydrate – a molecule containing carbon, hydrogen and oxygen, with the general formula $C_n(H_2O)_m$, where n and m are integers, but not necessarily having the same value. For glucose, m and n are both 6: $C_6H_{12}O_6$ can be written $C_6(H_2O)_6$. For sucrose, $C_{12}H_{22}O_{11}$, m is 12 and n is 11: $C_{12}(H_2O)_{11}$.

Plants consume glucose as a source of energy during respiration. Glucose is a monosaccharide molecule that can be polymerized into starch chains via condensation reactions, which the plant uses as an energy store (Figure 24.26).

Figure 24.26 Polymerization of glucose forming starch



Plants also use glucose as a structural 'building block': glucose molecules combine together to make cellulose, a tough, fibrous substance, which is used to make cell walls (Figure 24.27).



■ **Figure 24.27** Formation of cellulose from glucose

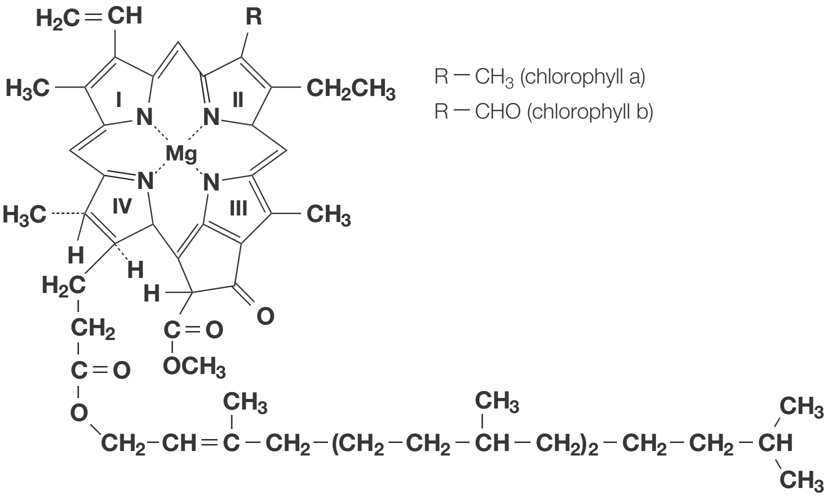
The structures of cellulose and starch are in many ways quite similar. However, they are sufficiently different that enzymes in the digestive systems of living organisms cannot break down both. Close inspection of the two structures reveals that the orientation of the glucose monomers is different in each structure.

Photosynthesis happens inside plant cell structures called chloroplasts, which are disc-shaped organelles (cell organs) adapted for harvesting light energy.

Absorption of light by chlorophyll

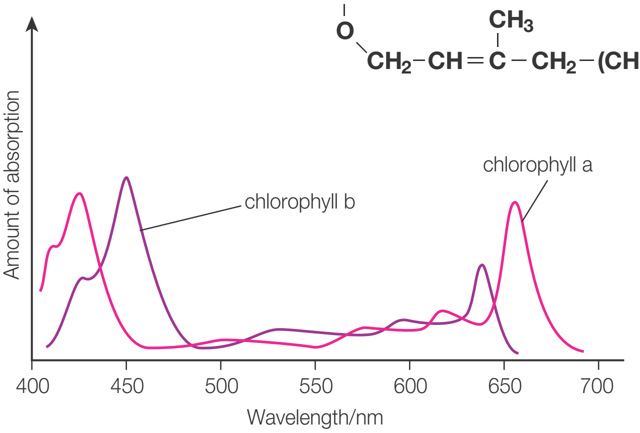
Chloroplasts contain a molecule called **chlorophyll**, which is a very efficient absorber of sunlight. Chlorophyll is a **pigment** – a coloured molecule. It is coloured because it absorbs some of the wavelengths present in white light, with the remaining wavelengths being reflected. Chlorophylls absorb light most strongly in the red and violet parts of the visible spectrum. The green part of the spectrum is poorly absorbed, so when white light (which contains the whole visible spectrum) shines on to leaves, the green light is reflected, meaning that the leaves appear green.

Chlorophyll actually exists in two slightly different forms (chlorophyll a and chlorophyll b; Figure 24.28), which absorb slightly different wavelengths. This means that they appear as slightly different shades of green.



R – CH_3 (chlorophyll a)
R – CHO (chlorophyll b)

■ **Figure 24.28** Structure of chlorophyll



■ **Figure 24.29** Absorption spectra of chlorophylls a and b

Both types of chlorophyll have strong absorption at each end of the spectrum (Figure 24.29). However, chlorophyll a has its largest absorbances around the 430 nm mark which corresponds to the blue–violet region and the 670 nm mark which is in the red region. Chlorophyll b has its strongest absorbances at about 455 nm (blue) and 640 nm (the orange end of the red region). The different proportions of blue, violet, red and orange absorbed leads to us perceiving chlorophyll a as blue–green and chlorophyll b as yellow–green.

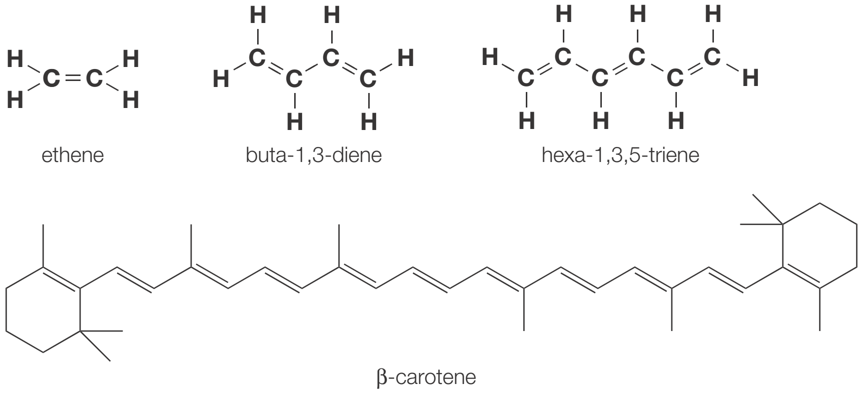
Identification of features of the molecules that allow them to absorb visible light

Chlorophylls are members of a family of molecules called **porphyrins**, which feature four **pyrrole** groups (five-member carbon rings containing a nitrogen atom) joined together by alternating carbon–carbon single and double bonds. This structure is called a **tetrapyrrole**. The four nitrogen atoms are held in place in the centre of the molecule, where they are able to bond to metal ions using dative bonds. The porphyrin molecule acts as a polydentate ligand (Chapter 13) forming a complex with the central metal ion. In chlorophyll the central metal ion is an Mg^{2+} ion.

Metal ions complexed by porphyrins feature in a number of biological systems. The iron(II) ion complexed by a heme group in hemoglobin is another example (Section 23.9).

Both forms of chlorophyll feature an extensive system of alternating single and double bonds, within the pyrrole rings and between them via the connecting linkages. Such a system of bonds is called a conjugated system, and it allows for delocalization of the π electrons over large parts of the molecule. Figure 24.30 shows a series of molecules with increasing amounts of conjugation. The chlorophyll porphyrin therefore features π molecular orbitals above and below the plane of the molecule. It is this π system which is responsible for the colour of chlorophyll and other porphyrin pigments.

Figure 24.30 Molecules showing increasing amounts of conjugation



When light is shone on to an unsaturated organic molecule, one of the π electrons may be excited to a higher energy level called π*. The energy needed to promote the electron to this higher level corresponds to a photon of a specific wavelength. For most π bonded molecules, these photons are in the ultraviolet region, so there is no effect on the visible light region, and the molecule is colourless. However, as more conjugated π bonds are added, the energy of the photons absorbs falls, as the π* levels decrease in energy. This means that the wavelengths absorbed increase until they appear in the visible region. Therefore, molecules with large conjugated systems absorb visible radiation, making them coloured.

Table 24.8 shows some examples of maximum absorption wavelengths (λ_{max}) for a series of organic molecules with increasing amounts of conjugation.

Table 24.8 Maximum absorption wavelength, λ_{max}, for a series of organic molecules

Molecule	λ_{max}/nm	Colour
Ethene	175	Colourless (absorbs in ultraviolet)
Buta-1,3-diene	217	Colourless (absorbs in ultraviolet)
Hexa-1,3,5-triene	258	Colourless (absorbs in ultraviolet)
β-carotene *(11 conjugated C=C bonds)*	approximately 450	Orange (absorbs blue–violet)
Chlorophyll a	approximately 650	Blue–green (absorbs red)

Chlorophyll a and chlorophyll b have different values of λ_{max} because of the slight differences in their structures. Figure 24.28 shows that chlorophyll a has a methyl ($–CH_3$) group attached to the second pyrrole ring, whereas chlorophyll b has an aldehyde group (–CHO) in the same position.

This slight difference changes the energy difference between the π energy level and the excited π* level, which changes the wavelengths of light absorbed, changing the colour of the molecule.

What does chlorophyll do?

Chlorophyll a is the molecule that most green plants use to capture the Sun's energy. Look again at the structure of chlorophyll a. In addition to the porphyrin group discussed above, it has a long hydrocarbon 'tail'. Recall that chlorophyll molecules are found inside cell structures called chloroplasts. These chloroplasts carry many chlorophyll molecules, grouped together.

Imagine the porphyrin part of a chlorophyll a molecule as the 'head'. The porphyrin is planar (flat) so if placed face-on towards the light source it can capture sunlight effectively. The hydrocarbon 'tail' helps to position the porphyrin, like the handle of an umbrella. The hydrocarbon tails are embedded into the chloroplast membranes and hold the porphyrins in place.

A chloroplast contains hundreds of thousands, possibly millions, of chlorophyll molecules embedded in protein scaffolds in its membrane, which 'harvest' light energy. Each chlorophyll a absorbs photons of red light. The energy is channelled along the chloroplast membrane, from chlorophyll to chlorophyll, until it reaches a 'reaction centre' where it powers photosynthesis.

Chlorophyll a fulfils a dual role:

1. It captures photons efficiently, owing to its shape.
2. It uses the energy from the light to initiate the photosynthetic process.

Photosynthesis is a redox reaction, which can be split into two half-reactions as follows:

Step 1: the oxidation of water:

$$2H_2O \rightarrow O_2 + 4H^+ + 4e^-$$

Step 2: the reduction of carbon dioxide to carbohydrate (eventually glucose):

$$4e^- + 4H^+ + CO_2 \rightarrow (CH_2O) + H_2O$$

Addition of the two half-reactions gives:

$$H_2O + CO_2 \rightarrow (CH_2O) + O_2$$

Which generalizes to:

$$nH_2O + nCO_2 \rightarrow (CH_2O)_n + nO_2$$

Each of the two half-reactions requires energy input from sunlight. Chlorophyll molecules in different regions of the chloroplast carry out the two processes.

The presence of chlorophyll b in plants broadens the range of light wavelengths that can be absorbed, so chlorophyll b tends to be present in higher concentration in 'shade-adapted' chloroplasts – those which function better at lower light levels.

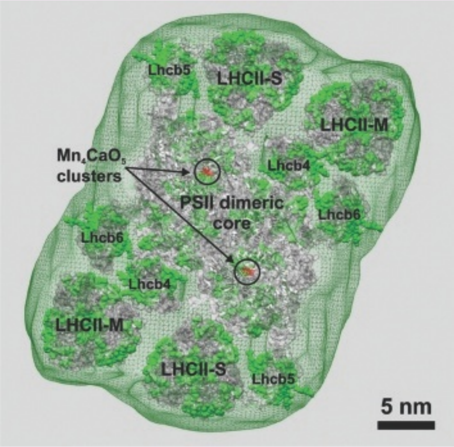
8 Photosynthesis involves the reduction of carbon dioxide to glucose. Deduce the oxidation numbers of carbon in carbon dioxide and glucose. Take the oxidation numbers of hydrogen as 1 and oxygen as –2.

Additional Perspective

Photosystems

Chlorophyll molecules act as 'light-harvesters' which provide energy to drive the reduction of carbon dioxide to carbohydrates. This requires the formation of an actual chemical species which will act as the reducing agent. This process occurs in reaction centres called photosystems, of which there are two types, called PSI and PSII. The two systems absorb light of slightly different wavelengths and are linked together by a series of reactions in which PSII transfers electrons to PSI. A PSII complex might contain 100–250 chlorophyll molecules, and there will be hundreds of photosystems on each of hundreds of membranes in each chloroplast, leading to millions of chlorophyll molecules in a chloroplast. Figure 24.31 shows a PSII supercomplex.

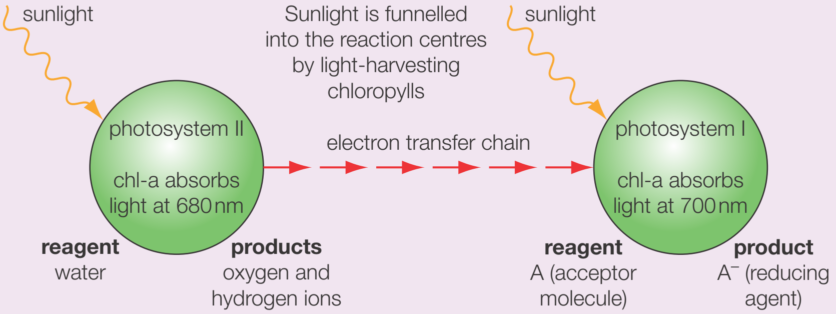
Figure 24.31 An electron micrograph of a PSII supercomplex. In this diagram, the individual chlorophyll molecules (about 220 of them) are shown as green spheres



During the electron transfer processes in the photosystems, water is oxidized to oxygen in PSII and the electrons released are fed into PSI. PSI then takes these electrons and uses them to reduce an electron acceptor molecule, A. The structure of molecule A is not fully known.

The reduced form of molecule A, A$^-$, is then able to reduce carbon dioxide to glucose elsewhere in the plant. Figure 24.32 shows the action of photosystems I and II.

Figure 24.32 Photosystems I and II; chl-a refers to chlorophyll a



Using the stored energy in plants: biofuels

Ethanol

Ethanol (C_2H_5OH) is a biofuel made by fermentation of glucose. It is most widely used in Brazil and the USA, where crops such as corn (maize) and sugar cane are used as the source of glucose.

Fermentation is carried out by single-celled fungi called yeasts, which use an enzyme called **zymase** to convert glucose molecules into ethanol molecules, with production of carbon dioxide:

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2$$

The glucose is added to the yeasts in a vessel called a **bioreactor**. The reaction is exothermic, so the temperature must be carefully controlled so that it does not get too hot, denature the enzyme and kill the yeasts.

Fermentation produces a solution of ethanol, along with side products such as aldehydes and other alcohols. The maximum achievable ethanol concentration is limited to about 15 per cent as concentrations higher than this will poison the yeasts.

The mixture is distilled to remove side products and remove the water, increasing the ethanol concentration.

Evaluation of the advantages and disadvantages of the use of ethanol

Advantages of ethanol

- *Lower greenhouse emissions:* ethanol has been enthusiastically adopted in some countries as a means of lowering greenhouse gas emissions. Ethanol does produce carbon dioxide when burned:

  $$C_2H_5OH + 3O_2 \rightarrow 2CO_2 + 3H_2O$$

  However, the carbon in the ethanol was absorbed from the atmosphere by the plants used to make the ethanol, as they grew, so theoretically the ethanol is 'carbon neutral'.
- *Reduced oil use:* ethanol is widely used in Brazil as sugar cane is a large-scale crop there. Since 1976 it has been mandatory to blend ethanol with gasoline, as a way of reducing oil use. Many vehicles sold in Brazil are designated 'flex-fuel', because they can run on a wide range of blends, ranging from almost pure ethanol to almost pure gasoline. In the USA, most new cars sold can run on gasoline containing 10 per cent ethanol, which means that in areas where ethanol is widely available, oil use (and so carbon footprint) per person will decrease.

Disadvantages of ethanol

- *Lower specific energy and potential damage to engines:* ethanol has a lower specific energy than gasoline. Approximately 1.5 litres of ethanol is required to provide the equivalent energy of 1 litre of gasoline. Unlike gasoline, ethanol absorbs moisture from the atmosphere, which can lead to engine corrosion. Blended fuels containing ethanol and gasoline are more prone to causing engine damage than pure gasoline.
- *High embodied energy:* although ethanol is a renewable fuel, as sugar crops can be quickly replenished, ethanol production is criticized for being energy intensive. Ethanol from bioreactors must be distilled which requires large amounts of energy. By some calculations, when the processing is taken into account, use of ethanol does not actually reduce greenhouse gas emissions relative to traditional gasoline.
- *Food versus fuel debate:* when maize (corn) is used as a source of ethanol, as in the USA, critics claim that the price of maize on world markets is forced up, making poorer countries more likely to export their staple food crops and leaving their own populations vulnerable to hunger. However, others argue that the majority of world maize production is used for animal feed, so the impact of maize-based ethanol on human food production is minimal.

Use of gasoline and ethanol mixtures in different countries – the E nomenclature

Most cars in the USA can run on 'E10' – gasoline containing 10 per cent ethanol by volume. In corn-producing regions of the USA, higher proportions of ethanol are common: E85 is widely available.

In Brazil, some vehicles run on 100 per cent ethanol (E100), but an ethanol content of at least 18 per cent (E18) is mandatory for all gasoline vehicles.

In Sweden, flex-fuel vehicles are available which run on a wide range of gasoline–ethanol blends, from pure gasoline (E0) up to E85. Such vehicles rely on electronic engine management systems, which adjust the spark and inlet valve timings according to the fuel composition, to prevent knocking (pre-ignition) in the engine.

Biodiesel

Diesel engines run on a much heavier crude oil fraction than gasoline. Whereas gasoline is composed of hydrocarbons with about eight carbon atoms, diesel fuels contain longer hydrocarbons, typically 14–20 carbon atoms. Diesel is therefore more viscous and less volatile than gasoline.

Ethanol is not a suitable fuel for vehicles with diesel engines because it is less viscous and much more volatile than diesel, so the vehicle's fuel system would need to be extensively modified. Also, ethanol has a much lower energy density than diesel.

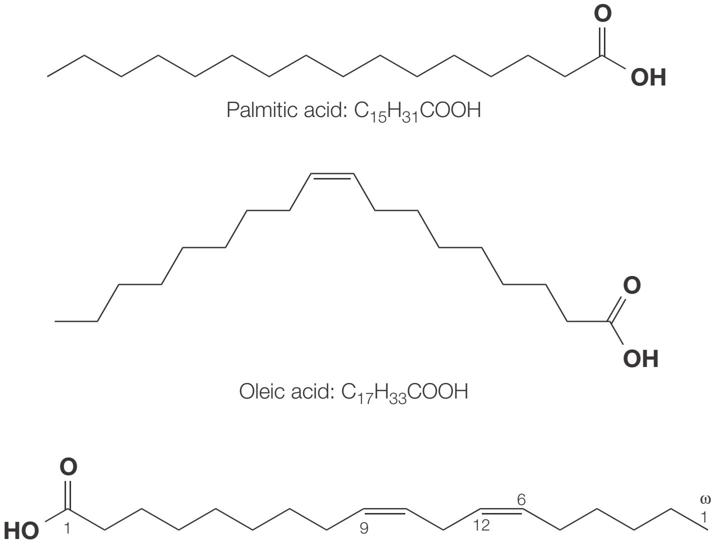
However, there are plant-based fuels that are much closer to diesel in their properties. These fuels are derived from vegetable oils and are commonly called biodiesels.

The structures of vegetable oils

Fats and oils are examples of **lipids**, naturally occurring biochemical molecules found in both plants and animals. In plants, lipids are more often liquids and are called oils. In animals, lipids are more often solid and are called fats.

Lipids are esters of propane-1,2,3-triol (glycerol) and long-chain carboxylic acids called **fatty acids** (Figure 24.33). Glycerol has three hydroxyl groups, each of which performs a condensation reaction with a fatty acid molecule, with a water molecule being eliminated. Figure 24.34 shows an example of the formation of a lipid.

In this example the three fatty acid chains are identical saturated molecules with 18 carbon atoms (octadecanoic acid). Such molecules are called **triglycerides**. However, the three fatty acids may be different from one another and may be unsaturated.



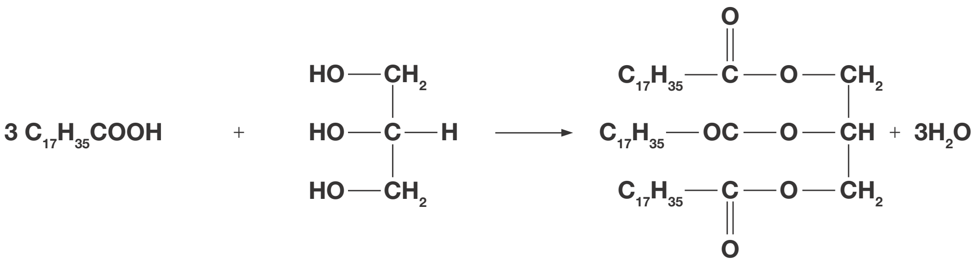
Palmitic acid: $C_{15}H_{31}COOH$

Oleic acid: $C_{17}H_{33}COOH$

Linoleic acid: $C_{17}H_{31}COOH$

Figure 24.33 Some commonly occurring fatty acids

Figure 24.34 The formation of a lipid



octadecanoic acid (stearic acid)

propane-1,2,3-triol (glycerol)

propane-1,2,3-triyl trioctadecanoate (tristearin)

In plant oils, the fatty acids are more commonly mono-unsaturated (with one carbon–carbon double bond) or poly-unsaturated (with several double bonds). Table 24.9 lists some examples.

Table 24.9 Composition of some vegetable oils

Oil	Approximate composition of triglyceride
Olive oil	Up to 80% palmitic acid Up to 20% linoleic acid Up to 20% oleic acid
Canola oil	Approximately 60% oleic acid Approximately 20% linoleic acid Plus others including up to 7% saturated fatty acids
Palm oil	Approximately 44% palmitic acid Approximately 39% oleic acid Approximately 10% linoleic acid Plus others

Lipids with unsaturated fatty acids have lower melting points than those with saturated fatty acids. The presence of the double bond introduces a ‘kink’ in the carbon chain. The fatty acid chains are less able to lie closely alongside one another, so the dispersion forces are weaker. Lipids with unsaturated fatty acids are therefore likely to be liquid at room temperature. Figure 24.35 illustrates the structures of saturated and unsaturated lipids:

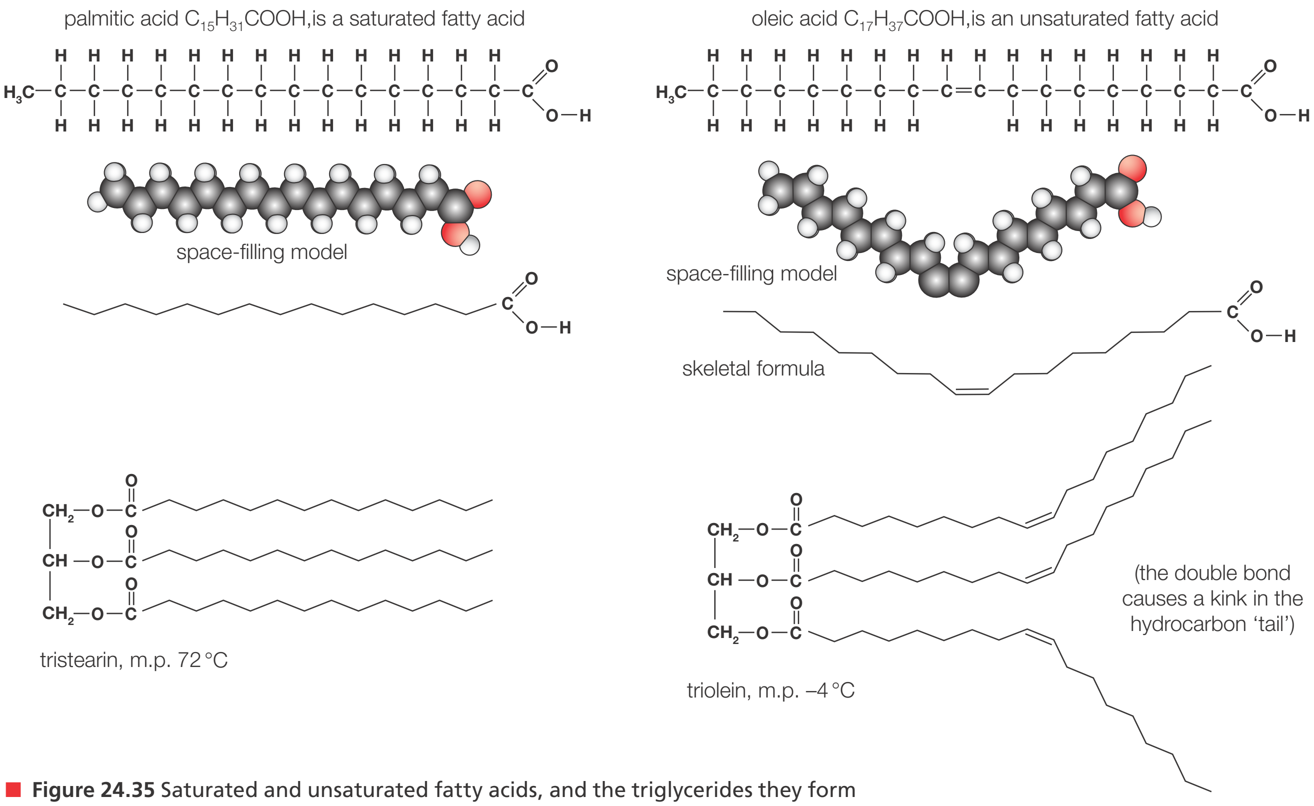

palmitic acid $C_{15}H_{31}COOH$,is a saturated fatty acid
space-filling model
oleic acid $C_{17}H_{37}COOH$,is an unsaturated fatty acid
space-filling model
skeletal formula
tristearin, m.p. 72 °C
triolein, m.p. –4 °C
(the double bond causes a kink in the hydrocarbon ‘tail’)

Figure 24.35 Saturated and unsaturated fatty acids, and the triglycerides they form

The properties of vegetable oils

Vegetable oils are extracted by pressing the fruit or the seeds of plants. For example, palm oil is made from the fruit pulp, and canola oil is made from the seeds of rapeseed or field mustard plants. Olive oil is made by pressing whole olives – the fruit of the olive tree.

When extracted, vegetable oils are more viscous, but their energy densities are close to fossil fuel diesel (Table 24.10). Some older diesel engines are able to use untreated vegetable oils directly. This is referred to as straight vegetable oil (SVO). In Germany it is possible to buy a kit to modify the fuel system of a diesel vehicle so that SVO is preheated to lower the viscosity before it enters the engine. A vehicle so modified is able to run on used cooking oil.

Table 24.10 Energy density of fossil fuels compared with plant-derived fuels

Fuel	Energy density/$MJ\,dm^{-3}$
Gasoline	46
Diesel	43
Sunflower oil	33
Commercial biodiesel	33–35

More modern turbocharged diesel engines cannot tolerate the higher viscosity, and so the vegetable oil must be chemically treated to lower the viscosity.

Transesterification of vegetable oils

The high viscosity of vegetable oils arises from their high relative formula masses and the long-chain fatty acid chains present, which leads to strong dispersion forces. To lower the viscosity of the oils, the approach taken is to break up the triglyceride into fatty acid esters and glycerol, by a process called **transesterification**.

Transesterification involves reacting the triglyceride with methanol or ethanol, replacing the glycerol in the esters with alcohol molecules and freeing the glycerol. The products are three methyl or ethyl ester molecules and a glycerol molecule (Figure 24.36).

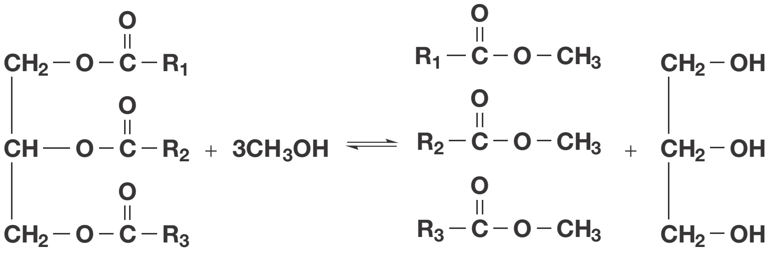


Figure 24.36 Transesterification of a triglyceride with methanol, producing methyl esters and glycerol

The reaction is catalysed by a strong acid such as sulfuric acid, or a strong base such as potassium hydroxide. Esterification reactions are covered in Chapter 10.

The equation in Figure 24.36 shows that three moles of methanol (or ethanol) should be used for the transesterification process. However, as the reaction is in equilibrium, in commercial manufacture about 6 moles of methanol are provided per 1 mole of triglyceride. By applying Le Châtelier's principle, we see that this has the effect of shifting the equilibrium to the right, which increases the yield of methyl esters. Excess methanol is then recovered and recycled into the reaction vessel.

The transesterification process is often incomplete, meaning that some of the triglycerides are broken down into fatty acids which do not then re-esterify with the added methanol or ethanol. The resultant mixture therefore has some fatty acids present in it.

Deduction of equations for transesterification reactions

If ethanol is used for the transesterification reaction, then ethyl esters will be formed. If methanol is used then methyl esters will be formed.

Worked example

Olive oil is largely composed of triglycerides of palmitic acid, $C_{15}H_{31}COOH$. Deduce the equation for transesterification with methanol of this triglyceride.

The triglyceride can be expressed as $(C_{15}H_{31}COO)_3CH(CH_2)_2$

Three methanol molecules are used in the transesterification process:

$$(C_{15}H_{31}COO)_3CH(CH_2)_2 + 3CH_3OH \rightleftharpoons \underset{\text{methyl palmitate}}{3C_{15}H_{31}COOCH_3} + \underset{\text{glycerol}}{CH_2OHCHOHCH_2OH}$$

Explanation of the reduced viscosity of esters produced with methanol and ethanol

The mixture of substances present after transesterification has broadly similar chemical bonds and functional groups within it as the original vegetable oil, so the energy density of the mixture remains close to fossil fuel diesel. However, the molecules within it are much smaller than the original triglycerides, meaning that the dispersion forces in the mixture are weaker.

9 Deduce the transesterification reaction, using ethanol, of a triglyceride of linoleic acid, $C_{17}H_{31}COOH$.

Evaluation of the advantages and disadvantages of biodiesel

Many of the advantages and disadvantages of ethanol fuel, discussed above, can also be applied to biodiesel. Briefly:

Advantages:

- decreased fossil fuel use
- potential for lower greenhouse emissions, but this is open to question.

Disadvantages:

- lower specific energy than the fossil fuel equivalent
- displacement of food crops by crops grown specifically for fuel
- high embodied energy in biodiesel, because extensive processing is required.

Another particular point of controversy surrounding biodiesel is **loss of biodiversity**. In South-East Asian countries such as Malaysia and Indonesia, large areas of primary tropical rainforest have been logged for timber and then replaced with a monoculture based on palm oil production (Figure 24.37). Palm oil is an extremely important oil for food uses around the world but can also be processed for biodiesel. Palm oil plantations consist of very large areas of land planted with palm trees in continuous rows. Primary rainforests contain an enormous range of trees and other plants, which support a complex ecosystem of insects, birds and mammals. Many of these organisms cannot survive in a palm oil plantation. As plantations grow and rainforests shrink, the numbers of organisms that can be supported will fall. Governments in Indonesia and Malaysia are under pressure from environmental groups to protect regions of primary rainforest in order to prevent widespread extinctions. However, the palm oil industry is hugely important economically. It provides employment not only on plantations but also in associated transport and infrastructure projects to support the industry, and it also provides export income for the countries concerned. The rise of biodiesel as a popular 'green' fuel in economically developed countries is potentially a driver for environmental harm in less developed countries.

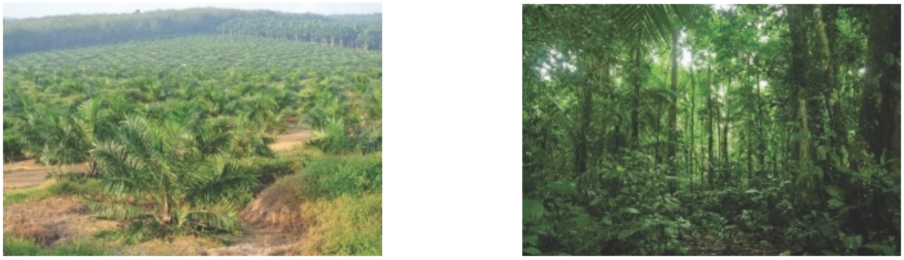


Figure 24.37 a Palm oil plantation in Malaysia; **b** primary rainforest in Borneo, Malaysia

ToK Link

Cold fusion

In 1989, two chemists at the University of Utah made a spectacular announcement: they had achieved nuclear fusion at room temperature. Had this claim proven true, the world might have entered a new era of unlimited clean energy.

Stanley Pons and Martin Fleischmann had devised an experiment involving the electrolysis of heavy water. Heavy water is water in which the hydrogen atoms are replaced with atoms of deuterium. During the electrolysis of water, hydrogen is liberated at the cathode (the negative electrode). However, in Fleischmann and Pons' experiment, the cathode was made of the element palladium, which has an unusual property: palladium is able to absorb hydrogen atoms. The metal absorbs atoms of hydrogen, and the hydrogen atoms fill up spaces in the palladium crystal structure. So effective is this process that a piece of palladium is able to absorb 900 times its own volume of hydrogen. When Fleischmann and Pons connected their apparatus, they reported spikes of intense heat in the apparatus: more energy than was being pumped in by the electrical supply.

The university recognized the implications immediately and encouraged the chemists to announce their findings in a press conference. Fleischmann and Pons speculated that the palladium was packing the deuterium atoms in so tightly that they were fusing together, producing enormous amounts of heat.

Although deuterium is a relatively rare form of hydrogen, there is so much water on Earth that 'cold fusion', as it was dubbed, could solve all the world's energy problems.

What Fleischmann and Pons had not done, however, was remove the inaccuracies in their experiment. The energy spikes were erratic and not reliably reproducible in subsequent experiments. Other research institutions around the world seized on cold fusion and tried to reproduce the results, with little success.

Within weeks of their press conference, Pons and Fleischmann's work had been comprehensively discredited, and the pair were castigated in the scientific press. Today, research in the area of cold fusion is extremely limited as funding institutions are reluctant to associate themselves with such a high-profile failure. This is despite the fact that Fleischmann and Pons' energy spikes have not been fully explained.

Scientific research runs into dead ends like this all the time. Most scientists are able to refine and improve their work, or move on to other projects. Unless results are deliberately falsified, it takes a lot to completely lose a scientific reputation. Why were Fleischmann and Pons treated so harshly? One factor is the way that cold fusion was announced. The normal way of presenting new scientific research is via a peer-reviewed journal. Articles are prepared and are scrutinized by experts in the field, before publication. The cold fusion press conference, although it made an enormous impact, was viewed as arrogant by other researchers – an attempt to circumvent the normal review process. Fleischmann has since claimed that he wanted to publish the research in the normal way, but the university pressured him into the press conference for publicity reasons.

Another problem was that Pons and Fleischmann were reluctant to issue all the technical details of their work. In peer review, a very detailed account of the experiment is provided for the express purpose of allowing others to attempt to replicate the data. This is seen as a necessary part of scientific endeavour. Fleischmann and Pons' approach left them open to further accusations of 'glory-seeking' – as if they wanted all the credit for cold fusion in the future.

The cold fusion affair raises some interesting questions about scientific enquiry:

- Peer review is sometimes criticized as a rather 'closed' way of recording scientific discoveries. In the internet age, some believe that all research findings should be freely available. What might the implications of this be?
- The lack of reproducibility of Fleischmann and Pons' work led to its being discredited. Why do scientists place such great emphasis on reproducibility?

24.5 Environmental impact – global warming – *gases in the atmosphere that are produced by human activities are changing the climate as they are upsetting the balance between radiation entering and leaving the atmosphere*

Nature of Science

Climate science as a transdisciplinary endeavour

The study of global warming and its likely effects is highly transdisciplinary, encompassing aspects of biology, chemistry, physics and geology.

Atmospheric chemists study the composition of the atmosphere. In order to do so they use a combination of direct measurement, laboratory work and computer modelling.

Direct measurement can be carried out by using aircraft to collect samples of the atmosphere or using spectroscopic methods to assess the concentration of different compounds. Laboratory experiments allow the kinetics of atmospheric chemical reactions to be studied. Atmospheric and experimental data can then be used to set the parameters for a computer model which will attempt to predict future changes in the composition of the atmosphere and its interaction with electromagnetic radiation.

Atmospheric physicists study these interactions with electromagnetic radiation, and also use statistics in developing more advanced climate models. The National Aeronautics and Space Administration (NASA) in the USA has some of the most advanced climate models in the world.

Biologists also play a part as they explore the natural processes involved in sustaining carbon dioxide levels, and the impact of global warming on different ecosystems. Geologists employ an understanding of volcanic emissions to the atmosphere.

Experts in all these fields, from countries all over the world, contribute to our understanding of climate change.

Climate treaties

The Intergovernmental Panel on Climate Change (IPCC) was established by the United Nations in 1988. Its purpose is to study anthropogenic (man-made) climate change, its likely impacts and possible solutions.

The IPCC is a collaborative endeavour. A volunteer team of scientists, numbering thousands, assesses the peer-reviewed literature on climate and prepares summary reports which are then reviewed by governments. The IPCC then publishes a summary report that sets out its recommendations based on this broad review of literature.

The IPCC report is used to inform action on global warming. The United Nations Framework Convention on Climate Change (UNFCCC) is an international treaty that draws on the conclusions and recommendations of the IPCC report. The most well-known outcome of the UNFCCC was the Kyoto Protocol of 1997 – a binding agreement to control carbon dioxide emission levels. Kyoto was extended at another conference in Doha, Qatar in 2012.

Although over 190 countries signed up to Kyoto, not all the major carbon-emitting states agreed to be bound by it. The most famous was the USA, which withdrew from Kyoto as a protest against what it saw as insufficient demands upon developing economies such as China.

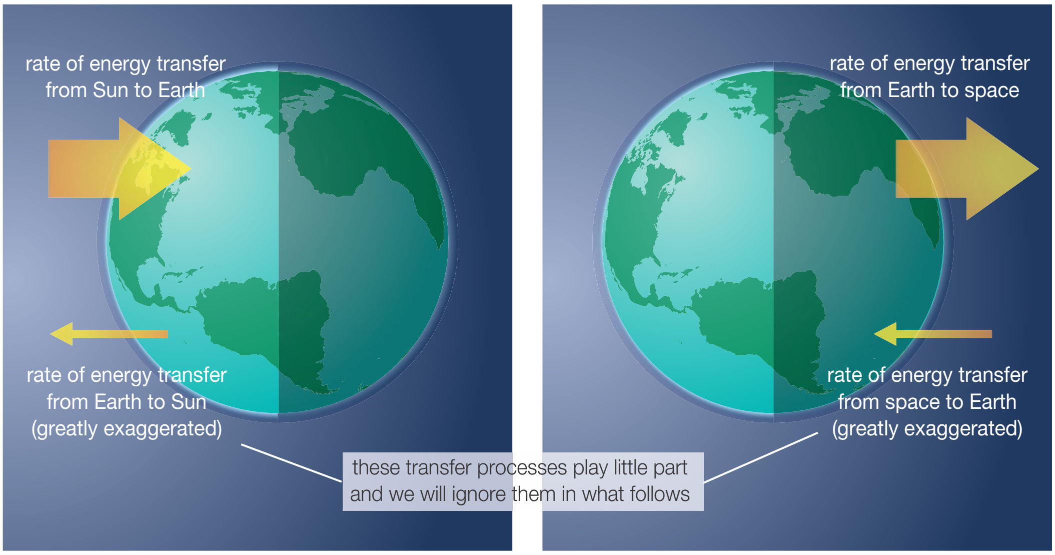
Another major climate treaty is planned for 2020, when the success or failure of Kyoto and Doha will be assessed in detail.

The greenhouse effect

The **greenhouse effect** is often referred to as the cause of the phenomenon of **global warming**, in which the average temperature of the Earth rises, causing various environmental disasters. In fact the greenhouse effect is a necessary mechanism for maintaining the Earth's temperature at a habitable level. However, it is thought that human activity is disrupting the natural equilibrium in the atmosphere, causing warming of the planet.

The average temperature of the troposphere at sea level is 14–15°C, but varies with location (the tropics are hotter than the poles, for instance). This average temperature is maintained because the energy incident on the Earth (coming from the Sun) is balanced by the energy leaving the Earth and escaping into space (Figure 24.38).

Figure 24.38 The greenhouse effect of the Earth's atmosphere


rate of energy transfer from Sun to Earth
rate of energy transfer from Earth to Sun (greatly exaggerated)
rate of energy transfer from Earth to space
rate of energy transfer from space to Earth (greatly exaggerated)
these transfer processes play little part and we will ignore them in what follows

Most of the radiation from the Sun is in the visible region, with some at invisible longer wavelengths close to the visible region (known as near infrared), and a smaller amount at shorter wavelengths close to the visible region (near ultraviolet).

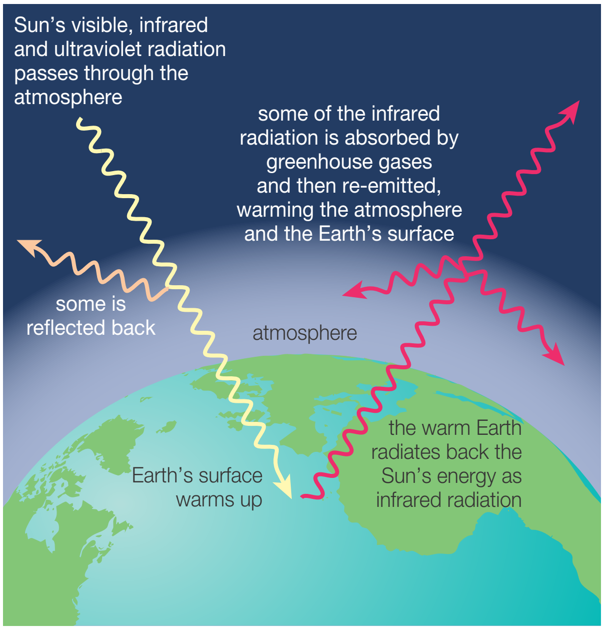
If all this radiation reached the Earth's surface, the average temperature would be intolerably hot. In fact only about 47 per cent of the energy reaches the surface; the remainder is reflected back into space by the atmosphere and clouds, or is absorbed and retained by the atmosphere.

The peak of the incoming radiation has a wavelength in the visible region at around 500 nm. This wavelength is not absorbed to a significant extent by any of the major atmospheric gases (oxygen, nitrogen or argon), or even the less abundant gases (carbon dioxide or water vapour).

This radiation is therefore mostly absorbed by the Earth's surface. This causes the surface temperature to rise. Since energy flows from hotter to colder regions, energy is transferred to the atmosphere, warming it. The atmosphere is hence mostly warmed from below. (This explains why it gets colder as you go higher up in the troposphere.)

The energy that is re-radiated from the surface, and transferred to the atmosphere, is no longer in the visible region. It is in the infrared region, with a much longer wavelength, with a maximum intensity at around 10 000 nm. This change in wavelength occurs because the Earth is radiating at a much lower temperature than the Sun. If all of this energy was allowed to escape into space, the Earth would cool to a temperature around −20°C. However, these wavelengths, while still not absorbed by nitrogen or oxygen, are strongly absorbed by water vapour and carbon dioxide.

The carbon dioxide and water molecules are able to both emit and absorb infrared radiation. They therefore emit (or re-radiate) the energy they have absorbed from the surface. Some of this energy is re-radiated in the direction of space, and the remainder is radiated back down towards the surface (Figure 24.39).


Sun's visible, infrared and ultraviolet radiation passes through the atmosphere
some of the infrared radiation is absorbed by greenhouse gases and then re-emitted, warming the atmosphere and the Earth's surface
some is reflected back
atmosphere
the warm Earth radiates back the Sun's energy as infrared radiation
Earth's surface warms up

■ **Figure 24.39** The different types of radiation involved in the greenhouse effect

The temperature at the surface is therefore maintained by a balance between the downward solar radiation and downward infrared radiation from the atmosphere, and the upward radiation of infrared from the surface.

There are other factors to consider. Some energy is carried by water as it vaporizes. It then rises in the atmosphere by convection. The vapour condenses in the cooler upper troposphere, producing clouds, and releasing the energy at this higher altitude. The energy may then escape the Earth in the form of infrared radiation. The atmosphere is therefore heated not only by radiation but also by convection.

Despite these complications, it can be seen that an increase in the concentration of infrared-absorbing gases, carbon dioxide and water vapour, will decrease the amount of energy escaping from the Earth by radiation (by absorbing the energy on its journey upwards) and will increase the amount of energy moving downwards towards the surface (by re-radiating this infrared). The upwards/downwards equilibrium is therefore disturbed, and the surface temperature will rise until the upwards energy flow again equals the downwards flow.

Global dimming

In Section 24.1 it was noted that emission of solid carbon particles from inefficient fossil fuel combustion or from open fires is a major health issue in developing countries. This so-called 'black carbon' emission contributes to high levels of respiratory disease in many parts of the world.

However, despite its public-health implications, the international community has only recently focused much attention on black carbon emissions, because these particulate emissions are responsible for the phenomenon of 'global dimming'.

The presence of particulates in the atmosphere leads to a small percentage of the incoming sunlight being reflected. As much as a 4 per cent reduction in sunlight intensity may be occurring. This leads to a cooling of the planet, as there is less energy reaching the Earth's surface.

Climate scientists are concerned about global dimming because the cooling effect may have masked the effect of greenhouse gases on global warming. If the measures being taken to reduce carbon emissions also lead to a reduction in black carbon, then the lack of a cooling effect may actually lead to an increase in global temperatures.

Using dimming to mitigate climate change

Some scientists have considered using dimming as a means to mitigate global warming. If we cannot successfully reduce carbon dioxide emissions in time to prevent a climate disaster, perhaps we should aim to cut down the amount of sunlight reaching the Earth. This might be achieved by introducing particulates into the stratosphere.

Current research has focused on introducing sulfur dioxide into the stratosphere, which will react with water, generating an aerosol, or fine mist, of sulfuric acid droplets. These would then reflect solar radiation. The process is thought to be reasonably cheap and fast-acting. Although it sounds drastic, it is actually a process that occurs naturally after volcanic eruptions.

However, computer modelling does suggest that there are risks associated with the process. A major issue is that sulfur dioxide can deplete stratospheric ozone, leading to a damaged ozone layer. Other risks include ocean acidification (as sulfuric acid falls from the atmosphere) and drought. Environmentalists point to the unpredictability of the climate, warning that we should not manipulate the climate without a very clear idea of the extent of the outcomes. In addition, some believe that using dimming to mitigate climate change will divert attention from the need to cut down on carbon dioxide emissions.

The history of climate modelling

Jean Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist who is best known for investigating the Fourier series. The Fourier transform is named in his honour. In 1824 he discovered that the gases in the atmosphere might increase the surface temperature of the Earth, later termed the greenhouse effect (see Figure 24.38). He established the concept of planetary energy balance – that planets obtain energy from a number of sources that cause temperature increase. Fourier recognized that the Earth primarily obtains energy from solar radiation. Later Svante Arrhenius (see Chapter 16) suggested that changes in the levels of carbon dioxide in the atmosphere could alter the surface temperature via the greenhouse effect. He predicted that the emission of carbon dioxide from the burning of fossil fuels would lead to a warmer Earth, but felt that a warmer Earth would be a positive change.

How do the greenhouse gases absorb and emit infrared radiation, and how does this warm the atmosphere?

Absorption and emission in the infrared range of wavelengths occur when molecules vibrate and rotate. Recall that the absorption and emission of ultraviolet radiation occurs when electrons move up and down between fixed energy levels (see Chapter 2). The vibrational and rotational energies of molecules are similarly quantized; that is, a molecule can absorb a photon and move to a higher vibrational energy level. If it falls to a lower energy level, a photon is emitted. These photons have wavelengths in the infrared region.

Absorption and emission of infrared radiation by molecules can only occur if the molecule has an electric dipole. Nitrogen and oxygen, being diatomic molecules, do not have permanent dipoles and create no temporary dipoles when they vibrate. Water absorbs strongly because it is an asymmetric molecule, and also the O–H bonds have a permanent dipole owing to the greater electronegativity of oxygen (see Chapter 3).

Inspection of the carbon dioxide molecule suggests that it has no permanent dipole (see Chapter 4), because the individual C=O bond dipoles cancel out, as the molecule is linear. However, one of the vibrational modes of carbon dioxide is a so-called asymmetrical stretching mode, in which the symmetry of the molecule is disrupted and a temporary dipole formed. Vibrational transitions of carbon dioxide can therefore occur, resulting in emission or absorption in the infrared at a wavenumber of 2360 cm^{-1}. (Wavenumber is related to frequency. For a full discussion see Chapter 11.) Carbon dioxide also has a second vibrational mode corresponding to a bending vibration. This mode absorbs much more weakly at a wavenumber of 670 cm^{-1}. Both vibrational modes can be seen in Figure 24.40.

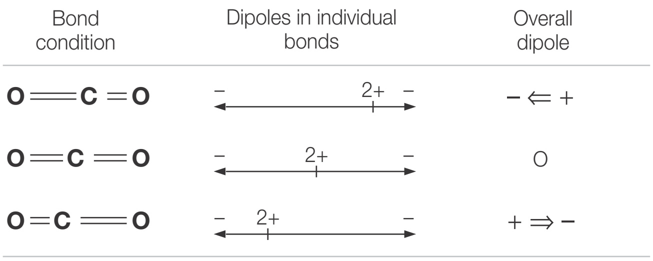


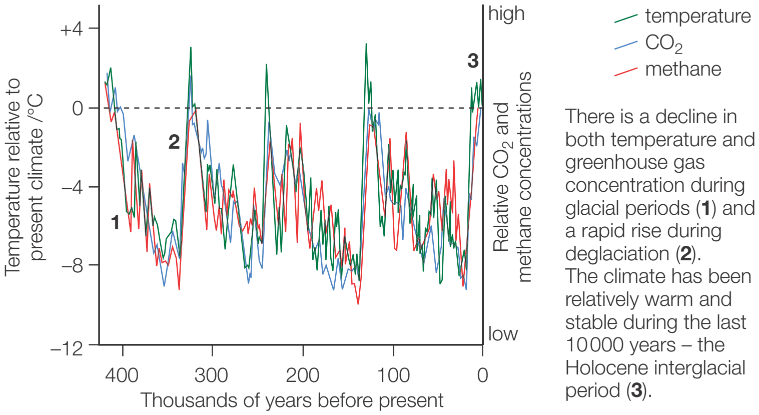
Figure 24.40 A diagram illustrating how the asymmetrical stretch (bending) leads to a change in the dipole moment of carbon dioxide

Both the emission and absorption of infrared are important to the greenhouse action of water and carbon dioxide. Collisions between molecules may ‘excite’ the molecules to higher energy levels. When the molecule ‘relaxes’ to a lower energy level, infrared radiation is emitted. This radiation may move upwards into space, or downwards towards the surface. Conversely, a molecule may absorb infrared radiation. Collisions pass this additional energy to the surrounding gas, warming it.

The evidence for the relationship between the increased concentration of gases and global warming

Evidence for the rise in global temperature came from the Greenland Ice Core Project. Between 1989 and 1992 this group drilled into the polar ice and analysed samples from different depths. The deeper the ice sample, the older the ice. The atmospheric temperature at the time the ice was laid down can be determined by finding the ratio of ^{18}O and ^{16}O atoms in the water molecules of the ice using mass spectrometry (see Chapter 2). Water molecules containing ^{18}O are slightly heavier than those containing ^{16}O, and hence require more energy to vapourize. A lower ratio of $^{18}O/^{16}O$ corresponds to a lower temperature. The project reported that there have been cyclic temperature changes throughout the past 200 000 years, but that the present period of warming has been mostly steady since industrialization in the 19th century (Figure 24.41).

In addition to the temperature studies, measurement of atmospheric carbon dioxide has taken place in isolated areas of Greenland and Hawaii (Figure 24.42), with similar steady increases shown. Comparison of carbon dioxide levels from the 19th century (again, obtained by analysis of ice cores) shows that this increase has continued for 150 years or more.



There is a decline in both temperature and greenhouse gas concentration during glacial periods (**1**) and a rapid rise during deglaciation (**2**). The climate has been relatively warm and stable during the last 10 000 years – the Holocene interglacial period (**3**).

Figure 24.41 The three types of data recovered from the Vostok ice cores over 400 000 years of Earth history

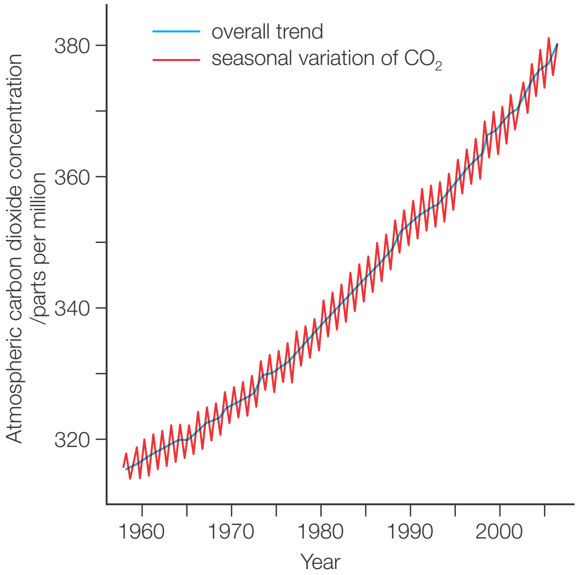


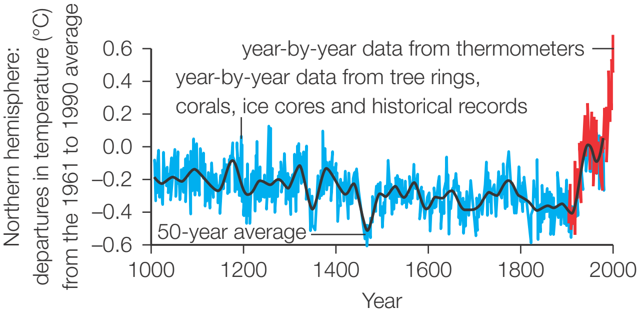
Figure 24.42 Atmospheric carbon dioxide levels recorded at Mauna Loa, Hawaii, 1957–2005

Both carbon dioxide levels and global temperatures have increased. However, has one caused the other? Is there a causal relationship? The IPCC and other scientific organizations believe so.

Nature of Science

Correlation, cause and understanding: the hockey stick controversy

Figure 24.43 shows a graph that first appeared in the IPCC report of 2001. On the graph, temperatures show minimal fluctuation between around 1000 and 1900. This section of the graph is likened to the shaft of a hockey stick. Then, over the past 100 years there is a marked increase in temperature – the blade of the stick. The graph, which summarizes data from many sources including ice cores, tree rings and lake sediments, seems to offer compelling evidence of global warming, outside the normal natural variation. Not only that: the warming coincides with the recent decades of industrialization and the increased carbon emissions that have accompanied it.



■ **Figure 24.43** The 'hockey stick' graph – variations in the Earth's surface temperature

It is widely accepted by scientists that the anthropogenic carbon emissions are the cause of the unprecedented warming.

However, the hockey stick graph is a source of disagreement among scientists, some of whom question the statistical analysis used to estimate the older temperature data. Some scientists and statisticians claim that the hockey stick oversimplifies the message and points to a misleading conclusion. They claim that the true shape of the temperature graph is more nuanced, with a pronounced bulge in the years 1000–1200. They argue that this analysis points to an earlier warm period in medieval times, and so, in theory, the current warming may not be unprecedented. Climate sceptics have seized on this idea, claiming that the current period of warming was matched some 1000 years ago, when there were essentially no anthropogenic carbon emissions to cause it.

However, for scientists keen to alert policymakers and the public to the current climate crisis (as they see it) the 'hockey stick' is convenient 'shorthand' for the temperature trend. Whatever the cause of the medieval warm period, there is sufficient evidence and cause for concern that anthropogenic warming is real and should be acted upon.

The graph shown does illustrate that year-by-year data sometimes reaches large extremes, which are de-emphasized by using the 50-year averages. However, the vast majority of climate scientists do accept that the current period of warming is both warmer and of longer duration than the possible medieval warm period.

ToK Link

Doubt surrounding climate change

The Intergovernmental Panel on Climate Change reported in 2001 'most of the observed increase in globally averaged temperatures since the mid-twentieth century is very likely due to the observed increase in anthropogenic (man-made) greenhouse gas concentrations'.

The national academies of science of all the major industrialized nations, and many others, are explicit in their belief that there is a scientific consensus regarding the existence of global warming, and that it is largely caused by anthropogenic emissions.

Why then is there widespread belief among the public that global warming may be a myth or that the scientific community is divided on the issue?

Some politicians, business leaders, scientists and journalists have claimed that climate scientists have a vested interest in presenting global warming as a forthcoming disaster. The scientists are motivated by a desire to secure funding for their work and have exaggerated the extent of its effects.

However, these same individuals may themselves have vested interests. If global warming is real, and action to mitigate it is urgent, then this will have a detrimental impact on the fossil fuel industry.

Some political parties are quite explicit in their belief that economic growth (involving rampant fossil fuel use) is more important than any environmental concern. However, in the business world, some organizations fund climate research designed to reach conclusions favourable to their interests.

These organizations with a vested interest in continued fossil fuel use, such as oil companies, have been able to exploit the doubt regarding the *extent* of global warming by equating it with doubt regarding the *existence* of global warming. This phenomenon is sometimes referred to as '*manufacturing doubt*'. (Similar techniques were previously used by the tobacco industry, which denied for many years the connection between smoking and chest diseases.) Selective presentation of evidence, such as the idea that the Earth's atmosphere undergoes cyclic variations in temperature, ignores ice core evidence that places current carbon dioxide levels far outside these natural variations.

Scientists do disagree about the likely effects of global warming, as they disagree about the details in other fields, such as evolution by natural selection. However, this is not to say that they reject the existence of anthropogenic climate change.

Another difficulty surrounding the science of global warming is that while it is believed to be occurring, the amount of future warming and the extent of its effects are difficult to predict. Climate modelling involves writing a computer simulation based on certain assumptions and parameters. This is complicated by uncertainty about how changes in one greenhouse gas impact on the concentration of another, for example, and also by unknown feedback effects. For example, the role of clouds in warming is not well understood: clouds add to warming by reflecting radiation back to the Earth and clouds reduce warming by reflecting the Sun's radiation away from the Earth. The uncertain issue here is the net result.

A possible positive feedback effect relates to methane gas locked up in ice in the Siberian tundra. If this ice were to melt, massive amounts of greenhouse-enhancing methane would be released into the atmosphere, accelerating warming further. Similarly, the loss of Arctic ice and glaciers reduces the amount of radiation reflected from the Earth, and increases the area of more absorbing land and sea water, leading to further warming. Some scientists are concerned that a point will be reached at which the greenhouse effect cannot be reversed, and runaway global warming will occur.

The Precautionary Principle (see also Section 24.7) states that if an action has a risk of causing harm to the public or to the environment, then the burden of proof that it is not harmful falls on those taking the action.

The Precautionary Principle is widely accepted and used by scientists. For example, climate scientists claim that there is sufficient consensus that anthropogenic climate change will lead to disaster so that the burden of proof lies with policymakers to prove that continued fossil fuel use is harmless. Fossil fuel proponents dispute this and claim that the burden of proof lies with climate scientists to prove that climate change will definitely lead to harm.

The main greenhouse gases and their contributions to global warming

The contribution of a greenhouse gas to the warming of the atmosphere depends on three factors:

- the abundance of the gas in the atmosphere
- its ability to absorb infrared radiation
- its their lifetime in the atmosphere, before being removed by chemical processes.

The second and third factors are often combined to give a figure called the **global warming potential (GWP)**. Note that this figure must specify the timescale over which it is measured. Some gases are extremely effective at absorbing radiation but are present in the atmosphere for a very short time, minimizing their contribution. Other gases are less effective but are present for many years, so their contribution to global warming is more significant. Carbon dioxide is assigned a GWP of 1 over all time periods. Other gases are assigned a relative value comparing their infrared absorption to that of the same mass of carbon dioxide.

Water vapour, H_2O

Water vapour is the most important greenhouse gas. The GWP of water is sometimes given as 0.1 but is often not calculated because water vapour is constantly cycling through the atmosphere, and its concentration varies according to temperature and location. Typically the percentage of water vapour in the troposphere ranges from 1 per cent to 4 per cent: far greater than any other greenhouse gas. In addition, water absorbs infrared over a broad range of frequencies. Increased atmospheric temperatures lead to more rapid evaporation of the oceans, and a larger capacity of the air to carry water vapour, leading to increased concentration of water in the atmosphere, which may lead to further warming. Estimates of water's contribution to global warming range from 36 per cent to 75 per cent.

Carbon dioxide, CO_2

The percentage of carbon dioxide in the atmosphere is only 0.039 per cent, or about one-thirtieth that of water. However, it is more efficient than water at absorbing infrared radiation (GWP = 1, by definition). Importantly, carbon dioxide absorbs infrared in a 'window' of wavelengths at which water does not absorb. Increases in carbon dioxide concentration therefore upset the equilibrium of absorption and transmission though the atmosphere. Atmospheric concentrations of carbon dioxide are rising for three reasons:

1. Combustion of fossil fuels releases carbon dioxide into the atmosphere.
2. Manufacture of cement and concrete involves the thermal decomposition of calcium carbonate to calcium oxide, releasing carbon dioxide:

$$CaCO_3 \rightarrow CaO + CO_2$$

3 Deforestation in the tropics leads to lower rates of photosynthesis, meaning that carbon dioxide is entering the atmosphere more quickly than it is being removed.

Estimates of carbon dioxide's contribution to global warming range from 9 per cent to 26 per cent. The Earth's oceans are warming, making them less able to absorb (dissolve) carbon dioxide emissions.

Methane, CH_4

The concentration of methane in the atmosphere is around 1.7×10^{-4} per cent. However, its GWP (100 years) is 25, meaning it absorbs infrared more much efficiently than carbon dioxide. In fact, methane's GWP (20 years) is 72, meaning that methane is a very powerful greenhouse gas in the short term but is removed from the atmosphere relatively quickly.

Methane is formed when cellulose (plant fibre) decomposes anaerobically by the action of bacteria. Cellulose is a long-chain carbohydrate made from glucose units (represented below as $(CH_2O)_n$).

$$(CH_2O)_n \rightarrow \frac{n}{2}CH_4 + \frac{n}{2}CO_2$$

This reaction occurs on a large scale in the following situations:

1 rice cultivation (paddy fields)
2 fermentation of grass in the stomachs of ruminants (cows), and methane produced from rotting manure; in the European Union around 10 million tonnes annually are produced this way – it is the largest source of methane
3 leaking gas pipelines
4 fermentation of organic material in covered landfills (waste tips).

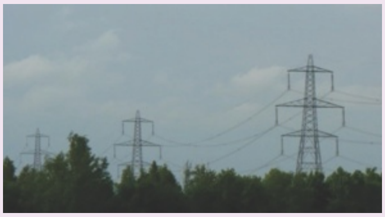
Estimates of methane's contribution to global warming range from 4 per cent to 9 per cent.

Additional Perspective

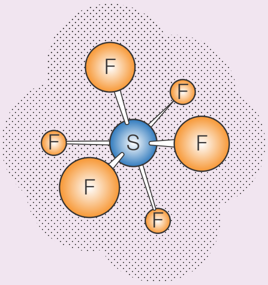
SF6, NOx, CFCs and ozone

Sulfur hexafluoride

Sulfur hexafluoride, SF_6, is a colourless, odourless, non-toxic and flammable gas. It is poorly soluble in water but soluble in non-polar organic solvents. It is prepared by direct synthesis from sulfur and fluorine. Over 8000 tonnes are produced per year. It is used in the electrical industry in circuit breakers on electrical transmission lines (Figure 24.44) and as a contrast agent for ultrasound imaging to improve the visibility of blood vessels. Sulfur hexafluoride is the most potent greenhouse gas with a global warming potential of 22 000 times that of carbon dioxide measured over 100 years. Its ability to strongly absorb infrared radiation is a consequence of the large number of vibrational modes accessible to the molecule due to the cage-like structure in which the sulfur atom is suspended (Figure 24.45). There are no natural sources of sulfur hexafluoride and it is one of six types of greenhouse gases to be regulated by the Kyoto Protocol. The emissions into the atmosphere are small, but it has an atmospheric lifetime of 3200 years.



■ **Figure 24.44** Overhead pylons: potential emitters of SF_6



■ **Figure 24.45** Sulfur hexafluoride molecule, SF_6

Nitrous oxide (nitrogen(I) oxide), N_2O

Nitrous oxide has a GWP (100 years) of 296. It is less efficient at absorbing infrared radiation than carbon dioxide; its high GWP arises from its long residence time in the atmosphere. Its concentration in the atmosphere stands at 0.031 per cent, but this figure is increasing. Despite its low concentration, its high GWP means that it is thought to account for about 5 per cent of global warming effects.

Nitrous oxide is produced naturally by bacteria in the oceans and the soil, and human activity only accounts for about 10–12 per cent of the nitrous oxide

■ **Figure 24.46** A tractor spreading fertilizer

produced each year. However, anthropogenic nitrous oxide is increasing for the following reasons:

- Industrialized agriculture, including the use of nitrogenous fertilizers (Figure 24.46), stimulates soil bacteria to produce more nitrous oxide.
- Industrialized livestock farming and poor handling of animal waste generate nitrous oxide.
- The chemical industry, especially nitric acid and nylon production, generates nitrous oxide.

Chlorofluorocarbons, CFCs

Chlorofluorocarbons (CFCs) are usually encountered in discussions about ozone depletion, but they are also important greenhouse gases. CFC-12, full name dichlorodifluoromethane, CCl_2F_2, has a GWP (100 years) of 8500. This gas was banned in the USA in 1994, but its long residence time in the atmosphere means it will contribute to global warming for centuries to come.

CFCs have been largely replaced as aerosol propellants and refrigerants by **hydrochlorofluorocarbons** (HCFCs) and **hydrofluorocarbons** (HFCs). These gases are less damaging to the ozone layer but still have GWP values much higher than carbon dioxide (for example HFC-23 (trifluoromethane, CHF_3) has a GWP (100 years) of 14 800) and are important contributors to global warming, some sources claiming a 25 per cent contribution.

Ozone, O_3

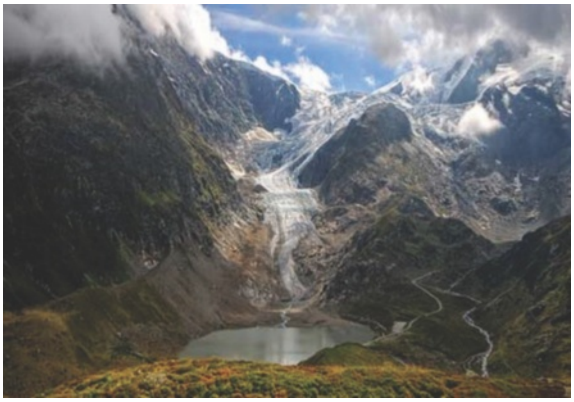
Ozone formed in the troposphere is a constituent of photochemical smog, and it also acts as a natural greenhouse gas. The production of ground level ozone has risen dramatically since the Industrial Revolution. This ozone is formed by the action of sunlight on hydrocarbons and nitrous oxide from the burning of fossil fuels. This ozone rises in the troposphere and acts as a greenhouse gas. However, ozone in the stratosphere is necessary to filter out harmful cosmic rays.

Effects of greenhouse gases

As the Earth's temperature increases there will be potentially catastrophic effects. However, the scientific community is uncertain as to the extent of these effects. Indeed, the IPCC published a report in 2001 in which they attempted to assess the risks, assigning categories ranging from 'risks to unique or threatened systems' to 'risk of irreversible large-scale and abrupt transitions'.

There is general agreement as to the types of effects that might be expected. These are:

- rising sea levels
- glacier retreat
- acidification of the oceans
- changing patterns of agriculture.



■ **Figure 24.47** The Aletsch glacier in Switzerland

Rising sea levels

As the temperature increases sea levels will rise for two reasons. Firstly, the increased temperature causes accelerated melting of polar ice and glaciers (Figure 24.47). This will deposit more water into the oceans. Note that the floating ice at the Arctic will not cause a sea-level rise on melting, as it already displaces water while it floats. However, melting land-based ice will add to the volume of the oceans. Secondly, as the oceans warm up, the water in them will expand, occupying more volume. The IPCC predicts a sea-level rise of up to 59 cm in the 21st century. However, other climate scientists, such as James Hansen, predict a much larger rise, measured in metres. Some low-lying countries, such as the Maldives, are faced with being almost completely submerged (Figure 24.48).

■ **Figure 24.48** The Maldives. What responsibility do the major carbon-emitting nations have towards this small island country, which is faced with being swamped by rising sea levels?

Glacier retreat

Glaciers undergo a seasonal melting and refreezing cycle as temperatures vary through the year. In the Himalayas glacial melt water is an important source of fresh water, feeding the rivers of South Asia. Increased melting increases erosion and risk of flooding downriver, a particular problem in low-lying countries such as Bangladesh. In addition, the retreating glacier will eventually disappear entirely, meaning that countries in South Asia will lose an important source of fresh water.

Changing patterns of agriculture

In temperate regions (e.g. Europe), yields of grain will most likely increase due to the higher temperature, longer growing season and increased concentration of carbon dioxide available for photosynthesis. However, the increased humidity and rainfall that accompanies the temperature rise may lead to increased incidence of fungal crop diseases and migration of tropical insects to higher latitudes. In addition, the increase in yield of useful crops will be accompanied by increased weed growth, leading to greater use of herbicides. Raised sea levels will reduce land availability both directly, as land becomes submerged, and indirectly, as salt concentrations in groundwater increase, damaging the land's ability to support crops.

At higher latitudes, more land may become available as now-frozen land thaws and becomes workable. However, in tropical regions, temperature increases will lead to now-fertile land becoming desert, with much lower crop yields and erosion of the existing soil.

Worldwide, the possibility of extreme weather increases the likelihood of ruined harvests due to soil erosion, flooding or storms.

Ocean acidification

Carbon dioxide dissolved in the upper part of the oceans and carbon dioxide in the atmosphere exist in dynamic equilibrium:

$$CO_2(g) \rightleftharpoons CO_2(aq)$$

As the concentration of carbon dioxide in the atmosphere rises, the rate of absorption of carbon dioxide in the oceans will increase until a new equilibrium is established in which the concentration of dissolved CO_2 is greater.

The oceans therefore act as a **carbon sink** – a means of storing carbon. Much of the additional carbon entering the atmosphere from human activities ends up in the oceans.

You might think that the 'sink' role of the oceans means that we need not be so concerned about adding carbon to the atmosphere. However, there are two reasons that the oceans are not the answer to climate change.

1 The solubility of carbon dioxide decreases as the temperature rises. As the atmosphere warms, the upper oceans will warm also, lowering their ability to absorb and retain carbon dioxide. A negative feedback loop is set up, in which increasing temperatures will harm the oceans' ability to absorb carbon and minimize the problem.
2 Dissolved carbon dioxide leads to the formation of carbonic acid, which lowers the pH of the oceans. This ocean acidification is one of the environmental effects of global warming that scientists are most concerned about.

pH change in the oceans

As carbon dioxide dissolves in the oceans it forms carbonic acid, H_2CO_3:

$$CO_2(aq) + H_2O(l) \rightleftharpoons H_2CO_3(aq) \quad [1]$$

Carbonic acid is a weak diprotic acid, which dissociates to form hydrogen carbonate ions (HCO_3^-):

$$H_2CO_3(aq) + H_2O(l) \rightleftharpoons HCO_3^-(aq) + H_3O^+(aq) \quad [2]$$

The hydrogen carbonate ion then dissociates (to a smaller extent) to form carbonate ions (CO_3^{2-}):

$$HCO_3^-(aq) + H_2O(l) \rightleftharpoons CO_3^{2-}(aq) + H_3O^+(aq) \qquad [3]$$

Both these reactions generate H_3O^+. As the reactions move to the right the solution becomes more acidic and the pH falls.

An increase in the concentration of dissolved carbon dioxide will lead to an increased concentration of carbonic acid in equilibrium [1]. Increasing the concentration of carbonic acid in equilibrium [2] leads to a shift to the right and therefore more hydrogen carbonate and oxonium (H_3O^+) ions. More hydrogen carbonate ions in equilibrium [3] causes a shift towards more oxonium ions again.

In summary, increased carbon dioxide concentrations in the atmosphere will lead to more carbon dioxide in the oceans and an increase in the oceans' acidity.

Increased acidity may kill sea creatures with calcium carbonate shells, an effect which will then move through the entire food chain. In addition, the larvae or eggs of sea creatures may be unable to survive.

A further effect may be the loss of coral. Coral reefs have formed over millions of years as corals (tiny marine organisms similar to anemones) secrete calcium carbonate. The accumulation of calcium carbonate eventually leads to a massive coastal formation that supports thousands of marine plants and animals. As the pH of the ocean falls, the loss of calcium carbonate may occur more quickly than the corals can replace. The result is a substantial decrease in ocean biodiversity.

Experimental work: modelling the equilibrium between $CO_2(g)$ and $CO_2(aq)$

Commercial soda water ('club soda') contains carbon dioxide dissolved under pressure. If the drink is placed in a sealed container, an equilibrium will be established between dissolved carbon dioxide and gaseous carbon dioxide in the 'head space' above the liquid.

$$CO_2(g) \rightleftharpoons CO_2(aq)$$

Some of this dissolved carbon dioxide will react with water, forming carbonic acid:

$$CO_2(aq) + H_2O(l) \rightleftharpoons H_2CO_3(aq)$$

The amount of CO_2 dissolved in the water could therefore be estimated by measuring the pH of the liquid.

Several variables could be investigated:

- *Stirring the solution*: this will increase the rate of diffusion of CO_2 towards the surface of the liquid and will increase the surface area of contact between the liquid and the atmosphere.
- *Raising the temperature of the solution:* this will tend to shift the equilibrium in the endothermic direction. CO_2 dissolving in water is exothermic, so we expect the concentration of dissolved CO_2 to decrease.
- *Adding sodium hydrogencarbonate to the solution:* this increase in hydrogencarbonate ions will influence equilibrium [3] above.
- *Adding dilute hydrochloric acid to the solution:* this will influence equilibria [2] and [3] above.

Approaches towards controlling carbon dioxide emissions

A number of approaches have been proposed to control carbon dioxide emissions in order to limit the progress of global warming.

Increased fuel efficiency

Many countries have introduced fuel efficiency standards in an attempt to reduce the amount of fossil fuels used and decrease carbon dioxide emissions. Government regulations are often based on the average fuel consumption of the various vehicles sold by a particular manufacturer. This encourages the manufacturer to provide a range of vehicles including some very efficient models. However, large powerful cars are still associated with high status, and in many markets larger cars are associated with greater safety, despite statistical evidence to the contrary.

Alternative energy sources (renewables)

Switching to alternative sources of energy such as solar, tidal or wind power is a potential way of lowering carbon dioxide emissions. However, in each case, expensive infrastructure is required, meaning that such sources are unlikely to form more than a small part of the energy mix in the near future. For more information about solar energy and its potential as a major energy source, see Section 24.8.

Carbon-neutral fuels

In theory, biofuels are carbon neutral, meaning that the carbon dioxide released into the atmosphere when they are burnt can be offset by the carbon dioxide absorbed when the plants used to make the fuel were grown.

In practice, such fuels have a large embodied energy, because energy is needed to process the plants and manufacture the fuels. Although biofuels most likely do not achieve carbon-neutral status, blending biofuels with gasoline does offer a way of lowering the net carbon emission per litre of fuel used.

However, another cause of concern for biofuels is that switching to fuel crops might displace food crops and lower biodiversity (see Section 24.4).

Market-based controls

Market-based emission controls seek to offer economic incentives to cut carbon emissions. There are two main models: ‘cap-and-trade’ and carbon taxation.

In **cap-and-trade** schemes, a cap, or overall limit on emissions, is proposed. Nations or companies that emit less than their assigned limit can sell ‘carbon credits’ to other nations or companies that might exceed their limits.

The Kyoto Protocol introduced emissions limits for the signatory nations. These nations could buy carbon credits from other, less polluting nations, or they could also win credits by sponsoring carbon-reduction schemes in other countries.

Cap-and-trade schemes only work if countries agree to be bound by them. The USA famously refused to ratify the Kyoto Protocol in protest at what it saw as lenient treatment of developing nations.

In **carbon taxation** schemes, governments decide on appropriate limits to carbon emissions and will tax emitters (e.g. energy companies, car manufacturers) directly in an attempt to force them to take steps to lower emissions. However, this may affect economic growth in the country, and so it is politically unpopular, especially during economic downturns.

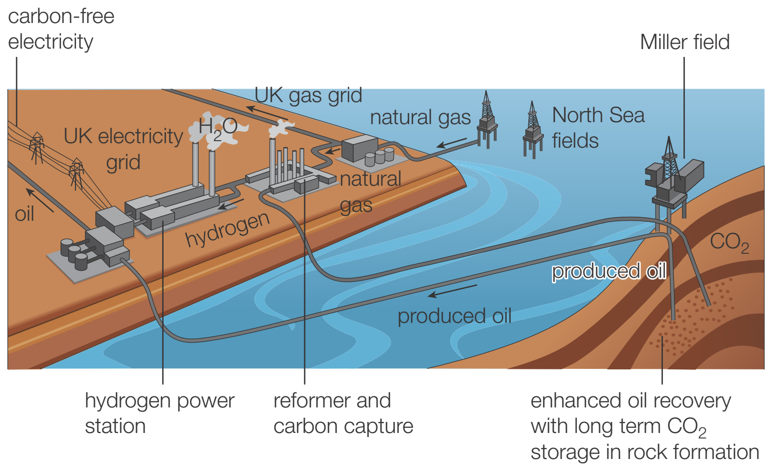


Figure 24.49 Carbon capture and storage

Carbon sequestration (capturing carbon)

One simple way of capturing carbon dioxide is to plant long-lived plants such as trees, which will absorb carbon dioxide as they grow. Eventually these trees will die and the carbon within them will return to the soil and eventually the atmosphere as microorganisms feed on the organic matter and use it for respiration.

Carbon capture and storage (CCS; Figure 24.49) is an approach to reducing global warming based on capturing carbon dioxide from fossil fuel power stations and storing it instead of releasing it into the atmosphere. The carbon dioxide could be pumped into the sea or into underground salt-water deposits and exhausted oil or natural gas fields. However, there are significant energy costs in capturing and compressing carbon dioxide. Carbon dioxide can be chemically removed from the air by the use of sodium hydroxide, but current approaches use zeolites (aluminosilicates) or ion-exchange resins.

24.6 Electrochemistry, rechargeable batteries and fuel cells (AHL) *– chemical energy from redox reactions can be used as a portable source of electrical energy*

Nature of Science

Risk/benefit analyses of batteries and fuel cells

Batteries and fuel cells are sometimes viewed as an essential part of moving us away from fossil fuels towards renewables. Motor vehicles rely on densely stored energy to provide them with adequate range. Fossil fuels have met this need superbly, but with the disadvantage that they produce pollution. If energy can be obtained from solar power, wind power or other renewables, then in theory an effective battery can store this energy, leading to a less polluting vehicle.

However, critics of electric vehicles claim that they are actually more environmentally damaging for three reasons:

1. Batteries contain harmful substances. If the vehicle is in use for 10 years or so, what happens to the batteries after this time?
2. Many of these modern batteries contain rare-earth elements (lanthanoids and actinoids) that must be mined. Mining these rare elements is itself environmentally damaging.
3. Production of the batteries (mining raw materials and manufacturing) uses a lot of energy. This energy must be factored in to the 'embodied' energy of the vehicle.

Car maker Renault released a report in 2011 that claimed that its electric vehicles were less polluting than its conventional petrol vehicles, measured over the whole lifecycle of the car. Although manufacturing the electric car was more polluting, it led to fewer emissions. This was true even if the electricity used to charge the batteries came from coal-fired power stations.

Car makers address point 1 above by undertaking to take the batteries back when the car's useful life is over, so that useful parts can be recycled and toxic substances disposed of appropriately.

■ Voltaic cells

When a voltaic cell is connected into a circuit it provides the energy for charge to flow around the circuit. The external circuit is the connecting wires and other electrical components, for example an electric motor or light bulb. In this, the flow of charge is electrons and this flow is called an electric current.

A voltaic cell (Chapter 9) has the following components:

- a positive electrode (cathode) that receives electrons from the circuit as the cell discharges and does work – a reduction reaction takes place at this electrode while current is provided to the circuit
- a negative electrode (anode) that donates electrons to the circuit as the cell discharges and does work – an oxidation reaction takes place at this electrode during discharge
- a conducting electrolyte which allows ions to migrate between the two electrodes
- sometimes a porous separator (analogous to a salt bridge) is inserted between the two electrodes to prevent them from making physical contact and causing a short circuit.

As the process continues, the active materials (the chemically reactive materials on the surface of the electrodes that participate in the redox reactions) gradually get used up and the reactions slow down until the battery is no longer able to move electrons around the external circuit. The battery is now described as being 'discharged' or 'flat'.

In this chapter, you will study some real-world applications of electrochemistry and explore ways in which electrochemical cells might contribute to new types of transport in the future.

Power, voltage and current

The power P generated by a battery is measured in watts (W) and is calculated as the product of the voltage V and the current I in amps:

$$P = V \times I$$

However, since $V = I \times R$ (Ohm's law), where R represents the resistance:

$$P = I^2 \times R \text{ and } P = \frac{V^2}{R}$$

Larger batteries (of the same type) deliver more power because, although their voltage is unchanged, the maximum current they can deliver increases proportionally.

The total energy E (in joules) generated by a battery is calculated as the product of the power P in watts and time t in seconds:

$$E = P \times t = V \times I \times t$$

A larger battery, containing more chemicals, will be capable of generating power for a longer period of time.

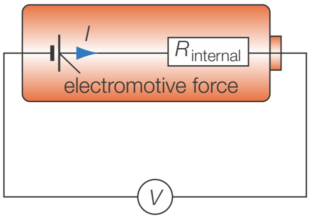
Many commercial batteries express their total energy content using non-SI units, most commonly kilowatt-hours (kWh). A 100 kWh battery is capable of delivering 1 kW of power for 100 hours.

Internal resistance

The potential difference set up by the chemical reaction in the voltaic cell is responsible for the electromotive force (EMF) that drives the current round the circuit.

If a high-resistance voltmeter is connected to the voltaic cell, no current can flow, and the potential difference measured is the same as the theoretical potential difference delivered by the chemical reaction.

However, if the cell is connected to a load (e.g. a light bulb) then a current will flow. The current flow is affected not only by the resistance of the light bulb, but also by the so-called 'internal resistance' of the cell itself. The current within the voltaic cell consists of ions diffusing to and from the electrodes, and this diffusion takes time. The maximum current delivered by the cell is limited by the rate of this diffusion.



■ **Figure 24.50** Internal resistance

In Figure 24.50 the circuit symbol for the cell (which can be thought of as an 'ideal' or theoretical cell capable of delivering infinite current) is accompanied by a resistor symbol, representing the internal resistance of the cell.

The internal resistance of the cell is affected by the temperature – at higher temperatures the internal resistance falls because the ion mobility increases. As the cell runs down, the build-up of chemical products around the electrodes causes an increase in internal resistance.

Anodes and cathodes

Throughout the following sections bear in mind the formal definitions of the terms anode and cathode.

- **Anode**: the site of **oxidation**. Electrons are lost, so they flow away from the anode.
- **Cathode**: the site of **reduction**. Electrons are gained, so they flow towards the cathode.

■ Cells and batteries

In Chapter 9 you studied the Daniell cell, which uses copper and zinc electrodes to generate an EMF of approximately 1.1 V. Most applications of cells require larger voltages than this, so a number of cells are connected together in series. An array of voltaic cells connected in series is called a **battery**. The word battery has come to mean any portable device which offers a self-contained source of electricity.

There are three main types of cells or batteries: primary cells, secondary cells, and fuel cells.

Primary cells

Disposable batteries used for household items such as torches, clocks and toys are primary cells. A primary cell contains a chemical mixture that uses an electrochemical reaction to generate an EMF when connected in a circuit When the chemicals have 'run out' the battery is exhausted and is thrown away. Such batteries are referred to as **dry cells** because they use an almost dry paste, rather than a liquid, as the electrolyte. This lowers the chance of toxic chemicals leaking from the battery.

Primary batteries cannot be efficiently recharged because the redox reactions that occur involve the formation of solids or gases. Sometimes these substances tend to build up around an electrode (and increase the internal resistance and as a consequence the voltage drops). However, the battery may 'recover' partially as these substances slowly diffuse away.

The alkaline dry cell (e.g. a Duracell or Energizer battery) contains an anode of powdered zinc and a cathode of manganese(IV) oxide. Each of these is mixed to a paste in an electrolyte of potassium hydroxide.

The electrode half-equations are:

Anode: $Zn(s) + 2OH^-(aq) \rightarrow Zn(OH)_2(s) + 2e^-$

Cathode: $2MnO_2(s) + 4H_2O(l) + 2e^- \rightarrow 2Mn(OH)_2(s) + 2OH^-(aq)$

Overall reaction: $Zn(s) + 2MnO_2(s) + 4H_2O(l) \rightarrow Zn(OH)_2(s) + 2Mn(OH)_2(s)$

Under the conditions of the cell, this generates an EMF of 1.5 V.

Alkaline dry cells are cheap and convenient, but they are not useful for large-scale applications such as motor vehicles, as they have a low voltage and cannot be re-used.

Secondary cells

The workings of rechargeable cells

Secondary cells are more commonly known as rechargeable batteries. They differ from primary cells in that once they are exhausted, a current can be passed through them which reverses the chemical reaction and returns the chemical mixture to its initial state. Secondary cells can be used many times and can be thought of as a way of storing electricity, rather than simply generating it.

Three common types of secondary cells are the **lead-acid cell**, the **nickel–cadmium cell** and the **lithium-ion cell**. These have different strengths and weaknesses and find different applications.

Lead-acid cells

■ **Figure 24.51** A car battery

Lead-acid cells are very widely used. They are commonly known as car batteries.

The car battery consists of six cells (shown by the six caps on the example in Figure 24.51). Each cell generates an EMF of approximately 2 V, leading to an overall voltage of 12 V for the battery.

The car battery must be able to generate a very high current for a short period of time, as it is used to drive the starter motor. The starter motor spins the engine to enable the internal combustion reaction to begin. It also generates the 'spark' for ignition in a gasoline engine. The car battery is recharged by a generator (called an alternator) attached to the car engine once the car is running.

Each of the six cells consists of an anode of lead–antimony alloy and a lead(IV) oxide cathode. The electrolyte is $6\,mol\,dm^{-3}$ sulfuric acid (a solution of 38% sulfuric acid by mass).

When the electrochemical reaction is allowed to proceed, generating a current, this is called the 'discharge' reaction, as it discharges or 'runs down' the battery.

The electrode half-equations for the discharge reaction are:

Anode: $Pb(s) + SO_4^{2-}(aq) \rightarrow PbSO_4(s) + 2e^-$

Cathode: $PbO_2(s) + 4H^+(aq) + SO_4^{2-}(aq) + 2e^- \rightarrow PbSO_4(s) + 2H_2O(l)$

Overall reaction: $Pb(s) + PbO_2(s) + 4H^+(aq) + 2SO_4^{2-}(aq) \rightarrow 2PbSO_4(s) + 2H_2O(l)$

In this scheme the lead anode is oxidized to lead(II) sulfate, and the lead(IV) oxide cathode is reduced to lead(II) sulfate. Both electrodes are coated with lead(II) sulfate.

However, when a current is passed through the cell, these reactions are reversed. (Imagine the electrochemical reaction running 'downhill' as the cell discharges. When the cell is recharged, the external power source drives it back 'uphill'.)

The half-equations for the 'recharge' reaction are the reverse of the above:

Anode: $PbSO_4(s) + 2e^- \rightarrow Pb(s) + SO_4^{2-}(aq)$

Cathode: $PbSO_4(s) + 2H_2O(l) \rightarrow PbO_2(s) + 4H^+(aq) + SO_4^{2-}(aq) + 2e^-$

Overall reaction: $2PbSO_4(s) + 2H_2O(l) \rightarrow Pb(s) + PbO_2(s) + 4H^+(aq) + 2SO_4^{2-}(aq)$

The charging and discharging reactions are summarized in Figure 24.52.

A side reaction that can limit the life of the battery is that the water in which the sulfuric acid is dissolved gets electrolysed to hydrogen and oxygen in the battery. These gases diffuse away, and the fluid level in the battery gradually drops. The battery can be topped up with distilled water by opening each cell in turn via the screw caps.

Some batteries are marketed as 'maintenance-free' batteries without the screw caps. They use anodes of lead–calcium alloy (replacing the antimony) which slows down the electrolysis of water.

Key questions about the lead-acid cell

- **Why is the lead-acid cell able to deliver such a high current?**

  All batteries experience a phenomenon known as 'internal resistance'. The amount of current that can be delivered is limited by the battery's ability to transfer electrons. In most cases, the internal resistance arises from the limited speed at which ions can diffuse through the electrolyte. However, the lead-acid cell benefits from a liquid electrolyte (rather than a paste, like many cells), and the very high conductivity of the sulfuric acid electrolyte.

- **Why is the lead-acid cell able to be recharged?**

  Lead(II) sulfate, which is the product of the discharge reaction at both anode and cathode, is a very insoluble substance. It adheres to the anode and cathode plates. This means that it remains close to the two electrodes, and is able to gain or lose electrons during the recharging process.

- **Why don't car batteries have an infinite lifespan?**

  As well as the possibility of losing water by electrolysis, if the battery is left in its 'run down' state the lead(II) sulfate continues to crystallize on the electrodes. This lowers the surface area of the electrodes, and then the crystals do not fully dissolve during the recharging process.

- **Why does each cell in the lead-acid battery have a voltage of approximately 2V?**

  The reduction and oxidation reactions that occur inside a simple battery each produce a fixed standard electrode potential E° measured as a voltage. The sum of the two electrode potentials is the voltage of the battery.

  For example, the discharge reaction and associated standard electrode potential at the positive electrode of a lead–acid battery is:

  $Pb(s) + SO_4^{2-}(aq) \rightarrow PbSO_4(s) + 2e^- \qquad E = +1.685\,V$

  The reaction at the negative electrode and associated standard electrode potential is:

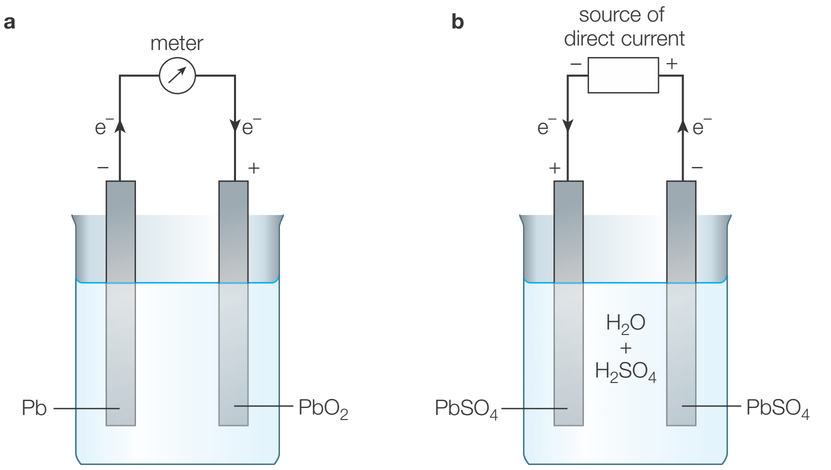
  $PbO_2(s) + 4H^+(aq) + SO_4^{2-}(aq) + 2e^- \rightarrow PbSO_4(s) + 2H_2O(l) \qquad E = +0.356\,V$

  Adding the two half-equations and cancelling electrons gives the overall cell reaction for the discharge process:

  $$Pb(s) + PbO_2(s) + 4H^+(aq) + 2SO_4^{2-}(aq) \rightarrow 2PbSO_4(s) + 2H_2O(l)$$
  $$E_{cell} = +1.685\,V + (+0.356\,V) = 2.041\,V$$

  Note that the concentration of the sulfuric acid electrolyte can also affect the voltage of a lead-acid battery – as its concentration decreases, so does the voltage.

■ **Figure 24.52** A lead-acid cell **a** discharging **b** recharging



Experimental work

In a school laboratory, it is possible to make a lead-acid battery and investigate the effect of a number of factors on its ability to hold its charge. Two lead plates are immersed in sulfuric acid. These plates will have a layer of lead(II) oxide on their surfaces. The battery can therefore be formed by connecting it to a power supply that causes the lead(II) oxide on the surface of one electrode to be reduced and the lead(II) oxide on the surface of the other electrode to be oxidized.

The relevant half-equations describing this forming process are:

$$PbO(s) + 2H^+(aq) + 2e^- \rightarrow Pb(s) + H_2O(l)$$

$$PbO(s) + 2H_2O(l) \rightarrow PbO_2(s) + 2H^+(aq) + 2e^-$$

Possible independent variables for investigation

Variables that may affect the ability of the lead-acid storage cell to hold a charge include:

1 surface area of the lead electrodes
2 acid temperature
3 sulfuric acid concentration
4 charging time
5 voltage used during charging
6 charge reversal during charging.

Charge reversal

If the power pack connections are reversed occasionally during the life of the battery it is thought that the surface of the lead electrodes will be roughened, which will increase their surface area. This might increase the useful life of the battery and enable it to retain its charge for longer.

Measurement of dependent variable

After charging, the battery is connected to a bulb. The time that this bulb remains lit is our measure of the ability of the battery to hold its charge. A series of different values of the independent variable can be tried, for example a series of different-sized electrodes (surface area). The time for the bulb to remain lit is measured and compared to the independent variable.

Nickel–cadmium cells

For most applications, the typical 1.5 V alkaline dry cell can be replaced with a similar-looking 1.25 V nickel–cadmium rechargeable cell.

Nickel–cadmium (or Ni–Cad) cells consist of a cadmium anode and a nickel(III) hydroxide cathode. As with the alkaline dry cell, the electrolyte is potassium hydroxide.

The half-equations for the discharge reaction are:

Anode: $Cd(s) + 2OH^-(aq) \rightarrow Cd(OH)_2(s) + 2e^-$

Cathode: $2Ni(OH)_3(s) + 2e^- \rightarrow 2Ni(OH)_2(s) + 2OH^-(aq)$

Overall reaction: $Cd(s) + 2Ni(OH)_3(s) \rightarrow Cd(OH)_2(s) + 2Ni(OH)_2(s)$

Ni–Cad batteries suffer from what is known as the 'memory effect'. This effect occurs when a battery is partially discharged and then recharged. The Ni–Cad battery 'remembers' the lower state and will not fully charge. Ni–Cad batteries should be completely discharged prior to being fully charged. This phenomenon occurs due to the formation of a passive (very unreactive) surface on the electrodes. This increases the internal resistance of the cell and forms a barrier to further cell reactions during the charging process.

Lithium-ion batteries

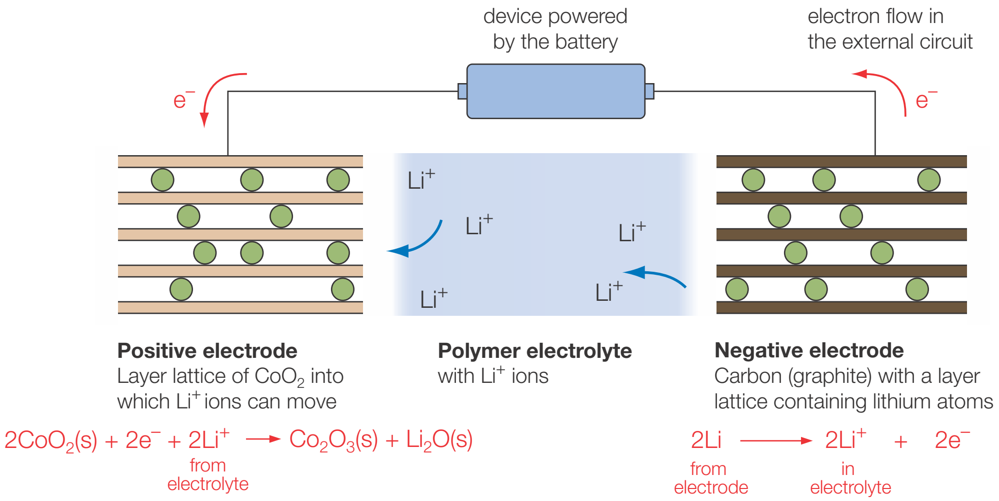
Lithium can be used to make primary cells. It is an attractive option for battery manufacture as it has a very high (very negative) reduction potential, meaning that it is a powerful reducing agent. However, lithium primary cells have certain disadvantages. The main problem arises from the high reactivity of lithium and the generation of potentially explosive hydrogen during the cell reactions. Lithium cells are therefore dangerous if punctured.

However, lithium-ion batteries do not contain any metallic lithium. Instead, they contain lithium ions which migrate between the two electrodes during charge or discharge. These lithium-ion batteries should be viewed as a separate category of battery. They are rechargeable and were introduced for use in mobile phones, video cameras and laptop computers. Lithium-ion batteries are secondary batteries, whereas lithium batteries are primary cells.

Reaction at the anode

The anode is made of graphite, into which lithium atoms are **intercalated** (Figure 24.53). This means that lithium atoms are inserted into the spaces between the graphite layers. During the cell reaction, lithium ions migrate out of the graphite electrode via the electrolyte, while their electrons enter the external circuit.

$$2\text{Li (graphite)} \rightarrow 2\text{Li}^+ \text{ (electrolyte)} + 2e^-$$


device powered by the battery
electron flow in the external circuit
e^-
e^-
Li^+
Li^+
Li^+
Li^+
Li^+
Positive electrode
Layer lattice of CoO_2 into which Li^+ ions can move
Polymer electrolyte
with Li^+ ions
Negative electrode
Carbon (graphite) with a layer lattice containing lithium atoms
$2CoO_2(s) + 2e^- + 2Li^+ \rightarrow Co_2O_3(s) + Li_2O(s)$
from electrolyte
$2Li \rightarrow 2Li^+ + 2e^-$
from electrode
in electrolyte

■ **Figure 24.53** Schematic diagram of a lithium-ion polymer battery showing the battery discharging. The electrode processes are reversible so that the battery can be recharged

Reaction at the cathode

The cathode is commonly made of cobalt(IV) oxide, CoO_2, which has a layer structure into which the lithium ions can move from the electrolyte. Electrons from the external circuit enter this lattice and reduce cobalt(IV) to cobalt(III).

$$2CoO_2(s) + 2e^- + 2Li^+(\text{electrolyte}) \rightarrow Co_2O_3(s) + Li_2O(s)$$

During recharging, the lithium ions are driven back though the polymer electrolyte to the graphite anode.

Lithium-ion batteries are much lighter than lead-acid batteries for a given capacity, and they are being used in the latest generation of electric vehicles (Figure 24.54) and new aircraft such as the Boeing 787 Dreamliner.

Lithium-ion batteries must be manufactured to very high standards. If the thin polymer film between the electrodes is damaged the battery can short circuit, leading to overheating and 'thermal runaway', in which the heat produced by the chemical reaction leads to further battery damage and more heat being generated. Lithium-ion batteries can catch fire if not carefully treated, which creates a serious hazard if used in cars or aircraft (Figure 24.55).

■ **Figure 24.54** The Tesla Model S claims a 300 mile range from its 85kWh lithium-ion battery pack

■ **Figure 24.55** The Boeing 787 has experienced a number of fires caused by thermal runaway of its on-board lithium-ion batteries

The advantages of different types of secondary cells

Table 24.11 summarizes the advantages and disadvantages of different types of secondary cells.

■ **Table 24.11** Advantages and disadvantages of different types of secondary cells

	Advantages	Disadvantages
Lead-acid battery	Very well-established technology Ability to deliver high current essential for vehicle applications (starter motor, ignition system) Battery can be recycled and lead re-used	Low power-to-weight ratio (i.e. they are heavy). They are suitable for small, slow electric vehicles such as golf carts but not higher-performance vehicles Not suitable for fast recharging, because it leads to electrolysis of water or break-up of the lead(II) sulfate layer, which shortens battery life Possibility of damage if left 'flat' for too long, owing to crystallization of lead(II) sulfate
Nickel–cadmium battery	High number of charge–discharge cycles possible (thousands) Long storage time – can retain a charge 'on the shelf' for months	Cadmium is highly toxic so batteries must be disposed of carefully Memory effect – if the battery is not discharged fully it 'remembers' the lower-charge state and will not fully charge on subsequent cycles
Lithium-ion battery	High energy density – batteries are light and powerful Minimal memory effect – batteries are more tolerant of incomplete charge–discharge cycles	Manufacturing flaws or battery damage can lead to thermal runaway and the battery catching fire

Fuel cells

A fuel cell converts the chemical energy of a fuel directly into electrical energy. Fuel cells are designed so that the substances to be oxidized and reduced at the electrodes are stored outside the cell and are continually supplied to the electrodes. A fuel cell is therefore a flow battery that continues to operate so long as reactants are introduced.

Explanation of the workings of fuel cells including diagrams and deduction of the relevant half-equations

Hydrogen–oxygen fuel cells

One of the most successful types of fuel cells uses the reaction between hydrogen gas and oxygen gas to form water, and is known as the hydrogen–oxygen fuel cell. A cross-section of a hydrogen–oxygen fuel cell is shown in Figure 24.56. The cell consists of porous carbon electrodes which are impregnated with catalyst (platinum, silver or cobalt(II) oxide). There are two types of hydrogen–oxygen fuel cells – alkaline cells and acidic cells.

- **Alkaline fuel cells:** hydrogen and oxygen gases are bubbled through the electrodes into an electrolyte of concentrated aqueous sodium hydroxide. The fuel cell runs continuously so long as the two gases are supplied at a relatively high temperature and pressure. The electrode reactions are:

  Anode: $H_2(g) + 2OH^-(aq) \rightarrow 2H_2O(l) + 2e^-$ (oxidation)
  Cathode: $O_2(g) + 2H_2O(l) + 4e^- \rightarrow 4OH^-(aq)$ (reduction)

  Multiplying the first half-equation through by 2, and then adding it to the second gives the overall (redox) reaction:

  $2H_2(g) + O_2(g) \rightarrow 2H_2O(l)$

  Note that hydroxide ions (OH^-) are consumed in the oxidation reaction but produced in the reduction reaction, so the pH of the electrolyte remains constant.

- **Acidic fuel cells:** in some fuel cells, an acidic electrolyte is used – the electrode reactions are:

  Anode: $H_2(g) \rightarrow 2H^+(aq) + 2e^-$ (oxidation)
  Cathode: $O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$ (reduction)
  Overall reaction: $2H_2(g) + O_2(g) \rightarrow 2H_2O(l)$ (redox)

  This time, hydrogen ions are released in oxidation but consumed in reduction.

The overall reaction is the same in both types of cell. In these fuel cell reactions the energy released when hydrogen and oxygen combine is transferred as electrical energy, rather than the thermal energy that is transferred when hydrogen burns in oxygen.

Proton-exchange membrane fuel cells

Proton-exchange membrane (PEM) fuel cells are currently in development. These replace the platinum electrodes with a very thin polymer membrane (i.e. a plastic film) into which much smaller amounts of platinum or other catalysts are embedded. The membrane needs to allow protons to flow through it with minimal resistance, while preventing electrons from passing through. It must also be impermeable to the gases entering the cell.

Methanol fuel cells

One disadvantage of hydrogen–oxygen fuel cells is that a supply of hydrogen must be available. Hydrogen is a flammable gas that must be stored under pressure.

An alternative to hydrogen is methanol, which is a liquid at room temperature and pressure and so is far more easily stored and transported. Methanol can be stored in sealed cartridges which have an energy density far higher than hydrogen or even lithium-ion batteries.

Methanol fuel cells rely on the catalytic oxidation of methanol to form carbon dioxide. The electrode reactions are:

Anode: $CH_3OH + H_2O \rightarrow 6H^+ + 6e^- + CO_2$ (oxidation)

Cathode: $\frac{3}{2}O_2 + 6H^+ + 6e^- \rightarrow 3H_2O$ (reduction)

Overall reaction: $CH_3OH + \frac{3}{2}O_2 \rightarrow CO_2 + 3H_2O$ (redox)

Once again, methanol fuel cells rely on platinum as a catalyst at both electrodes, which contributes to the high cost of these cells.

Although convenient, methanol fuel cells produce carbon dioxide, which is a greenhouse gas, although in small quantities. There may be a risk associated with carrying a laptop computer, for example, powered by a methanol cartridge, as the fuel is flammable and the reaction produces gas that must be allowed to diffuse away.

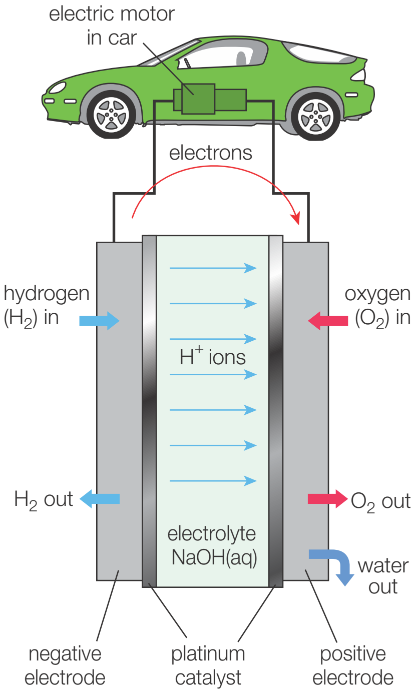
The use of methanol fuel cells to power military equipment such as satellite communications units and computers is being explored.

The design of the fuel cell

The traditional fuel cell shown in Figure 24.56 shows the hydrogen passing over an anode made of finely divided platinum, which splits the hydrogen into H^+ ions (protons) and electrons. The protons move into the electrolyte (in this case, sodium hydroxide, so a neutralization reaction occurs, forming water). The electrons enter the external circuit. Excess hydrogen gas passes out of the cell and can be re-used.

The platinum cathode similarly catalyses the reaction of oxygen gas and water, forming hydroxide ions. Platinum is a much less effective catalyst for this reaction, which means that superior alternatives are sought.

Fuel cells are very useful as power sources in remote locations – such as spacecraft, remote weather stations, large parks and rural locations – and in certain military applications. Potential household uses include smart phones (with high power consumption due to large displays) and notebook computers where alternating current (AC) charging may not be available for weeks at a time. Fuel cells are used in electric (Figure 24.57) and hybrid vehicles.



■ **Figure 24.56** A hydrogen–oxygen fuel cell providing power to drive a car

■ **Figure 24.57** A Mercedes–Benz Ecobus uses a hydrogen-based fuel cell

Microbial fuel cells

A microbial fuel cell typically uses waste water containing organic compounds, which are acted upon by bacteria to generate electricity. Instead of using inorganic catalysts such as platinum to enable the redox reactions, the cell taps into the electron transfer chains within the microorganism, using the microbe's biochemical pathways to catalyse the reaction.

Design of a microbial fuel cell

Like a hydrogen–oxygen fuel cell, the microbial fuel cell consists of two compartments separated by a membrane.

At the anode, bacteria consume the organic substrate, generating carbon dioxide, hydrogen ions (protons) and electrons. The electrons enter the external circuit while the protons pass through the membrane (Figure 24.58).

At the cathode, the protons combine with electrons and oxygen, forming water.

One particular reaction that shows promise is the use of *Geobacter sulfurreducens* to oxidize ethanoate ions.

Possible electrode reactions are:

Anode: $CH_3COO^- + H_2O \rightarrow 2CO_2 + 5H^+ + 6e^-$ (oxidation)

Cathode: $\frac{3}{2}O_2 + 6H^+ + 6e^- \rightarrow 3H_2O$ (reduction)

Overall reaction: $CH_3COO^- + H^+ + \frac{3}{2}O_2 \rightarrow 2CO_2 + 2H_2O$

Microbial fuel cells offer a number of advantages: they produce useful energy from a waste product, they help to clean up waste water by removing organic pollutants, and they decrease the demand for expensive catalysts such as platinum.

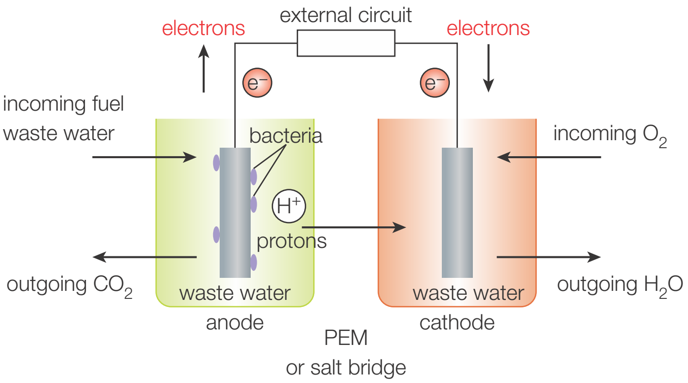


Figure 24.58 Microbial fuel cell

Calculating thermodynamic efficiency (ΔG/ΔH) of a fuel cell

The thermodynamic efficiency of a fuel cell is given by: efficiency $= \frac{\Delta G}{\Delta H} \times 100$ where ΔG is the Gibbs free energy change of the cell reaction and Δ*H* is the enthalpy change of the reaction.

Worked example

Calculate the thermodynamic efficiency of the methanol fuel cell reaction. Use the data in Table 24.12.

$CH_3OH(l) + \frac{3}{2}O_2(g) \rightarrow CO_2(g) + 3H_2O(l)$

Substance	Enthalpy of formation/ ΔH_f, kJ mol^{-1}	Gibbs free energy of formation/ ΔG_f, kJ mol^{-1}
$CH_3OH(l)$	−239	−166
$O_2(g)$	0	0
$CO_2(g)$	−394	−394
$H_2O(l)$	−286	−237

Table 24.12 Enthalpy and Gibbs free energy data

Standard Gibbs free energy change is found by:

$\Delta G^\ominus = \sum \Delta G^\ominus_f(\text{products}) - \sum \Delta G^\ominus_f(\text{reactants}) = [(-394) + (-237 \times 3)] - [-166] = -939\,\text{kJ mol}^{-1}$

Standard enthalpy change of the cell reaction is found by:

$\Delta H^\ominus = \sum \Delta H^\ominus_f(\text{products}) - \sum \Delta H^\ominus_f(\text{reactants}) = [(-394) + (-286 \times 3)] - [-239] = -1013\,\text{kJ mol}^{-1}$

Efficiency $= \frac{\Delta G}{\Delta H} \times 100 = \frac{-939}{-1013} \times 100 = 92.7\%$

What is the significance of the thermodynamic efficiency of the fuel cell?

In the efficiency equation, the standard Gibbs free energy change tells us how much electrical energy can be theoretically generated by this chemical reaction, per mole of methanol.

The standard enthalpy change tells us how much heat can be released when methanol is burnt, per mole.

Why are the two values not the same? The second law of thermodynamics essentially tells us that 'you cannot break even'; in other words, some proportion of the energy released in a chemical process will be released as heat (wasted, for our purposes). You can think of this in terms of entropy (Chapter 15). The methanol combustion equation shows a decrease in the moles of gas, so the degree of disorder in the system falls. The reaction is exothermic though, so it heats up the surroundings, causing an increase in entropy. The Gibbs free energy is the proportion of the energy change that can be usefully employed to do work (in our case, the electrical energy generated by the cell) while still releasing sufficient heat to the surroundings that they warm up enough to offset the decrease in the entropy of the system.

In this example, 92.7 per cent of the energy released by cell reaction can be used to do useful work. The remainder is released as heat.

10 Calculate the thermodynamic efficiency of the hydrogen fuel cell. Use the data in Table 24.13.

Substance	Enthalpy of formation/ ΔH_f, kJ mol^{-1}	Gibbs free energy of formation/ΔG_f, kJ mol^{-1}
$H_2(g)$	0	0
$O_2(g)$	0	0
$H_2O(l)$	–286	–237

Table 24.13 Enthalpy and Gibbs free energy data

Comparison between primary cells, fuel cells and rechargeable batteries

Table 24.14 offers a comparison between primary cells, fuel cells and rechargeable batteries.

Table 24.14 Advantages of primary cells, fuel cells and rechargeable batteries

	Advantages	Disadvantages
Primary cells	Cheap and widely available Small and highly portable Low toxicity compared to most secondary cells Long shelf life	Wasteful – they are used once and thrown away
Fuel cells	Do not require recharging – they can operate continuously as long as new fuel is supplied Fuel cells are well suited to remote locations	Expensive catalysts required (e.g. platinum in methanol fuel cells) Some fuel cells produce carbon dioxide Cannot be used in enclosed spaces due to fire risk Hydrogen fuel cells require a source of hydrogen stored under pressure
Rechargeable batteries (secondary cells)	Highly convenient for use in portable devices Potentially long lifetime with many charge cycles	Often contain toxic substances that can enter the environment if thrown away

Energy from batteries and fuel cells

We have already studied how batteries and fuel cells enable an EMF to be generated from redox reactions. But what determines the value of that EMF? In Chapter 9 you studied electrochemical cells operating under standard conditions, in which metals are suspended in solutions of their own ions, at concentrations of 1 mol dm^{-3}at 298 K.

As we have seen, commercial batteries do not use standard electrolytes or concentrations, so how can EMF be determined?

As a starting point to a discussion of non-standard electrode systems, we will consider the effect of changing electrolyte concentration.

Electrode potentials can only be used to predict the feasibility of a redox reaction under standard conditions. Electrode potentials for oxidizing agents in acidic conditions refer to $1.0\,\text{mol}\,\text{dm}^{-3}$ concentrations of hydrogen ions, H^+ (pH 0), as in the standard hydrogen electrode.

Consider the following reaction:

$$MnO_2(s) + 4H^+(aq) + 2Cl^-(aq) \rightleftharpoons Mn^{2+}(aq) + 2H_2O(l) + Cl_2(g)$$

This reaction can be thought of as the combination of these two half-reactions:

$$MnO_2(s) + 4H^+(aq) + 2e^- \rightleftharpoons Mn^{2+}(aq) + 2H_2O(l) \quad E^\ominus = +1.23\,V$$

$$Cl_2(g) + 2e^- \rightleftharpoons 2Cl^-(aq) \quad E^\ominus = +1.36\,V$$

The reaction as written reverses the second reaction (i.e. chloride ions are oxidized to chlorine gas molecules) so the electrode potential sign is reversed.

$$E^\ominus_{cell} = 1.23 + (-1.36) = -0.13\,V$$

Since the cell potential, $E^\ominus_{cell}$ is negative, this reaction is not spontaneous under standard conditions. However, when *concentrated* hydrochloric acid is heated with manganese(IV) oxide the cell potential becomes positive and the reaction can occur: chlorine gas is evolved.

This happens because the pH conditions are no longer standard ($1\,\text{mol}\,\text{dm}^{-3}$) and neither is the temperature (298 K, 25 °C).

In addition, the loss of chlorine gas shifts the equilibrium to the right, favouring the formation of more chlorine.

In general, for a redox equilibrium:

$$Ox + ne^- \rightleftharpoons Red$$

Increasing the concentration of the oxidized species, [Ox], or decreasing the concentration of the reduced species, [Red], will shift the equilibrium to the right, reducing the number of electrons transferred and hence making the cell potential more positive. Similarly, the cell potential will become more negative if the concentration of the oxidized species, [Ox], is decreased or the concentration of the reduced species, [Red], increased. These shifts can be predicted from application of Le Châtelier's principle.

Consider the Daniell cell, consisting of a zinc half-cell connected to a copper half-cell:

$$Zn^{2+}(aq) + 2e^- \rightleftharpoons Zn(s) \quad E^\ominus = -0.76\,V$$

$$Cu^{2+}(aq) + 2e^- \rightleftharpoons Cu(s) \quad E^\ominus = +0.34\,V$$

Under standard conditions the cell reaction is:

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$$

and

$$E^\ominus_{cell} = +0.34 - (-0.76) = +1.10\,V$$

Now imagine that we decrease the concentration of Zn^{2+} ions in the zinc half-cell:

$$Zn^{2+}(aq) + 2e^- \rightleftharpoons Zn(s) \quad E^\ominus = -0.76\,V$$

This will shift the half-cell equilibrium to the left, increasing the tendency of the zinc electrode to release electrons, making the electrode potential more negative.

Going back to the expression for $E^{\ominus}_{\text{cell}}$, if we increase the magnitude of the zinc half-cell potential (make it more negative) this will lead to E_{cell} becoming more positive, so the cell will have a larger EMF.

If the concentration of Zn^{2+} ions in the zinc half-cell is increased, the half-cell reaction shifts to the left, decreasing the tendency of zinc atoms to release electrons, and making the electrode potential more positive (less reducing, and hence more oxidizing). A more positive value of E for the zinc system leads to a lower E_{cell}.

The Nernst equation

The Nernst equation allows chemists to calculate the cell potential of non-standard half-cells where the concentrations of ions are not 1 mol dm^{-3}. The mathematical relationship between the electrode potential and the concentration of aqueous ions is called the Nernst equation. It describes the relationship between cell potential and concentration (at constant temperature). It also describes the relationship between cell potential and temperature (at constant concentration).

For the generalized case of a redox system:

$$\text{Ox} + ne^- \rightleftharpoons \text{Red}$$

$$E = E^{\ominus} + \frac{RT}{nF}\ln\frac{[\text{Ox}]}{[\text{Red}]}$$

where R represents the gas constant (8.31 J mol^{-1} K^{-1}), F the Faraday constant (96 485 C mol^{-1}, which is the product of the charge on an electron and the Avogadro constant), T the absolute temperature (in Kelvin, often 298 K) and n the number of electrons transferred.

Solution of problems using the Nernst equation

For a metal/metal ion system:

$$M^{n+}(aq) + ne^- \rightleftharpoons M(s)$$

$$E = E^{\ominus} + \frac{RT}{nF}\ln\frac{[M^{n+}(aq)]}{[M(s)]}$$

In these circumstances, the concentration of a solid, [M(s)], is taken as 1.

Worked examples

What is the electrode potential of a half-cell consisting of zinc metal suspended in a 0.01 mol dm^{-3} solution of zinc ions, at 298 K?

$$E = E^{\ominus}_{Zn} + \frac{RT}{nF}\ln\frac{[Zn^{2+}(aq)]}{[Zn(s)]}$$

$$E = -0.76 + \frac{8.31 \times 298}{2 \times 96485}\ln\frac{0.0100}{1.00}$$

$$E = -0.76 + (-0.0600) = -0.82\text{V}$$

Note that the E value is more negative when the zinc solution is diluted, as predicted in the qualitative example above.

Note that if $R = 8.31$, $T = 298$ and $F = 96485$, then the equation simplifies to

$$E = E^{\ominus} + \frac{0.0257}{n}\ln\frac{[M^{n+}(aq)]}{[M(s)]}$$

Nernst equation applied to a voltaic cell (two half-cells)

In a voltaic cell in which both half-cells have non-standard concentrations, we can combine the Nernst equations for each half-cell to devise a single equation which allows calculation of the cell EMF.

Consider the Daniell cell. The standard EMF is calculated by:

$$E^{\ominus}_{\text{cell}} = E^{\ominus}_{\text{Cu}} - E^{\ominus}_{\text{Zn}}$$

Substituting the non-standard equations for the copper and zinc half-cells gives:

$$E_{\text{cell}} = \left(E^{\ominus}_{\text{Cu}} + \frac{RT}{nF}\ln\frac{[\text{Cu}^{2+}(\text{aq})]}{[\text{Cu(s)}]}\right) - \left(E^{\ominus}_{\text{Zn}} + \frac{RT}{nF}\ln\frac{[\text{Zn}^{2+}(\text{aq})]}{[\text{Zn(s)}]}\right)$$

Remembering that the concentrations of the solids are both 1, we can simplify this to:

$$E_{\text{cell}} = E^{\ominus}_{\text{cell}} + \frac{RT}{nF}\ln\,[\text{Cu}^{2+}(\text{aq})] - \frac{RT}{nF}\ln\,[\text{Zn}^{2+}(\text{aq})]$$

$$E_{\text{cell}} = E^{\ominus}_{\text{cell}} - \frac{RT}{nF}\ln\frac{[\text{Zn}^{2+}(\text{aq})]}{[\text{Cu}^{2+}(\text{aq})]}$$

Or, at 298 K:

$$E_{\text{cell}} = E^{\ominus}_{\text{cell}} - \frac{0.0257}{n}\ln\frac{[\text{Zn}^{2+}(\text{aq})]}{[\text{Cu}^{2+}(\text{aq})]}$$

Note the *negative* sign between the two terms.

The expression $\frac{\text{Zn}^{2+}(\text{aq})}{\text{Cu}^{2+}(\text{aq})}$ generalizes as the 'reaction quotient', sometimes called 'Q' (see Chapter 7) as the spontaneous reaction here is:

$$\text{Zn(s)} + \text{Cu}^{2+}(\text{aq}) \rightarrow \text{Zn}^{2+}(\text{aq}) + \text{Cu(s)}$$

A general form of the Nernst equation is therefore:

$$E_{\text{cell}} = E^{\ominus}_{\text{cell}} - \frac{0.0257}{n}\ln Q$$

Worked example

Use the Nernst equation to calculate the cell potential at 298 K of a Daniell cell where the zinc ion concentration is 0.00500 mol dm^{-3} and the copper(II) ion concentration is 1.50 mol dm^{-3}.

$$E_{\text{cell}} = E^{\ominus}_{\text{cell}} - \frac{0.0257}{n}\ln\frac{[\text{Zn}^{2+}(\text{aq})]}{[\text{Cu}^{2+}(\text{aq})]}$$

$$E_{\text{cell}} = +1.1\,\text{V} - \frac{0.0257}{2}\ln\frac{0.00500}{1.50}$$

$$E_{\text{cell}} = +1.1\,\text{V} - (-0.070\,\text{V})$$

$$E_{\text{cell}} = +1.17\,\text{V}$$

Use the Nernst equation to calculate the cell potential at 298 K of a voltaic cell consisting of a silver electrode suspended in 0.0100 mol dm^{-3} aqueous silver nitrate and a zinc electrode in 1.50 mol dm^{-3} aqueous zinc sulfate.

The cell reaction here is: $\text{Zn(s)} + 2\text{Ag}^+(\text{aq}) \rightarrow \text{Zn}^{2+}(\text{aq}) + 2\text{Ag(s)}$

The reaction quotient Q is therefore $\frac{\text{Zn}^{2+}(\text{aq})}{\text{Ag}^+(\text{aq})^2}$ (note the Ag^+ concentration raised to second power, because Ag^+ has a coefficient of 2).

Under standard conditions:

$$E^{\ominus}_{\text{cell}} = E^{\ominus}_{\text{Ag}} - E^{\ominus}_{\text{Zn}}$$

$$E^{\ominus}_{\text{cell}} = +0.80\,\text{V} - (-0.76\,\text{V}) = +1.56\,\text{V}$$

Under these non-standard conditions:

$$E_{\text{cell}} = E^{\ominus}_{\text{cell}} - \frac{0.0275}{2}\ln\frac{[\text{Zn}^{2+}(\text{aq})]}{[\text{Ag}^+(\text{aq})]^2}$$

$$E_{\text{cell}} = +1.56\,\text{V} - \frac{0.0275}{2}\ln\frac{150}{0.0100^2}$$

$$E_{\text{cell}} = +1.56\,\text{V} - (0.120\,\text{V})$$

$$E_{\text{cell}} = +1.44\,\text{V}$$

Concentration cells

As the concentration affects the electrode potential of a half-cell, this means that it is possible to generate a cell EMF using two electrodes of the same material, and the same electrolyte, but at different concentrations.

In such cases, the standard electrode potential $E^{\ominus}_{cell}$ equates to zero, as this would have both electrolytes at 1 mol dm^{-3}. However, when the two concentrations are changed, a measurable EMF is produced.

For concentration cells at 298 K, the Nernst equation takes the form:

$$E = \frac{0.0257}{n} \ln \frac{C_2}{C_1}$$

where C_2 and C_1 represent the concentrations of electrolyte in the half-cells containing the anode and cathode.

Worked example

An example of a concentration cell is a nerve cell. Different concentrations of potassium ions inside and outside the cell can result in the generation of an electrical potential. Estimate this potential if the concentration of potassium ions (K^+) outside the cell is known to be about 20 times that inside the cell.

$$E = \frac{0.0275}{1} \ln \frac{20}{1}$$

$E = 0.077$ V

11 Use the Nernst equation to calculate the cell potential at 298 K of a voltaic cell consisting of a copper electrode suspended in 0.100 mol dm^{-3} aqueous copper(II) sulfate and an iron electrode suspended in 0.0200 mol dm^{-3} aqueous iron(II) nitrate.

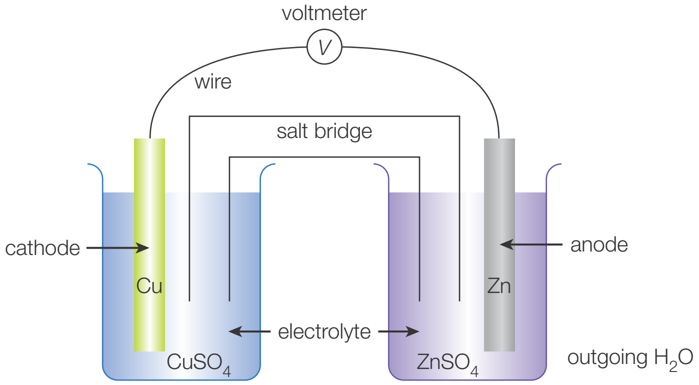
12 Use the Nernst equation to calculate the cell potential at 298 K of a concentration cell consisting of two zinc electrodes – one suspended in a solution of 0.100 mol dm^{-3} zinc sulfate and the other in 0.000500 mol dm^{-3} zinc sulfate.

Experimental work

The Nernst equation offers an excellent source of experimental investigations. A simple Daniell cell (copper/zinc) can be set up in the school laboratory (Figure 24.59) and the effect of changing the concentration of either the copper(II) ion solution or the zinc ion solution can be investigated. Alternatively the temperature of one or both half-cells could be changed using a water bath and the voltage measured.

An alternative investigation might involve a concentration cell, in which two copper or zinc half-cells are connected (with an EMF of zero expected) and the concentration of ions in each half-cell is then varied.

Figure 24.59 A simple Daniell cell. The voltage reading on the voltmeter can be investigated as the concentrations of the copper(II) sulfate or zinc sulfate solutions is changed at constant temperature



Battery disposal

Primary cells contain elements such as zinc and manganese. Secondary cells contain elements such as nickel, cadmium and cobalt. How do we ensure that these substances do not enter the environment?

Responses to this question vary widely around the world. In more economically developed countries, battery recycling and careful disposal has gained more attention recently. The European Union passed a directive in 2009 requiring battery producers to collect and dispose of their products. In the USA, many state laws require that rechargeable batteries are recovered and recycled, owing to their toxicity, although few primary cells are recycled.

Even within the same country, laws vary. For example in Western Australia, all used batteries are required to be encased in concrete so that harmful substances cannot leach out in landfill areas. Queensland has no such requirement.

Laws regulating recycling of lead-acid batteries in rich countries can actually do more harm than good. Manufacturers often export the batteries to less economically developed countries such as Mexico, India and Brazil where environmental and labour laws are less stringent. This may result in lead pollution entering the local environment and lead poisoning of employees and their families.

ToK Link

The language of electricity

You have probably studied electricity during school science using a series of metaphors. Electric current is likened to a flow of water, or traffic on a road; voltage or electromotive force is the 'push' given to the charged particles by a power supply; resistance is like the narrowing of a river or a decrease in the lanes of traffic, making it more difficult for current to flow.

These metaphors are helpful because they render the atomic world inside a wire (for instance) as comprehensible, and in many cases the metaphoric explanation does lead to predictable outcomes. For example, a thinner wire (like a narrower road) has a higher resistance.

Upon considering the nature of 'current' more carefully we might say that likening it to a fluid is incorrect, when it is actually a series of electrons, moving from valence shell to valence shell of a series of atoms. But even at this point, we are using metaphors such as 'shells' to represent the location of electrons, while also assuming that the electron is a discrete particle, when it also exhibits wave-like behaviour.

This example illustrates that it is difficult to eradicate metaphor entirely from scientific discourse. Some philosophers of science express a wish to do so, proposing that scientific theories should be literal and precise. One the other hand, some of the greatest minds in science appreciated the power of metaphor. James Clerk Maxwell used 'lines of force' to explain magnetism. Richard Dawkins described genes as 'selfish'. When advanced scientific theories can in fact only be described accurately using mathematics, a simple metaphor can be incredibly useful.

24.7 Nuclear fusion and nuclear fission (AHL) – *large quantities of energy can be obtained from small quantities of matter*

Nature of Science

Theory and experiment

The development of our understanding of nuclear processes, and its application in weapons and power generation, demonstrates the interplay between theory and experiment in the sciences.

In 1933, Hungarian physicist Leó Szilárd conceived the possibility of a nuclear chain reaction driven by the recently discovered neutron. However, it was only after failed experiments on various elements by many scientists that he and Enrico Fermi established that uranium was able to sustain a nuclear chain reaction.

Later, during the Manhattan Project, it was established by theoreticians led by Seth Neddermeyer that one way to create the critical mass needed for detonation of a bomb was to compress a mass of plutonium using an explosive shockwave. However, the technical difficulties in creating a symmetrical shockwave that would compress the mass of plutonium evenly and not simply blow it apart were enormous. It required dozens of experimental tests to develop the shaped explosive charges needed to achieve this, as well as theoreticians calculating the shapes of the explosive charges needed.

A much more modern example is the development of nuclear fusion. We know that nuclear fusion is possible: it takes place in the Sun. However, a self-sustaining fusion reaction on Earth has yet to be achieved. One major technical hurdle is maintaining a ‘plasma’ of hydrogen nuclei at a temperature high enough that they are able to collide and fuse. In 2012, researchers at the Princeton Plasma Physics Laboratory made a theoretical breakthrough in the understanding of the behaviour of this plasma. They realized that the formation of ‘clumps’ in the plasma led to it cooling down and collapsing. Experimental physicists now have a basis for further work in establishing how to design reactors to prevent this cooling effect.

Energy from nuclear processes

In Section 24.3 we explored the applications of nuclear fission and fusion. The main point is that these processes are able to generate enormous quantities of energy.

Two major questions that remain to be answered are these:

1 Why is so much energy produced by these processes (nuclear fuels have energy densities hundreds of thousands of times higher than chemical fuels)?
2 How can it be that fission (splitting the nuclei of large atoms) and fusion (combining the nuclei of small atoms) both result in release of energy?

To answer these two questions requires an understanding of the mass–energy relationship proposed by Einstein, described in his famous equation $E = mc^2$, and an appreciation of the strong forces that bind nuclei together.

Einstein’s equation

In Einstein’s equation (known as the mass–energy equivalence relationship), E represents energy (in J), m represents mass (in kg) and c is the speed of light in a vacuum.

What the equation shows is that energy and mass are interchangeable quantities. During a chemical reaction, it is assumed that energy is conserved and that mass is conserved separately. However, during nuclear processes, we need only assume that the total quantity of mass and energy are conserved.

The numerical value of c is very high (approximately $3.00 \times 10^8\,\mathrm{m\,s^{-1}}$) so a small mass, if converted to energy, releases enormous amounts of energy.

This is the basis of the enormous energy density of nuclear fuels. During fission and fusion processes, the masses of the nuclei involved are *not quite* conserved. Small amounts of mass (fractions of an atomic mass unit) are converted to energy during the processes.

Nuclear binding energy and mass defect

Section 24.3 includes a qualitative discussion of the energy changes during fission or fusion processes. To briefly recap, a nuclear process will occur if the binding energy of the nuclei of the products is greater than the binding energy of the reactant nuclei.

The binding energies of nuclei depend on a balance between the strong nuclear force (which dominates for small nuclei) and the electrostatic repulsion between the protons (which becomes more important in large nuclei, with many protons).

- *Binding energy* (ΔE): the energy required to separate a nucleus into protons and neutrons. It is expressed in kilojoules per mole of nuclei or megaelectronvolts (MeV).

  If the masses of the protons and neutrons in a particular nucleus are added up it is found that the total is slightly less than the actual mass of that nucleus. This arises because the binding energy of a nucleus has a small amount of mass associated with it (from $E = mc^2$) so the actual nuclear mass is slightly higher than its substituent protons and neutrons. This difference is called the mass defect (Δm).

- *Mass defect* (Δm): the difference between the mass of a nucleus and the total mass of its substituent nucleons.

Calculating the mass defect and binding energy of a nucleus

If we sum the masses of the nucleons in a given nucleus, and compare the total to the mass of the nucleus, we can use the mass defect to find the binding energy.

In the following examples, the atomic mass unit (amu) refers to a mass equal to one-twelfth that of a carbon-12 nucleus.

Worked example

Find the mass defect and binding energy of a helium-4 nucleus (^4_2He). Use the data in Table 24.15.

Actual mass of helium-4 nucleus	4.002602 amu
Mass of proton	1.007276 amu
Mass of neutron	1.008665 amu
Mass of 1 amu in kg	1.660539×10^{-27} kg
Speed of light in a vacuum	299 792 500 m s^{-1}

Table 24.15 Data for helium-4

A helium-4 nucleus consists of two protons and two neutrons.

Total mass of helium nucleus = (no. of neutrons × mass of neutron) + (no. of protons × mass of proton)

Mass = (2 × 1.008665) + (2 × 1.007276) = 4.031882 amu

Mass defect = mass of constituent particles – mass of nucleus

= 4.031882 amu – 4.002602 amu = 0.029280 amu

Hence the mass of the actual nucleus is smaller than the mass of its substituents. The 'missing mass' represents the energy released when the nucleus is formed.

To find the binding energy:

First, convert the mass defect into kilograms:

Mass defect = $0.029280 \times 1.660539 \times 10^{-27} = 4.862058 \times 10^{-29}$ kg

Apply Einstein's equation to convert this into energy (J):

Energy = mass defect × (speed of light)^2

$E = 4.862058 \times 10^{-29} \times (299\,792\,500)^2 = 4.368927 \times 10^{-12}$ J per nucleus

To convert to energy change in kilojoules per mole (kJ mol^{-1}), multiply by the Avogadro constant to obtain the joules per mole, then divide by 1000:

$4.368927 \times 10^{-12}\,\text{J} \times 6.022141 \times 10^{23}\,\text{mol}^{-1} = 2.631029 \times 10^{12}\,\text{J mol}^{-1}$

$= 2.631029 \times 10^{9}\,\text{kJ mol}^{-1}$

Megaelectronvolts as a unit of energy:

The binding energy graph shown in Section 24.3 (see Figure 24.17 on p.27 or Section 36 of the IB *Chemistry data booklet)* uses the unit MeV, that is, megaelectronvolts.

The electronvolt (eV) is defined as the energy required to move one electron across a potential difference of one volt.

$$1\,\text{eV} = 1.6022 \times 10^{-19}\,\text{J}$$

In the example above, the binding energy of a helium nucleus is calculated as 4.368927×10^{-12} J. To convert into electron volts:

$$\frac{4.368927 \times 10^{-12}\,\text{J}}{1.6022 \times 10^{-19}\,\text{J eV}^{-1}} = 27\,268\,000\,\text{eV} = 27.3\,\text{MeV}$$

How much energy is produced in a fusion reaction?

The mass difference between the products and reactants can be used to calculate the energy released during a fusion process.

Application of the Einstein mass–energy equivalence relationship to determine the energy produced in a fusion reaction

Consider this fusion step in the nucleosynthesis of $^{4}_{2}\mathrm{He}$, which takes place in the Sun:

$$^{3}_{2}\mathrm{He} + {}^{3}_{2}\mathrm{He} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H}$$

The masses of the species concerned are provided in Table 24.16.

Table 24.16 Mass data for helium and hydrogen isotopes

Particle	$^{3}_{2}\mathrm{He}$	$^{4}_{2}\mathrm{He}$	$^{1}_{1}\mathrm{H}$
Mass (amu)	3.016029	4.002602	1.007940

Total mass of reactants: $2 \times 3.016029 = 6.032058\,\mathrm{amu}$

Total mass of products: $4.002602 + (2 \times 1.007940) = 6.018482\,\mathrm{amu}$

Mass difference: mass of reactants − mass of products $= 6.032058\,\mathrm{amu} - 6.018482\,\mathrm{amu}$

$= 0.013576\,\mathrm{amu}$

Mass difference in kilograms: $0.013576\ \mathrm{amu} \times 1.660539 \times 10^{-27}\,\mathrm{kg\,amu^{-1}}$

$= 2.254348 \times 10^{-29}\,\mathrm{kg}$

Apply Einstein's equation to convert this mass difference into energy (J):

Energy = mass difference × (speed of light)^2

$E = 2.254348 \times 10^{-29} \times (299\,792\,500)^2 = 2.026107 \times 10^{-12}\,\mathrm{J}$ per nucleus

Expressed in electronvolts this becomes:

$$\frac{2.026107 \times 10^{-12}\,\mathrm{J}}{1.6022 \times 10^{-19}\,\mathrm{J\,eV^{-1}}} = 12\,646\,000\,\mathrm{eV} = 12.6\,\mathrm{MeV}$$

To convert to kilojoules per mole ($\mathrm{kJ\,mol^{-1}}$), multiply by the Avogadro constant to obtain the answer in joules per mole, then divide by 1000:

$2.026107 \times 10^{-12}\,\mathrm{J} \times 6.022141 \times 10^{23}\,\mathrm{mol^{-1}} = 1.22015 \times 10^{12}\ \mathrm{J\,mol^{-1}}$

$= 1.22015 \times 10^{9}\,\mathrm{kJ\,mol^{-1}}$

How much energy is produced in a fission process?

Similarly, the mass difference between the products and reactants can be used to calculate the energy released during a fission process.

Application of the Einstein mass–energy equivalence relationship to determine the energy produced in a fission reaction

Consider the following fission reaction in which a neutron causes fission of a uranium-235 nucleus, leading to the formation of krypton-92 and barium-141, with three neutrons being released. Necessary data is provided in Table 24.17.

$$^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{92}_{36}\mathrm{Kr} + {}^{141}_{56}\mathrm{Ba} + 3{}^{1}_{0}\mathrm{n}$$

Table 24.17 Mass data for uranium-235 fission

Particle	$^{1}_{0}\mathrm{n}$	$^{235}_{92}\mathrm{U}$	$^{92}_{36}\mathrm{Kr}$	$^{141}_{56}\mathrm{Ba}$
Mass (amu)	1.008665	235.043930	91.926156	140.914411

Total mass of reactants: 1.008665 + 235.043930 = 236.052595 amu

Total mass of products: 91.926156 + 140.914411 + (3 × 1.008665) = 235.866562 amu

$$\text{Mass difference: mass of reactants} - \text{mass of products} = 236.052595\,\text{amu} - 235.866562\,\text{amu}$$
$$= 0.186033\,\text{amu}$$

Mass difference in kilograms: $0.186033\,\text{amu} \times 1.660539 \times 10^{-27}\,\text{kg amu}^{-1} = 3.089151 \times 10^{-28}\,\text{kg}$

Apply Einstein's equation to convert this mass difference into energy (J):

$$\text{Energy} = \text{mass difference} \times (\text{speed of light})^2$$

$$E = 3.089151 \times 10^{-28} \times (299\,792\,500)^2 = 2.776391 \times 10^{-11}\,\text{J per nucleus}$$

Expressed in electronvolts this becomes:

$$\frac{2.776391 \times 10^{-11}\,\text{J}}{1.6022 \times 10^{-19}\,\text{J eV}^{-1}} = 173\,290\,000\,\text{eV} = 173\,\text{MeV}$$

To convert to kilojoules per mole (kJ mol^{-1}), multiply by the Avogadro constant to obtain the energy change in joules per mole, then divide by 1000:

$$2.776391 \times 10^{-11}\,\text{J} \times 6.022141 \times 10^{23}\,\text{mol}^{-1} = 1.671981 \times 10^{13}\,\text{J mol}^{-1}$$
$$= 1.671981 \times 10^{10}\,\text{kJ mol}^{-1}$$

One can compare this with the enthalpy of combustion of octane, which is −5470 kJ mol^{-1}, smaller by a factor of a million.

13 Calculate the mass defect in kg and the binding energy in MeV of this fusion reaction which is part of the nucleosynthesis of nitrogen: ${}^{12}_{6}\text{C} + {}^{1}_{1}\text{H} \rightarrow {}^{13}_{7}\text{N}$. Necessary data is provided in Table 24.18.

Particle	${}^{1}_{1}\text{H}$	${}^{12}_{6}\text{C}$	${}^{13}_{7}\text{N}$
Mass (amu)	1.00794	12.00000	13.00574

Table 24.18 Mass data for nucleosynthesis of nitrogen

Producing nuclear fuel

In Section 24.3 it was noted that nuclear fission weapons contain a relatively small quantity of uranium-235 of high purity. Since U-235 is fissionable, a 'critical mass' of this substance can lead to a chain reaction, with the instantaneous release of enormous amounts of energy. However, nuclear power stations instead rely on the slow release of energy from the fission reaction. A runaway chain reaction leads to reactor meltdown. So in a power station, uranium containing a much smaller proportion of fissionable U-235 is used, with the remainder being mostly uranium-238.

In both situations however, the proportion of U-235 is greater than that which occurs naturally (see Table 24.19). It is therefore necessary to **enrich** natural uranium by a process of isotope separation, in order that the proportion of U-235 is sufficient for the stated purpose.

Table 24.19 Different enrichment grades of uranium

	Proportion of U-235
Natural uranium	0.7%
Low-enriched uranium (for reactors)	3–4%
High-enriched uranium (for weapons)	>90%

First stages of uranium production

Uranium is mined as U_3O_8, a yellow solid ore commonly known as yellowcake. This solid undergoes a series of reactions, via uranium dioxide (UO_2), ending with oxidation with fluorine gas, to form uranium hexafluoride (UF_6; uranium(VI) fluoride; commonly known as *hex*). Uranium hexafluoride is a solid at room temperature, but it sublimes at only 56.5 °C at 1 atmosphere pressure (Figure 24.60). The isotope separation of uranium uses UF_6 in the gas phase.

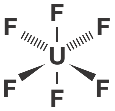


Figure 24.60 Uranium hexafluoride

UF_6 is a non-polar covalent molecular substance. In the solid phase, the molecules are held together in a lattice by dispersion forces. These require relatively little thermal energy to overcome them, leading to a low sublimation temperature.

Like other metallic halides with large degrees of covalency in their bonding, such as aluminium chloride, UF_6 hydrolyses when added to water, forming uranyl fluoride UO_2F_2 and hydrogen fluoride, according to the following equation:

$$UF_6(g) + 2H_2O(l) \rightarrow UO_2F_2(aq) + 4HF(g)$$

The different properties of UO_2 and UF_6 in terms of bonding and structure

Uranium dioxide is an ionic compound, containing uranium(IV) ions (U^{4+}) and oxide (O^{2-}) ions, arranged in a similar manner to the ions in calcium fluoride, CaF_2. As an ionic lattice comprised of highly charged ions (uranium(IV), 4+ and oxide, 2−) the melting point is very high, at 3140 K.

Uranium hexafluoride has a very high molar mass (349 g mol^{-1}) but is a gas at temperatures only slightly above room temperature. This might seem unusual but can be explained by considering the bonding within the substance. Fluorine forms short strong bonds with the central uranium atom, and the resulting molecule is a symmetrical octahedral molecule with no permanent dipole. The bonding within the UF_6 molecules is strong but the intermolecular forces binding it to its neighbours are weak dispersion forces.

Isotope separation

The different masses of the uranium atoms within uranium hexafluoride means that they can be separated:

If U-235 is present, the mass of a UF_6 molecule is 349 amu.

If U-238 is present, the mass is 352 amu.

This represents about a 0.8 per cent difference in mass between these two molecules. Although small, this mass difference affects the physical properties of UF_6 sufficiently that the molecules can be separated by physical means.

Separation by effusion

Effusion is a method by which gaseous substances can be separated according to their molecular masses. British scientists devised this method in 1940. When the USA entered World War II, the British atomic weapons effort was absorbed into the US Manhattan Project. The uranium used in the 'Little Boy' weapon dropped on Hiroshima was enriched by effusion.

The relationship between Graham's law of effusion and the kinetic theory

Effusion is defined as the process in which gas molecules pass through a small hole in their containment vessel. The rate at which the molecules pass through the hole is related to their mass. Heavier molecules pass through the hole at a lower rate. If a mixture of two gases is placed in the vessel, the gas with lighter molecules will escape through the hole at a higher rate, so the gas inside the vessel becomes richer in the heavier component.

This is explained by recalling that the kinetic energy is given by the expression $\frac{1}{2}mv^2$, where m is the mass and v is the average velocity of the gas molecules.

In a mixture of gases of different masses, at the same temperature, the average kinetic energy of the different gases must be the same. Since the lighter gas has a lower value of m, it must therefore have a higher value of v, so that $\frac{1}{2}mv^2$ reaches the same value. In short, the lighter gas has a higher average velocity of its molecules, so more of them will travel through the hole in the container wall, in a given time.

The relationship between mass of gas molecules and effusion rate is given by **Graham's law**:

$$\frac{\text{rate}_1}{\text{rate}_2} = \sqrt{\frac{\text{mass}_2}{\text{mass}_1}} = \sqrt{\frac{\text{density}_2}{\text{density}_1}}$$

In this equation rate$_1$ and rate$_2$ refer to the rates of effusion of two gases (measured in moles per unit time), and mass$_1$ and mass$_2$ are their formula masses. The equation can also be expressed in terms of the densities of the gases, density$_1$ and density$_2$.

Figure 24.61 shows how an effusion vessel operates. UF_6 vapour containing a mixture of the two uranium isotopes enters the vessel. The vessel is divided in two by a porous membrane. UF_6 containing uranium-235 effuses through the membrane more rapidly than that containing uranium-238 so the two isotopes are gradually separated.

Problems on the relative rate of effusion using Graham's law

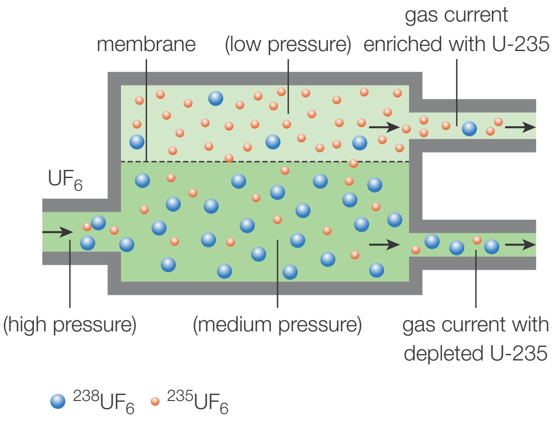
The masses of UF_6 molecules are slightly different, so the effusion rate will be slightly different also. The difference in effusion rate can be calculated using the molar masses cited above:

$$\frac{\text{rate}_{235UF6}}{\text{rate}_{238UF6}} = \sqrt{\frac{\text{mass}_{238UF6}}{\text{mass}_{235UF6}}} = \sqrt{\frac{352}{349}} = 1.004$$

Hence the rate of effusion of $^{235}UF_6$ is 1.004 times the rate of effusion of $^{238}UF_6$ (or 0.4 per cent faster). If a mixture of $^{235}UF_6$ and $^{238}UF_6$ is allowed to effuse through a hole in a vessel, the mixture of gases passing through the hole will be slightly richer in $^{235}UF_6$.

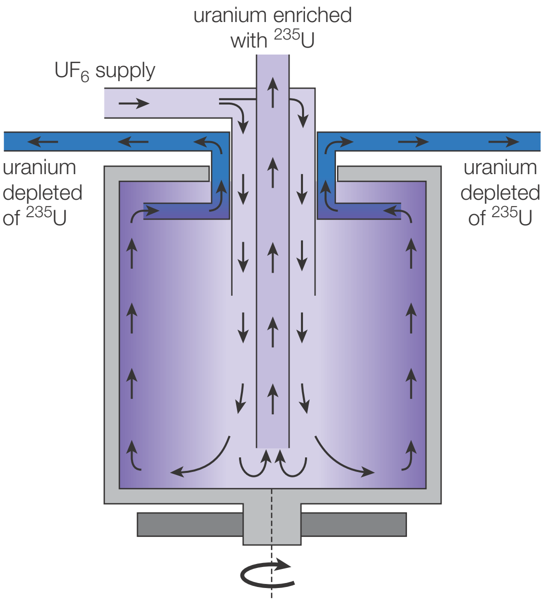
Figure 24.61 shows how an effusion vessel operates. UF_6 vapour containing a mixture of the two uranium isotopes enters the vessel. The vessel is divided in two by a porous membrane. UF_6 containing uranium-235 effuses through the membrane more rapidly than that containing uranium-238 so the two isotopes are gradually separated.

Figure 24.61 Separation of uranium isotopes by effusion


gas current enriched with U-235
membrane
(low pressure)
UF_6
(high pressure)
(medium pressure)
gas current with depleted U-235
● $^{238}UF_6$ ● $^{235}UF_6$

In practice, uranium hexafluoride is passed through a series of membranes, with the mixture becoming slightly richer in U-235 at each stage.

Isotope separation requires enormous amounts of energy and is incredibly expensive. Although this method dominated during the first decades of the Cold War, when both the USA and the USSR were determined to build up their atomic arsenals at all costs, it is now rarely used, having been mostly supplanted by centrifugation methods.


uranium enriched with ^{235}U
UF$_6$ supply
uranium depleted of ^{235}U
uranium depleted of ^{235}U

■ **Figure 24.62** Gas centrifuge used for isotopic separation of uranium

Centrifugation methods

A centrifuge is often used in science laboratories to separate mixtures. Biology laboratories use them to separate blood cells from plasma for example.

A centrifuge consists of a rotating cylinder into which a mixture is placed. When the cylinder is spun at high speed, more dense components of the mixture are forced outwards, away from the centre of rotation, while less dense components remain nearer the centre.

A gas centrifuge can therefore be used to separate a sample of uranium hexafluoride into its isotopic components. Heavier $^{238}UF_6$ molecules are forced away from the centre of rotation with more force than lighter $^{235}UF_6$.

Figure 24.62 illustrates the process. UF$_6$ richer in uranium-235 is extracted from the vessel close to the axis of rotation (i.e. the centre of the diagram), and UF$_6$ richer in uranium-238 is extracted further from the axis of rotation (i.e. the edges of the diagram).

Gas centrifugation requires much less energy than the effusion-based method outlined above – only about one-fiftieth as much. It also requires fewer stages, as the separation ratio at each stage is about 1.3 rather than 1.004.

Uranium enrichment

At present, only a small number of countries have nuclear weapons, but many others claim to want to pursue peaceful nuclear energy. A country wishing to develop nuclear power will need to either buy enriched uranium from other countries, or develop their own centrifuges and process their own. The problem here is that the basic principle is the same whether you wish to enrich uranium to 4 per cent for a reactor or to 90 per cent for a nuclear bomb.

In 2010 it was discovered that a computer worm, called 'Stuxnet', had been introduced into the control systems of the Iranian uranium enrichment facilities. This worm, which is harmless to most other computer systems, appeared to have been specifically developed to damage the Iranian enrichment effort. The difficulty of getting this advanced piece of malware into Iran and into the software systems involved led many to believe that governments were involved, specifically those of Israel and the USA. These countries are highly suspicious of Iran's peaceful nuclear ambitions, citing the aggressive statements made by Iran towards Israel in the past.

Do countries in possession of nuclear technology have the right to deny it to other countries? International agencies exist to deter nuclear proliferation, but the 'nuclear club' is one that many nations still aspire to join. Should countries be sharing nuclear technology with one another in order to reduce reliance on fossil fuels around the world?

The dangers of nuclear energy

In Section 24.3 we outlined the three types of ionizing radiation and their potential for harm to humans. How does ionizing radiation actually damage human tissues? Why does radiation lead not only to tissue damage but also cancers later on?

Alpha, beta and gamma radiation are *ionizing*; that is, they have sufficient energy to promote electrons to higher energy levels and, ultimately, out of an atom altogether, forming an ion.

Direct effects of radiation

If the atom being ionized is part of a DNA molecule, then the ionization of the atom can affect the bonding within the DNA molecule. This may affect the DNA's ability to encode proteins and to replicate itself. This will then lead to the failure of cell processes and ultimately the death of the cell.

Indirect effects of radiation

The chances of a gamma ray photon or a beta particle, for example, interacting with a DNA molecule are actually quite small, because only a very small part of each cell, and therefore the larger organism, is actually composed of DNA. However, an enormous proportion of the tissues of the organism is composed of water.

Ionizing radiation, upon collision with a water molecule, can generate hydroxyl radicals:

$$H_2O \rightarrow HO^{\bullet} + H^{\bullet}$$

Hydroxyl radicals, $HO^{\bullet}$, having an unpaired electron, are highly reactive species, and their presence in a cell in elevated concentrations is highly dangerous. They can collide and react with DNA molecules, causing damage that leads to an inability to replicate the DNA, or to mutations in the DNA that might result in tumours later on.

Another highly dangerous radical is superoxide, $O_2^{-\bullet}$ (Figure 24.63).

This species is actually employed by the body itself as part of the immune response – it is generated by phagocytes (white blood cells) when attacking invading pathogens.

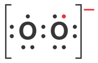


Figure 24.63 Lewis structure of a superoxide ion

Acute radiation syndrome (radiation sickness)

People exposed to large doses of radiation in a short time, such as victims of the atomic bombs in Japan or workers involved with nuclear accidents such as Chernobyl (1986) or Fukushima (2011), experience a range of symptoms.

The most sensitive system to the effects of ionizing radiation is the blood system. A large dose of radiation can cause rapid damage to bone marrow, meaning that victims experience symptoms related to their blood. A lack of platelets means that the blood cannot clot, leading to internal and external hemorrhaging. A lack of red blood cells leads to anemia, and a lack of white blood cells means that victims are highly susceptible to infections.

40
FORMER MARALINGA NUCLEAR TEST SITE
This land is part of the Maralinga Tjarutja Lands. Access requires the consent of Maralinga Tjarutja.
TELEPHONE: (08) 8625 2946
The Land contains artefacts of the Nuclear Test Era, including items contaminated at low levels with radioactivity. Any enquiries about such items should be directed to:
Maralinga Tjarutja or the South Australian Environment Protection Agency, Radiation Protection Branch.
Telephone: (08) 8463 7826

Figure 24.64 Nuclear testing at Maralinga, Australia

Longer-term effects of radiation

The damage to DNA and enzymes caused by ionizing radiation can have long-term health effects.

Cancer is the uncontrolled growth of cells. Normally, control processes ensure that cell repair and replacement occur at an appropriate rate to keep the body healthy. Radiation damage can interfere with these control processes, meaning that cells become 'immortal' and undergo uncontrolled cell division, leading to tumours. A tumour is a body of cells growing uncontrollably. These cells have their own blood supply and draw resources from the body. Cancer cells can then spread around the body leading to further tumours elsewhere. British, US and Australian servicemen and local islanders exposed to weapons tests in the Pacific in the 1950s experienced higher rates of cancers in later life (Figure 24.64).

In addition, ionizing radiation can cause a mutation in the DNA, meaning that the genetic sequence is changed. When the cells divide, they are no longer perfect healthy copies of the original. Sometimes, these mutations can lead to visible effects in the organism. In the aftermath of the atomic bombs on Japan, there was an increase in the number of fetal abnormalities, caused by the mothers' exposure to radiation.

Measuring radioactive doses

Doses of ionizing radiation are measured in sieverts (Sv). One sievert represents the effect on the body of one joule of energy from radiation on one kilogram of tissue. In SI unit terms, $1\,\text{Sv} = 1\,\text{J}\,\text{kg}^{-1}$.

The sievert can be used to estimate the increased risk of disease as radiation dose increases. For example, with each additional sievert of radiation exposure, the risk of cancer is estimated to increase by 5.5 per cent.

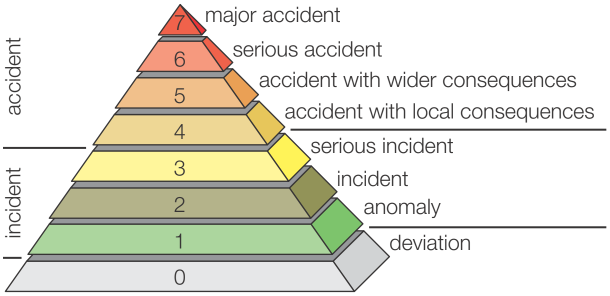
A typical radiation dose per year is 1 mSv (1 millisievert). Some activities increase the dose but are unlikely to lead to dangerous exposure. For example, aircrew receive more radiation due to time spent at high altitude, but this still only leads to about 1.5 mSv per year. A mammogram (X-ray) adds about 0.5 mSv.

Victims of the Hiroshima attack received doses ranging from 1 Sv to 10 Sv. Those receiving the higher doses all died quickly. Those receiving doses around 1 Sv experienced higher rates of cancers in subsequent years.

In 1946, Louis Slotin, a nuclear scientist, slipped while assembling the core of an experimental nuclear weapon and allowed the components to touch. For a fraction of a second, the mass went critical, and Slotin received an estimated dose of 21 Sv. He died 9 days later. Slotin gave permission for his illness to be filmed and used to educate others about the dangers of working with radioactive materials.

Nuclear accidents – cross-border disasters?

Despite strict safeguards, nuclear accidents do occur. They have the potential to be disastrous because radioactive material might leave the reactor and enter the oceans or the atmosphere, meaning that dangerous levels of radiation could be spread over a wide area, with no respect for international borders.



■ **Figure 24.65** The International Nuclear and Radiological Event Scale (INES)

In order that the severity of nuclear accidents can be rapidly communicated, in 1990 the International Atomic Energy Agency (IAEA) introduced the International Nuclear and Radiological Event Scale (INES; Figure 24.65). The INES is a logarithmic scale, so a category 7 event is ten times worse than a category 6 event.

In 1979 the Three Mile Island nuclear plant in Pennsylvania, USA, underwent a partial meltdown owing to a coolant leak. Radioactive gases were released into the atmosphere. INES did not exist at the time, but this event is now considered to be a level 5 event – an accident with wider consequences. Although there was no serious contamination on a national or international scale, some radioactive material was released into the atmosphere, and its spread was unpredictable.

The Chernobyl disaster of 1986 (Ukraine) was much more serious and is now categorized as a level 7 event. An explosion and fire at the plant led to enormous amounts of radioactive material entering the environment. The nearby cities of Chernobyl and Pripyat were completely abandoned. A 30 km exclusion zone was set up around the site, which still exists today, and elevated levels of radioactivity in air and soil were detected throughout Europe.

In 2011 a tsunami caused major damage to the reactor complex at Fukushima, Japan. Leaks of radioactive material into the sea and the atmosphere led to the establishment of a 20 km exclusion zone. The combined impact of damage to the six reactors in the complex warranted a level 7 event rating.

Kinetics of radioactive decay processes

In Section 24.3 we discussed that the time taken for the radioactivity of a sample to be halved is called the **half-life**. Radioactive decay processes are first-order processes – the half-life remains constant, but the number of radioactive atoms remaining in the sample decays exponentially.

Unlike chemical processes, radioactive decay rates are not affected by conditions such as temperature or pressure.

In first-order reactions, the half-life, $t_{\frac{1}{2}}$ is related to the decay constant λ by the following equation:

$$t_{\frac{1}{2}} = \frac{\ln 2}{\lambda} \text{ (where } \ln 2 = 0.693)$$

Hence the decay constant can be determined if the half-life is known:

$$\frac{\ln 2}{t_{\frac{1}{2}}}$$

The equation $N = N_0 e^{-\lambda t}$ is given in Table 1 of the *IB Chemistry data booklet* and can be used to determine the amount of radioactivity in a sample (N) given the decay constant (λ), the time elapsed (t), and the original amount of radioactivity. Rearranging gives

$$\frac{N}{N_0} = e^{-\lambda t} \text{ where } \frac{N}{N_0} \text{ is the fraction of the orginal radioactivity present.}$$

In Section 24.3 the decay equation was also demonstrated to rearrange to:

$$t_{\frac{1}{2}} = t\frac{\ln 2}{\ln\left(\frac{N_0}{N}\right)}$$

Worked examples

Calculate the decay constant λ for the beta decay of iodine-131 (^{131}I), which has a half-life of 8.02 days. Find the fraction of the original radioactivity remaining after 30 days.

8.02 days = 8.02 × 24 × 60 × 60 seconds = 692 928 seconds

$$\lambda = \frac{0.693}{692928\ \text{s}} = 1.00 \times 10^{-6}\,\text{s}^{-1}$$

Time elapsed = 30 days = 30 × 24 × 60 × 60 = 2 592 000 seconds

$$\frac{N}{N_0} = e^{-\lambda t} = e^{-1.00 \times 10^{-6} \times 2592000} = 0.0749$$

So the fraction of the original radioactivity present is 0.0749 or 7.49 per cent.

Strontium-90 is a radioisotope found in fallout from nuclear weapon tests. It is particularly dangerous because its chemistry is similar to that of calcium, meaning that it is readily absorbed by the body and incorporated into tissues such as bone. It has a half-life of 28.8 years. If 50 g of Sr-90 was produced in a weapon test in 1954, how many grams will remain in 2014?

$$\lambda = \frac{0.693}{28.8\ \text{y}} = 0.0241\,\text{y}^{-1}$$

$$N = N_0 e^{-t}$$

Where $N_0 = 50\,\text{g}$ and $t = 60$ years:

$$N = 50e^{-0.0241 \times 60} = 11.8\,\text{g}$$

ToK Link

Predictions

'There is no likelihood that humans will ever tap the power of the atom.'

(Robert Millikan, Nobel Laureate Physics 1923, quoted in 1928)

'It's hard to make predictions, especially about the future.'

(attributed to various people, but often Niels Bohr, Nobel Laureate Physics 1922, quoted in 1918)

With hindsight, Millikan's quote seems ridiculous, as the scientific progress that led first to atomic weapons and then nuclear power seems inevitable. However, Millikan was a Nobel Prize-winning physicist – an expert in his field. His best judgement from the facts available was that harnessing atomic energy would be impossible. When Millikan was quoted in 1928, Leó Szilárd had yet to make his theoretical breakthrough conceiving the neutron-driven nuclear chain reaction in 1933.

The above story illustrates the difficulties faced by scientists in making predictions. Although scientists make and test predictions about the outcomes of their carefully designed scientific experiments all the time, predicting the impact that a new discovery will have is incredibly difficult, as there are so many variables at work. Refer back to the Nature of Science section on pages 77–8 although Leó Szilárd conceived that a nuclear chain reaction was theoretically possible, it took hundreds of scientists and millions of dollars to actually achieve a chain reaction.

The use of quotes such as Millikan's (and there are many others) to criticize the wisdom of experts is fun but it ignores a consideration of historical context. In the study of history, the injection of modern perspectives into a historical analysis is called presentism (meaning 'based in the present'). When applied to the history of science, this leads to a fallacious way of thinking in which those who promoted new ideas and instigated 'paradigm shifts' are seen as the 'heroes', and those who offered criticism and were reluctant to accept the new paradigm are the 'villains'. This view profoundly underserves many very important scientific figures in history, who in applying the scientific method were quite right to demand strong and clear evidence before rejecting an existing scientific theory.

ToK Link

Science and moral responsibility

'The only reaction I had…was a very considerable elation and excitement, and going to parties…It would make a tremendously interesting contrast between what was going on in [America] at the same time as what was going on in Hiroshima.'

(Manhattan Project physicist Richard Feynman on his reaction after the successful detonation of an atomic bomb over Hiroshima in 1945, quoted in 1981).

In the interview, originally broadcast by the BBC, Feynman goes on to express regret for not having considered the wider implications of his work on the bomb. He explains that when he started work, it was felt that the USA was in a race with the Nazi regime in Germany to develop the weapon. He felt there was a moral imperative to develop the bomb first to deter the Nazis. Feynman's regret was that when it became clear that the Germans were not going to reach their goal, he was so caught up in the research that he did not then stop to reconsider why he was doing his research and how the weapon might be used.

Does Feynman bear a moral responsibility for the deaths in Hiroshima? Some would say no – the decision to use the weapon lay with the US president, Harry S Truman. Others suggest that the scientists knew they were developing a weapon of mass destruction (they were more knowledgeable than anyone about the potential power of the bomb) and they should have protested.

Fritz Haber, the German scientist who pioneered poison gas research in World War I (see Chapter 7), claimed to believe that his new weapon was so horrible and indiscriminate that it would lessen people's willingness to wage war, leading to fewer deaths. Alfred Nobel, inventor of dynamite, thought the same.

Weapons research is an extreme case – it is hard for the scientists concerned to deny the ultimate purpose of their work (although in the case of the Manhattan Project, the US was at war at the time, which might influence your moral view).

What about examples in which the scientists intend to do good? The development of genetically–modified crops could prove invaluable in feeding the growing global population, but it could lead to unforeseen consequences in nature. The use of geo-engineering to cool down the planet (see in Section 24.5) could mitigate global warming, or it could lead to a worse climate disaster. Ostensibly the scientists here are working towards something they believe to be beneficial, but which nonetheless might present a risk.

Is the scale of the possible consequence a factor in the moral calculus? In the 19th century a factory machine or a steam engine could kill a small number of people. In the 21st century a nuclear war using hydrogen bombs could annihilate mankind. A climate disaster could displace millions and lead to flooding and famine. A genetically engineered virus escaping from a research facility could lead to a global pandemic.

The Precautionary Principle (see also Section 24.5) might be applied in these cases. This states that if an action has a risk of causing harm to the public or to the environment, the burden of proof that it is not harmful falls on those taking the action.

This means that in the absence of a scientific consensus that geo-engineering is safe, for example, it is up to the geo-engineers to prove that it is. It is their responsibility to examine the consequences before taking action, and in the event of a disaster they must accept the moral responsibility.

So are there any areas of research which can never be morally justified? This is a difficult question. It might depend upon the ethical framework that is applied. Animal experimentation is regarded as abhorrent by many, but can be justified on utilitarian grounds (saving the lives of humans is more important than harming animals). Could utilitarian arguments be used to justify experimenting on humans, such as convicted criminals? This idea would induce revulsion in most people, suggesting that there are elements of moral principle (a view of one's duty towards others, aside from the utilitarian calculation) at work as well. In 1796 Edward Jenner deliberately exposed an 8-year-old boy to smallpox to test his hypothesis that cowpox inoculation offered protection against the disease. Would this be acceptable today?

24.8 Photovoltaic and dye-sensitized solar cells (AHL) *– when solar energy is converted to electrical energy the light must be absorbed and charges must be separated. In a photovoltaic cell both of these processes occur in the silicon semiconductor, whereas these processes occur in separate locations in a dye-sensitized solar cell (DSSC)*

Solar cells are designed to capture sunlight and convert it into electricity. Although they were invented in the late 19th century, solar cells became very important in the 1950s, when it was realized that a solar cell would reduce the reliance of early satellites on on-board batteries.

Solar cells now find a range of applications, from large rooftop solar panels to small cells designed for powering calculators and phones.

Alternative energy advocates like to point out that there is an enormous amount of sunlight energy incident on the Earth. If the US state of West Virginia was covered in solar cells, and these captured 10 per cent of the Sun's energy, this would be enough to meet all our energy needs. However, the high cost of solar cells, combined with the logistical difficulties of capturing sunlight on such a scale and then distributing the electricity, mean that solar energy is likely to form only a small part of the energy mix for some time yet.

■ How do solar cells work?

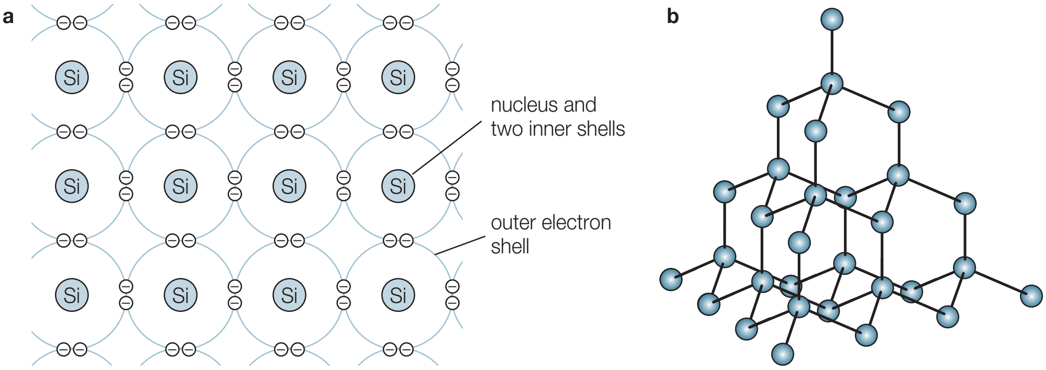
There are several types of solar cells. Two important types are the silicon photovoltaic cell and the dye-sensitized solar cell (DSSC).

Silicon photovoltaic cells

Silicon is a non-metal in group 14 of the periodic table (Figure 24.66). It has a giant covalent structure closely related to that of diamond. Each silicon atom (electron arrangement 2,8,4) forms four single covalent bonds to four other silicon atoms (Figure 24.67a). These bonds are in a tetrahedral arrangement leading to a very rigid three-dimensional network (Figure 24.67b).

Silicon is near the boundary between the metals and non-metals in the periodic table (Chapter 3). Silicon is therefore a metalloid and has some properties intermediate between metals and non-metals. In particular, the allotrope of silicon used to make solar cells is a semiconductor – this is a class of substances that have electrical resistivity properties between those of electrical conductors and insulators. Their conductivity rises with temperature because their resistance falls with an increase in temperature.

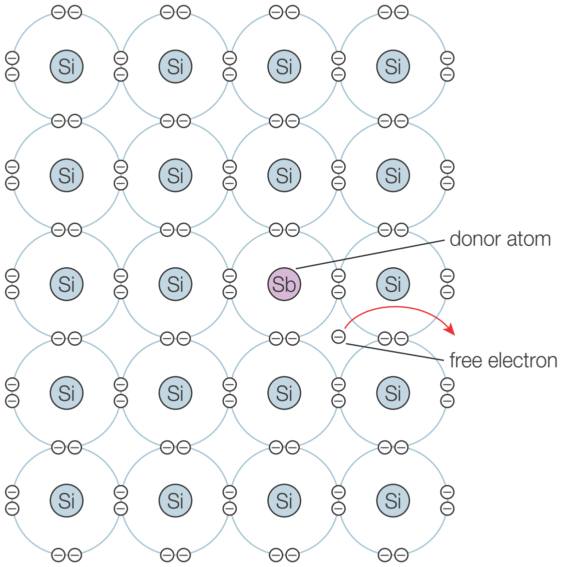
■ **Figure 24.66** A sample of the element silicon



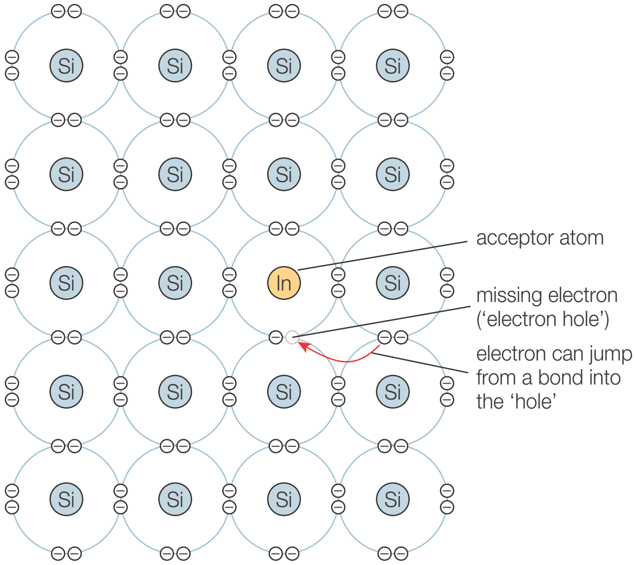
■ **Figure 24.67** Representation of **a** the two-dimensional structure; **b** the three-dimensional structure of silicon

Semiconductors

Semiconductors that are composed of pure elements or compounds are described as intrinsic semiconductors. Silicon is the most important intrinsic semiconductor and is used for the production of a wide range of semiconductor devices. These include the transistor, many kinds of diodes, the silicon-controlled rectifier and integrated circuits. Solar photovoltaic panels are large semiconductor devices that convert light energy directly into electrical energy. However, the conductivity of pure silicon can be greatly increased by adding small amounts of other substances in a process known as 'doping'. Doping is normally done by exposing silicon, in a furnace, to the vapour of the substance to be added. The substance is added in tiny carefully controlled amounts. This ensures that the atoms of the dopant will be well spaced out in the silicon lattice so that its structure is not weakened. There are two types of extrinsic semiconductor: n-type and p-type.



■ **Figure 24.68** A representation of an n-type semiconductor


Si
In
acceptor atom
missing electron ('electron hole')
electron can jump from a bond into the 'hole'

■ **Figure 24.69** A representation of a p-type semiconductor

n-Type semiconductors

In an n-type semiconductor (Figure 24.68) the dopants are group 15 elements – such as phosphorus and antimony – which have five valence electrons. When a dopant atom replaces an atom of silicon in the lattice, it uses four of the five valence electrons to form single covalent bonds with silicon atoms. The fifth valence electron is delocalized into the lattice creating a mobile negative charge. Group 15 elements are called 'donor atoms' because they donate electrons to the silicon atoms. The presence of a small number of delocalized electrons makes the extrinsic semiconductor a significantly better electrical conductor than pure silicon (the parent intrinsic semiconductor). Note that the crystal of the extrinsic semiconductor is uncharged because the additional electrons are associated with a positive charge on the donor nucleus.

p-Type semiconductors

In a p-type semiconductor (Figure 24.69) the dopants are group 13 atoms – such as boron, indium and aluminium – which have three valence electrons. When a group 13 dopant atom replaces a silicon atom, it forms three single covalent bonds with three silicon atoms, but the fourth bond is incomplete – the adjoining silicon atom contributes an electron to the bond, but the group 13 atom does not. The resulting bond consists of only one shared electron. The vacancy is termed an 'electron hole'. The added group 13 atoms are termed 'acceptor atoms' because they can accept electrons to fill the holes in the bonds. Electrons can jump from nearby bonds into the holes. This leads to the hole reappearing somewhere else – in the location the electron came from. As electrons move to fill the hole, the hole effectively 'moves' through the conductor, like a bubble moving through a liquid. This moving hole is said to be positive.

The n in n-type semiconductor stands for negative – the dopants add electrons, which of course are negatively charged. The p in p-type semiconductor stands for positive – the dopants create positive holes. However, as noted, these semiconductors do not actually carry excess negative or positive charge.

Metal conductivity compared to semiconductors

Within any given period, the metals (on the left side of the periodic table) have the largest atomic radii. As the period is traversed left to right, the atomic radii decrease. The ionization energies of the atoms get larger, as the valence electrons are closer to the nucleus, and the nucleus contains more protons.

Metals have relatively low ionization energies compared to other elements, so are more able to give up their electrons into the delocalized 'sea of electrons', making them good conductors.

Non-metals hold on tightly to their electrons, making conduction very difficult. Semiconductors are somewhere in between – under certain conditions the electrons can be made to move to higher energy levels and then on to other atoms.

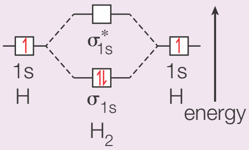
For more detail on how semiconductors conduct, see Band theory below.

Additional Perspective

Band theory

Band theory uses the idea of energy 'bands' to explain the conduction of electricity by semi-conductors.

For individual atoms, electrons occupy energy levels. These are the levels at which it is possible for the electrons to exist. The quantum model of the atom states that electrons can occupy these discrete energy level and nowhere else – they cannot exist in levels 'between' the energy levels or shells.



■ **Figure 24.70** Molecular orbital diagram for the H_2 molecule

The molecular orbital theory of bonding (Chapter 14) tells us that when two atoms bond with each other, their atomic orbitals combine to form two new 'molecular' orbitals (MOs) with different energy levels. In the case of a simple sigma (σ) bond between two atoms, two MOs are formed – one at an energy lower than the original atomic orbitals (this is called σ), and another at a higher energy (this is called σ* ; see Figure 24.70).

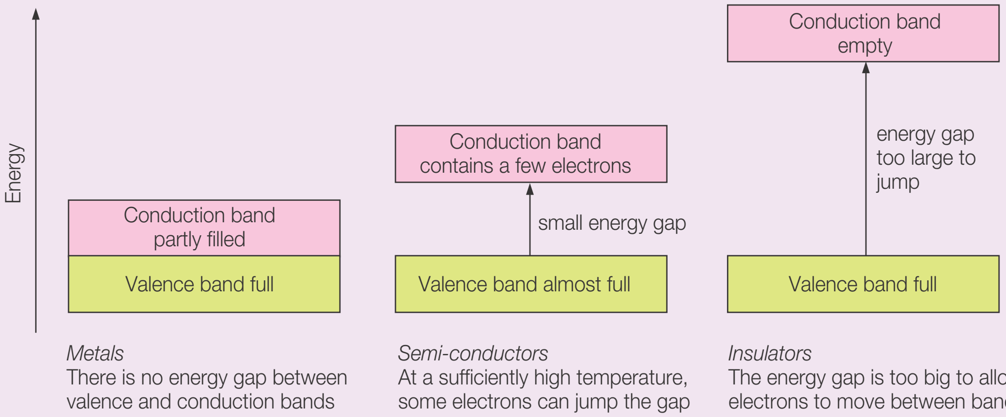
The larger the overlap, and the energy difference, between the atomic orbitals, the greater the energy separation between σ and σ*.

As more and more atoms are added (for example, to a giant molecular structure), each new atom's orbital overlaps with the others and new molecular orbitals are formed. In general *n* atomic orbitals will form *n* molecular orbitals. If *n* is large, as it is in a giant lattice-type structure, the *n* molecular orbitals are so close together that they create a continuous 'band' of energies, rather than a series of discrete levels. (If *n* is small, as in the molecular orbital diagram shown, there are few energy levels, with a clear separation between them.)

Depending on the degree of overlap between orbitals of atoms within the structure, there may be 'gaps' where no energy levels exist (Figure 24.71). These are called 'band gaps'.

In metals, there is no band gap. The atomic orbitals in the relatively large atoms in metals overlap with each other more than the smaller atoms within non-metals. This means that the energy levels 'spread out' more when the molecular orbitals are formed. The energy levels are so spread out that they overlap with each other and there are no 'gaps' in energy. In smaller non-metal atoms, the atomic orbitals overlap less. The bands are narrower and there are clear gaps in energy between higher and lower energy bands.

■ **Figure 24.71** Band theory



Metals
There is no energy gap between valence and conduction bands

Semi-conductors
At a sufficiently high temperature, some electrons can jump the gap

Insulators
The energy gap is too big to allow electrons to move between bands

In metals, electrons in the valence band can move to an unfilled energy level in a band of higher potential energy – this band is called the conduction band. Electrons in the conduction band are delocalized – they are not associated with any atom and move through the lattice when a potential difference (voltage) is applied across it.

In metals, there is no energy gap between the valence band and the conduction band. In insulators the energy gap is too big to allow electrons to move between the bands. In semiconductors the energy gap is smaller, and the probability of electrons jumping the gap increases as the temperature rises, or if light of suitable energy is absorbed.

This illustrates a key difference in the behaviour of metals and semiconductors. When the temperature of a metal increases, there is an increase in resistance, owing to the vibration of the atoms impeding the flow of electrons. However, in semiconductors, a temperature increase increases the likelihood that electrons can jump the energy gap between the valence and conduction bands, meaning that semiconductors are better conductors at high temperatures.

Photovoltaic cells

The band gap in pure silicon is 1.11 electronvolts (see Section 24.3) or 1.78×10^{-19} joules. This corresponds to a maximum absorption wavelength of 1100 nm which is in the infrared region. However, doping the silicon with other elements broadens the range of the absorbed frequencies so that many wavelengths are absorbed throughout the visible down to the infrared. When light is absorbed, valence electrons from the silicon atoms become temporarily delocalized. However, they eventually become valence electrons again – unless some process prevents them from doing this.

Photoconduction

According to band theory, when photons with sufficient energy are absorbed, valence electrons can move into the conduction band, leaving vacancies in the lower valence band (Figure 24.72). The electrons move around in the conduction band, but will eventually re-enter the valence band and occupy a vacant site – releasing energy in the process.

The operation of the photovoltaic solar cell

A photovoltaic cell consists of a pn-junction. When photons hit the n-type layer, electrons move across the junction from the n-type material into the p-type material (Figure 24.73). This creates a separation of charge on each side of the junction – a potential difference – which prevents any further movement of electrons.

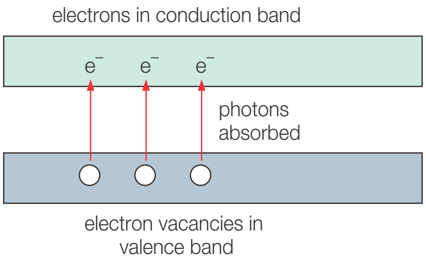
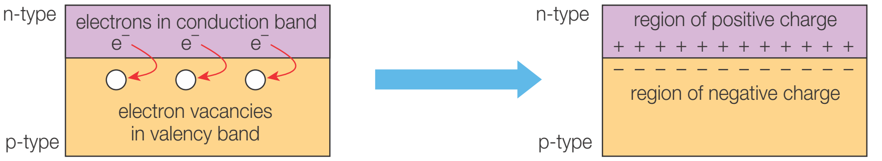

electrons in conduction band
e^-
e^-
e^-
photons absorbed
electron vacancies in valence band

Figure 24.72 Photoconduction in a semiconductor


n-type
electrons in conduction band
e^-
e^-
e^-
electron vacancies in valency band
p-type
n-type
region of positive charge
+ + + + + + + + + + + +
– – – – – – – – – – – –
region of negative charge
p-type

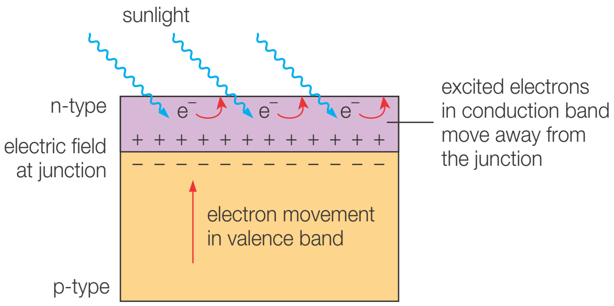
Movement of electrons across pn-junction…

… creates an electric field across the junction

Figure 24.73 Formation of an electric field at a pn-junction

If sunlight falls on the n-type material, electrons are excited to the conduction band. They are repelled away from the junction by the electric field and move towards the upper surface of the silicon wafer (Figure 24.74).

Figure 24.74 The effect of light on a photovoltaic cell


sunlight
n-type
electric field at junction
p-type
e⁻
+ + + + + + + + + + + + +
− − − − − − − − − − − − −
electron movement in valence band
excited electrons in conduction band move away from the junction

This leaves vacancies in the valence band in the n-type material, and electrons move across the junction from the valence band of the p-type material. If the upper and lower surfaces of the silicon wafer are connected through an electric circuit as shown in Figure 24.75, then electrons will flow from the top surface to the bottom surface to restore the balance of charge. The light energy of sunlight has been converted directly into electricity.

A typical photovoltaic cell is shown in cross section in Figure 24.76. The cell has an anti-reflectant coating over the upper n-type surface to increase the efficiency of light absorption. When the sunlight shines on the photovoltaic cell, the cell produces a small voltage and a current flows. The electric current is collected by strips of metal on the upper surface of the cell and channelled into an external circuit, which is connected to a metallic layer on the base of the cell. Cells are mounted in series and incorporated into a solar module.

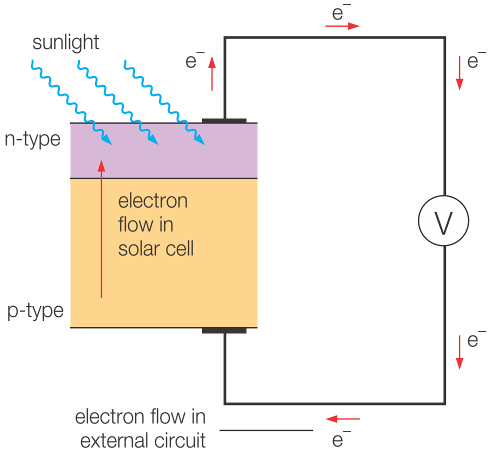

sunlight
n-type
p-type
e⁻
e⁻
e⁻
e⁻
e⁻
V
electron flow in solar cell
electron flow in external circuit

Figure 24.75 A silicon-based photovoltaic cell producing electric current in an external circuit

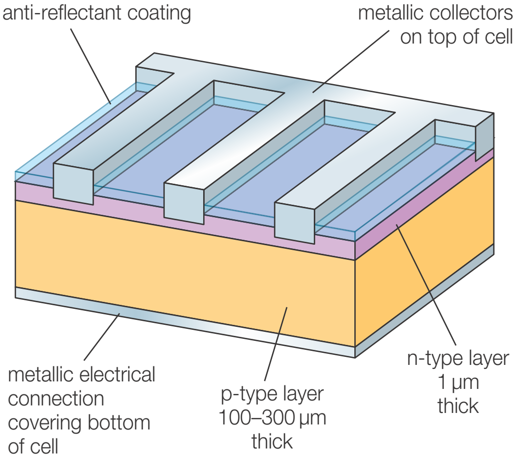

anti-reflectant coating
metallic collectors on top of cell
metallic electrical connection covering bottom of cell
p-type layer 100–300 μm thick
n-type layer 1 μm thick

Figure 24.76 Cross section of part of a photovoltaic cell

Dye-sensitized solar cells

The dye-sensitized solar cell (DSSC) is an attempt to mimic the photochemical processes of photosynthesis in a solar cell (see Section 24.4). An organo-metallic dye captures photons, whose energy is used to drive electrons through an external circuit.

In a silicon photovoltaic cell, the silicon pn-junction converts the sunlight directly into electrical energy. Electrons leave the cell from the n-type silicon and return via the p-type silicon.

In a DSSC, there are two major components needed to allow the cell to operate. Firstly, a dye is used to absorb light and use the energy to drive electrons through the external circuit. Secondly, a chemical electrolyte is needed to provide an electron to reduce the dye molecule back to its original state.

Construction of the DSSC

The Grätzel DSSC is named for Michael Grätzel (Figure 24.77), who along with Brian O'Regan developed a cheap and easily constructed cell (Figure 24.78). It consists of a transparent electrode to which a highly porous layer of titanium dioxide (TiO_2) nanoparticles coated in the organic dye is attached. This assembly has a very large surface area, which makes it much more efficient at capturing photons.

Figure 24.77 Michael Grätzel pictured with his record-breaking 15% efficient DSSC in 2013

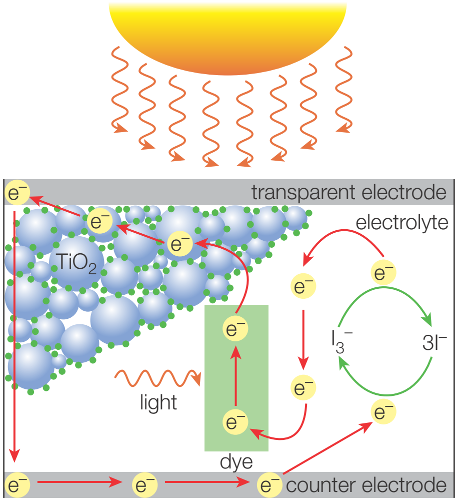


Figure 24.78 Schematic of a Grätzel DSSC

The dye absorbs photons which excite electrons. When the electron reaches a sufficient level of excitation, it enters the conduction band of the titanium dioxide semiconductor and then an external circuit. The electrons are re-introduced to the cell via a platinum electrode on the other side and electrons are returned to the dye via an electrolyte. The electrochemistry of the cell is complex, involving several pathways. One possible reaction scheme involves a mixture of iodide (I^-) and triiodide (I_3^-) ions in organic solvent. The iodide–triiodide mixture transfers electrons according to this half-equation:

$$I_3^- + 2e^- \rightarrow 3I^-$$

In the reaction scheme below, this equation is divided by 2, as the scheme follows the progress of a single electron through the electron transfer chain.

The operation of the DSSC

In the following reaction scheme, S represents the dye molecule and $h\nu$ represents a photon from the sunlight.

First, the dye molecule, S, absorbs a photon and an electron becomes excited (raised to a higher energy level, indicated by the asterisk *):

$$S \text{ (on TiO}_2\text{)} + h\nu \rightarrow S^* \text{ (on TiO}_2\text{)}$$

Next, the excited dye molecule gives up an electron to the TiO_2 network:

$$S^* \text{ (on TiO}_2\text{)} \rightarrow S^+ \text{ (on TiO}_2\text{)} + e^-$$

This electron flows through the TiO_2 and into the external circuit where it can drive a load such as a motor or a lamp. An electron re-enters the cell via the platinum counter electrode.

The iodide–triiodide electrolyte gains an electron from the counter electrode:

$$\frac{1}{2}I_3^- + e^- \rightarrow \frac{3}{2}I^-$$

The iodide ion reduces the oxidized form of the dye, S^+, back to S:

$$\frac{3}{2}I^- + S^+(\text{on TiO}_2) \rightarrow \frac{1}{2}I_3^- + S\,(\text{on TiO}_2)$$

This cycle of reactions thus uses sunlight to drive a flow of electrons through an external circuit. The dye, S, is first oxidized to S^+, but is then reduced back to S, meaning that it is not consumed by the reaction.

How nanoparticles increase the efficiency of DSSCs

Inside the DSSC, the three components (dye, TiO_2 particles and electrolyte) are in close contact. The TiO_2 nanoparticles are tiny transparent spheres that pack together, touching each other. (Imagine a box full of polystyrene spheres.) The touching spheres therefore offer a continuous pathway for electrons to flow through the cell. The spheres are coated in dye everywhere they are not touching. However, spheres do not pack perfectly. The spaces between them are full of electrolyte solution.

The DSSC therefore consists of a continuous network of touching TiO_2 particles (so electrons have a pathway from one side of the cell to the other), coated in dye. This network has a huge surface area covered in dye (owing to the tiny size of the TiO_2 particles), which in turn is in contact with a large area of electrolyte.

If the nanoparticles are too large, the area of dye (and hence absorption of light) decreases, and the area of contact with the electron-transferring electrolyte is also lower. If the nanoparticles are too small, the cell has a larger internal resistance, as it is more difficult for the electrons to find a pathway through the network. Nanoparticles for DSSCs fall in the range of 20–100 nm in diameter.

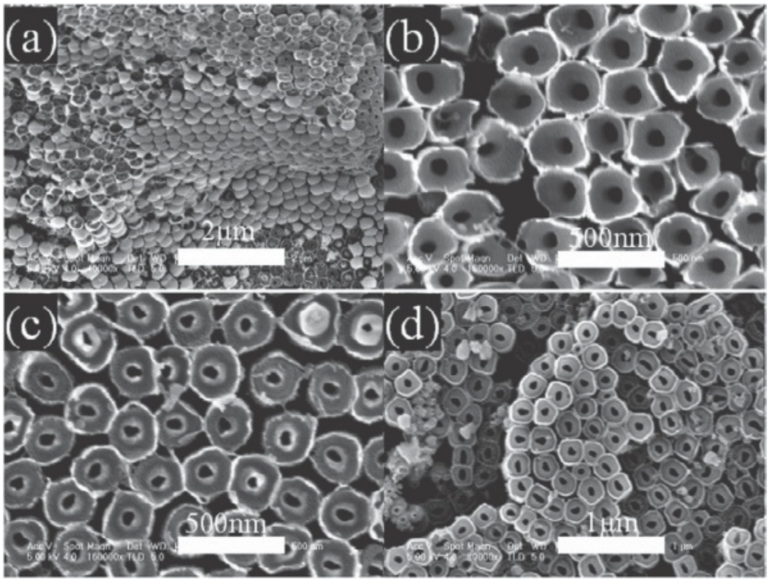


Figure 24.79 Researchers at Shanghai Jiao Tong University in China have devised a material made of TiO_2 nanotubes that they hope will offer a good compromise between conductivity and surface area for use in DSSCs

The size of the nanoparticles is optimized to allow the right compromise between the area of light-absorbing dye in contact with the electrolyte, and the ability of the electrons to find a direct path through the network. Figure 24.79 shows a network of TiO_2 nanotubes which could have application in DSSCs.

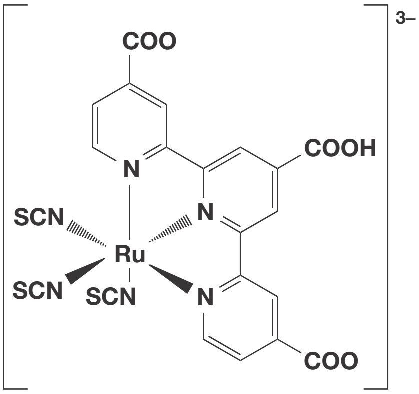


Figure 24.80 The Black Dye is a ruthenium-based complex ion dye with a π-conjugated system of pyridine rings. It absorbs across a wide range of visible light wavelengths, hence its name

The structure of the dye in a Grätzel DSSC

In Section 24.4, we discussed the importance of a conjugated system of double bonds to the absorption of visible light photons.

A conjugated system consists of a series of overlapping p-orbitals, often including alternating double and single bonds, or benzene rings. The wavelength of maximum absorption, λ_{max}, is greater when there is more conjugation. In the DSSC the choice of metal and structure of the dye is crucial in determining the range of frequencies at which the cell absorbs radiation.

An example of a dye used in a Grätzel DSSC is shown in Figure 24.80. It can be seen that it has a series of delocalized rings, similar to benzene, but with a nitrogen atom substituted for a carbon atom. These are called pyridine rings. The ring system is complexed with a central ruthenium atom, a transition metal. Conjugation is possible across the entire molecule. This dye is able to absorb photons over a wide range of frequencies in the visible region, all the way down the red end of the spectrum.

The advantages and disadvantages of photovoltaic cells and dye-sensitized solar cells

DSSCs have been under development for over 20 years and the technology is now rather mature. They are highly promising owing to their relative ease of manufacture and the high efficiencies already achieved. Silicon photovoltaic cells are an even more established technology which still receives a lot of attention and investment.

The advantages and disadvantages of each type of cell are summarized in Table 24.20.

Table 24.20 Comparison of silicon photovoltaics and DSSCs

	Advantages	Disadvantages
Silicon photovoltaic cell	Well-established technology is manufactured widely Efficiency of silicon photovoltaics is usually still higher – typically around 23%; so-called tandem cells, using several layers and multiple pn-junctions, have achieved 48%	Although economies of scale apply as they are widely made, they are still relatively expensive owing to the cost of processing silicon If they get hot (as they might on a sunny day), the silicon cells' efficiency decreases. For this reason they are often protected by a glass cover, which adds cost and weight
Dye-sensitized solar cell	The early stage of development means that significant advances are still being made to address the disadvantages of these cells. For example, gel or solid electrolytes are under development, which will remove the problem of the electrolyte solution DSSCs do not require any complicated manufacturing steps, and the dyes, although quite advanced, are cheap to make The dyes are more effective at capturing sunlight even at low light levels, whereas silicon cells need a minimum light intensity to work effectively	The electrolyte contains potentially harmful volatile organic compounds, which cannot be allowed to leak The electrolyte solution in iodide–triiodide cells makes the cell less resistant to temperature changes. At low temperatures it can freeze, which shuts the cell down. At high temperatures it can expand, causing sealing problems. However, more modern cells use solid-state or organic electrolytes which removes this problem DSSCs are typically less efficient than silicon-based cells, at around 11%, although 15% efficiency has been achieved in the laboratory. They are more useful for low-density applications such as rooftop solar panels, where physical size can be slightly larger with no real disadvantage. However, at present the cost of installation actually exceeds the cost of production, and this must be considered when calculating the payback time for the solar cell

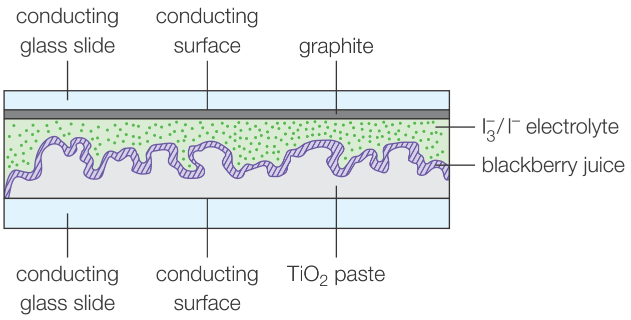
Experimental work

A Grätzel DSSC can be constructed in a school laboratory using relatively simple equipment (Figure 24.81).

1. Grind titanium dioxide powder to a fine paste with dilute ethanoic acid. Apply this paste in a thin layer to the conducting side of a conducting glass slide and allow to dry.
2. Add a few drops of a brightly coloured dye to the TiO_2 paste on the slide. Fruit juices such as raspberry or blackberry are suitable.
3. Place a second conducting slide on top of the first, conducting side down. The conductivity of the cell can be improved by adding a thin layer of pencil graphite to the conducting side of the glass plate.
4. Prepare a solution of I_3^- and I^- by dissolving iodine in potassium iodide solution. I_3^- is formed by combination of $I_2 + I^-$.
5. Use binder clips to hold the cell together, and add the solution of I_3^- and I^- to the edge of the cell using a pipette. It will creep between the two layers by capillary action.
6. The cell is now ready. Use a multimeter to measure the voltage generated by the cell in bright sunlight.

The DSSC constructed in this way is much less efficient than commercial DSSCs but it illustrates the concept and the construction is similar.

Figure 24.81 Schematic diagram of Grätzel DSSC using simple materials



Solar energy and worldwide energy needs

In his book *Sustainable Energy – Without the Hot Air* (Cambridge, 2009) David Mackay examines the feasibility of paving the Sahara desert with solar panels and using this to address the energy needs of European countries.

Mackay estimates that approximately 15 watts of electrical power can be generated from a square metre of land. This involves using a solar concentrator – a series of mirrors or lenses that concentrates sunlight into a photovoltaic collector. Given that world power consumption today is 15 000 gigawatts (1.5×10^{13} watts) then a land area of 1 million square kilometres (a 1000 km by 1000 km square) would be necessary. To supply just Europe and North Africa would require a 600 km by 600 km square, or an area the size of Germany. However, this would lead to a significant drop in the quantity of carbon added to the atmosphere from combustion of fossil fuels.

Assuming the logistical difficulties of actually constructing a solar power facility on this scale could be overcome, questions remain about the geopolitical implications of such a scheme:

Who would own the energy? Although the sunlight is incident on the countries concerned (possibly Algeria, Tunisia and Libya), it is likely that other countries or multinational energy corporations would develop the infrastructure. The African countries would have to be compensated for the energy, just as oil-rich countries are currently. What about the land itself? What is its value? Being optimistic, such projects could change the economic fortunes of hot, sunny countries around the world, with potential to raise living standards and lift many people out of poverty. However, good governance is required to ensure that the newfound income is used wisely to improve the countries' long-term economic prospects.

Mackay's book is highly theoretical and intended to illustrate the sheer scale of solar infrastructure needed to match the energy provided by fossil fuels. Realistically, the geopolitical situation does not allow for such projects. However, in 2006, Nathan Lewis and Daniel Nocera attempted a realistic analysis of the potential for solar energy to provide a useful proportion of worldwide energy needs.

Their main conclusions were that biologically derived solar cells, rather than those requiring silicon, are more likely to come down in cost sufficiently that they can offer large energy contributions. In addition, if solar energy is to be useful at higher latitudes, effective energy storage is needed, because solar energy on demand cannot be relied upon. An organic solar cell, which generates 'solar fuel' (perhaps by splitting hydrogen and oxygen from water) for use in a fuel cell, might offer a potential route to lowering our reliance on fossil fuels. Such cells would be less location-dependent, as even countries receiving much less sunlight than Saharan Africa would be able to make and store solar fuel during daylight hours.

Nature of Science

Transdisciplinary research

In Section 24.5 the transdisciplinary nature of climate science was discussed. The development of dye-sensitized solar cells also illustrates the overlap of different scientific fields.

The mechanism of action of the DSSC attempts to mimic the processes occurring during photosynthesis, so work done by biologists is incorporated. The use of TiO_2 nanoparticles requires the input of chemists with understanding of the kinetics of processes at the nanoscale. Chemists have also developed the dyes used, which aim to capture a wide range of solar radiation frequencies with a high efficiency.

Nature of Science

Funding

Scientific work has to be paid for. Sources of funding might include governments, charitable foundations or private companies. Private companies will naturally wish to see a return on their investment, and this will impact on the kinds of work done – it is likely to be focused on practical applications. Charitable foundations or governments might allocate some funding to purely theoretical research without any immediately obvious application – so-called 'blue skies research'. Government-funded research is susceptible to political concerns – for example, some political parties oppose stem cell research on moral grounds.

Sometimes research funding is justified in terms of one application, and then other potential applications are proposed. For example, much of the early work into fuel cells was carried out as part of the NASA Gemini programme in the 1960s and was funded as part of the enormous budget allocated to space flight at that time. The possibility of using fuel cells on Earth arose much later as part of a move away from fossil fuels.

ToK Link

Metaphors in science

The conjugated system in the dye molecules used in the dye-sensitized solar cell has some similarities with a violin string. The longer the conjugated system, the lower its natural frequency, meaning it absorbs radiation of longer wavelength. A violin string also has a natural frequency at which it vibrates, and this frequency is lower the longer the string. The behaviour of the violin string is a useful metaphor or memory aid at least, for the behaviour of the conjugated system.

However, is this metaphor really helpful in understanding why the conjugated system behaves as it does? Perhaps not. After all, the electrons in the conjugated system absorb light of specific wavelengths, but are they physically vibrating in the same way as the violin string? Some models propose the electron as a 'standing wave' vibrating at a specific frequency. Does the violin metaphor therefore oversimplify the complicated behaviour of the conjugated system, or overstate the similarity?

What are the purposes of metaphor in the acquisition of scientific knowledge? They are twofold: firstly, to help us get a 'sense' of the meaning of a scientific concept. As our understanding of science has increased, we have encountered many enormously complex phenomena that are not easily understood in terms of our everyday experience. Some phenomena such as quantum mechanics or the relativistic theory of gravitation can only be accurately described by mathematics. However, metaphors can be helpful in conveying the basics of an idea. The metaphor helps us to 'bridge the gap' between our own experience and the reality which might be impossible to conceive. For example, if the distortion of space–time by massive objects is likened to bowling balls resting on a rubber sheet, causing it to bend, this gives us an impression of how space–time might behave in the presence of larger or smaller objects.

This leads to metaphor's second purpose, which is to aid us in communicating our scientific knowledge. Your chemistry teacher may, at different times, have compared an electron to a particle, wave or a 'cloud' of charge. The electron is not exactly any of these things, but these metaphors can help us to make predictions about chemical behaviour, for example the formulae of compounds, their shapes or their properties.

You might argue that chemical understanding acquired on the basis of inaccurate metaphors is false understanding, and the knowledge gained is not actually knowledge at all, as it does not reflect reality. In practical terms, however, simplified ideas do have their uses. It is not necessary to use quantum mechanics to predict the formula of sodium chloride or to decide whether a metal or a non-metal will be a better conductor of electricity. In just the same way as simpler ideas have been augmented and sometimes replaced by more complex ideas in the history of science, you will learn new theories in chemistry which will augment and sometimes contradict the simpler ones you learnt in your pre-Diploma course.

Examination questions – a selection

Q1 Coal can be used to generate heat in our homes. Alternatively, it could be used in a power station to generate electricity, which can then be used for heating. Which approach is preferred? Outline some advantages and disadvantages of each option. [3]

Q2 Kerosene is used as fuel for jet aircraft. It is composed of hydrocarbons ranging from 6 to 16 carbon atoms.

a Assume kerosene is composed mainly of decane, $C_{10}H_{22}$. What is the specific energy of kerosene? Necessary data: ΔH_c: 6779 kJ mol^{-1}; M_r: 142.28 g mol^{-1}. [2]

b A Boeing 747 carries approximately 100 tonnes of fuel. How much energy is available in this mass of decane? [2]

Q3 Deduce equations for the production of decane $C_{10}H_{22}$ by:

a Direct hydrogenation of coal. [2]

b Liquefaction of coal gas, carbon monoxide, CO (g). [3]

Q4 A possible fission product of uranium-235 is xenon-137. This then undergoes beta decay forming caesium-137, releasing a gamma photon also. Caesium-137 also undergoes beta decay, and can be used as a beta radiation source to treat tumours. It has a half-life of 30 years. Caesium-137, although useful, is a hazardous isotope as it forms highly soluble salts which can enter the environment.

a Write an equation for the fission of uranium-235 to xenon-137, and identify the other element formed. [2]

b Write equations for the beta decay of xenon-137 and the subsequent beta decay of caesium-137. [2]

c Calculate the percentage of the original radioactivity left in a sample of caesium-137 after 180 years. [2]

d Why does the high solubility of caesium-137 salts make it a particularly dangerous radioisotope? [2]

Q5 Naturally occurring uranium-238 undergoes a series of radioactive decay processes terminating in lead-206. The first part of the decay chain takes uranium-235 to radium-226 in a series of steps.

a In the equation below, identify the type of radioactive particle (alpha, beta or neutron) emitted by the nucleus at each step:

$^{238}_{92}U \rightarrow {}^{234}_{90}Th \rightarrow {}^{234}_{91}Pa \rightarrow {}^{234}_{92}U \rightarrow {}^{230}_{90}Th \rightarrow {}^{226}_{88}Ra$ [5]

Radium-226 is an alpha-emitting gas with a half-life of 1600 years.

b Calculate the time taken for the radioactivity of a sample of radium-226 to decay to 20 per cent of its original value. [2]

c Why does the fact that radium is an alpha-emitting gas make it particularly dangerous? [2]

Q6 Ethanol and biodiesel are examples of biofuels.

a Outline the advantages and disadvantages of biofuels in general. [4]

b Discuss the advantages and disadvantages of ethanol versus biodiesel made from palm oil. [3]

c Explain why vegetable oils are not usually suitable for use as fuels for cars. [2]

d Outline the process of transesterification and explain why the resultant mixture is more suitable as a fuel for motor vehicles. [2]

e One of the components of palm oil is a triglyceride of oleic acid, $(C_{17}H_{33}COO)_3CH(CH_2)_2$. Write a balanced equation for the transesterification of this triglyceride with methanol. [2]

Q7 Outline the possible environmental effects on the oceans of increasing atmospheric CO_2 levels. Include equations in your discussion. [5]

Q8 This question is about the greenhouse effect and global warming. Atmospheric warming arises because the amount of radiation incident on the Earth from the Sun is no longer matched by the amount of radiation leaving the Earth into space.

a What types of electromagnetic radiation are most prevalent in sunlight? [3]

b How does this radiation result in warming of the atmosphere? [2]

c Explain how a greenhouse gas such as carbon dioxide results in increased atmospheric warming. [3]

d What do you understand by the term 'global warming potential' (GWP)? [4]

e Explain how methane, a much less abundant atmospheric pollutant than carbon dioxide, is thought to make a significant contribution to anthropogenic global warming. [2]

f Describe the likely effects of anthropogenic global warming. [4]

Q9 This question is about the hydrogen fuel cell.

a Outline the function of each component in the hydrogen fuel cell:

i Platinum catalyst

ii Aqueous sodium hydroxide electrolyte. [4]

b The proton exchange membrane fuel cell replaces the sodium hydroxide with a polymer membrane. Suggest an advantage of using the membrane instead of the aqueous solution. [1]

c Write the half-cell equations for the reactions occurring at each electrode in an alkaline cell. State which is the negative and which the positive terminal of the cell. [4]

d State one disadvantage of using hydrogen as a fuel for fuel cells. [1]

Q10 Lithium-ion batteries and nickel–cadmium batteries are both examples of secondary cells.

a Define the term 'secondary cell'. [1]

b For each cell type, describe the electrolyte and identify the charge carrying species in the electrolyte during discharge. [4]

c Write the half-equations for the anode and cathode of each cell type. [8]

d Suggest two advantages of each cell type. [4]

Q11 Methanol fuel cells offer some advantages over hydrogen fuel cells.

a Suggest two advantages of a methanol fuel cell over a hydrogen fuel cell. [2]

b **i** Deduce the oxidation state of carbon in methanol, CH_3OH. [1]

ii Deduce the oxidation state of carbon in carbon dioxide. [1]

iii Hence write the half-equation for the oxidation of methanol at the anode of the fuel cell. [2]

iv Write the half-equation for the reduction of oxygen gas at the cathode. [2]

c Some fuel cells run on ethanol as an alternative to methanol. Calculate the oxidation number of carbon in ethanol and hence write the anode half-equation for an ethanol fuel cell. [3]

Q12 The overall equation for the microbial fuel cell (MFC) can be expressed as:

$$CH_3COOH(aq) + \frac{3}{2}O_2(g) \rightarrow 2CO_2(g) + 2H_2O(l)$$

a Calculate the thermodynamic efficiency of this cell, using the following data: [2]

Substance	Enthalpy of formation/ΔH_f kJ mol^{-1}	Gibbs free energy of formation/ΔG_f, kJ mol^{-1}
CH_3COOH (aq)	−486	−390
O_2(g)	0	0
CO_2(g)	−394	−394
H_2O(l)	−286	−237

b Give two advantages of the MFC. [2]

Q13 **a** For each of the following nuclear reactions, classify them as fusion or fission processes, and calculate the mass defect (in atomic mass units) and the energy released (in MeV). [12]

i ${}^{235}_{92}U + {}^{1}_{0}n \rightarrow {}^{144}_{56}Ba + {}^{89}_{36}Kr + 3{}^{1}_{0}n$

ii ${}^{2}_{1}H + {}^{3}_{1}H \rightarrow {}^{4}_{2}He + {}^{1}_{0}n$

iii ${}^{6}_{3}Li + {}^{1}_{0}n \rightarrow {}^{4}_{2}He + {}^{3}_{1}H$

iv ${}^{235}_{92}U + {}^{1}_{0}n \rightarrow {}^{95}_{42}Mo + {}^{139}_{57}La + 2{}^{1}_{0}n + 7{}^{0}_{-1}e$

b For reaction (i) above, convert the energy value obtained to a value in kJ mol^{-1} of uranium-235. Comment on the value obtained. Necessary data:

Species	Mass (amu)
${}^{0}_{-1}e$	5.48580×10^{-4}
${}^{1}_{0}n$	1.00866
${}^{2}_{1}H$	2.01410
${}^{3}_{1}H$	3.01604
${}^{4}_{2}He$	4.00260
${}^{6}_{3}Li$	6.01512
${}^{89}_{36}Kr$	88.91764
${}^{95}_{42}Mo$	94.90584
${}^{139}_{57}La$	138.90635
${}^{144}_{56}Ba$	143.92295
${}^{235}_{92}U$	235.04393

[3]

Q14 This question is about the energy produced during nuclear reactions.

a Sketch a graph with nucleon number on the *x*-axis and energy on the *y*-axis. On these axes draw a curve showing the relationship between binding energy and nucleon number. [3]

b Explain why the graph in (a) has this shape, with reference to the electrostatic force and the strong nuclear force. [3]

c With reference to the fusion reactions of hydrogen forming helium, explain why the fusing of these nuclei leads to a release of energy. *(You need not calculate specific values in your answer.)* [2]

d With reference to the fission reaction of uranium-235 forming barium-139 and krypton-89, explain why the fission of uranium releases energy. *(You need not calculate specific values in your answer.)* [2]

Q15 Naturally occurring uranium consists mainly of uranium-238, with only 0.7 per cent uranium-235. It is necessary to increase the proportion of uranium-235, a process called enrichment.

a Why is it necessary to enrich uranium? [1]

b Uranium occurs naturally in an oxide ore called yellowcake, U_3O_8, which is converted to uranium dioxide, UO_2, and then to UF_6 in a two-step process.

i Write a balanced equation for the reaction of uranium dioxide with hydrogen fluoride, forming UF_4. State the type of reaction taking place here. [3]

ii Write a balanced equation for the conversion of UF_4 to UF_6 with fluorine gas. State the type of reaction taking place here. [3]

iii Explain why it is necessary to convert UO_2 to UF_6 for enrichment of uranium. [2]

c The uranium for the first atomic weapons was enriched by effusion of UF_6 containing each of the two isotopes of uranium in their natural abundances.

i Use Graham's law to calculate the relative rate of effusion of the two forms of UF_6. [2]

ii Given an original abundance of 0.7 per cent uranium-235, how many stages of effusion are necessary to reach an abundance of 90 per cent, which is viable for a weapon? [2]

iii Gas centrifugation methods achieve a 1.3 separation ratio. How many stages of this process are required to reach 90 per cent abundance? [1]

Q16 This question is about the dangers of ionizing radiation. A water molecule in a cell can be split into two radicals by a high energy gamma ray photon:

$H_2O \rightarrow HO^{\bullet} + H^{\bullet}$

a Write an equation for the reaction of a hydroxyl radical, $HO^{\bullet}$, with a water molecule, H_2O. [1]

b A hydrogen radical, $H^{\bullet}$, can combine with an oxygen molecule O_2, forming a hydroperoxy (or hydroperoxyl) radical, which is dangerous as it has a longer lifetime than a $H^{\bullet}$ radical.

i Write an equation for this reaction. [2]

ii Draw the Lewis structure of the hydroperoxy radical. [1]

c Deprotonation of the hydroperoxy radical leads to formation of a superoxide radical, O_2^-. Draw the Lewis structure of the superoxide radical. [1]

d Explain why the presence of radicals in the body might lead to formation of tumours. [2]

Q17 This question is about radioactive decay and half-lives.

a Americium-241 emits alpha particles and gamma rays. It is used in smoke detectors and has a half-life of 432 years. A typical smoke detector contains 0.29 micrograms of Americium-241.

i Americium-241 decays by alpha emission. What does it decay into? [1]

ii A 20-year-old smoke detector is found in a landfill. How many grams of americium would you expect it to contain? [3]

b Spent nuclear fuel typically contains about 0.8 per cent plutonium-239. Plutonium-239 is an extremely hazardous substance. It is thought that 1 milligram of plutonium-239 is sufficient to kill a person. Its half-life is 23 400 years.

i Calculate the decay constant for the decay of plutonium-239. [2]

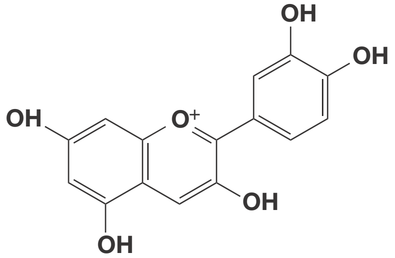
ii How long would it take for a sample of 1 g of spent nuclear fuel to decay to such a level that it was below that necessary to kill a person (i.e. less than 1 milligram of plutonium-239 present)? [2]

c Cerium-144 exists in small amounts in fallout from atomic weapons and radioactive waste from power plants. Its half-life is 285 days. It decays by beta emission.

i What product is formed from the beta emission of cerium-144? [1]

ii A sample of radioactive waste is stored in a cooling pond for 12 weeks before being transferred to long-term storage. What proportion of the cerium has transmuted into the new element proposed in c(i)? [2]

Q18 A simple home-made dye-sensitized solar cell (DSSC) can be made using dyes obtained from natural sources such as fruits or flowers. The dye pictured is called cyanidin and is obtained from the red flowers of *Anthurium*.



a What structural feature of cyanidin makes it suitable for making a DSSC? [1]

b Cyanidin is placed on to a glass slide coated in a thin layer of finely divided titanium dioxide powder.

i What is the purpose of the titanium dioxide powder? [1]

ii Why is the titanium dioxide finely divided? [2]

c An electrolyte is also required, which is usually a mixture of iodide and triiodide ions in equilibrium. Give equations showing how the iodide–triiodide system enables:

i Electrons to be transferred from the cell to the external circuit, by oxidizing the dye. [2]

ii The reduction of the dye in the cell as electrons re-enter the cell. [2]

d A homemade DSSC such as this might reach an efficiency of 0.7 per cent. The best DSSCs in the laboratory have reached 15 per cent efficiency. Suggest ways in which the best DSSCs might differ from those you could construct from simple materials. [3]

25 Option D Medicinal chemistry

ESSENTIAL IDEAS

- Medicines and drugs have a variety of different effects on the functioning of the body.
- Natural products with useful medicinal properties can be chemically altered to produce more potent or safer medicines.
- Potent medical drugs prepared by chemical modification of natural products can be addictive and become substances of abuse.
- Excess stomach acid is a common problem that can be alleviated by compounds that increase the stomach pH by neutralizing or reducing its secretion.
- Antiviral medications have recently been developed for some viral infections while others are still being researched.
- The synthesis, isolation and administration of medications can have an effect on the environment.
- Chiral auxiliaries allow the production of individual enantiomers of chiral molecules.
- Nuclear radiation, whilst dangerous owing to its ability to damage cells and cause mutations, can also be used to both diagnose and cure diseases.
- A variety of analytical techniques is used for detection, identification, isolation and analysis of medicines and drugs.

25.1 Pharmaceutical products and drug action – *medicines and drugs have a variety of different effects on the functioning of the body*

Drugs

A **drug** (Figure 25.1) may be defined as any substance (natural or synthetic) that, by its chemical nature, has an influence on the physical or mental functions of the body. Its effects may include one or more of the following: altering incoming sensory perceptions, for example hallucinogens such as LSD ('acid') and mescaline; altering mood or emotions, for example Valium and Mogadon; altering the physiological state of the body, including consciousness, for example alcohol and LSD; altering activity levels, for example amphetamines ('speed'), or muscular coordination, for example alcohol.

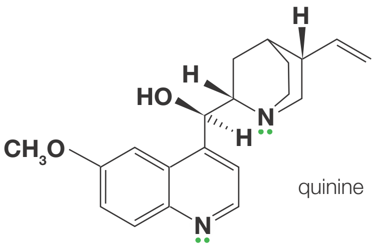
Figure 25.1 Propecia: a medicinal drug used to treat male pattern baldness (by reducing testosterone levels)

A **medicine** or pharmaceutical is defined as a drug that leads to an improvement in health. These include mild **analgesics**, or painkillers, such as aspirin, ibuprofen and paracetamol (acetaminophen); powerful analgesics such as morphine and codeine; drugs used for cancer treatment, such as cisplatin and Taxol; antivirals, such as zidovudine (Retrovir) used to treat people with HIV; **antibiotics** such as penicillin; and cimetidine (Tagamet) and ranitidine (Zantac) to treat gastric ulcers.

Not all drugs are pharmaceuticals: for example, caffeine, present in coffee and tea, and nicotine, present in cigarette smoke. Some former pharmaceuticals are no longer prescribed in many countries because of their highly addictive nature. These include cocaine and heroin, both formerly used as painkillers in the 1930s. Many drugs are described as OTC (over-the-counter) medications and do not require a prescription. Examples include the painkillers aspirin, ibuprofen and paracetamol (acetaminophen).

Many drugs fall into the following classes: **depressants**, such as ethanol (alcohol) at low concentrations which leads to relaxation; **stimulants**, which result in wakefulness and a sense of well-being, for example amphetamines; narcotics and analgesics, for example morphine; hallucinogens which produce altered perceptions, for example LSD; and psychotherapeutics which are used for the control of mental problems. Some drugs, such as caffeine, nicotine and marijuana, fit into more than one category.

Natural products are chemical compounds that are isolated from living organisms, for example plants and trees, using chemical techniques such as solvent extraction and

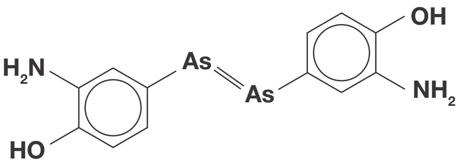


■ **Figure 25.2** Structure of quinine

chromatography. Some natural products are limited to a single species of plant or tree. For example, the anti-malarial drug quinine (Figure 25.2) is a natural product that is found only in the bark of the cinchona tree. Taxol (see section 25.7) is another natural product from the Pacific yew tree.

Chemotherapy, in its most general sense, refers to the treatment of disease by chemicals. The German microbiologist Paul Ehrlich (1854–1915) is regarded as the 'father of chemotherapy'. His early interests in the staining of tissues and cells with azo dyes caused him to consider the idea that certain chemicals might have a particular affinity for certain types of cells. In particular, he was keen to find chemicals, which he dubbed 'magic bullets', that would bind to bacterial cells and be toxic to them, but leave the cells of the host organism unaffected. One of his most famous 'magic bullets' was Salvarsan (Figure 25.3), used to treat syphilis and sleeping sickness until it was replaced by antibiotics in the 1940s.

■ **Figure 25.3** Structure of Salvarsan



International availability of drugs

Over-the-counter (OTC) drugs are medicines sold directly to a customer without a prescription from a healthcare professional (GP or doctor), while prescription drugs may be sold only to customers with a valid prescription. In many countries, OTC drugs are selected by a regulatory agency to ensure that the ingredients are safe and effective when used without a doctor's supervision. OTC drugs are usually regulated by active pharmaceutical ingredients (APIs), not final products. By regulating APIs instead of specific drug formulations, governments allow manufacturers freedom to formulate ingredients, or combinations of ingredients, into proprietary mixtures.

The term over-the-counter may be misleading, since, in many countries, these drugs are often located on the shelves of stores like any other packaged product. In contrast, prescription drugs are almost always passed over a counter from the pharmacist to the customer. Some drugs may be legally classified as over-the-counter (that is, no prescription is required), but may only be dispensed by a pharmacist after an assessment of the patient's needs or the provision of advice to the patient. In many countries, a number of OTC drugs are available in establishments without a pharmacy, such as general stores, supermarkets and petrol (gas) stations. Regulations detailing the establishments where drugs may be sold, who is authorized to dispense them, and whether a prescription is required vary considerably from country to country.

For example, Arcoxia, which is a strong anti-inflammatory, is non-FDA (Food and Drug Administration) approved in the USA but marketed by Merck outside the USA. It is prescription only in the UK but available OTC in Thailand. It can be ordered online in New Zealand.

Nature of Science

Classification of pharmaceuticals

Although there are thousands of different drugs, all drugs marketed in the USA fall under one or more of the first tier of the American Hospital Formulary Service (AHFS) Pharmacologic-Therapeutic Classification System. The classification was developed and is maintained by the American Society of Health-System Pharmacists (ASHP), a national association of pharmacists. One important tier of drugs under this classification are the anti-infective agents which includes the penicillins and antivirals.

In the USA, the legal classification of drugs was initiated under the Controlled Substances Act. Drugs fall within different schedules based on their potential for abuse. Some of the drugs are available only by prescription and some are available OTC.

- *Schedule 1 drugs* have high potential for abuse, are not recognized for medical use and pose a safety risk. These drugs include heroin, lysergic acid diethylamide (LSD), MDMA (ecstasy) and marijuana.
- *Schedule 2* drugs have a high potential for abuse, have medical effects and pose a high risk for dependency (addiction). Drugs categorized as Schedule 2 include opium, morphine, cocaine, methadone and methamphetamine.
- *Schedule 3* drugs have a lower potential for abuse, have medical uses and pose a moderate risk for dependency. Amphetamine, barbiturate, Valium, anabolic steroids and codeine are Schedule 3 drugs.
- *Schedule 4* drugs have limited potential for abuse, have high medical uses and pose a limited risk for dependency. This category includes chloral hydrate and phenobarbital.
- *Schedule 5* drugs pose minor problems and are generally preparations of drugs containing limited amounts of Schedule 1 through 4 drugs. Cough medicines with codeine are an example of Schedule 5 drugs.

Pharmacology

Pharmacology is the scientific study of the interactions of drugs with the various different types of cells found in the human body. Pharmacologists seek to understand the various biochemical changes produced by drugs. Many of the effects of drugs are a result of changes in intercellular communications.

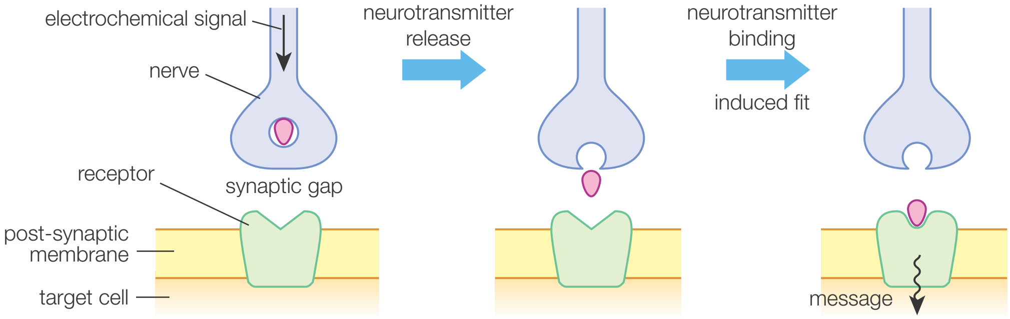
There are three modes of communication between cells:

- neurotransmission, where a nerve cell (neuron) passes a chemical signal on to another nerve cell or to a muscle or gland cell;
- hormonal, where hormones (‘chemical messengers’) are released from endocrine glands and carried by the blood to a distant ‘target’ organ;
- **autacoid**, where ‘local hormones’ are released and act on nearby cells.

Neurotransmission

A nerve impulse involves the movement of sodium and potassium ions across the cell membrane of the axon (nerve fibre). However, there is no direct connection between axons or between axons and muscle or gland cells. At the junctions or ‘gaps’ between these cells, the electrochemical impulse of the axon is converted to a chemical signal. A synapse is a junction between two neurons (see Figure 25.4).

Figure 25.4 The structure of a synapse

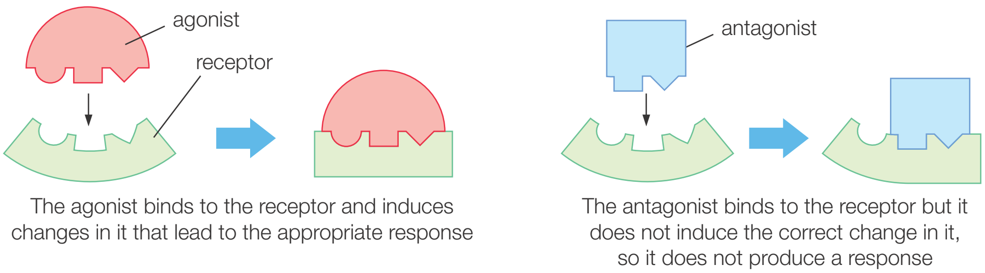


The impulse causes the release of a **neurotransmitter**, a chemical that diffuses across the synapse and then binds to the binding site of a specific receptor (compare this with the lock-and-key hypothesis of enzyme action in Chapter 23). These receptors are usually sugar-containing proteins (glycoproteins) located on the surface of the post-synaptic membrane.

Many drugs, for example nicotine and caffeine, also fit into the various receptors. Those that produce a response are called **agonists**, and those that bind tightly to the receptor and block it without producing a response are called **antagonists** (Figure 25.5). Neurotransmitters function by causing changes in the permeability of the target cell membrane to sodium, potassium, calcium and chloride ions. Once a new impulse has been generated the neurotransmitter is broken down by enzymes or ‘recycled’ back to the neuron from where it was released.

Many drugs function by reducing or increasing the release of neurotransmitters, by binding to their receptors, or by changing the permeability of neuronal cell membranes, mimicking or copying the action of the neurotransmitter, or by altering the re-uptake or ‘recycling’ by the neuron.

Figure 25.5 An illustration of an agonist and an antagonist binding to a receptor site on a protein molecule


agonist
receptor
The agonist binds to the receptor and induces changes in it that lead to the appropriate response
antagonist
The antagonist binds to the receptor but it does not induce the correct change in it, so it does not produce a response

Autacoids

Two well-studied local hormones, or **autacoids**, are histamine and serotonin. Histamine is released from cells in the gastric mucosa, or stomach lining, in response to eating and is a powerful stimulant of gastric juice production. It is also released by certain white blood cells in the lungs, called mast cells, in response to an allergic reaction, for example to pollen grains. Breathlessness, a runny nose and the watery eyes that accompany hay fever are two consequences of its action inside the body. Serotonin (5-hydroxytryptamine) is also produced in the gastric mucosa, but it is also produced by certain neurons in the central nervous system (brain and spinal cord) where it is involved with the control of sleep and vomiting. A deficiency within the brain may lead to migraine and depression.

Other local hormones are the prostaglandins produced in tiny quantities by many cells, which are involved in controlling gastric acidity, labour, blood platelet aggregation during blood clotting and bronchoconstriction (a reduction in the diameter of tiny air tubules, or bronchioles, inside lung tissue). They are also involved indirectly in inducing pain, fever and inflammation (redness and swelling).

Signal transduction

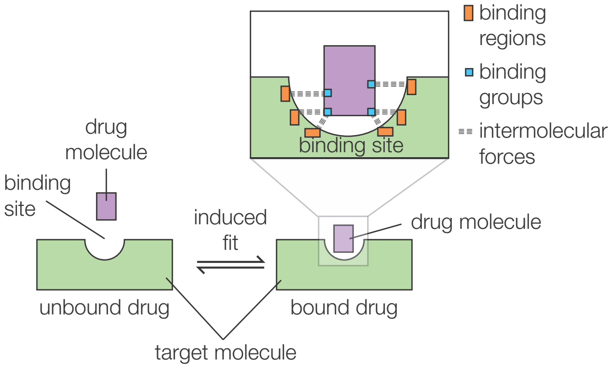
All these chemical messengers, whether they be neurotransmitters, hormones or autacoids, carry what is termed a primary signal to a cell. A process called **signal transduction** then occurs, where the primary signal induces some kind of response by the cell.

For example, many of the neurotransmitters bind to receptors in the cell membrane and this results in the opening of ‘tunnels’ across the membrane that allow sodium ions to pass into the cell. Some hormones, for example insulin, bind to receptors that activate enzymes in the cell membrane. Other primary signals may cause the production of a so-called secondary messenger, for example cyclic AMP (cAMP), which inhibits or activates enzymes inside the cell. Finally, steroid hormones, for example oestrogen, pass through the cell membrane and enter the nucleus. Here they bind to a protein receptor, which then binds to DNA, ‘switching on’ a variety of genes.

Receptors and their interactions

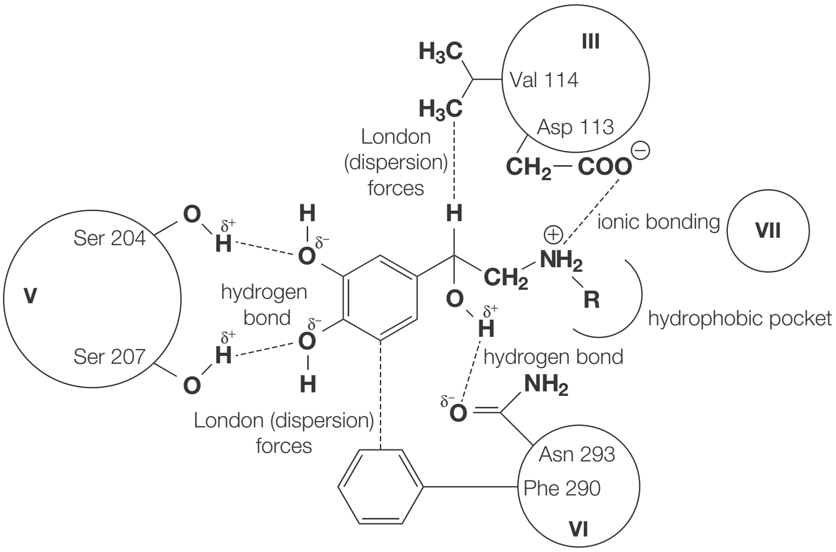
A **receptor** is a protein on the cell membrane or within the cytoplasm or cell nucleus of the target cell that binds reversibly to a specific molecule to initiate a biochemical or cellular response. Receptors interact and bind to a specific substrate to induce a change. The interaction is based upon the drug and receptor having complementary shapes that allow a variety of bonds and intermolecular forces to operate. The substrate can be a molecule or a hormone (steroid- or peptide-based). The drug enters the active site of the receptor and binds to it, often causing a change in shape of the receptor (Figure 25.6). If the receptor is an enzyme many drugs act as inhibitors, reducing the activities of enzymes via competitive or non-competitive inhibition.

■ **Figure 25.6** The equilibrium of a drug being bound and unbound to its target molecule



The forces responsible for receptor interactions include covalent bonding, ionic interactions, ion–dipole interactions, hydrogen bonding and London (dispersion) forces (Figure 25.7).

■ **Figure 25.7** Various electrostatic interactions between a drug molecule and its protein receptor (target). The drug is noradrenaline, a hormone and neurotransmitter



Covalent bonding is not a common force of interaction between receptors and drugs because covalent bonds are usually strong bonds. Once formed, they are usually difficult to break, and irreversible bonding occurs. A number of poisons, such as cyanide ions, function in this way.

Ionic interactions are strong attractive forces, involving the electrostatic forces of attraction between opposite charges. These interactions usually involve charged amino acid residues, such as negatively charged aspartic acid or positively charged lysine.

An ion–dipole interaction is the electrostatic attraction between a charge and an electron-deficient or electron-rich centre. It is weaker than an ionic interaction since a partial rather than a full charge is involved.

Hydrogen bonding is the relatively strong intermolecular force operating between a hydrogen atom bonded to a small and highly electronegative element (fluorine, oxygen or nitrogen) and the lone pair of that element. In proteins and receptors, most of the hydrogen bonds are due to the presence on molecules of hydroxyl groups, –OH, carboxylic acid groups, –COOH, peptide bonds (–CONH–) or the primary amine group (–NH_2).

The final receptor interactions are London (dispersion) forces. In non-polar molecules, the temporary dipoles are generated from random fluctuations in the distribution of electron clouds of molecules. These temporary dipoles then induce another dipole in the neighbouring molecules, which spreads through the sample of molecules. London (dispersion) forces are only significant at short molecular distances and when the molecules involved in receptor binding are relatively large. Polar molecules, in contrast, have permanent dipoles because of an uneven distribution of electrons due to a difference in electronegativity. The permanent dipoles of polar molecules attract each via electrostatic forces of attraction.

The placebo effect

A **placebo** is an inert chemical used as a control when testing a drug clinically. The **placebo effect** is the pharmacological effect on a person, or group of people, who have been given a placebo rather than an active drug.

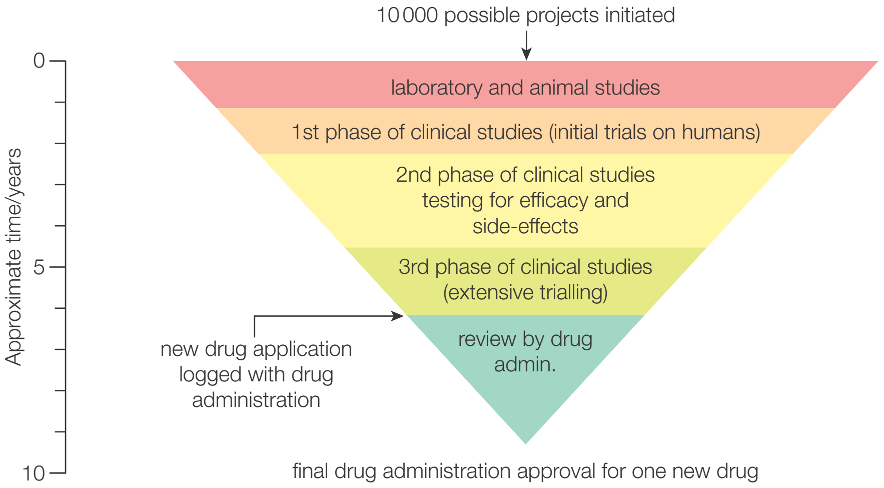
It is not understood why chemically inert substances can be effective in treating illness, but it is thought that a person's beliefs and hopes about a treatment can have a biochemical effect on the body, presumably via the endocrine and immune systems. This means that a person's mental attitude may be very important in determining whether he or she recovers from injury or illness. It is thought that the placebo effect triggers natural healing processes in the body.

Developing new drugs

Medicinal drugs have to be tested in clinical trials (Figure 25.8) to determine if they are safe. Clinical trials involve recording the results of treating a group of patients with the drug and comparing them to a group of similar patients receiving a different treatment or a placebo.

Before clinical testing begins, researchers analyse the drug's physical and chemical properties in the laboratory and study its pharmacological and toxicological properties on animals and cells grown in culture.

Figure 25.8 Summary of the steps in the development of a new drug


10 000 possible projects initiated
0
5
10
Approximate time/years
laboratory and animal studies
1st phase of clinical studies (initial trials on humans)
2nd phase of clinical studies testing for efficacy and side-effects
3rd phase of clinical studies (extensive trialling)
review by drug admin.
new drug application logged with drug administration
final drug administration approval for one new drug

In vivo ('in the living') tests involve the use of live animals, whereas *in vitro* ('in glass') tests do not. *In vitro* tests may involve testing potential drugs on cells grown in culture. *In vitro* testing is generally quicker and cheaper than *in vivo* testing. However, many drugs that are active *in vitro* turn out to be inactive *in vivo*. This is because a drug may not be able to reach its target in the human body.

If the laboratory and animal study results are promising, then the drug company will apply to conduct initial studies where the drug is administered to a small number of healthy volunteers to assess its safety, dosage and **side effects**. A side effect of a drug is any effect that is in addition to its intended effect, especially an effect that is harmful or unpleasant. If there are no major problems, such as unacceptable toxicity or severe side effects, then clinical trials will begin and the drug is tested in patients who have the condition it is intended to treat.

The principles of a clinical trial are simple: groups of patients are recruited and a drug is administered to see if there is an improvement in their health or survival rate. However, drugs do not usually miraculously reverse fatal illnesses and although they reduce the risk of death, they do not eliminate it. In addition, many viral diseases, such as influenza (flu), are spontaneously dealt with by the body's immune system, and some, such as asthma (caused by environmental and genetic factors), follow an unpredictable path. Some drugs, however, are easy to evaluate since they alter a clinical measurement, for example captopril reduces blood pressure.

A further difficulty in assessing the effectiveness of a drug is that some measurements of a disease, such as pain, are subjective, relying on the judgement of both doctor and patient. Such measurements can also be influenced by the doctor's or patient's expectations – the placebo effect.

To minimize these problems, controlled clinical trials are conducted where one group receives the drug under investigation and a group of similar patients receives either a different dose of the drug, another drug, a placebo or no treatment at all. The test and **control groups** are studied at the same time, but in some clinical trials 'historical controls' are conducted where patients given the drug are compared with similar patients treated with the control drug at a different time and place.

It is important that the treatment and control groups should be as similar as possible in characteristics such as age, weight, sex and general health. An important method used to achieve this is called randomization, where patients are assigned randomly to either the treatment or the control group. Randomization helps to reduce the possibility of 'selection bias' where healthier patients are selected to receive the new drug.

In conjunction with randomization, a design feature known as 'blinding' helps ensure that bias does not distort the results of the clinical trials. Single-blind studies involve not informing the patients whether they are receiving the new drug or a placebo. The more common **double-blind studies** are where the patients, the doctors and the drug company analysts do not know which patients received the drug. The study is 'unblinded' at the completion of the study, when assignment codes are deciphered to reveal the treatment and control patients.

The first stage of clinical trials, known as phase 1, is carried out on healthy human volunteers to provide a preliminary evaluation of the drug's safety and its dose levels. During the study, the volunteers do not take medication, caffeine, alcohol or cigarettes. This is to avoid any complications that might arise due to drug–drug interactions. In some cases, phase 1 studies are carried out not on healthy volunteers but on volunteer patients. This occurs when the drug is potentially toxic and is used to treat AIDS or cancer.

For ethical and economic reasons, pharmaceutical trials and studies must be kept to a minimum, which normally includes the determination of therapeutic doses, therapeutic window and specific drug interactions.

ToK Link

Drugs trials use double-blind tests. When is it ethically acceptable to deceive people?

The first placebo-controlled trial was probably conducted in 1931, when sanocrysin was compared with distilled water for the treatment of tuberculosis (a bacterial infection of the lungs). Ever since then, placebo-controlled trials have been controversial, especially when patients randomly assigned to receive a placebo have forgone effective medical treatments. Recently, the debate has become polarized. One view, termed 'placebo orthodoxy' by its opponents, is that methodology considerations make placebo-controlled trials necessary. The other view, which might be called 'active-control orthodoxy', is that placebo orthodoxy sacrifices ethics and the rights and welfare of patients to the gain of scientific knowledge.

Supporters of placebo-controlled studies argue that it is ethical to conduct such trials even in the case of medical conditions for which there are drugs known to be effective, because of the limitations of the methodology in the trials in which active treatment is used as the control group. Sometimes drugs that are known to be effective are no better than placebo in particular trials because of variable responses to drugs in particular populations, unpredictable and small effects, and high rates of spontaneous improvement in patients. Hence, without a placebo group to ensure validity, the finding that there is no difference between the investigational and standard treatments can be misleading or uninterpretable. New treatments that are no better than existing treatments may still be clinically valuable if they have fewer side effects or are more effective for particular subgroups of patients. However, no drug should be approved for use in patients unless it is clearly superior to placebo or no treatment. Supporters of such studies acknowledge that they are unethical in some circumstances, especially when withholding an effective treatment might be life-threatening.

Because of these problems, some people have attacked placebo orthodoxy as unethical. Supporters of active controls propose that whenever an effective drug for a condition exists, it must be used in the control group. Furthermore, they argue that placebo controls are inappropriate because the medical question is not whether a new drug is better than nothing but whether it is better than standard treatments.

Thalidomide

Thalidomide was a drug introduced in 1957 in West Germany by the pharmaceutical company Chemie Grünenthal. It was used as a hypnotic and sedative and was found to give relief from morning sickness in pregnant women.

It was subjected to some types of toxicity tests before it was marketed, but critics have stated these were superficial and incomplete. In particular, the drug was not tested for birth defects, because this was not required by the law at that time.

In early 1961 a sudden increase in the incidence of phocomelia (Figure 25.9), a failure in the development of the long bones of the arms and legs, was noticed. The drug was withdrawn by the end of 1961, but by then an estimated 10 000 deformed babies had been born.

However, soon after thalidomide was launched in West Germany reports were noted of a condition called peripheral neuropathy, which is a result of damage to the nervous system. It is characterized by a prickly feeling followed by numbness and coldness, followed by severe muscular cramps and a lack of coordination. This note of peripheral neuropathy was sufficient for the US FDA to refuse approval for the drug in the USA.

Thalidomide is an example of a teratogenic drug, indicating that it produces malformed fetuses. The drug has no effect in rats and mice even at very high doses. However, effects in rabbits are found at very high concentrations. Aspirin is a potent teratogen in rats, mice and hamsters and not recommended for use during pregnancy. Synthetic products such as thalidomide and aspirin are chemical compounds that are made by chemists in a laboratory.

Figure 25.9 The long bones in this little boy's arms failed to develop as a result of his mother taking thalidomide during pregnancy

Administering drugs

Oral administration

This method of drug delivery is the most popular as it is simple and cheap. A drug taken by mouth enters the digestive system before being absorbed into the bloodstream. However, the stomach contains strongly acidic gastric juices which will hydrolyse many drugs. These acid-labile drugs can be protected to some extent by covering them with an acid-resistant polymer called an enteric coating.

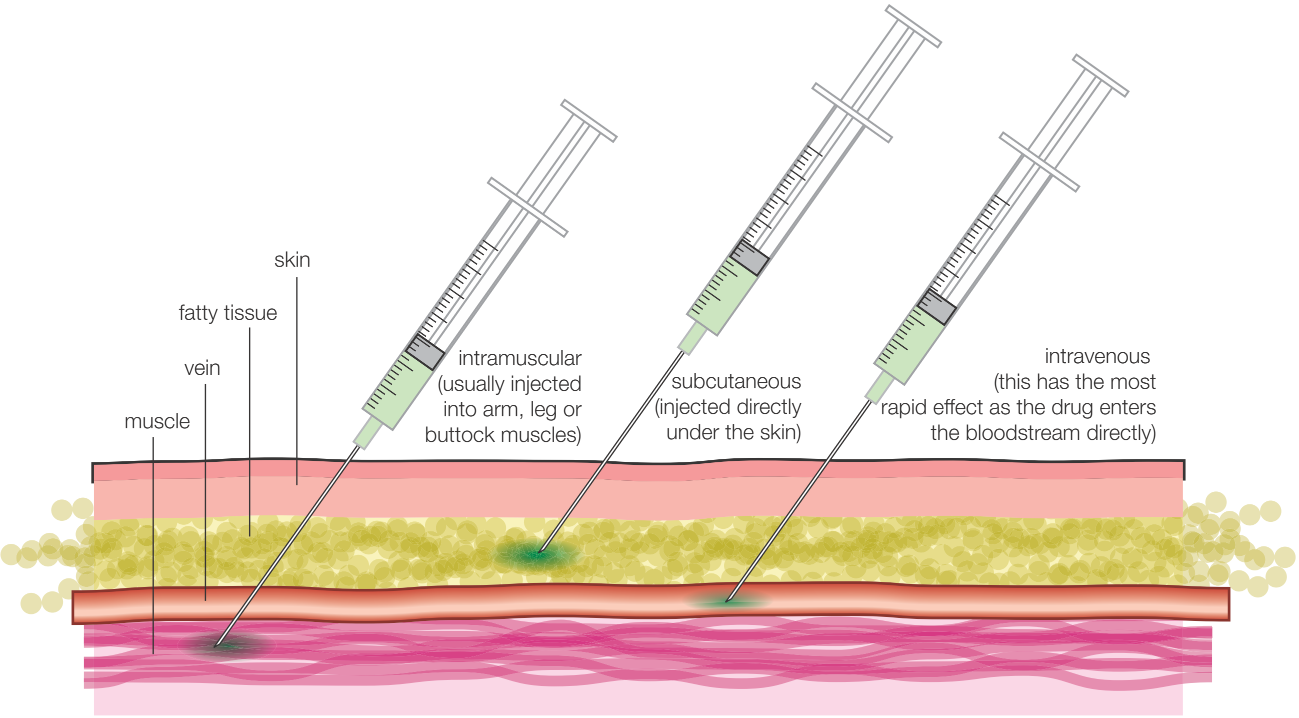
To minimize the problem of drug decomposition, many drugs are taken at mealtimes. When food is present in the stomach the pH is higher, so acid attack on the drug and its coating is reduced. Few drugs, except ethanol ('alcohol'), are absorbed in the stomach since the stomach wall is lined with an insoluble mucous layer that prevents absorption.

Most drugs are absorbed in the small intestine, which has a large surface area due to the presence of millions of tiny finger-like projections called villi. Hydrogen carbonate ions are also secreted from the wall of the intestine to neutralize the acidic stomach contents. Enteric coatings are designed to dissolve in the neutral or slightly alkaline environment of the small intestine. Many drugs, for example aspirin, are designed to ionize and become more soluble as they pass from the stomach to the small intestine.

Drugs absorbed from the small intestine into the bloodstream pass via the hepatic portal vein to the liver. This organ produces a variety of enzymes designed to protect the body from poisons. Many drugs are slowly decomposed by the liver, but this can be remedied by ensuring a high oral dose is given, thereby ensuring a therapeutically or medically active amount remains in the blood after passing through the liver. However, some drugs are designed to be swallowed in an inactive form called a **prodrug** which is activated by a liver enzyme into an active form. Such an approach is used particularly for toxic anti-cancer drugs.

Injection

Some drugs cannot be given by mouth, for example insulin used to treat diabetes, which is a protein and would be broken down by the acid and enzymes present in gastric juice. Others may be broken down rapidly by the liver. Such drugs are given **parenterally**, or by injection. There are three methods of injection: **intravenously**, by injection into a vein; intramuscularly, into a muscle; and subcutaneously, under the skin (Figure 25.10). For example, insulin is injected subcutaneously.


skin
fatty tissue
vein
muscle
intramuscular (usually injected into arm, leg or buttock muscles)
subcutaneous (injected directly under the skin)
intravenous (this has the most rapid effect as the drug enters the bloodstream directly)

■ **Figure 25.10** Summary of common methods of drug delivery by injection

Suppositories

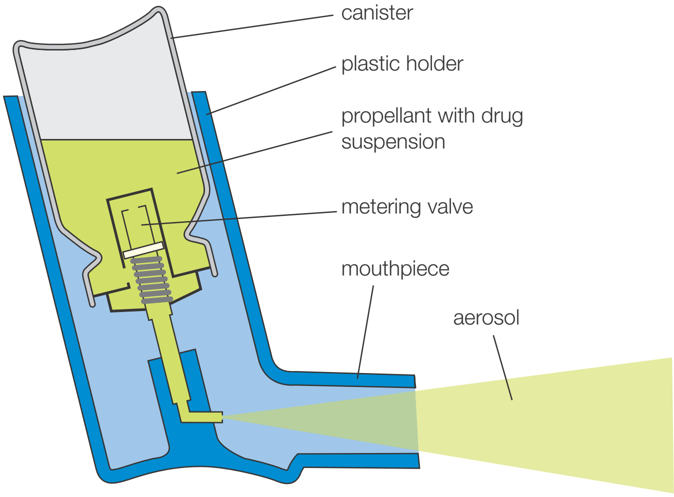
To minimize metabolism in the liver, some drugs are inserted into the vagina or rectum, since veins in these organs do not carry absorbed food and hence do not pass directly through the liver – most of the blood and dissolved drug is transported to other body tissues.

Drugs administered via the vagina and rectum are given in the form of pessaries or **suppositories**. They are made from waxy polymers that slowly melt at body temperature, releasing the drug and allowing it to be absorbed across the mucous membranes into the bloodstream.

Fat-soluble drugs can be slowly released through the skin by means of a 'patch' where the drug is suspended in a polymer matrix behind a protective coat. Such an approach is used for female contraceptives (Chapter 23), hormone replacement therapy (HRT) and anti-smoking aids. Some topical medications (ointments, pain-relieving gels) are absorbed through the skin.

Inhalation

An increasingly important method of administering drugs is via inhalation. A familiar example is the drug called Ventolin which is given to asthma patients to dilate their bronchioles. A pressurized delivery device (Figure 25.11) ensures that a fine mist of droplets of the correct size reaches the lungs.



■ **Figure 25.11** Cross section of metered dose inhaler (MDI)

Types of membranes

There are a number of membranes that drugs may pass when present in the human body. They differ in their degree of resistance to the movement of drugs.

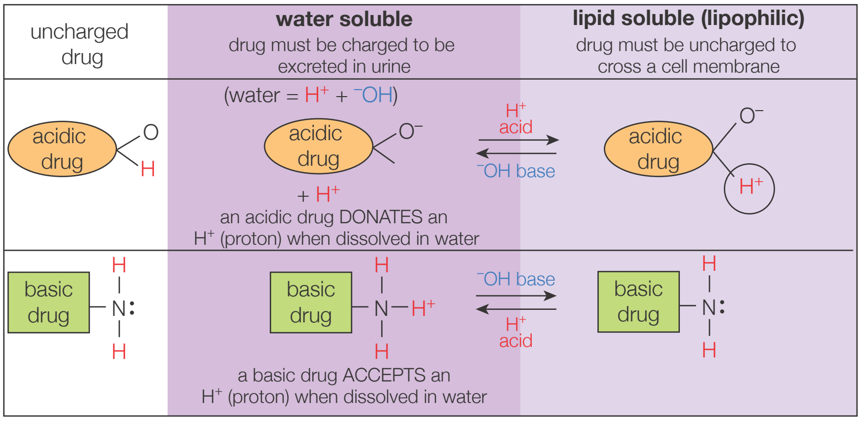
- *Cell membranes:* this barrier is permeable to many drug molecules but not to others, depending on their lipid solubility. Small pores permit small molecules such as alcohol (ethanol) and water to pass through.
- *Walls of capillaries:* pores between the epithelial cells are larger than most drug molecules, allowing them to pass freely, without lipid solubility being a factor.
- *Blood–brain barrier:* this barrier provides a protective environment for the brain. The rate of transport across this barrier is limited by the lipid solubility of the psychoactive molecule.
- *Placental barrier:* this barrier separates the mother from her baby but is very permeable to lipid-soluble drugs.

Solubility of drugs

Drugs that are water soluble are often composed of molecules that are ionized and hence have an electrical charge. They are able to cross through pores in capillaries, but not cell membranes. Lipid-soluble drugs are uncharged and not ionized.

The dissociation constant, K_a or pK_a indicates the pH where 50 per cent of the drug (X) is ionized (water soluble) and 50 per cent non-ionized (lipid soluble); pK_{eq} = pH + $\log_{10}$[X]ionized/[X]non-ionized. This affects a drug's solubility, permeability, binding and other characteristics (Figure 25.12).

Figure 25.12 Water-soluble and lipid-soluble drugs


uncharged drug
water soluble
drug must be charged to be excreted in urine
lipid soluble (lipophilic)
drug must be uncharged to cross a cell membrane
(water = H⁺ + ⁻OH)
acidic drug
O
H
acidic drug
O⁻
+ H⁺
H⁺ acid
⁻OH base
acidic drug
O⁻
H⁺
an acidic drug DONATES an H⁺ (proton) when dissolved in water
basic drug
H
N:
H
basic drug
H
N—H⁺
H
⁻OH base
H⁺ acid
basic drug
H
N:
H
a basic drug ACCEPTS an H⁺ (proton) when dissolved in water

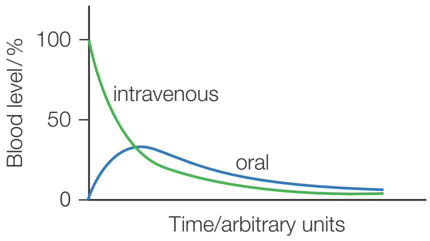
Drug distribution

Depending upon its route of administration and target area, every drug has to be absorbed, by diffusion, through a variety of bodily tissues. Tissue is composed of cells that are surrounded by membranes, consisting of three layers, two layers of water-soluble complex lipid molecules (phospholipids) and a fluid layer of lipids, sandwiched within these layers. Suspended within the layers are large proteins, with some, such as receptors, spanning all three layers.

The permeability of a cell membrane for a specific drug depends on a ratio of its water to lipid solubility. Within the body, drugs may exist as a mixture of two interchangeable forms, either water-soluble (ionized-charged) or lipid-soluble (non-ionized). The concentration of two forms depends on characteristics of the drug molecule (pK_a, pH at which 50 per cent of the drug is ionized) and the pH of fluid in which it is dissolved. In water-soluble form, these drug molecules cannot pass through lipid membranes, but to reach their target area, they must permeate a variety of types of membranes.

Drug administration methods

Slow absorption methods for drugs include the following methods of drug administration: orally, through mucous membranes, topical/transdermal (through the skin) and rectally (suppository). Faster absorption is achieved through the various types of injections and inhalation (via the lungs). The fastest method of absorption is directly into the brain. The general principle of drug



■ **Figure 25.13** Drug concentration in the bloodstream as a function of time

delivery to the brain is the faster the absorption, the quicker the onset and the higher the addictiveness, but the shorter the duration.

In order to reach the target organ, most drugs have to enter the bloodstream, which may be an issue if a drug is lipophilic (non-polar) and has limited solubility in water or has a slow absorption rate from the stomach or small intestine when it is administered orally. When a drug is administered intravenously (via injection) its bioavailability is 100 per cent, but other routes of administration, such as oral, decrease its bioavailability (Figure 25.13) due to incomplete absorption and chemical degradation. The bioavailability refers to the fraction or percentage of drug that is available in the bloodstream.

ToK Link

Naming of drugs

The same drug can be identified by different names. Are names simply labels or do they influence our other ways of knowing?

One of the greatest human skills is the ability to name things. We name and classify the members of the natural world into categories. We saw a tall plant and we defined that plant a 'tree'. Then we looked at some of the appendages sticking out from the tree and we defined them as branches. And the flat green parts that sprout out of the branches we called leaves. A tree is a living organism that includes branches and leaves. Branches sprout leaves and leaves grow on branches and trees.

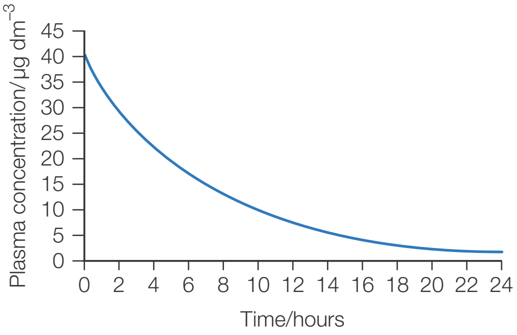
Trees, branches and leaves already existed before humans named them. We are not creating but just labelling biological objects that already exist. Did the things that we named exist before we named them? Or more precisely, in what sense did they exist before they were named, and how did their existence change after being named? If we think about trees it is easy to assume that our naming it a tree does not change anything. It is not an act of creation.

However, consider a newly synthesized anti-cancer drug. Many of these drugs have reassuring complex and unfamiliar chemical names, for example, sunitinib (technical name) or Sutent (commercial name). Patients might be less willing to take this drug if it was renamed as 'last chance lung-cancer treatment'. Emotion has a powerful effect on drug names.

The names we give to drugs and many other objects are inevitably not just labels – they are interpretations. And we act on the basis of those interpretations, and those interpretations define reality for us. We act as if the distinctions we make in the world are real and once these distinctions become fixed in our minds it is very hard to view the world in any other way.

■ Drug kinetics

Drugs do not remain in the body. They may be broken down or converted to an unreactive product. The drug and its metabolites are excreted via the lungs (as with inhaled anaesthetics), through the skin (rarely and to a small extent) or in the urine and faeces.



■ **Figure 25.14** Exponential decay of a drug with a half-life of five hours

As time passes the concentration of drug in the body fluids, such as the blood, will fall. Many drugs are lost (eliminated) in a simple first-order reaction, in which the rate of the loss of the drug is proportional to its concentration:

$$\text{Rate} = k[\text{drug}]$$

where square brackets denote concentration in moles per cubic decimetre and k the rate constant. This results in an exponential decay of the drug concentrations with time. Figure 25.14 plots the concentrations of a drug in the blood at various times after a single intravenous injection. A characteristic feature of first-order reactions is their constant half-life; the curve in Figure 25.14 has a half-life of five hours.

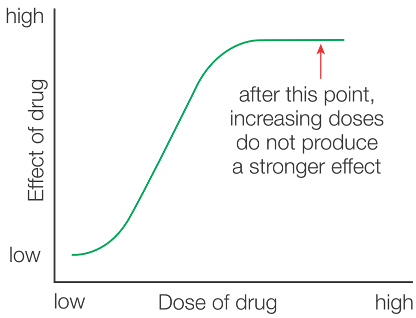


Figure 25.15 Dose–response curve

Dosage

The dose–response curve shows the relationship between the drug dose and the magnitude of the drug effect (Figure 25.15). It should be noted that drugs can have more than one effect and drugs vary in effectiveness: they have different sites of actions and different affinities for receptors. The effectiveness of a drug is considered relative to its safety (therapeutic index).

Lethal dosage

Tests are performed on all pharmaceuticals (medicinal drugs) to establish their toxicity. Such data is normally reported in the form of LD_{50} values (**lethal dose, 50%**) which indicate the quantity of the drug or other substance that, if administered to a group of animals, will kill 50 per cent of them.

LD_{50} values are normally expressed in milligrams (mg) of drug per kilogram (kg) of the test animal's body weight. The method of administration, for example injection into the blood or via the mouth (oral), should also be recorded, as should the subject animal.

A large LD_{50} value means that the substance is relatively non-toxic and that a large quantity of the substance is required to cause a toxic response. Conversely, a small LD_{50} value means that the substance is relatively toxic and that only a small quantity of the substance is needed to cause a toxic response. The important point to note is that all substances are potentially poisonous: it is only the dose that determines whether a substance is poisonous. The concepts of toxicology imply that no drug is 100 per cent safe.

Table 25.1 shows some examples of LD_{50} values. In the case of caffeine the LD_{50} value is achieved by taking about 50 caffeine pills or drinking between 100 to 200 normal cups of coffee in one hour. The LD_{50} value for humans is estimated from the value for rats, allowing for difference in body weight and sensitivity to caffeine.

Table 25.1 LD_{50} values for selected substances

Chemical	LD_{50}/mg kg^{-1}	Subject, route
Sucrose	29 700	Rat, oral
Ethanol	7060	Rat, oral
Sodium chloride	3000	Rat, oral
Caffeine	192	Rat, oral
Sarin (a nerve gas)	25	Human, subcutaneous (via skin)
Sodium cyanide	6.4	Rat, oral

Toxic dosage and effective dosage

In toxicology, the toxic dose, TD_{50}, is the dose at which toxicity occurs in 50 per cent of humans in a test sample of a drug or chemical. The type of toxicity needs to be specified for this value to have any practical meaning. Since toxicity need not be lethal, the TD_{50} value is generally loss than the lethal dose, LD_{50}. The LD_{50} value can be regarded as an upper boundary on the TD_{50} value. An effective dosage, ED_{50}, in pharmacology is the amount of drug that produces a therapeutic effect in 50 per cent of the human subjects taking it.

Therapeutic index

The ED_{50} of a drug is the dose required to produce the desired effect, for example reduce blood pressure, in 50 per cent of test animals or patients. The ratio of LD_{50} to ED_{50} is known as the **therapeutic index**. A therapeutic index of 10 indicates an LD_{50}:ED_{50} ratio of 10 to 1. This means a tenfold increase in the dose corresponding to the ED_{50} would result in a 50 per cent death rate.

The dose–response curves for a drug's therapeutic and lethal effects can be compared (Figure 25.16) to determine whether the therapeutic ratio is safe. The curves should not overlap on the *x*-axis, which means that the shallower the slopes of the curves, the higher the risk attached to use of the drug.

Figure 25.16 shows the therapeutic and lethal dose–response curves for a drug that cures a particular animal illness. A 50 mg dose of the drug will cure 95 per cent of the test animals, but will kill 5 per cent of the animals. Obviously, this drug would not be tested on human patients because of its high therapeutic window (10) (see below on this page). A safer drug would have to be developed that has a considerably lower therapeutic window, for example in the region of 1.

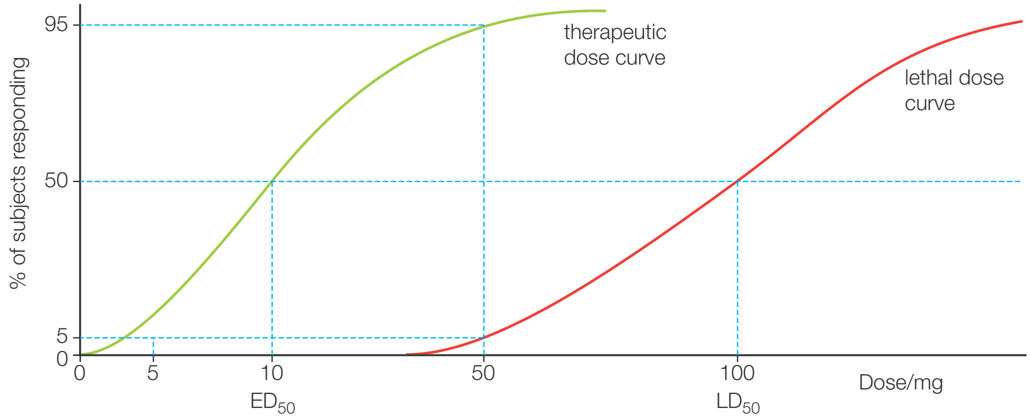


Figure 25.16 Comparison of therapeutic and lethal dose curves for a pharmaceutical

Therapeutic windows are relevant to animal testing of drugs, but they do not take into account any non-lethal side effects. Hence, the therapeutic dose levels that are used on patients should have a minimum risk of side effects or long-term toxicity.

Therapeutic window

Drugs are usually taken at regular time intervals, and the doses are designed to keep the blood levels of the drug within a maximum and minimum level such that they are not too high to be toxic but not so low as to be ineffective. The **therapeutic window** is the range of dosages between the minimum amount of the drug that produces the desired effect and a medically unacceptable adverse effect.

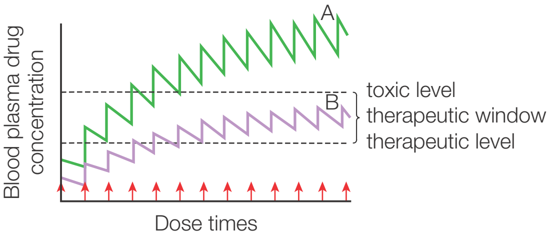


Figure 25.17 Approaches to the dosing of a drug

In general, the concentration of free drug in the blood is a good indication of the availability of a drug at its target. Figure 25.17 shows two different types of dosing. The dosing approach labelled A quickly reaches the therapeutic level but continues to rise to a steady state which is toxic. The dosing approach labelled B involves half the amount of the drug provided with the same time frequency. The time taken to reach the therapeutic level is longer, but the steady-state levels of the drug remain between the therapeutic and toxic levels – the therapeutic window (Figure 25.17).

Drug dosing involving regular administration of a drug works well in most cases, especially if the size of each dose is less than 200 mg and doses are taken once or twice a day.

The experimental foundations for therapeutic index and therapeutic window through animal and human studies

Drugs have therapeutic effects, toxic side effects and, in some cases, lethal effects. Side effects and lethal effects are typically dose-dependent and can be quantified by defining the dose that produces a toxic effect in 50 per cent of the population (TD$_{50}$) and (at least in animals) lethal effects in 50 per cent of the population (LD$_{50}$). All of these measurements are based upon experiments with large numbers of controlled experiments in animals and humans.

The dose relationships for toxic and lethal effects may have different slopes compared to the therapeutic dose–response relationship because they produce effects by different receptors or mechanisms (Figure 25.18). It is of value to know the relative difference between the average toxic dose and the average therapeutic dose. The most common way to define this relationship is using the therapeutic index, which is defined as the ratio between the TD_{50} and ED_{50}. Figure 25.18 also shows the therapeutic window which is a well-defined range of a drug's concentration in the blood at which a desired effect occurs, below which there is little effect, and above which toxicity occurs; the therapeutic window differs among patients and is determined from experimental studies.

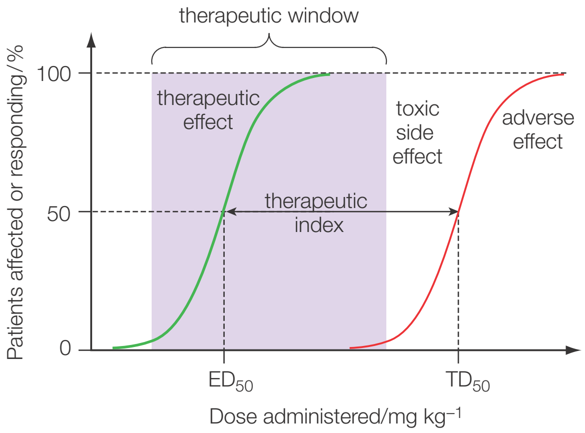


Figure 25.18 The therapeutic index and therapeutic window

Pharmacokinetics

The study of how a drug reaches its target cell, tissue or organ in the body is known as pharmacokinetics. This information is important in the pharmaceutical industry when a new drug is being developed. There is no benefit for medicinal chemists trying to optimize the structure of a new compound to bind and interact with a specific protein receptor, if little or none of the drug reaches the target site.

There are four issues to consider:

- Drug absorption is the route by which the drug reaches the bloodstream from the point at which the drug entered the body.
- Drug distribution refers to the transport of the drug in the bloodstream and then the slower transport to the organs, tissues and cells.
- Drug metabolism refers to the chemical breaking down of a drug due to the action of enzymes, often in the liver.
- Drug excretion refers to the removal of the drug from the body, usually via the urine. The main route for excretion is via the kidneys and bladder.

The aim is to design a drug with an optimum lifetime in the body, long enough to reach its target and be effective, but short enough to ensure its concentration does not build up in the body.

Tolerance

Tolerance towards a particular drug occurs when a larger dose of the drug is required in order to achieve the effect originally obtained by a smaller dose. A classic example is alcohol: a steady consumption of alcohol in alcoholic drinks leads to tolerance.

There are two different time frames of tolerance: acute tolerance and chronic tolerance. Acute tolerance occurs following a single exposure to the drug; chronic tolerance occurs gradually as the drug is taken repeatedly. Both types of tolerance can occur with alcohol, though the latter is more common.

There are two mechanisms by which tolerance to alcohol can develop: metabolic tolerance involving an increase in liver activity, and functional tolerance produced by changes in the central nervous system (brain and spinal cord). Tolerance of the functional type is also observed with nicotine. Smokers develop an increased number of nicotine receptors in the brain. While alcoholics may have a higher tolerance for alcohol than social drinkers or abstainers (those who do not drink), the size of a lethal dose is not very different between these two groups.

Cross-tolerance can also develop between alcohol and many other sedative or hypnotic drugs. This means that tolerance to alcohol also produces tolerance to other drugs, so that they do not have the required effect. Once the body develops dependence on alcohol, a sudden stopping of drinking is likely to produce withdrawal symptoms. These can be life-threatening and include severe anxiety, tremors, hallucinations and convulsions.

Side effects

The term **side effect** is used to describe any effect other than the intended effect of drug treatment. Side effects may be negative, neutral or positive for the patient. The drug Viagra was originally screened as a potential drug to treat heart disease. Prolonged penile erections were noted in patients as an unexpected side effect.

■ **Figure 25.19** Seed pods of the opium poppy

Morphine is a **narcotic** (pain-relieving and sleep-inducing drug) that is extracted from the opium poppy (*Papaver somniferum*) (Figure 25.19) grown mainly in parts of Afghanistan. It is a very powerful (and addictive) painkiller that is prescribed only to alleviate severe pain, frequently in patients who are terminally ill. It also causes a sense of euphoria or great happiness and well-being.

However, unwanted serious side effects of the drug include a feeling of drowsiness, slow weak breathing, mental confusion, seizures, nausea and constipation. (**Laxatives** such as Dulcolax or Senokot, which stimulate contractions of the bowel, are often used to relieve the constipation caused by morphine.)

Morphine's most dangerous side effect is its desensitization of the medulla oblongata in the brain which is concerned with the regulation of breathing. This causes the ventilation of the lungs to slow down. Moderate doses of morphine may become fatal if taken in conjunction with alcohol.

Bioavailability

Bioavailability refers to how quickly and how much of a particular drug reaches the blood supply following absorption, distribution, metabolism and excretion. Oral bioavailability is the fraction of the dose taken by mouth that reaches the blood stream. This is an important property when it comes to designing new drugs and should be considered with the pharmacokinetics of the drug (how effectively the drug interacts with its target molecule).

Bioavailability can depend on a variety of factors including the crystal form of the drug, whether the drug is given (administered) as a tablet or capsule, or a variation in the components of a tablet or capsule. It is important during phase 1 studies for researchers to check that bioavailability remains the same should there be any change to the manufacturing process or how the drug is stored. Powder-filled capsules are often used in phase 1, whereas tablets are used in phases 2 and 3. Therefore it is necessary that these formulations show bioequivalence in healthy volunteers involved in the drug testing.

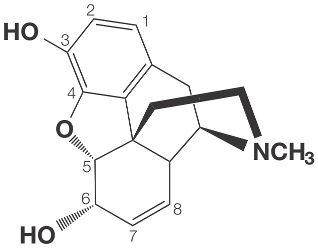
Comparison of how functional groups, polarity and medicinal administration can affect bioavailability

The bioavailability is the fraction of an administered dose of drug that reaches the bloodstream. A number of properties affect bioavailability, including the chemical structure of the drug, which determines physical properties such as solubility, pK_a and hydrophobicity (degree of lipid solubility) and method of administration.

In developing a drug, medicinal chemists often chemically modify the structures of lead compounds. A lead compound is a molecule that shows pharmaceutical activity in a screening process for active molecules.

This relationship between the molecular structure of a drug and its effect on animals or plants is termed the structure–activity relationship. It may be investigated by synthesizing a series of structurally related molecules, followed by measuring the change in the therapeutic effect and bioavailability in animals.

Morphine (Figure 25.20) acts as an agonist and binds to the analgesic (pain) receptors in the brain and turns them on, causing an analgesic effect. Since the molecule has a complex structure, small changes in the molecular structure may result in large differences in the activity and bioavailability of the drug. Since modification of structure leads to changes in physical properties, such as solubility, acidity and polarity, it results in changes in properties of absorption, distribution, metabolism and excretion of the drug.



■ **Figure 25.20** Morphine structure with atom labels

A free phenol group (position 3) is essential for analgesic activity. Substituting the phenol group (–OH) in the benzene ring with a methoxy group (as in codeine) reduces the activity of the analogue 15 per cent relative to morphine. The activity is further reduced if an ethoxy group is used. This is due to the fact that the ether group significantly decreases the binding to the receptor. In fact when codeine is directly injected into the brain, no analgesic activity is observed. The activity of the drug is derived from the metabolic demethylation process (conversion of methoxy group (–O–CH_3) to hydroxyl group) brought about by the liver. Only if codeine passes through the liver can its activity be observed in the body.

In contrast, methylating the alcohol group at position 6 increases analgesic activity. It is believed that the ether group in position 6 reduces the polarity of the analogue (the ether functional group is less polar than the alcohol group), leading to a more efficient crossing of the blood–brain barrier. Hence the compound may accumulate in the brain in higher concentrations, enhancing the analgesic effect. The charge on the nitrogen atom and the size of the alkyl group attached to it are crucial to the analgesic activity of the compound.

Bioavailability is also determined by the method of administration. Morphine administered nasally to humans as a simple solution is only absorbed to a limited degree, with a bioavailability of the order of 10 per cent compared with intravenous administration.

Nature of Science

Drug testing and risk/benefit analysis

It is extremely difficult to develop an ideal medicinal drug, but an ideal drug should have the following characteristics: no toxicity, high effectiveness and selectivity (lack of side effects), ease of administration, high chemical stability, low cost, high purity and no tendency for the patient to develop addiction or tolerance. No such drug exists but the list serves as an objective in pharmaceutical research and development. In deciding whether the drug is safe enough to be on the market, regulators would evaluate the risk-to-benefit ratio. It would be ideal to have low risk, but if an effective treatment is found for cancer or human immunodeficiency virus (HIV), it may still be acceptable to release the drug even if there is a high risk of harmful side effects.

Before a drug is released to the market it must go through a wide variety of tests, starting with *in vitro* tests – laboratory tests involving tissues, cells or enzymes. These tests aim to study the biochemical functions of the target molecule as a result of binding to a potential drug. *In vivo* tests give more knowledge about the action of the drug in animals. Drug candidates with the desired activity undergo further tests to assess the risk of the drug upon administration. Preclinical trials involve toxicity, drug metabolism and other biological tests in animals. Potential drugs are often chemically modified based on the results of the tests. Preclinical trials help researchers to estimate the toxic dose of a potential drug, preparing for later clinical trials on normal healthy volunteers and patients. Clinical trials take place in various phases (Table 25.2) and often involve blind and double-blind trials.

■ **Table 25.2** Clinical (human) trials

Phase	Subjects	Test results
1	Small number of healthy consenting volunteers	Toxicity and safety dosage (TD_{50}) and side effects
2	Small number of patients	Effectiveness and effective dosage (ED_{50}) and short-term side effects
3	Large number of patients	Comparison of effectiveness with other drugs available in the market, further data on effectiveness and side effects

■ Drug action

Isomerism

Compounds that have the same molecular formula but different structural formulae are described as isomers. Isomerism is most common in organic chemistry, but structural, geometric and optical isomerism are all shown by transition metal complex ions (see Chapter 13), some of which are used as pharmaceuticals.

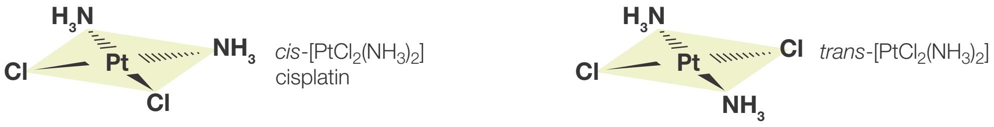


Figure 25.21 Geometrical isomerism in diamminedichloroplatinum(II)

Geometrical isomerism occurs when transition metal complex ions have the same numbers and types of ligands, but their arrangement around the central metal ion varies. This is often loosely called *cis–trans* isomerism, referring to the relative positions of two of the selected ligands. This type of isomerism is only possible where the number of ligands is greater than or equal to four.

Diamminedichloroplatinum(II), [Pt(NH$_3$)$_2$Cl$_2$], a square planar complex, exhibits geometrical isomerism (Figure 25.21). There are two different ways of arranging the four ligands: ammonia molecules adjacent to one another – a so-called '*cis*' arrangement – and a '*trans*' arrangement where the two ammonia molecules are opposite one another.

The *cis* isomer of [Pt(NH$_3$)$_2$Cl$_2$] is used medicinally as an anti-cancer drug and is known as cisplatin (see Chapter 13). The *trans* isomer does not show any anti-cancer properties.

Cisplatin is a highly effective drug commonly used to treat cancer of the ovaries and testes, as well as tumours in the lungs, head, neck, bladder and cervix. Cisplatin is injected directly into the blood where it easily diffuses through the cell membranes of the tumour cells, since it is electrically neutral. Inside the cell, cisplatin exchanges a chloride ion for a molecule of water to form [Pt(NH$_3$)$_2$(Cl)(H$_2$O)]$^+$ which is the active principle (the molecule that has the anti-tumour properties).

This positively charged molecule then passes through the nuclear membrane where it binds to DNA. Once the complex [Pt(NH$_3$)$_2$(Cl)(H$_2$O)]$^+$ has bound to DNA it exchanges another chloride ion to form [Pt(NH$_3$)$_2$(DNA)(H$_2$O)]$^{2+}$, which will then bond at a second site on the double helix of the DNA. Bonding occurs between the platinum ion at the centre of cisplatin and the nitrogen or oxygen atoms of the four bases of DNA (Figure 25.22). The second bond formed can be within the same base, between two adjacent bases on the same strand of DNA or between two bases on opposite strands of DNA. As a consequence, the helical turns of DNA are shortened and bent, thereby preventing DNA replication during cell division.

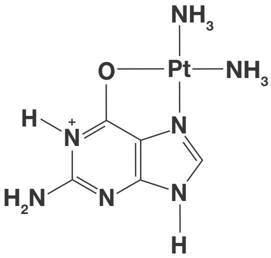


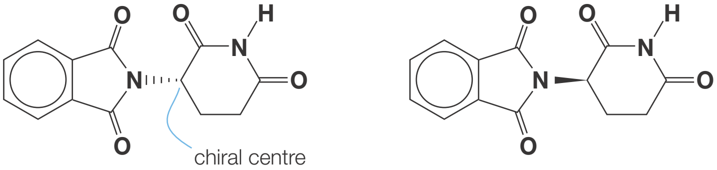
Figure 25.22 Cisplatin bonded to guanine

The importance of chirality in the action of drugs

Enantiomers that act as drugs have different biological properties because they interact and bind with receptor molecules. These receptors are frequently cell membrane proteins whose surface contains a groove or cavity that is capable of interacting and binding only with the enantiomer that has a complementary structure (similar to the lock-and-key hypothesis of enzyme action).

Thalidomide is a chiral molecule that was prescribed as a racemic mixture. It contains one chiral centre based at the carbon atom of the six-membered ring attached to the nitrogen (alpha-amino position) on the five-membered glutarimide ring. There are therefore two possible enantiomers (Figure 25.23).

Figure 25.23 Structures of the two enantiomers of thalidomide



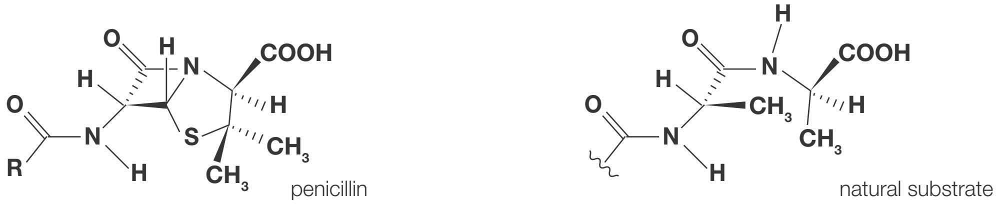
After thalidomide was withdrawn from the pharmaceutical market it was reported that the isomer on the right of Figure 25.23 was **teratogenic**. It has been suggested that if thalidomide was marketed only as the isomer shown on the left the thalidomide tragedy could have been avoided. However, this long accepted hypothesis has become controversial since recent studies have shown that the two enantiomers undergo rapid conversion in human blood plasma.

Penicillin

The various penicillins are a group of antibiotics that belong to a large class of antibiotics termed beta-lactam antibiotics. All the antibiotics in this class contain the *beta*-lactam ring (Figure 25.24). It is this strained four-membered ring that is responsible for the antibacterial activity of penicillins and other beta-lactam antibiotics. The bond angles of the carbon and nitrogen atoms of the amide group in this ring are 90°, instead of 109° and 120° for sp^3 and sp^2 hybridized atoms, respectively.

Penicillins inhibit the final step in the synthesis of bacterial cell walls. The final step involves the enzyme-controlled cross-linking of peptidoglycan strands. Penicillin resembles the dipeptide moiety (alanine–alanine) incorporated by the transpeptidase enzyme to perform the cross-linking reaction (Figure 25.24). The penicillin mimics the enzyme's normal substrate and hence enters the active site. The enzyme becomes covalently attached to the penicillin molecule and is then not capable of catalysing any further reactions.

■ **Figure 25.24** Structures of penicillin and alanine–alanine, the substrate for transpeptidase

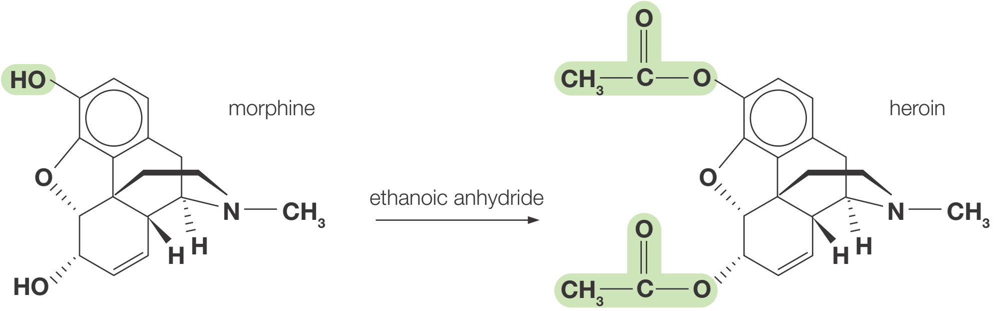


The high chemical reactivity of the amide group within the four-membered ring structure is a result of ring strain. The bond angles in the four-membered ring will be close to 90°, considerably less than the favoured bond angle of 109° for sp^3 hybridized carbon atoms. The beta-lactam ring opens so that the penicillin molecule becomes covalently bonded to the transpeptidase.

Heroin and morphine

Semi-synthetic drugs are drugs synthesized from a naturally occurring chemical rather than from simple starting substances. An early example of a semi-synthetic drug was diamorphine (heroin), which was synthesized from morphine. The conversion is performed by reacting morphine with excess ethanoic anhydride which converts an alcohol functional group (Figure 25.25) into an ester functional group. This is an example of an acylation, specifically ethanoylation (acetylation).

■ **Figure 25.25** Synthesis of heroin



The consequence of the acylation is that a polar alcohol group (–OH) has been converted to a less polar ethanoyl group. This functional group conversion results in an increased potency by a factor of between 5 and 10 for heroin relative to morphine. Morphine, heroin and other opiates act on the brain but they have to pass through the blood–brain barrier (Figure 25.26). This is a membrane-based structure that protects the delicate brain cells (neurons and glial cells) from toxic chemicals in the blood, but still allows the absorption of nutrients and dissolved oxygen.

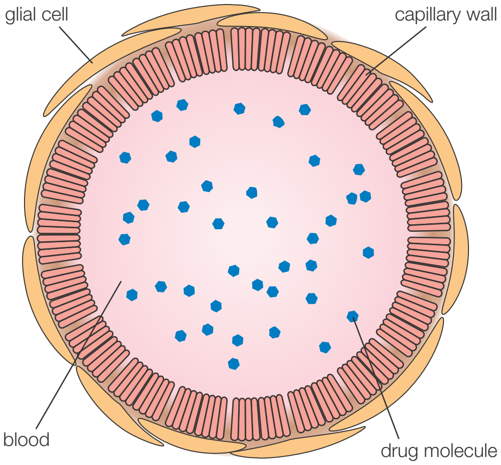


Figure 25.26 The blood–brain barrier: many large drug molecules cannot pass through the thick capillary wall and glial cells

Drug design

Traditionally, researchers first identified a disease target, usually an enzyme or structural protein, and then searched for a chemical that would bind and interact with the **drug target** molecule and inhibit its action. Many of these compounds were discovered by 'screening' natural products, often from plant extracts, and testing them with cells grown in culture. The cells are frequently cancer cells of various types or, more recently, white cells infected with HIV. The cells were then observed for any changes in growth, specifically inhibition.

However, once a compound had been identified to be active against a particular target, for example cancerous cells, it rarely had all the necessary properties required to become a drug. For example, it might be poorly soluble in water. Therefore, chemical analogues of the active compound were synthesized in order to optimize desirable properties and, equally, minimize any harmful effects. These chemical analogues constitute a **chemical library**: they will share common structural features but will have one or more different functional groups, for example Taxol (paclitaxel) and Taxotere (docetaxel). The various chemical analogues were then re-tested until a suitable molecule, known as the **drug candidate**, with the best balance of chemical, pharmacological and physical properties was obtained.

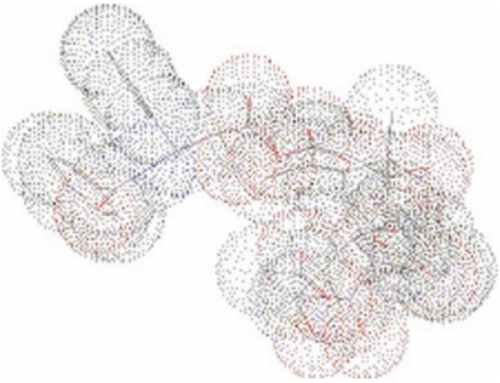
The problem with this approach is that chemists using traditional methods of synthesis can typically synthesize only one different compound a day. Synthesizing a range of compounds to test against a disease target could have taken several years. Optimizing an active compound and discovering a drug candidate could take several more years.

The term **pharmacophore** refers to the particular group of atoms or functional groups required for a drug to bind to a target. The knowledge of a drug's pharmacophore is critical to drug optimization.

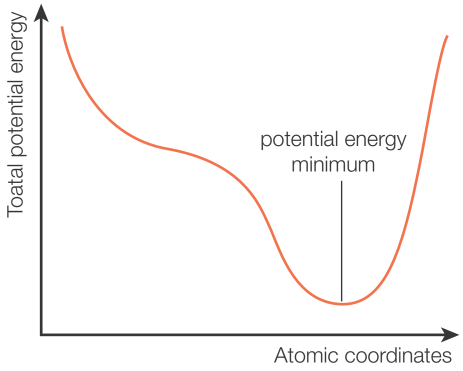
Computer modelling

The modern approach to drug design is target orientated, aiming to improve the binding between the drug and the target molecule. It also aims to increase the drug's selectivity for the target molecule. If these aims are achieved then activity should be increased, while side effects should be decreased.

At a molecular level this means that the drug molecule must be of the correct size and shape to fit into the active site of the target molecule. The necessary functional groups must be present and in the required orientation so that binding can occur.



■ **Figure 25.27** A computer-generated model of Taxol (using Chem 3-D from Cambridgesoft Software)



■ **Figure 25.28** The technique of energy minimization

■ **Figure 25.29** Disprin, a popular brand of soluble aspirin from the UK

Many drug targets have been identified and their genes can be easily cloned and inserted into a bacterium using genetic engineering techniques. Large quantities of a particular receptor or enzyme can then be grown. The structures of the drug target and its active site can then be established using X-ray diffraction (see Chapter 4) and NMR (see Chapter 21).

In theory it is then possible to design a drug based on the structure of the target binding site using computer-based molecular modelling software (Figure 25.27). This is known as *de novo* drug design, but to date no clinical drugs have been designed by this method. Generally, molecular modelling is used to optimize **analogues** derived from a drug candidate or identify which analogues are most likely to bind to the active site of the drug target.

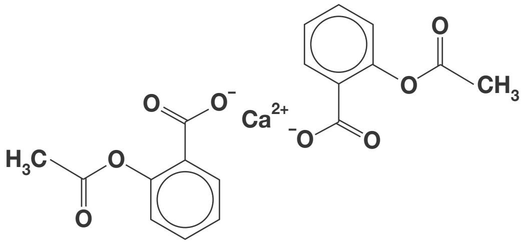
Molecular modelling software allows medicinal chemists to create an accurate model of a molecule in three dimensions. However, the molecule may be distorted and not have the correct bond lengths and, in particular, bond angles. There will be many shapes that the molecule can adopt known as **conformations** (see Chapter 20).

The software will then perform a process called energy minimization (Figure 25.28) where the program modifies the bond lengths and angles in the computer model of the molecule and calculates the steric energy of the new conformation. If the energy increases significantly, it means that the new structure is unstable. The process is repeated many times, until new modifications carried out have little effect on the total energy of the molecule. This corresponds to a stable structure or an energy minimum.

Improving the solubility of drugs

Drugs need to be soluble in both polar and non-polar environments. The solubility of a drug in these environments determines its absorption, distribution and excretion. Modifying the polarity of a drug or its acid/base properties will vary its solubility and may improve its medicinal properties.

Drugs that contain a carboxylic acid functional group can be made more polar by converting them to a salt. This is the case for soluble aspirin (Figure 25.29) which is the calcium or sodium salt of aspirin (Figure 25.30). Once the aspirin anion reaches the strongly acidic environment of the stomach it reverts back to the un-ionized or molecular form (free acid).



■ **Figure 25.30** Structure of the calcium form of soluble aspirin

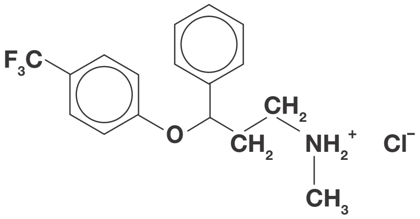


Figure 25.31 Structure of Prozac (fluoxetine hydrochloride)

Many drugs that contain an amine group are similarly administered as their hydrochloride salt as this increases their solubility in polar environments and hence their transport around the body. The antidepressant drug Prozac (Figure 25.31) is administered as the hydrochloride.

Chiral auxiliaries

There are many examples of drugs where the biological activities of the two isomers are very different. Until recently many synthetic drugs were produced as racemic mixtures consisting of equimolar mixtures of both enantiomers. Such mixtures are difficult to resolve (separate) and frequently the two enantiomers crystallize into a racemic crystal. However, various techniques have been developed that allow the production of single enantiomers. One such approach is the use of a reagent or reactant that contains a **chiral auxiliary**, a group located near the reaction site that controls the stereochemistry and is easily removed afterwards. This approach is an example of true asymmetric synthesis because chirality is created in the course of the reaction. This approach is used in the total synthesis and semi-synthesis of Taxol (see Section 25.7).

Figure 25.32 shows the stereoselective conversion of propanoic acid (an achiral compound) to a specific enantiomer of 2-hydroxypropanoic acid (a chiral compound). A chemical synthesis that did not employ a chiral auxiliary would result in the formation of a racemic mixture.

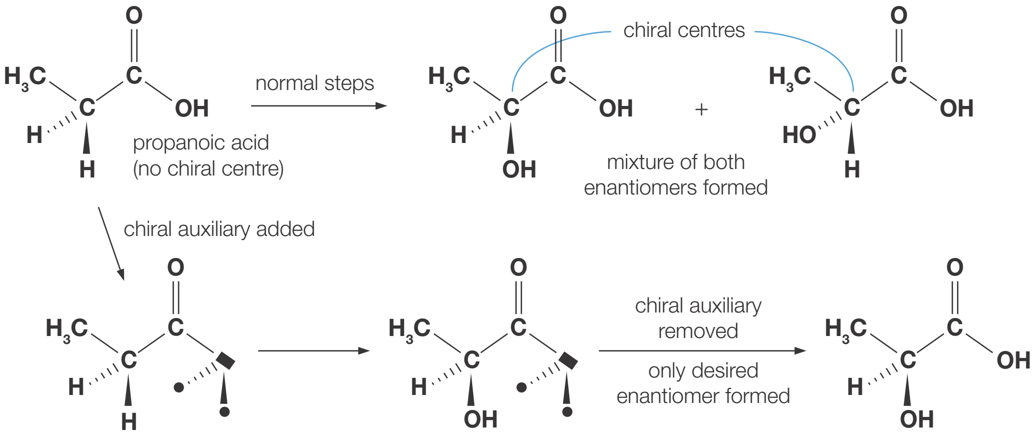
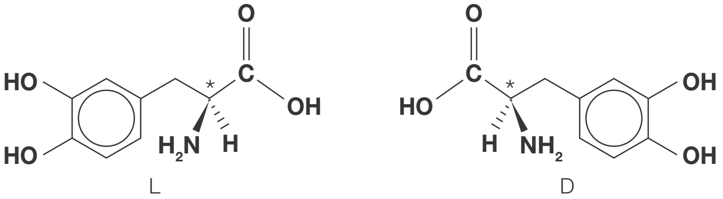


Figure 25.32 The use of a chiral auxiliary in the synthesis of 2-hydroxypropanoic acid

A chiral auxiliary is employed in the synthesis of the drug L-DOPA, used to control Parkinson's disease. This is a degenerative disease of the brain that impairs motor skills and speech. It is characterized by muscle rigidity, shaking and a slowing of physical movement. 3,4-Dihydroxyphenylalanine (DOPA) exists in two enantiomers (Figure 25.33), but only L-DOPA is active. The D-form is inactive.

Figure 25.33 The two enantiomers of DOPA. The * marks the chiral centre



The key step in the synthesis of L-DOPA is the use of a homogeneous rhodium catalyst to hydrogenate a carbon–carbon double bond. It is this step that creates the chiral centre. By using a catalyst which itself contains only one stereoisomer of a chiral phosphorus ligand, the catalyst is able to direct the addition of hydrogen to the carbon–carbon double bond so that only one enantiomer is formed. Hydrolysis of this product forms a mixture containing 97.5 per cent L-DOPA and 2.5 per cent D-DOPA.

Rational drug design – enzyme inhibitors as medicines

In rational drug design, biologically active compounds are specifically designed to interact with a particular drug target. Rational drug design often involves the use of molecular design software, which researchers use to create three-dimensional models of drugs and their biological targets.

Receptors and enzymes are very similar. Both are globular proteins with precise structures that have evolved (by natural selection) to accommodate specific arrangements of atoms on another molecule such as a neurotransmitter or a substance.

Just as antagonists are used to block receptors, so medicinal drugs can function by inhibiting the action of enzymes. A good example is captopril – which is widely used to treat high blood pressure (hypertension). If not treated, this can lead to serious consequences such as stroke or heart attack.

A small protein, angiotensin II, formed from eight amino acids is known to be a key factor in raising blood pressure. The body synthesizes angiotensin II for an inactive 10-amino-acid protein called angiotensin I. The enzyme which brings about the loss of the two amino acids in this conversion is called angiotensin converting enzyme (ACE).

$$\text{Angiotensin I} \xrightarrow{\text{ACE}} \text{Angiotensin II} + \text{dipeptide}$$

An imbalance in the production of angiotensin II results in high blood pressure, so one method of treatment is to inhibit the enzyme which catalyses its formation. The medicine needs to be an ACE inhibitor and captopril works in this way.

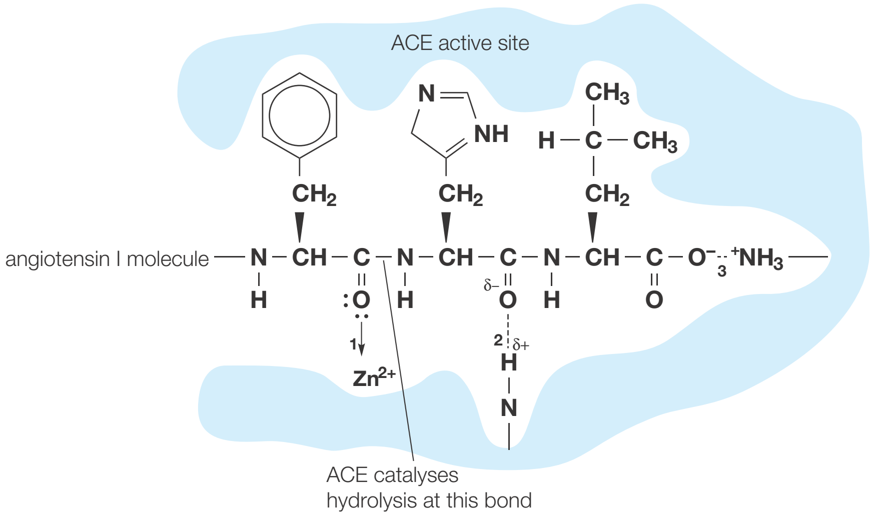
The lead for the development of captopril was the discovery that the venom of the Brazilian arrowhead viper (Figure 25.34) brought about its toxic effect by causing a decrease in the blood pressure. The venom is a complex mixture, but separation, analysis and further study of the components led to the identification of several small proteins which were powerful ACE inhibitors. Unfortunately, proteins are difficult to administer as medicines by mouth because they are readily broken down by pepsin and hydrochloric acid present in the stomach.

Figure 25.34 Brazilian arrowhead viper

The snake venom proteins could not therefore be used medically as ACE inhibitors, but their structures gave medicinal chemists information about the shape and structure of the enzyme's active site. They learned more by studying the way the enzyme changes angiotensin I into angiotensin II. From this they realized that ACE must be similar to another, well-studied zinc-containing enzyme, carboxypeptidase. The combination of these results led to a model in which the ACE enzyme binds to its substrate (angiotensin I) in three places by three different types of interaction:

1. a metal–ligand coordinate bond between a zinc ion on the enzyme and an atom with a partial negative charge on the substrate
2. a hydrogen bond between an N–H group (from an amino acid) on the enzyme and a negatively charged atom on the substrate (angiotensin I)
3. an ionic interaction between a positively charged $-NH_3^+$ group on the enzyme and a negatively charged $-COO^-$ group at one end of the substrate.

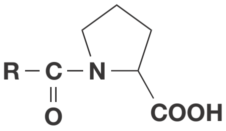
These interactions (labelled 1, 2 and 3) between ACE and the three amino acid residues at the –COOH end of angiotensin I are shown in Figure 25.35. Note that the enzyme's active site has a precise shape that also fits other groups on the amino acid side-chains.



■ **Figure 25.35** The structure proposed for the ACE–angiotensin I complex

It is essential for the biological action of ACE, and all enzymes, that the products are bound less strongly than the substrate. This ensures that the products will leave the enzyme and allow more substrate molecules to bind, thus continuing the catalysis. In this case, angiotensin I (the substance) is bound by three interactions, but angiotensin II (one of the products) is only bound by the metal–ligand bond (interaction 1 in Figure 25.35), and the dipeptide (the other product) is only held by the hydrogen bond and the ionic interaction (interactions 2 and 3 in Figure 25.35).

A successful ACE inhibitor needs to bind but not react, so it remains attached by all three interactions. The use of interactive computer graphics to study the shape and charge density in the active site played an important role in the design of possible compounds.



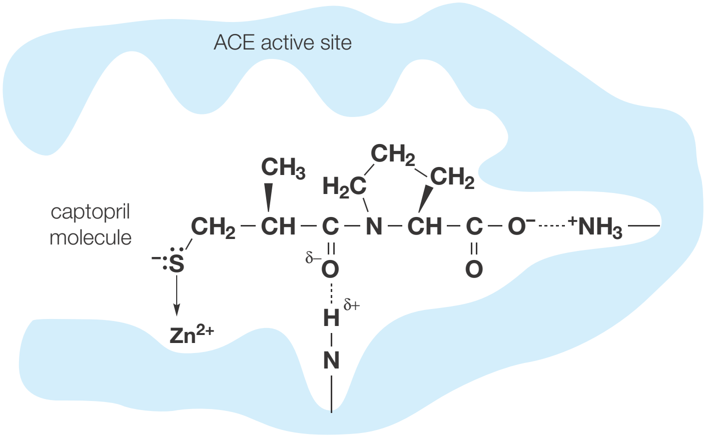
■ **Figure 25.36** Proline with an acyl group

All the active proteins which were isolated from the snake venom had proline as the amino acid at the –COOH end of the chain. This was chosen as a good lead for the development of an ACE inhibitor. Medicinal chemists set about synthesizing hundreds of compounds with the general structure as shown in Figure 25.36, but with different R groups, each with a different side-chain.

They were all capable of interacting at positions 2 and 3 with the ACE enzyme. The breakthrough came when it was discovered that derivatives which contained an –SH group in the right position to interact with the Zn^{2+} were particularly effective.

Figure 25.37 shows how captopril is thought to bind to ACE. Captopril is not a protein and is not broken down by digestive enzymes when it is taken orally (by mouth). The peptide link which is broken in angiotensin I is not present in captopril. At the corresponding position on the captopril–enzyme complex, there are S–C and C–C bonds which cannot be hydrolysed. The development of captopril is an excellent example of the rational design of a new molecule. The drug went through clinical trials and was marketed in 1980 and is now in widespread use.

■ **Figure 25.37** Captopril in the ACE active site



ToK Link

Risks and benefits of drugs

All drugs carry risks as well as benefits. Who should ultimately be responsible for assessing these? Public bodies can protect the individual but also limit their freedom. How do we know what is best for society and the individual?

Regulatory bodies are responsible for accessing the scientific and clinical data, and approving whether a drug can proceed to clinical trials or the marketplace.

In the USA, pharmaceutical companies submit an IND (Investigational Exemption to a New Drug Application) to the FDA (Food and Drug Administration) when they want to start clinical trials on a drug.

The IND should contain information regarding the chemistry, manufacture and quality control of the drug. It should also include information regarding the drug pharmacology, distribution and toxicology in animals, and any clinical information that may be known about the drug. The IND is a confidential document and is not released to the public.

A new drug application has to be submitted to the FDA if a drug is to be marketed in the USA. This contains all the scientific and clinical information regarding the drug's manufacture and testing. Fast tracking is permissible for drugs that are effective against life-threatening diseases such as ebola where existing drugs are either lacking or not effective.

Labelling must be approved by the regulatory authority and should give full details on the drug, its side effects and its administration.

Good laboratory, manufacturing and clinical practice (GLP, GMP and GCP, respectively) are regulations designed to ensure high professional, scientific and clinical standards in laboratory, manufacturing and clinical settings, respectively. Detailed documentation must be kept to prove that the company is adhering to these standards. Regulatory bodies such as the FDA must approve any marketing claims made about new drugs.

However, local regulations vary greatly between countries so the same drug may be available over the counter in some countries but require a doctor's prescription or even be unavailable in other countries. Such national differences restrict international trade and raise a number of ethical issues, such as the balance between the freedom of individuals and the right of public and regulatory bodies to protect the physical and mental health of their citizens.

25.2 Aspirin and penicillin *– natural products with useful medicinal properties can be chemically altered to produce more potent or safer medicines*

Analgesics

Analgesics (Figure 25.38) are drugs used to relieve pain without causing loss of consciousness. Strong analgesics include the opium derivatives morphine and codeine and are used to provide relief from pain. Morphine is a much stronger analgesic than codeine. Mild analgesics are able to provide relief from mild pain and, on occasions, severe pain. They are also frequently **antipyretic**: they reduce fever to normal body temperature. Mild analgesics include aspirin, paracetamol (acetaminophen), ibuprofen and naproxen. Local anaesthetics are also analgesic in function.

Figure 25.38 Counterpain: a commercial mild analgesic containing methylsalicylate and the essential oils menthol and eugenol

By the 1970s medical researchers had discovered that aspirin reduces pain by acting on inflamed tissues and the associated nerves. This mechanism of action is in contrast to the action of morphine, a very powerful painkiller that acts directly on the brain. However, aspirin has other side effects that could not be easily accounted for, such as its ability to inhibit platelet function. Platelets are small fragments of cells concerned with the clotting of blood.

In 1971 John Vane (now Sir John Vane and a winner of the 1982 Nobel Prize in chemistry) suggested that aspirin functions by inhibiting production of **prostaglandins**. These are 'local hormones' or autacoids that control a number of processes, such as blood pressure and platelet function. Unlike hormones they are present in all cells and are released when the cells are damaged or stimulated by hormones.

Prostaglandins are responsible for the classic symptoms of inflammation, namely redness (caused by dilation of blood vessels) and heat (leading to fever). Vane and his colleagues demonstrated in 1973 that aspirin inhibits the production of several prostaglandins from arachidonic acid (Figure 25.39), a fatty acid component of cell membranes. The enzyme responsible for this conversion is prostaglandin H synthase, which has an amine group that undergoes a chemical reaction with aspirin (Figure 25.40).

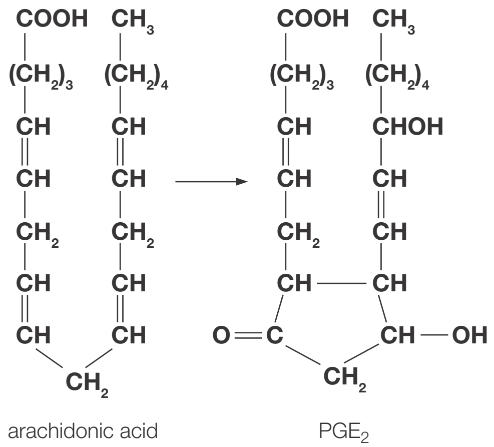


Figure 25.39 The formation of prostaglandin (PGE$_2$) from arachidonic acid

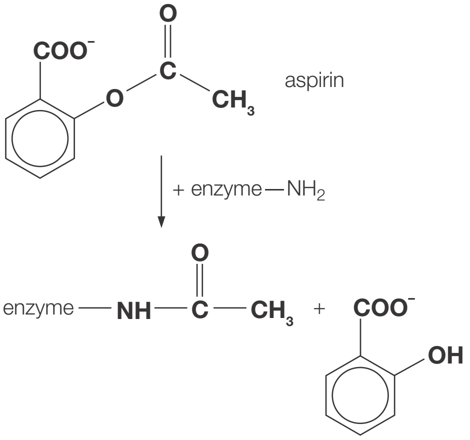


Figure 25.40 The reaction between aspirin and prostaglandin H synthase

Vane's hypothesis also explained some of the side effects of aspirin, such as its tendency to cause stomach irritation and ulcers. He suggested that aspirin causes these problems because it blocks the synthesis of prostaglandins. The stomach lining releases these to regulate the overproduction of hydrochloric acid and to synthesize the protective mucous barrier that prevents self-digestion. In addition, aspirin prevents the body from excreting salt and water properly by preventing the kidney from synthesizing a prostaglandin concerned with dilating blood vessels.

Other side effects of aspirin use may include fluid accumulation and interference with the clotting of blood. The latter has been exploited in the use of aspirin as an anti-coagulant to prevent and reduce strokes (preventing blood clotting inside the blood vessels of the brain) and heart attacks. An unwanted side effect of aspirin is aspirin sensitivity syndrome, a rare genetic condition resulting in constriction of the trachea (windpipe), which can cause death.

Paracetamol does not prevent prostaglandin H synthase from synthesizing prostaglandins within body tissues, but is effective at inhibiting the enzyme in the brain, which perhaps explains its antipyretic activity. Some of the effects of aspirin and paracetamol are believed to depend on the direct interaction of these molecules with cell membranes.

Aspirin

Aspirin is a very widely used mild analgesic (painkiller). It is derived from salicylic acid (2-hydroxybenzoic acid). It has a range of useful medicinal actions. In 1758, an English vicar, the Reverend Edward Stone, discovered that a crude (impure) extract from the bark of the willow tree (Figure 25.41) was effective in reducing the fever associated with various conditions, for example malaria (which was present in England during those times). The active ingredient was later identified as salicin (Figure 25.42), a glycoside. It was first isolated in significant amounts in 1828 by a French pharmacist, Leroux, and takes the form of bitter-tasting yellow crystals.

Figure 25.41 A weeping willow tree: its bark contains salicin

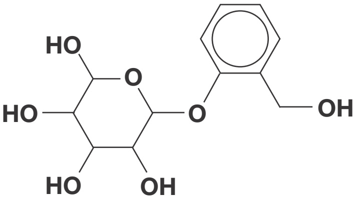


Figure 25.42 Structure of salicin

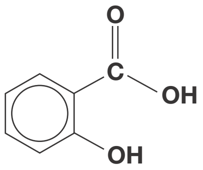


Figure 25.43 Structure of salicylic acid (2-hydroxybenzoic acid)

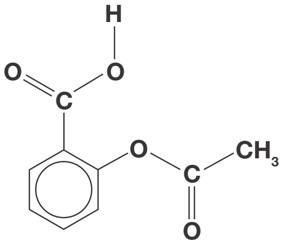


Figure 25.44 Structure of aspirin (acetylsalicylic acid)

In 1860 the German chemist Hermann Kolbe synthesized salicylic acid (Figure 25.43) by reacting sodium phenoxide ($C_6H_5O^-Na^+$) and carbon dioxide. In 1876 two German doctors reported the successful treatment of the symptoms of acute rheumatic fever, an auto-immune disease where the body's own immune system attacks bacteria in the joints between bones. In 1888 salicylic acid was used to alleviate the symptoms of arthritis.

Acetylsalicylate, or aspirin (Figure 25.44), was introduced in 1898 by the German chemist Felix Hoffmann who found that his arthritic father could not tolerate the large daily doses of bitter sodium salicylate (a soluble salt of salicylic acid). Aspirin was much more palatable than sodium salicylate and was equally effective, but at lower doses. Aspirin was initially sold as a powder, but then later sold in tablet form bound together by starch. It not only proved effective at treating fevers and pain, but also reduced swelling and inflammation.

A potentially fatal condition linked to aspirin usage is **Reye's syndrome** which, although primarily a children's disease, can occur at any age. It affects all organs of the body, but those most at risk are the brain, which swells due to an increase in fluid pressure, and the liver, due to a massive accumulation of fat.

The cause of Reye's syndrome is not known but a risk factor is the use of aspirin and related drugs during a viral illness. Recovery from Reye's syndrome depends critically on the amount of swelling that has occurred within the brain. A complete recovery is possible, but permanent brain damage, coma or death from a heart attack are also possibilities.

Aspirin shows a synergistic effect with alcohol. This is an effect between two or more drugs that produces an effect greater than the sum of their individual effects. Drinking alcohol and daily use of aspirin may result in torn stomach lining and internal bleeding (gastrointestinal hemorrhaging).

Synthesis of aspirin

Aspirin (2-acetoxybenzoic acid) can be synthesized in the laboratory by reacting salicylic acid (2-hydroxybenzoic acid) with ethanoic anhydride in the presence of concentrated sulfuric acid, with hydrogen ions acting as the catalyst (Figure 25.45). The reaction mixture is heated under reflux and cooled in ice to induce crystallization. The crude (impure) aspirin is removed by suction filtration and washed with ice-cold water. The crystals are dried and recrystallized and the melting point determined to establish purity. A sharp melting point suggests the aspirin is pure; impurities will often lower the melting point.

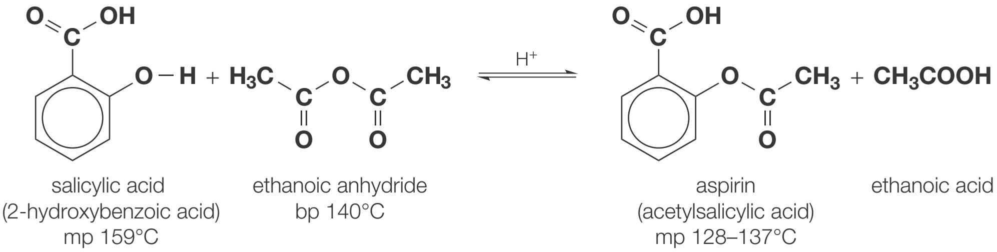


Figure 25.45 Reaction between salicylic acid and ethanoic anhydride to form aspirin (acetylsalicylic acid)

Aspirin still has its side effects because the carboxylic acid functional group remains intact. This may result in hemorrhaging of the stomach walls even with normal dosages. The acidic irritation can be reduced through the use of buffering agents, like antacids, in the form of magnesium hydroxide, magnesium carbonate and aluminium glycinate when mixed with aspirin (known commercially as Bufferin).

An alternative synthetic route to aspirin in the laboratory involves ethanoyl chloride and a base catalyst (for example pyridine) (Figure 25.46).

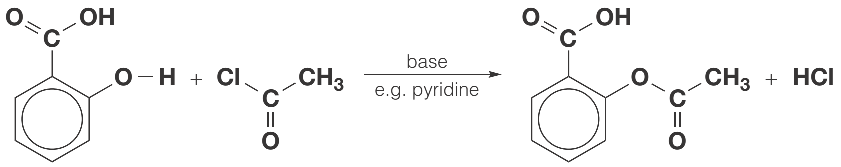


Figure 25.46 Reaction between salicylic acid and ethanoyl chloride to form aspirin (acetylsalicylic acid)

1 A student produced a sample of aspirin by the esterification of 9.40 g of 2-hydroxybenzoic acid with excess ethanoic anhydride. After purification by recrystallization, 7.77 g of aspirin was obtained. Calculate the percentage yield obtained.

Analysis of salicylic acid and aspirin

Chemical analysis

Salicylic acid can be extracted from willow bark by refluxing fresh chopped bark with sulfuric acid and potassium manganate(VII). Thin-layer chromatography can be used to show that the pharmacologically active chemical in willow bark is salicylic acid.

A knowledge of some relatively simple test-tube experiments can often be used effectively in the identification of unknown organic substances. For example, you will be familiar with the use of bromine solutions for detecting double bonds between carbon atoms in alkenes (see Chapter 10). Three chemical tests are helpful in confirming the structure of salicylic acid:

1 An aqueous solution of the compound is weakly acidic.
2 Salicylic acid reacts with alcohols (like ethanol) to produce compounds called esters. Esters have strong odours, often of fruit or flowers.
3 A neutral solution of iron(III) chloride turns an intense purple colour when salicylic acid is added. This is a complex formation.

Tests 1 and 2 are characteristic of carboxylic acids (compounds containing the –COOH functional group); test 3 indicates the presence of a phenol group (an –OH group attached directly to a benzene ring). Aspirin would give positive results for tests 1 and 2, but not for test 3, since it has an acetyl group rather than a phenol group. Iron(III) chloride can be used as a test of aspirin purity prepared by a laboratory synthesis.

Although chemical tests provide evidence for the presence of carboxylic acid and phenol groups in salicylic acid, instrumental techniques such as mass spectrometry (MS), infrared (IR) spectroscopy and nuclear magnetic resonance (NMR) are today's most efficient analytical tools used in drug research and analysis (see Chapter 21).

Evidence from infrared spectroscopy

One of the very first tasks which would be done with any undefined new organic substance is to record its infrared (IR) spectrum. An IR spectrum measures the extent to which electromagnetic radiation in part of the infrared region is transmitted through a sample of a substance.

The wavenumber ranges absorbed provide important clues about the functional groups present. The functional groups absorb at similar frequencies in many different compounds so an IR absorption pattern provides a kind of fingerprint of the molecule.

The IR spectrum of salicylic acid (Figure 25.47) shows clear evidence of the presence of the C–O and –OH groups. The narrow peak centred at 1700 cm^{-1} is due to the carbonyl group present in the carboxylic acid functional group. The IR spectrum suggests an –OH group (free or more likely internally hydrogen bonded) at about 3600 cm^{-1} and an –OH group (hydrogen bonded) at about 3250 cm^{-1}. The C–H (arene) absorption is at about 3100 cm^{-1}.

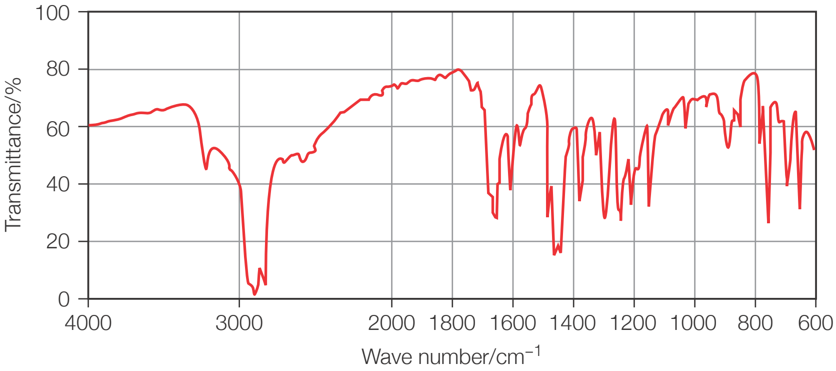


Figure 25.47 The IR spectrum of salicylic acid

Evidence from nuclear magnetic resonance spectroscopy

A second instrumental technique which would be applied to an unidentified compound is nuclear magnetic resonance (NMR) spectroscopy. The high-resolution and low-resolution proton (^{1}H) NMR spectra for salicylic acid are shown in Figure 25.48.

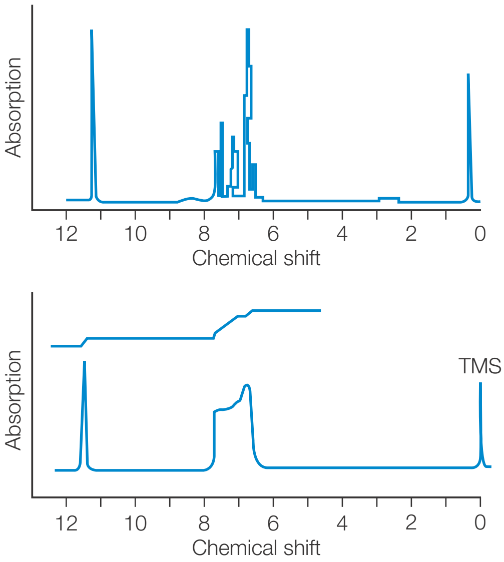


Figure 25.48 The high-resolution and low-resolution ^{1}H NMR spectra of salicylic acid

The NMR spectra show that salicylic acid contains:

- one hydrogen in a –COOH environment
- one hydrogen in a phenolic –OH environment
- four hydrogens attached to a benzene ring.

A combination of IR and NMR spectroscopy shows that salicylic acid has an –OH group and a –COOH group both attached to a benzene ring. However, there are three possible isomeric hydroxybenzoic acids: 2-hydroxybenzoic acid, 3-hydroxybenzoic acid and 4-hydroxybenzoic acid. A decision about which isomer salicylic acid is can be made by analysis of the mass spectrum of salicylic acid.

Figure 25.49 shows the mass spectrum of salicylic acid. Chemists can use information contained in spectra like this to find the structures of compounds.

The mass spectrum shows signals which correspond to positively charged fragmentation ions formed from the parent compound fragmentation. The molecular ion signal at m/z = 138 confirms that the substance has an empirical formula of $C_7H_6O_3$.

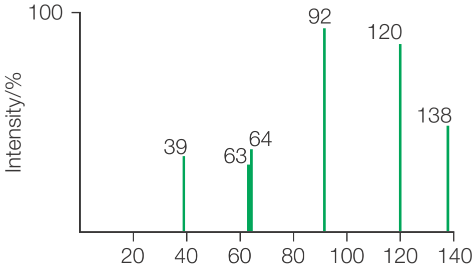
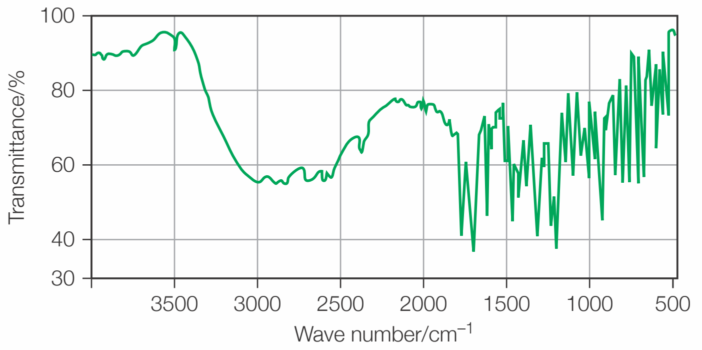


Figure 25.49 The mass spectrum of salicylic acid

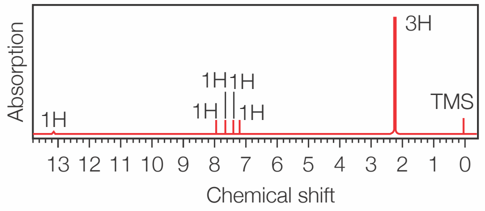
The way in which a molecular ion breaks down is characteristic of that compound. In this case, comparison with a database of known mass spectra identifies salicylic acid as 2-hydroxybenzoic acid. Considering the way that the 3-hydroxybenzoic acid and the 4-hydroxybenzoic acid isomers would break down also leads to the conclusion that these isomers could not form some of the fragments observed in the mass spectrum of salicylic acid.

2 Identify the bonds responsible for the broad peak around 3000 cm^{-1}, and the sharp peaks at 1750 cm^{-1} and 1690 cm^{-1} in Figure 25.50 showing the IR spectrum of aspirin.



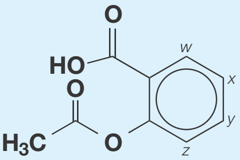
■ **Figure 25.50** IR spectrum of aspirin

3 Identify the hydrogen atoms responsible for the signals at chemical shifts 2.2. and 13.1 in Figure 25.51 showing the NMR spectrum of aspirin.



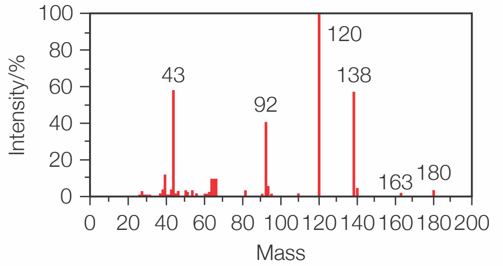
■ **Figure 25.51** ^{1}HNMR spectrum of aspirin

The hydrogen atoms responsible for the cluster of signals in the chemical shift range 7.5 ± 0.5 are indicated as *w*, *x*, *y* and *z* on the structure of aspirin shown in Figure 25.52.



■ **Figure 25.52** The aromatic hydrogen atoms responsible for the cluster of downfield signals in the NMR spectrum

4 Identify the ions responsible for the peaks at mass 180, 163, 120 and 43 in Figure 25.53 which shows the mass spectrum of aspirin.



■ **Figure 25.53** Mass spectrum of aspirin

Nature of Science

Discovery of aspirin

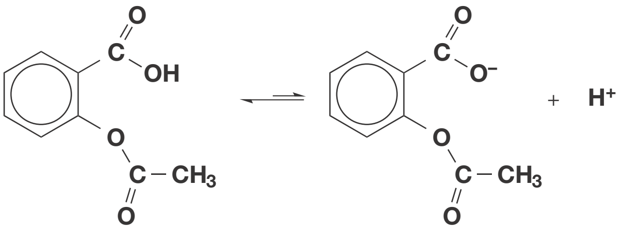
Aspirin is the most widely used medicinal drug in the world and a good example of a medicine that bridges the gap between natural and synthetic medicines. Over 2000 years ago in Greece, Hippocrates recorded that the Ancient Greeks used powdered willow bark to reduce fever. This knowledge was rediscovered by the Revered Edward Stone in 1763 who lived in a village in Oxfordshire, UK. He used an extract of willow bark to cure fevers in more than 50 residents. He was influenced by the 'doctrine of signatories'. Willow grew in damp places and because fevers were associated with damp places, willow might produce a medicine to treat them. The active ingredient in the willow bark is salicin which can be metabolized to salicylic acid.

Soluble aspirin

Aspirin has limited solubility in water due to the presence of a hydrophobic benzene ring. Aspirin is a weak monoprotic acid ($K_a = 2.8 \times 10^{-4}\,\mathrm{mol\,dm^{-3}}$ at 25 °C), so very little of the molecular aspirin dissociates to form acetylsalicylate anions.

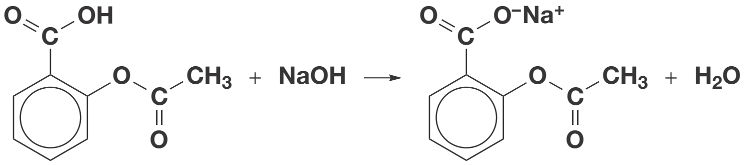
For the equilibrium dissociation reaction (Figure 25.54), the equilibrium position lies on the left, strongly favouring molecular aspirin.

Figure 25.54 Dissociation of aspirin in water



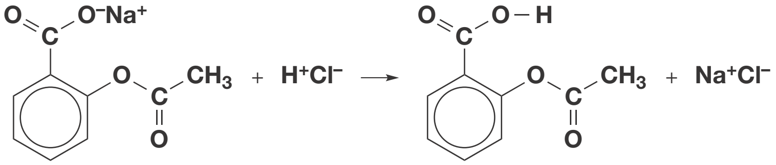
Because of the low solubility of aspirin in water, its bioavailability in the blood is limited. The solubility and bioavailability of pharmaceutical drugs can be increased by converting them into ionic salts. In the case of aspirin, the carboxyl group can be neutralized with sodium hydroxide solution, producing the water-soluble sodium salt of acetylsalicylic acid, also known as soluble aspirin (Figure 25.55). Salts of calcium and lysine are also used to prepare 'soluble aspirin'.

Figure 25.55 Formation of the soluble sodium salt of aspirin



In aqueous solution the sodium salt of aspirin completely dissociates into sodium cations and acetylsalicylate anions, which form ion–dipole forces and hydrogen bonds with surrounding water molecules. However, the sodium salt is rapidly converted back into less soluble aspirin by the reaction with hydrochloric acid in the stomach (Figure 25.56), so the bioavailability of soluble aspirin is only slightly higher than that of aspirin.

Figure 25.56 Reaction between the sodium salt of aspirin and hydrochloric acid



Analysis of aspirin

The acid-neutralizing capacity of a tablet is the amount of hydrochloric acid that it can neutralize. It is the quantity that is referred to in some advertisements when it is stated that the tablet 'neutralizes *x* times its mass in stomach acid.'

This capacity can be determined by a back titration (see Chapter 1). A known amount of antacid is dissolved in an excess of hydrochloric acid, and then the excess acid is back titrated with standardized sodium hydroxide solution.

When the end-point is reached with a suitable indicator, the amount of acid that was added to the antacid sample is equal to the amount of base present, sodium hydroxide plus the antacid.

Therefore, the amount of hydrochloric acid that was neutralized by the antacid is equal to the total amount of hydrochloric acid added minus the amount that was neutralized by the sodium hydroxide:

$$\begin{aligned}\text{amount of acid neutralized} &= (\text{amount of HCl added}) \\ &\qquad - (\text{amount of NaOH required for back titration}) \\ &= (\mathrm{MHCl} \times V\mathrm{HCl}) - (\mathrm{MNaOH} \times V\mathrm{NaOH})\end{aligned}$$

where M = molarity (mol dm^{-3}) and V = volume in cubic decimetres (dm^{3}).

Nature of Science

Natural products

Some drugs, such as penicillin and cisplatin, have been discovered by accident. Taxol was discovered as the result of a systematic search, but there was a long period of time between its discovery and its medical use. In 1958 the American National Cancer Institute launched a program to discover new pharmaceuticals by investigating natural sources such as plants and marine organisms.

In 1962 the botanist Arthur Barclay collected bark samples from Pacific yew trees in Washington state. In 1967 Munroe Wall and Manuskh Wani found that a chemical in the samples killed leukemia cells. They isolated the white crystals of the molecule responsible and named it Taxol after the Pacific yew tree (*Taxus brevifolia*) and determined its complex structure in 1971. There was little progress because the optimum yield from the bark is 0.014 per cent and the drug had not been tested against solid tumours.

However, in 1978 Susan Horwitz found that Taxol stabilized microtubules which are involved in the movement of chromosomes during cell division. The drug inhibits mitosis and the cells were unable to divide and grow. Taxol was tested against human tumour cells (breast, large intestine and lung) transplanted to mice and found to cause regression (reduction in size).

Uses of aspirin

Aspirin is the most widely used medication in the world. Worldwide there are over 200 analgesic formulations that contain aspirin. There are so many because the medicine may come in different forms, for example solids, soluble substances and syrups. Other compounds may be present to help relieve other symptoms which occur along with the one being treated. There may be other substances present to help the action of the principal compound. For all these different formulations, there are many companies each producing their own brand-name equivalent.

Aspirin has a wide variety of beneficial effects. It is a mild analgesic (pain reliever) and antipyretic (reduces fever). In addition it has anti-clotting properties that could help prevent heart attacks and limit the brain damage from a stroke. Hence at low doses aspirin is described as cardio-protective. It is a prophylactic for reducing the risk of heart disease. A prophylactic is a medicine used to treat or prevent the occurrence of a disease or condition. However, the potential benefits, especially in patients with no signs of heart disease, must be balanced against the possible harm from side effects, such as bleeding and gastrointestinal symptoms.

Aspirin is thought of as a 'safe' medicine, but like all medicines it if only safe if taken in the recommended dose. The lethal dose is 30 g for an adult of average size. A typical tablet contains 0.3 g, so 100 aspirin tablets could be a lethal dose. Unpleasant symptoms would be experienced with far fewer tablets than this. The recommended dose of aspirin is no more than 12 tablets a day, and it is not recommended for children under 12 years old.

Paracetamol

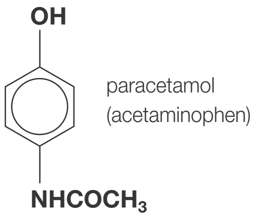


Figure 25.57 Structure of paracetamol (*N*-(4-hydroxyphenyl) ethanamide)

Paracetamol (UK) or acetaminophen (USA) (Figure 25.57), like aspirin, relieves pain and reduces fever, but unlike aspirin does not reduce stiffness, redness and the swelling symptoms of arthritis. It has very few side effects when taken as directed for a short period, but it should not be taken with alcohol, nor by patients with kidney or liver disease. It is the preferred treatment for patients with aspirin allergy, ulcers or clotting disorders. However, there is some evidence that people who take high doses of paracetamol and drink large amounts of alcohol will have an increased risk of liver damage.

The differences between aspirin and paracetamol (Figure 25.58) are summarized in Table 25.3.

Effect	Aspirin	Paracetamol
Causes gastrointestinal tract blood loss?	Yes	No
Increases the tendency to bleed?	Yes	No
Thins the blood and prevents clots from forming?	Yes (effective against heart attacks and strokes)	No
Can cause an allergic reaction?	Yes	No, because it is not a salicylate
Potential for causing nausea and vomiting?	More	Less
Possesses anti-inflammatory action?	Yes and hence used in treating rheumatoid arthritis	Weakly anti-inflammatory
Potential to act as a suicide drug?	Yes	Yes
Antipyretic	Yes	Yes

Table 25.3 A summary table of the differences between aspirin and paracetamol

Figure 25.58 Paracetamol and aspirin

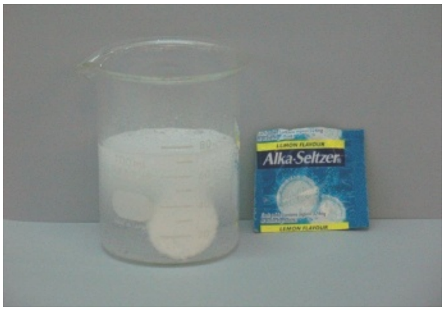


Figure 25.59 Alka-Seltzer tablet reacting with water

Alka-Seltzer (Figure 25.59) is a name owned by the German Bayer Corporation for a line of medications sold over the counter. Alka-Seltzer is a combination of aspirin, sodium hydrogencarbonate ($NaHCO_3$) and citric acid, designed to treat pain and simultaneously neutralize excess stomach acid (the *Alka* being derived from the word alkali). It is provided in the form of large effervescent tablets, which are dissolved in a glass of water. As the tablets dissolve, the base (hydrogencarbonate) and the acid (citric acid) react vigorously, producing carbon dioxide gas (hence the *Seltzer*).

ToK Link

Pain and perception

Different painkillers act in different ways. How do we perceive pain, and how are our perceptions influenced by the other ways of knowing?

Sensory perception is one of the eight ways of knowing identified by the IB's Theory of Knowledge Program. It is an important source of knowledge and our sense organs provide a means of communication between us and the physical world. Pain is an unpleasant but important type of perception. It often plays an important role in survival and protecting our bodies from being hurt. Consider the withdrawal reflex when you touch a hot object. Information about pain is brought to the brain from free nerve endings in the skin.

Opioid-based painkillers such as morphine bind to brain receptors that modulate pain. Opioids are used to control moderate to severe pain, but are highly addictive. Non-opoid-based painkillers are mild analgesics, such as aspirin, which act locally to reduce pain at the site of the injury.

Research has shown that our brains can regulate, control, determine and even produce pain. This has been demonstrated most clearly in patients with phantom limbs. Many people have the ability to feel pain, pressure, temperature, and many other types of sensations including pain in a limb that does not exist (either amputated or born without).

These observations suggest that 'reality' is 'simulated' inside the brain and that the brain creates the impression that the limb exists and functions. The majority of scientists and philosophers would agree that pain is generated within the brain. However, how do the electrochemical signals generated within the neurons of brain tissue generate the sensation of pain? How do brain states generate mental states? Are brain states the same as mental states?

There are factors that can affect the way perception influences our thinking. The most common is logic or reason. Human reason can filter out or accept new perceptions. Human values and culture can also filter or accommodate new perceptions regardless of the logicality of new information. These values may come from religion or faith, education or superstitions.

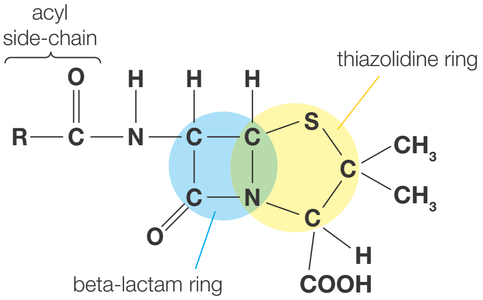
Antibacterials

Alexander Fleming was a bacteriologist who had worked in a hospital in France during World War I. On his return to England he took up a post in St Mary's Hospital in London where he studied an antibacterial agent present in tears and nasal mucus. Unfortunately, neither of these secretions was particularly active. However, egg white was identified as a better source of an antibacterial agent he called **lysozyme**.

His discovery of penicillin occurred when he left a pile of used culture plates containing a growth of a *Staphylococcus* bacterium on the end of a bench when he went on holiday. In his absence a rare strain of a fungus called *Penicillium notatum* entered through a window and landed on one of the plates. During a warm spell later in the month the fungus multiplied and inhibited the growth of the surrounding *Staphylococcus* bacteria. In nature *Penicillium* produces penicillin when its food resources are limited, to kill surrounding bacteria.

Fleming cultured the fungus and found it would inhibit the growth of a range of bacterial strains. Once the fungus had been identified Fleming named the extract **penicillin**. Fleming arranged for two of his graduate students to grow large quantities of the mould so that it could be characterized and later given to patients with a variety of bacterial infections with mixed success. Early research indicated that penicillin was not a protein and was not stable in the presence of acid. During the next decade the group at St Mary's continued to culture *P. notatum*.

Howard Florey was appointed as a professor in the Pathology Department of Oxford University in 1935 where he continued his studies of the antibacterial properties of lysozyme begun in 1927. In 1937, one of his students, Ernst Chain, showed that it attacked the cell walls of susceptible bacteria, causing their disintegration.



■ **Figure 25.60** General structure of the penicillins

In 1938 Chain, Abraham and Healey repeated Fleming's earlier work on the isolation of penicillin, but significantly improved the purity of their extract. They then began to produce large quantities of penicillin and conduct medical trials.

The Nobel Prize in Physiology or Medicine was jointly awarded to Fleming, Florey and Chain in 1945. The structure of a penicillin (penicillin G from *Penicillium chrysogenum*) was established using X-ray crystallography by Dorothy Hodgkin at Oxford. Penicillin molecules consist of a four-membered beta-lactam ring fused to a five-membered sulfur-containing thiazolidine ring (Figure 25.60).

ToK Link

Serendipitous discoveries

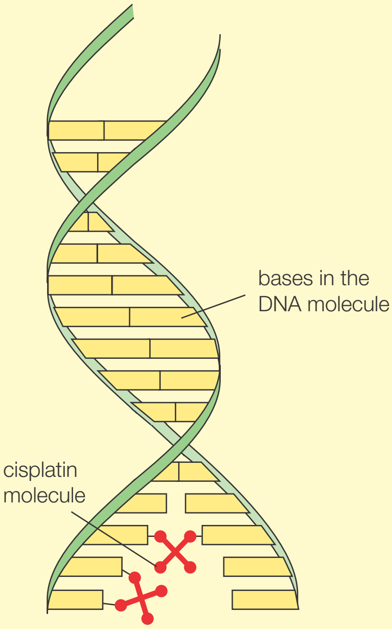
'Chance favours only the prepared mind' (Louis Pasteur). Fleming's discovery of penicillin is often described as serendipitous but the significance of his observations would have been missed by non-experts. What influence does an open-minded attitude have on our perceptions?

Serendipity refers to the accidental discovery of something important while looking for something else. The word derives from an old Persian fairy tale (The Three Princes of Serendip) and was introduced by Horace Wallace in 1754. The history of science has many examples of serendipitous discoveries and inventions. Pharmacology and chemistry are probably the fields where serendipity is most common. Although luck is obviously a factor one should bear in mind Louis Pasteur's maxim: 'in the field of observation, chance favours only the prepared mind'.

Fleming's discovery of penicillin provides an example of scientific serendipity. He failed to disinfect cultures of bacteria when leaving for his holiday, only to later find them contaminated with penicillium moulds. However, it should be noted that he had previously done extensive research into antibacterial substances.

Antibiotics are compounds produced by bacteria and fungi that are capable of killing, or inhibiting, competing microbial species. This phenomenon has long been known; it may explain why the ancient Egyptians had the practice of applying a poultice of mouldy bread to infected wounds.

The initial discovery of the anti-cancer drug cisplatin was also serendipitous. The American chemist Barnett Rosenberg (1926–2009) was looking into the effects of an electric field on the growth of bacteria. He noticed that bacteria ceased to divide when placed in an electric field and eventually pinned down the cause of this phenomenon to the platinum electrode he was using. A chemical reaction between the electrode and nutrients in the solution caused production of small amounts of cisplatin. Cisplatin was later shown to inhibit the replication of DNA (Figure 25.61).



■ **Figure 25.61** A diagram showing how cisplatin might disrupt DNA replication. It forms intra-strand DNA cross-links between adjacent bases on the same strand of DNA and inter-strand DNA cross-links between DNA strands.

Science involves more than common sense, logic and pure observation. Although reason and sharp powers of observation can lead to knowledge, they have limitations. Often break-throughs in science begin with questioning and rejecting authority – including scientific authority. The first attitude of science, therefore, is to be open-minded and sceptical and question authority, including scientific authority. Scientists are always open to accepting whatever theory the data (evidence) reveals, however strange or unexpected it may prove to be.

■ **Figure 25.62** Erythromycin tablets

Types of antibiotics

A **broad-spectrum antibiotic** is active against a wide range of bacterial diseases or bacterial strains. A **narrow-spectrum antibiotic** is active against only a small range of bacterial diseases or bacterial strains.

Examples of broad-spectrum antibiotics include ampicillin and amoxicillin (both are examples of penicillins), streptomycin, gentamicin, erythromycin (Figure 25.62), tetracycline and chloramphenicol. Examples of narrow-spectrum antibiotics include penicillin G, polymyxin and bacitracin.

How penicillins work

The bacterial cell wall, or envelope, contains a protective **peptidoglycan** 'mesh' made up of many long chains of alternating sugars (*N*-acetylmuramic acid (NAM) and *N*-acetylglucosamine (NAG)), cross-linked by short peptides (Figure 25.63). Penicillin resembles the dipeptide moiety (alanine–alanine) and functions by preventing proper cross-linking of the peptidoglycan layer when the bacteria are creating new cell walls, during periods of growth or cell division. Penicillin has no effect on peptidoglycan strands that are already cross-linked. Penicillins do not enter the bacterial cell to be active as long as they can access the periplasmic space between the cell membrane and the cell wall. The peptidoglycan cell wall is not a barrier either since it is so porous. However, additional outer barriers beyond the cell wall can prevent penicillins reaching the periplasmic space.

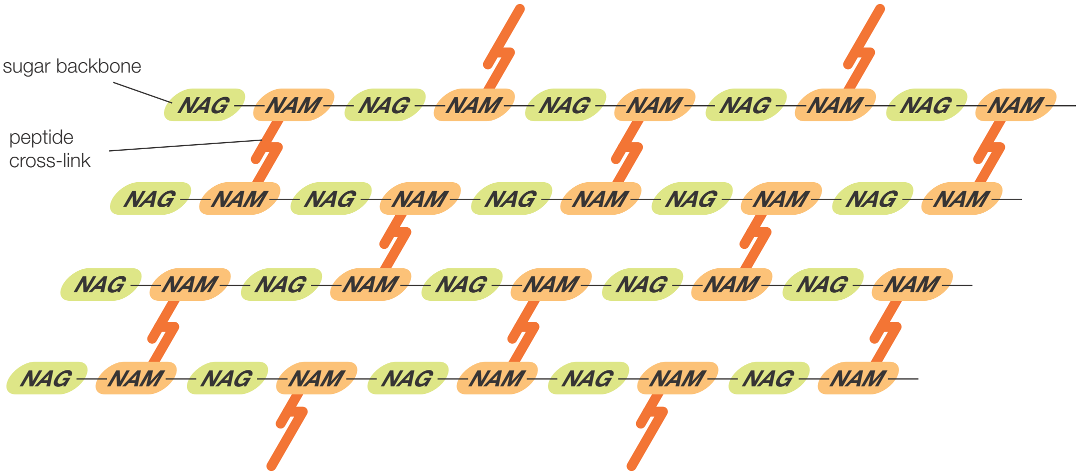


Figure 25.63 Simplified structure of the bacterial cell wall

The original *P. notatum* mould was found to produce a mixture of different penicillins that were given the letters G, X, F and K. The most useful of these was penicillin G because of its high effectiveness and the higher yield from the fermentation process.

Attempts were then made in the 1950s to modify the 'natural' penicillins chemically to vary the nature of the R group in Figure 25.60. Three new semi-synthetic penicillins were produced: S (R = $CH_3(CH_2)_3SCH_2-$), O (R = $CH_2{=}CHCH_2SCH_2-$) and V (R = $C_6H_5OCH_2-$). The only one still in use is penicillin V, which is acid stable. This means that the drug can be given orally rather than by injection. It also allows doctors to prescribe the drug for home use. However, the spectrum of antibacterial activity is the same as for penicillin G.

Bacteria possess enzymes named **penicillinases** that can hydrolyse penicillin to the inactive penicilloic acid. Penicillinase hydrolyses the four-membered beta-lactam ring present in penicillins. Bacteria resistant to penicillin due to penicillinase production have become widespread in hospitals.

A number of drugs are mixtures of penicillins with compounds that increase their potency. One such combination is Augmentin, which is amoxicillin with clavulanic acid (Figure 25.64). Clavulanic acid is a beta-lactamase inhibitor; that is, it stops the action of enzymes that degrade beta-lactam compounds, and thus protects the amoxicillin molecule from destruction.

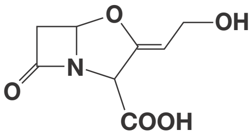


Figure 25.64 Structure of clavulanic acid

Vancomycin is often described as the antibiotic of 'last resort'. It is a glycopeptide antibiotic used only after treatment with all other antibiotics has failed. Vancomycin was first isolated in 1956 from a soil sample from the jungles of Borneo (Indonesia) by a missionary. It is administered by injection and, like penicillin, inhibits bacterial cell wall synthesis by interfering with peptidoglycan synthesis.

Bacterial resistance to antibiotics

Penicillin quickly became a widely used antibiotic against a variety of bacteria. During World War II it helped to conquer the biggest wartime killer: infected wounds. It was also used to treat sexually transmitted diseases that were common in soldiers. However, just four years after the drug companies began mass-producing penicillin in 1943, penicillin-resistant bacteria began to appear.

One reason why bacterial resistance to penicillin (and other antibiotics) has arisen is that they are often overprescribed. If people take penicillin for a minor bacterial infection that their immune system would be able to deal with, they are increasing the chances of a resistant strain emerging. When the non-resistant bacteria are destroyed, the small numbers of resistant forms will be able to multiply due to an absence of competition. A process of natural selection has occurred in the bacterial population. Similar reasoning can also be applied to people who do need to take antibiotics but who do not complete their course of treatment.

Bacterial resistance to penicillin comes in many forms. A mutation in bacterial DNA can limit the permeability of the cell wall, preventing uptake of penicillin into the periplasmic space between the cell wall and cell membrane. Or it may cause the bacteria to produce enzymes called penicillinases, which break down the penicillin molecule, or alter cross-linking enzymes involved in cell wall production so that penicillin can no longer inhibit them.

Bacteria can acquire antibiotic-resistant genes in a number of ways. The bacterial DNA may mutate spontaneously. Such a process involves a change in the sequence of bases along the length of DNA. In a form of sexual reproduction called conjugation or transformation, one bacterium may acquire DNA from another bacterium. Antibiotic resistance spreads most quickly when bacteria acquire small circular pieces of DNA called **plasmids**. A single bacterial plasmid can confer resistance to several different antibiotics.

Antibiotics have been widely used in the livestock industry to treat sick animals and, more significantly, to promote growth. Other drugs prescribed to livestock include hormones to fatten and promote growth steroids, to build up bulk and weight, and tranquilizers to reduce stress. Perhaps the most dangerous aspect of the use of antibiotics in livestock feeds is that they promote the emergence of antibiotic-resistant strains in the animals.

Tuberculosis (TB) is a common and often deadly infectious disease caused by *Mycobacterium tuberculosis*. Tuberculosis (Figure 25.65) is spread through the air by droplets when people who have the disease cough, sneeze or spit. TB usually attacks the lungs (pulmonary TB) but can also affect the central nervous system and the circulatory system. The classic symptoms of TB are a chronic cough (with blood in the sputum), night sweats and weight loss. People with HIV are especially prone to TB infection because of their compromised immune systems. TB treatment is difficult and requires long courses (6–12 months) of multiple antibiotics, often involving rifampicin and isoniazid. Antibiotic resistance is a growing problem in the treatment of TB.

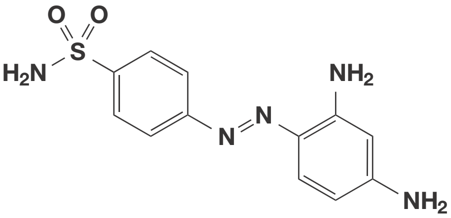
■ **Figure 25.65** TB service centre in Tanjong Pagar, Singapore, 1948

The use of antibiotics carries a number of other risks. Penicillins can cause allergic reactions in some people, and some antibiotics, such as nitrofuran, are carcinogenic at high concentrations. Other antibiotics, such as streptomycin (the first effective antibiotic against TB), can have serious side effects due to toxicity, such as loss of hearing (leading to deafness) and kidney damage.

Antibacterials

In 1924 the 16-year-old son of the president of the USA got a blister on the toe of his right foot while playing tennis. He contracted septicaemia and was dead within one week. Septicaemia is caused when certain virulent strains of bacteria get into the bloodstream. Twelve years later, in 1936, the son of president Franklin Roosevelt developed a septic sore throat; he was injected with a new antibacterial drug called Prontosil and made a complete recovery. This was the first commercially available antibacterial and also helped to reduce the death rate due to bacterial infections among women giving birth.

Prontosil (Figure 25.66) was found to be effective against some bacteria by a German chemist Gerhard Domagk (subsequently awarded the Nobel Prize in Medicine in 1939) who tested a large number of dyestuffs on bacteria. In 1932 he found that Prontosil worked well against streptococcal infections in mice. He later treated a 'strep' infection in his own daughter who was facing a possible amputation of her arm. Prontosil and then later penicillin in 1943 hugely reduced the mortality rate of soldiers in World War II.



■ **Figure 25.66** The structure of Prontosil, 4-[(2,4-diaminophenyl)azo]benzene sulfonamide

Prontosil is rarely used because some bacteria are resistant to its action and it can cause damage to the liver and other side effects. However, it was the first widely used and successful example of chemotherapy: using chemicals to treat diseases. It was also the first antibacterial and was widely used to save many lives from bacterial infection. Prontosil and penicillin transformed the treatment of bacterial diseases.

25.3 Opiates – *potent medical drugs prepared by chemical modification of natural products can be addictive and become substances of abuse*

Opiates

Morphine is a natural opiate found in **opium,** which is extracted from the latex in the seed capsule of the opium poppy. **Opiates** are natural **alkaloids** found in opium. Alkaloids are a diverse group of nitrogen-containing bases extracted from plants. The two main opiates present in opium are morphine and codeine. Heroin is readily synthesized from morphine via a simple single-step synthesis. Heroin, morphine and codeine are all powerful analgesics that act on proteins known as opioid receptors located on the surfaces of nerve cells in the brain, spinal cord and intestines. The molecules of these opiates have a common structure, and hence shape, as shown in Figures 25.67, 25.68 and 25.69. Opiates are also known as narcotics since they have sleep-inducing properties.

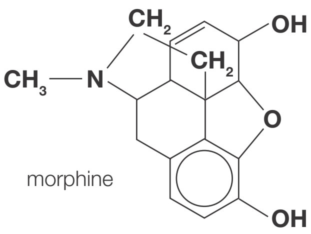


Figure 25.67 Structure of morphine

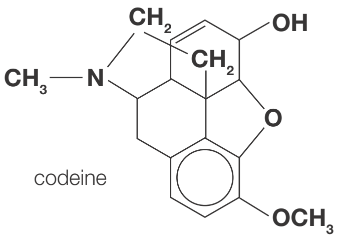


Figure 25.68 Structure of codeine

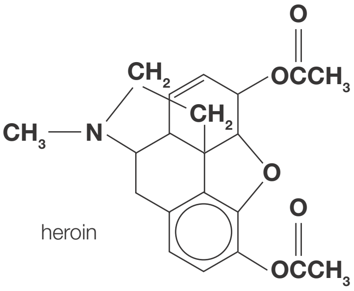


Figure 25.69 Structure of heroin

(The structures are also in the *IB Chemistry data booklet*.) The binding sites of the various opiate receptors have specific shapes and therefore allow only certain molecules to bind at the sites. The opiates fit into these receptors because they have the correct structure and molecular shape.

Codeine

Codeine (Figure 25.68) is an alkaloid found in opium, but most codeine used in medicine is synthesized from morphine (via the attachment of a methyl group). It is an analgesic (although significantly less potent than morphine) but it also has antitussive and antidiarrheal properties: it suppresses the urge to cough and restores normal bowel movements.

Codeine is sometimes marketed in combination preparations with paracetamol (acetaminophen) as co-codamol, or with aspirin as co-codaprin. These combinations provide greater pain relief than either agent alone due to the operation of the synergistic effect.

Like all opiates, codeine is addictive unless used infrequently. However, the withdrawal symptoms are relatively mild and as a consequence codeine is considerably less addictive than the other opiates.

Codeine is an example of a prodrug: a drug which is converted to a more active form inside the body. A small proportion of the codeine is converted by liver enzymes to morphine and a range of other compounds.

Diamorphine (heroin)

Diacetylmorphine (diamorphine) or heroin (Figure 25.69) is a semi-synthetic opiate, or opioid, derived from morphine. In its pure form it is a white powder that readily dissolves in water. Heroin can be swallowed or dissolved in water and injected into a vein – this method produces the most intense effect. Alternatively, it can be sniffed, or the fumes from the heated powder inhaled – this method is sometimes called 'chasing the dragon'.

When heroin was first synthesized in 1874 it was marketed as a safe non-addictive substitute for morphine. However, dependency or addiction quickly developed in some individuals and heroin and other opiates were made illegal in many countries, for example in 1920 in the USA.

As well as acting as a strong analgesic, heroin produces a range of side effects such as depression of the activity of the nervous system, including breathing and heart rate. Blood vessels are also widened (vasodilation) which reduces bowel activity and hence causes constipation.

Diamorphine can be prepared from morphine by reaction with ethanoic anhydride in a simple acetylation reaction (Figure 25.70). This is analogous to reaction between salicylic acid and ethanoic anhydride to form aspirin (acetylsalicylic acid) and ethanoic acid.

During the acetylation of morphine both hydroxyl groups are substituted with ester groups which significantly reduces the polarity of the molecule. This increases its lipophilicity (ability to dissolve in lipids) and hence its ability to cross the blood–brain barrier.

In the brain diamorphine is rapidly metabolized into morphine, which binds to the opioid receptor. This mechanism of action makes diamorphine about five times more potent an analgesic than morphine when injected into the blood.

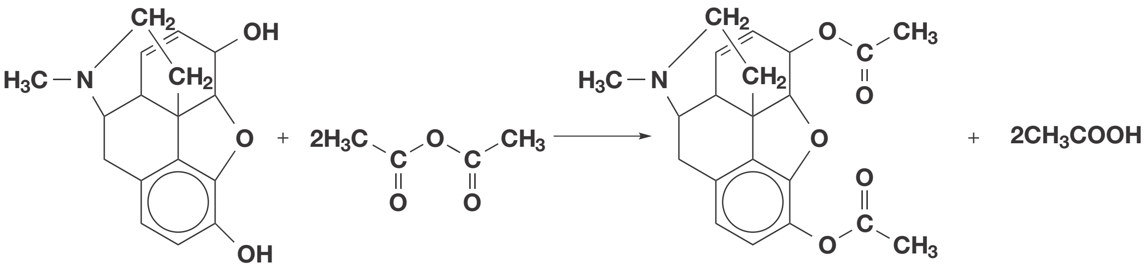


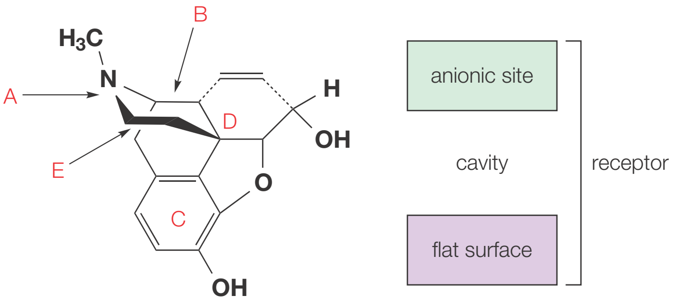
Figure 25.70 The formation of diamorphine

Comparison of the structures of morphine, codeine and diamorphine (heroin)

The term **pharmacophore** defines the important functional groups for the binding to the receptor and the activity of a drug molecule. Some pharmacophores are described by linking the important functional groups by a common carbon skeleton.

The opioid receptor is believed to have three main areas that bind morphine, codeine, diamorphine and related opioids. There is an ionic site which binds the tertiary amine nitrogen (A), which can be become positively charged via protonation; a cavity (hole) which accepts the piperidine ring (B); and a flat surface that binds the benzene ring (C) by London (dispersion) forces (Figure 25.71).

Figure 25.71 Opioid receptor and opioid molecule



5 List all the functional groups present in morphine, codeine and diamorphine (heroin).

A number of structural features have to be present for the opioid painkiller to be active, including the tertiary amine nitrogen with a small alkyl group (A), the presence of a piperidine ring (B), a quaternary carbon (D), a benzene ring (C) and a C2 'spacer' (E), part of the piperidine ring, between the tertiary nitrogen and the quaternary carbon. The molecules have a rigid structure which helps them bind to the active site of the opioid receptor.

Nature of Science

Discovery and uses of opium

Plants make molecules known as secondary metabolites using a series of enzyme-controlled reactions. The unripe seed pods of *Papaver somniferum* (see Figure 25.19) exude a creamy coloured sap, which contains a mixture of perhaps 50 different alkaloids – molecules made by the plant from amino acids. The sap dries to a yellow-brown paste known as opium. One of the most abundant compounds present in opium is the powerful painkiller morphine.

Opiates are alkaloid molecules containing nitrogen atoms in the form of amine functional groups. Since amines are bases, alkaloids form salts when reacted with acids just as alkalis do: they were originally referred to as plant alkalis and later the term alkaloid was created. It is the ability of alkaloids to readily form crystalline salts when treated with acids that makes them relatively easy to isolate in pure form.

Over 5000 years ago, the Mesopotamians (in southern Iraq) had discovered some of the euphoric properties of opium. Euphoria is a feeling of great happiness or well-being. By the time of the Greek and Roman civilizations, people were also familiar with its painkilling effects. It was probably Paracelsus, a famous 16th century alchemist, who first prepared the alcoholic solution of opium known as laudanum, a widely used painkiller.

At the start of the 19th century, several chemists isolated pure morphine from opium. Morphine exerts its effects in the body by binding to the mu-opioid receptor in the brain, causing analgesia and sedation. It is because of its sedative properties that morphine is named after *Morpheus*, the Greek god of dreams.

Diamorphine, better known today as heroin, was first marketed as a cough suppressant and, ironically, as a cure for morphine addiction. Although heroin is inactive in the body, it crosses the blood–brain barrier faster than morphine and is subsequently metabolized there to produce morphine. This gives the user the intense 'rush' that makes heroin so addictive. Today, heroin addiction affects millions of people around the world and claims thousands of lives every year.

Problems associated with heroin use

Heroin causes users to feel drowsy, warm and contented. It also relieves stress and discomfort by creating a relaxed detachment from pain. However, users quickly become physically dependent and tolerant to the drug: larger and larger doses must be taken to achieve the same state of euphoria (extreme happiness).

A sudden withdrawal from high doses of heroin produces a variety of withdrawal symptoms similar to the flu. Although they rarely last more than a week, feelings of weakness and anxiety can last for several months.

Many heroin addicts inject the drug (Figure 25.72) and the sharing of unsterilized needles can spread diseases such as tetanus, AIDS and hepatitis. There is also a risk of using impure drugs which have been mixed, or 'cut', with unknown and potentially harmful substances. Apathy and the reduced appetite caused by the drug can lead to a poor diet, and dependency on the drug can lead to financial problems which people try to resolve by theft or prostitution.

Methadone is a synthetic opioid frequently used to treat heroin addicts. Although chemically different from morphine and heroin, it acts on the same receptors and produces many of the same effects. Methadone's usefulness in treating heroin addicts is due to its long duration of effect and its ability to block the heroin withdrawal symptoms. At high concentrations it can block the euphoric effects of heroin and morphine.

Figure 25.72 Heroin which will be cooked in a spoon

History of opium

Alexander the Great took opium to Persia (modern day Iran) and India, from where it reached China. It was used medicinally as a painkiller. After the Chinese banned tobacco smoking in 1644, many Chinese turned to opium as an alternative and the British East India Company made large profits exporting to China. Chinese governments were unhappy with this and the Opium Wars of the 18th century were a result. One consequence of the Opium Wars was that Hong Kong became a British Crown Colony until 1997. Opium is a cash crop across a region stretching from the Middle East to East Asia. Ninety per cent of the world's opium and heroin originate in Afghanistan.

Morphine, heroin and cocaine and many other addictive substances of abuse are illegally produced in a small number of countries and then distributed globally by criminal organizations. According to the UN World Drug Report over 80 per cent of illegal opiates are produced in Afghanistan. Myanmar, Thailand, Laos and Latin America also produce opium.

Only two per cent of these drugs are used by the local population; the remaining 98 per cent are exported to Europe, Asia, Africa and North America. This amounts to an export value of about 4 billion US dollars. Afghanistan is also the largest producer of cannabis (mostly as hashish) in the world.

This is a reflection of differences in cultural and economic viewpoints. The problem of drug abuse is an international issue and can only be solved by recognizing and dealing with the differences through education, economic development, law enforcement and international cooperation.

ToK Link

Knowledge and culture

Cultures often clash over different perspectives and ideas. Is there any knowledge which is independent of culture?

Cultures often clash over different historical, religious, racial and political ideas and perspectives. There is often a perceived clash between science and different faiths or religions. However, it is worth noting the Harvard biologist Stephen Jay Gould's views on this matter. By nature, according to Gould, science and religion do not and cannot conflict, because their respective concerns are entirely distinct:

'Science tries to document the factual character of the natural world, and to develop theories that coordinate and explain these facts. Religion, on the other hand, operates in the equally important, but utterly different, realm of human purposes, meanings, and values – subjects that the factual domain of science might illuminate, but can never resolve.' http://skepticaljew.blogspot.sg/2010/05/stephen-jay-gould-on-science-and.html

Some knowledge and ways of knowing, such as emotion, may transcend culture. For example, the psychologist Paul Ekman showed that isolated New Guinean Highlanders could recognize the facial expression in the photographs of his American college students. Charles Darwin observed that children who are born blind and deaf from birth display virtually the full range of facial expressions, including happiness, sadness, anger, fear, disgust and surprise.

Apparently Eskimos have no word for anger and Tahitians do not recognize guilt, sadness or longing. Cultures may differ in how their members talk and write about their emotions, but may not reflect how people feel at the level of biological sense perceptions. The observation that a language does or does not have a word for an emotion may have very little significance. Whether a language appears to have a word for an emotion depends on the skill of the translator and the structure and history of the language.

It is likely that our basic emotions and our ability to recognize them and other basic sense perceptions of time, distance and space in other humans may be universal and to some extent influenced by genes.

25.4 pH regulation of the stomach *– excess stomach acid is a common problem that can be alleviated by compounds that increase the stomach pH by neutralizing or reducing its secretion*

Stomach acid

The process of digestion involves a series of enzyme-controlled catabolic reactions that transform large insoluble molecules into smaller soluble molecules. Many of these hydrolysis reactions take place in the stomach, where the food is mixed with a digestive fluid. This fluid, also known as gastric juice, is composed of water, salts (mostly KCI and NaCl, which are dissociated into their ions K^+, Cl^- and Na^+), hydrochloric acid (HCl) and a protease (pepsin) secreted by the parietal cells in the stomach lining. Pepsin and hydrochloric acid are responsible for the breakdown of proteins into peptides and individual amino acids. Other cells produce hydrogen carbonate ions (HCO_3^-) and gastric mucus to buffer the acid and prevent the gastric juice from digesting the delicate stomach tissues and leading to an ulcer.

The molar concentration of hydrochloric acid in the stomach varies from approximately 0.003 to 0.1 mol dm^{-3} (0.01–0.4%) which corresponds to a pH range of 2.5 to 1.0. The hydrochloric acid in gastric juice provides an optimum pH for pepsin in the gastric juice. In addition, hydrochloric acid acts a disinfectant, killing nearly all harmful microorganisms, especially bacteria, that are ingested with the food.

Worked example

A 10.0 cm^3 sample of gastric juice, obtained several hours after a meal, was titrated with 0.100 mol dm^{-3} NaOH to neutrality; 7.20 cm^3 of NaOH was required. Assume that no buffers were present. Deduce the pH of the gastric juice.

Amount of NaOH = $0.100\,\text{mol dm}^{-3} \times 7.20 \times 10^{-3}\,\text{dm}^3 = 7.20 \times 10^{-4}\,\text{mol}$

Amount of H$^+$(aq) in 10.0 cm^3 = $7.20 \times 10^{-4}\,\text{mol}$

$$[\text{H}^+(\text{aq})] = \frac{7.20 \times 10^{-4}\,\text{mol}}{0.010\,\text{dm}^3} = 0.072\,\text{mol dm}^{-3}$$

$\text{pH} = -\log_{10}(0.072) = 1.14$

Nature of Science

Early studies of stomach and gastric acid function

The presence of hydrochloric acid was discovered in 1838 by a US army surgeon William Beaumont, who was observing a patient with a gastric fistula (an unhealed hole in the stomach) left by a gunshot. By taking samples of gastric juice and using them to 'digest' pieces of protein containing foods in cups, Beaumont discovered that digestion was a chemical rather than a mechanical process. He also investigated the effects that temperature, exercise and even emotions have on the digestive process.

Further experiments revealed the negative effects of excess stomach acid, for example ulcers, which led to the development and uses of weak bases as antacids. The dose levels required for neutralization were large and caused unpleasant side effects. The study of how hydrochloric acid is generated by the cells lining the stomach led to the development (by rational drug design) of new pharmaceuticals. These drugs, such as ranitidine and omeprazole, regulate the acidity of the stomach by suppressing the secretion of hydrochloric acid. Ranitidine (Zantac) and cimetidine (Tagamet) function by binding to histamine receptors; proton pump inhibitors such as omeprazole (Prilosec) and esomeprazole (Nexium) function by inhibiting the proton pumps in parietal cells.

Antacids

Peptic ulcer disease

Ulcers of the stomach (gastric ulcers) and the first part of the small intestine (duodenal ulcers), known collectively as **peptic ulcers**, affect large numbers of people. Between 10 per cent and 20 per cent of adult men in Western countries will suffer from a peptic ulcer during their lives. The disease is very painful and can be fatal.

Ulcers are small damaged areas of the mucous membranes of the stomach or small intestine which expose the underlying muscle layers of the gut wall to hydrochloric acid and the enzyme pepsin. These are both produced and released by the cells lining the stomach, and are involved in the digestion of proteins into polypeptides. However, the acid is also concerned with killing bacteria and other invading pathogens.

For many years the main medical treatment for peptic ulcers was the use of **antacids** (Figure 25.73) to neutralize the gastric juice in the stomach. The earliest antacids were sodium hydrogen carbonate ('sodium bicarbonate') and calcium carbonate (chalk), which although efficient and rapid, are not recommended for long-term use.

Figure 25.73 Mylanta: a commercial antacid containing magnesium and aluminium hydroxides

$$\text{NaHCO}_3(\text{s}) + \text{HCl}(\text{aq}) \rightarrow \text{NaCl}(\text{aq}) + \text{H}_2\text{O}(\text{l}) + \text{CO}_2(\text{g})$$

$$\text{CaCO}_3(\text{s}) + 2\text{HCl}(\text{aq}) \rightarrow \text{CaCl}_2(\text{aq}) + \text{H}_2\text{O}(\text{l}) + \text{CO}_2(\text{g})$$

Excessive use of sodium hydrogen carbonate may lead to alkalosis (a rise in the pH of blood) and fluid retention ('bloating'). Repeated use of chalk as an antacid may lead to excessive amounts of calcium ions being absorbed into the body.

Both antacids, but in particular sodium hydrogen carbonate (bicarbonate), suffer from 'acid rebound' where acid levels rise after neutralization as the stomach senses the change and stimulates acid secretion via production of a hormone called gastrin. Calcium ions can cause constipation so they are frequently combined with magnesium ions in commercial preparations of antacids.

Aluminium hydroxide and magnesium hydroxide (milk of magnesia) are two other compounds used as antacids. Aluminium-containing antacids are slow acting and buffer the pH at acidic pH values (3–5). Aluminium ions have an astringent taste and may cause constipation.

$$Al(OH)_3(s) + 3HCl(aq) \rightarrow AlCl_3(aq) + 3H_2O(l)$$

$$Mg(OH)_2(s) + 2HCl(aq) \rightarrow MgCl_2(aq) + 2H_2O(l)$$

Aluminium ions can also prevent uptake of phosphate ions, due to precipitation of aluminium phosphate, and will bind to certain drugs because of their large charge density (a function of aluminium's small ionic radius and high charge). Magnesium hydroxide (milk of magnesia) is a rapid and efficient antacid, but it does also act as a laxative. Aluminium and magnesium ions are frequently combined together in commercial antacids. Calcium hydroxide is sometimes also used as an antacid.

■ **Figure 25.74** Macgel (chewable antacid): contains aluminium hydroxide, magnesium hydroxide and polydimethylsiloxane (dimethicone)

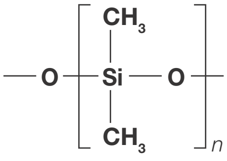
Some antacids are very weak bases and act very slowly but over a long period of time. Magnesium trisilicate slowly reacts with hydrochloric acid to form magnesium chloride and gelatinous silicon dioxide, which protects the mucosa (lining) of the stomach.

Sucralfate (Carafate) is a hydrated complex of aluminium hydroxide with a derivative of sucrose frequently prescribed to people with peptic ulcer disease. It has no neutralizing capability, but forms a sticky viscous gel that protects the damaged membrane from further attack by hydrochloric acid and pepsin.

Dyspepsia is derived from the Greek word *pepsis* meaning digestion. Dyspepsia is commonly known as indigestion and is frequently due to expulsion of gastric juice from the stomach – acid reflux.

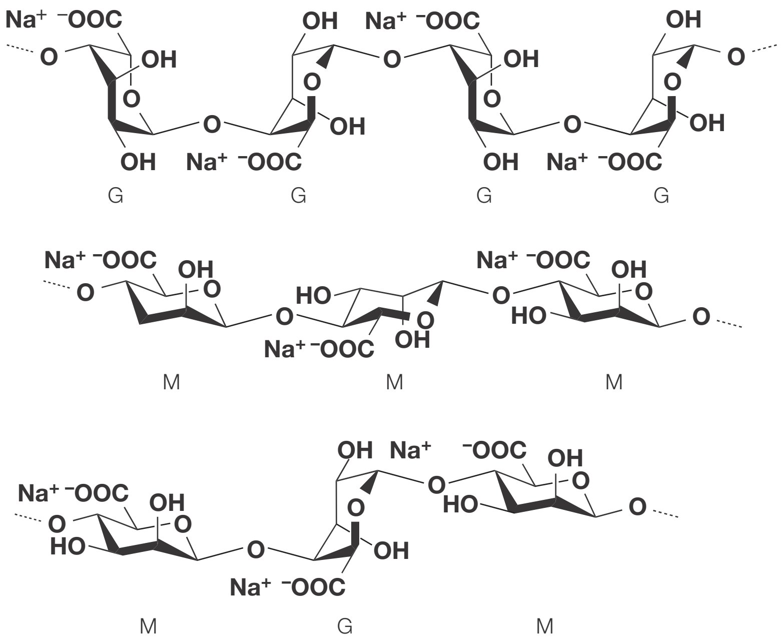
Dimethicone is an anti-foaming agent commonly present in antacid tablets (Figure 25.74) and other pharmaceuticals. It is also used as a defoaming agent in the food processing industry. It is a surface active ingredient that reduces the surface tension of large bubbles, causing them to coalesce and produce a foam.

It is a colourless viscous liquid with a low melting point. Its anti-foaming and anti-flatulent properties are effective even when it is present at low concentrations. In addition it is chemically inert and does not interfere with the process of neutralization and is non-toxic to humans. Dimethicone is a polydimethylsiloxane whose structure is shown in Figure 25.75.



■ **Figure 25.75** Structure of a polydimethylsiloxane

Alginates (Figure 25.76) are derived from alginic acid, a slimy polysaccharide that prevents brown seaweeds from excessive drying and dehydration when exposed to air during low tides. Alginic acid is used in industry as a thickening agent and emulsion stabilizer and is found in whipped toppings, sauces, dressings, gravy, ice cream, milk shakes, jellies and custards. Alginic acid also readily produces soft fibres which are used in light-weight wound dressings. They keep the wound moist and are able to absorb large amounts of fluid. Alginates can bind with toxic 'heavy metals' and remove them from the body, since alginates are not absorbed into the body from the intestines.



■ **Figure 25.76** Three sections of sodium alginate polymer molecules showing the two different monomers, 'G' and 'M'

6 Although it is a large molecule, sodium alginate is quite soluble in water. Explain why you would expect this.

Alginates are frequently present in antacids. They produce a neutralizing layer on top of the contents of the stomach and prevent 'heartburn' when gastric juice moves up from the stomach through the pyloric sphincter (a ring of circular muscle at the top of the stomach) into the oesophagus. Certain foods and drinks, and even lying down to sleep, can cause heartburn.

Worked example

The antacid tablet in a glass of water contains 810.0 mg of magnesium oxide per tablet. Determine the volume of 0.1000 mol dm^{-3} stomach acid (HCl) that one tablet could neutralize.

Amount of magnesium oxide = (810.0/1000) g × (24.31 + 16.00) g mol^{-1} = 0.02025 mol

$MgO(s) + 2HCl(aq) \rightarrow MgCl_2(aq) + H_2O(l)$

Amount of HCl = 2 × 0.02025 mol = 0.04050 mol

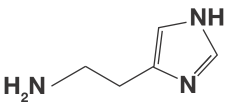
Volume = Amount/molarity; volume = 0.04050 mol/0.1000 dm^3; volume = 0.4050 dm^3 = 405 cm^3

■ The causes of peptic ulcers

Modern treatments of peptic ulcer disease have depended on an understanding of the various mechanisms that control gastric secretions. Before food arrives in the stomach it stimulates the secretory gastric cells in the stomach lining to produce acetylcholine and histamine (Figure 25.77).

When the food enters the stomach the organ becomes enlarged, which stimulates the gastric cells to release a hormone called gastrin. Gastrin, together with acetylcholine and histamine, stimulate the release of gastric juice (hydrochloric acid and pepsin) from parietal gastric cells. Another type of gastric cell secretes a glycoprotein called mucin, which coats the stomach to protect it against the actions of hydrochloric acid and pepsin. Mucin is continually produced but only slowly digested. However, if the production of mucin is inhibited then part of the stomach wall or mucosa could be digested, leading to a gastric ulcer.

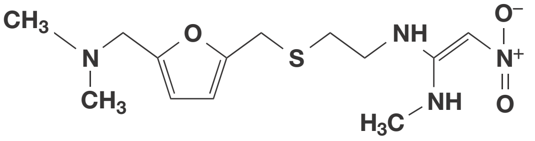
The histamine acts directly on the acid-secreting parietal cells where it stimulates the potassium–hydrogen ion 'pump'. This is a membrane protein that uses the energy from respiration to remove potassium ions from the gastric juice in the stomach and replace them with hydrogen ions.



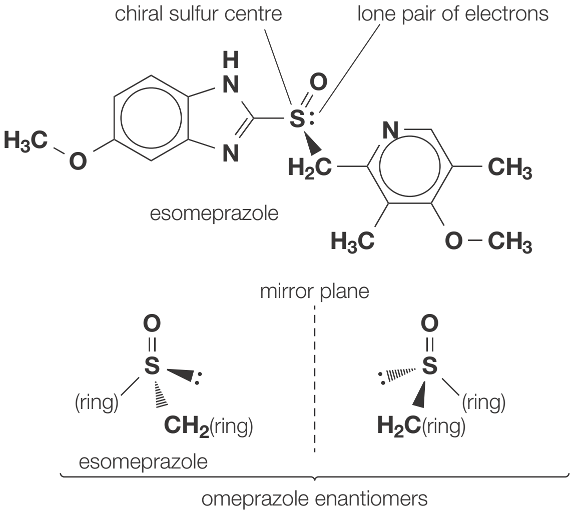
■ **Figure 25.77** Structure of histamine

One of the most common and successful drugs for the treatment of stomach ulcers is ranitidine (Zantac) (Figure 25.78). It is classified as an anti-secretory agent because it inhibits the secretion of gastric juice. Its molecular structure has some similarities to histamine and ranitidine acts as an antagonist of the histamine receptor. Therefore, this blocks the process by which the potassium–hydrogen ion pump is stimulated.

The drugs and histamine are not binding directly to the proton pump. They bind to the histamine receptor. If the histamine receptor is activated by histamine, the potassium–hydrogen ion proton pump is deactivated. This means that the pump no longer operates and the secretion of gastric juice in the stomach is significantly reduced, allowing the stomach to repair the damage caused by the ulcer.



■ **Figure 25.78** Structure of ranitidine (Zantac)



■ **Figure 25.79** Structures of omeprazole

Proton pump inhibitors were the next step in drug development to reduce acid secretion in the stomach and tests have shown that they are very effective compared with Zantac and placebos. They work by directly inhibiting the H^+-K^+-ATPase enzyme, a membrane-based protein that uses ATP to pump acid (protons) into the stomach.

Omeprazole (Prilosec) (Figure 25.79) is a well-known example of this type of drug. It is a substituted benzimidazole and binds to a specific amino acid of the enzyme, forming a covalent disulfide link and thereby inhibiting acid secretion. Omeprazole is a racemic mixture of two enantiomers. It is optically active because the molecule is chiral due to the presence of three different substituents and a lone pair of electrons on the chiral sulfur atom.

Omeprazole exists as two enantiomers. The R-form is inactive, though it converts into the active S-enantiomer *in vivo* (in the body). The pharmaceutical company AstraZeneca sells the S isomer drug as Nexium (esomeprazole).

In contrast to many drugs, both enantiomers of omeprazole show very similar pharmacological properties. In their original forms (known as prodrugs) they are biologically inactive and do not interact with the gastric proton pump directly.

Due to their low polarity, omeprazole and esomeprazole readily cross the lipid bilayer of cell membranes and enter the cytoplasm of parietal cells containing hydrochloric acid. They are actually activated once they exit the parietal cells into the stomach. The proton pump is pumping out protons as soon as they are generated, and so the increased acid conditions are near the outer surface of cell and not within the cell itself. Therefore, the drugs are activated once they depart the cell and then react with the pumps on the outer surface of the parietal cell. If the drugs were activated inside the cell, they would have adverse effects on other proteins within the cell.

■ **Figure 25.80** Electron micrograph of *H. pylori* possessing multiple flagella (negative staining)

In this acidic environment near the proton pump both enantiomers undergo a series of acid-catalysed chemical changes and produce the same active metabolites, which bind to the proton pump of the parietal cell (via a sulfur–sulfur bridge) and inhibit the secretion of hydrochloric acid. Omeprazole also becomes ionized which makes it polar and hence unable to cross back into the cell through the cell membrane This mechanism of action increases the efficiency and bioavailability of both drugs and allows a reduced frequency of administration. Omeprazole and proton pump inhibitors have few side effects because of their selectivity and mechanism of action.

Helicobacter pylori

For many years the ultimate cause of ulcers was not known and it was widely believed that hard work, high stress and a poor diet were the causes of ulcers. However, in the 1980s evidence accumulated that ulcers are caused by a bacterium, *Helicobacter pylori* (Figure 25.80). *H. pylori* was first isolated by two Australian scientists, Robin Warren and Barry Marshall, in 1982. They were awarded the Nobel Prize in Physiology or Medicine in 2005 for their work on *H. pylori* and peptic ulcer disease.

H. pylori is found under the layers of mucus in the stomach lining, where the pH is only 4 compared to the value of 1 for the gastric juice. The bacterium produces large quantities of an enzyme called urease which hydrolyses urea and leads to the production of ammonia. This protects the *H. pylori* bacteria from the surrounding hydrochloric acid. A breath test has been developed to detect the ammonia from *H. pylori* infections.

$$\underset{\text{urea}}{CO(NH_2)_2(s)} + H^+(aq) + 2H_2O(l) \rightarrow HCO_3^-(aq) + 2NH_4^+(aq)$$

The treatment of ulcers has therefore changed and currently the treatment involves antacids, drugs such as ranitidine (Zantac), as well as antibiotics, such as amoxicillin and metronidazole. Recent work has also shown that the presence of *H. pylori* increases the risk of gastric (stomach) cancer sixfold, and accounts for about half of all gastric cancers. *H. pylori* is believed to be transmitted orally.

Nature of Science

Dyspepsia

The term 'dyspepsia' describes a clinical problem referring to a cluster of upper gastrointestinal symptoms that has been defined in many ways. Symptoms need to be present for 4 weeks and include upper abdominal pain or discomfort, heartburn, acid reflux, nausea and vomiting.

Dyspepsia is a symptom, not a diagnosis. Patients with dyspepsia can be divided into subgroups on the basis of an endoscopic diagnosis, where a camera is introduced into the oesophagus (esophagus).

Gastro-oesophageal reflux disease refers to patients with characteristic heartburn and acid regurgitation. Approximately half of these patients will also have oesophagitis, an inflammation of the lining of the oesophagus. Peptic ulcer disease can be subdivided into gastric and duodenal ulcers. *Helicobacter pylori* and the use of non-steroidal anti-inflammatory drugs, such as aspirin, are the major causes. Oesophageal and gastric cancer can also cause dyspepsia.

Dietary and lifestyle choices may contribute to dyspepsia. Certain foods and medications, for example ibuprofen, may trigger relaxation of the stomach sphincter muscle, causing acid reflux and heartburn. Studies have shown that cigarettes, spicy food, stress, fatty food, carbonated drinks, alcohol, obesity, pregnancy and lifting can also cause dyspepsia. Eating smaller and more frequent meals may relieve some of the symptoms.

ToK Link

Problem solving

Sometimes we utilize different approaches to solve the same problem. How do we decide between competing evidence and approaches?

When a hypothesis explains scientific evidence or another hypothesis, the two cohere with each. This implies both should be accepted or rejected. In contrast, if two hypotheses contradict each other, or if they offer some competing explanations of some scientific evidence, then they are incoherent with each other. This should lead to accepting one and rejecting the other.

In the early 1980s there were two competing hypotheses for the cause of peptic ulcers: excess acidity and the newly discovered *H. pylori.* Excess acidity as the cause of ulcers was a well-established theory with supporting evidence from the healing of ulcers using drugs such as cimetidine. The hypothesis that gastric bacteria exist was incompatible with the accepted assumption that the stomach was too acidic for permanent bacterial growth. Only a few researchers had observed the bacteria and there was only a small amount of evidence connecting ulcers and bacteria. It was understandable that Marshall's hypothesis about *H. pylori* being the cause of ulcers was not accepted. The bacteria that Warren observed could be explained as the result of contamination. Later in the 1990s, many researchers had studied *H. pylori* and the idea of the stomach as being bacteria free had been discarded. Many studies from different researchers using different techniques had shown that ulcers could be cured using antibiotics that eradicated the *H. pylori* bacteria. Researchers also suggested that *H. pylori* increases stomach acidity and that the removal of *H. pylori* decreases acidity. Maximizing coherence therefore required acceptance of the hypothesis that ulcers are caused by bacteria.

Acid–base buffers

Gastric juice undergoes large changes in the concentration of hydrogen ions, $H^+(aq)$, by a factor of 100, which corresponds to a change of 2 pH units. The pH of gastric juice will vary due to the nature of the food being digested and whether the stomach is empty or full of digesting food (chyme).

In contrast, the pH of other biological fluids (e.g. blood) remains relatively constant because of the presence of buffers. The pH of blood is buffered at around 7.4.

Buffers (see Chapter 18) are aqueous acid–base systems that can neutralize small amounts of strong acids and strong bases without their pH changing significantly. Each acid–base buffer contains a conjugate acid–base pair which differ by a single proton (H^+). The conjugate acid is the protonated species and the conjugate base is able to accept a proton from the conjugate acid.

For example, a mixture of propanoic acid and sodium propanoate behaves as an acidic buffer whose pH will be less than 7. The propanoic acid molecule, CH_3CH_2COOH, can donate a proton (from the –COOH group) and the propanoate anion, $CH_3CH_2COO^-$, can accept a proton. Hence, propanoic acid is the conjugate acid and the propanoate anion is the conjugate base.

In buffer solutions, the conjugate acid and the conjugate base exist in equilibrium. The conjugate acid is strong and the conjugate base is weak, so the equilibrium lies to the left. For example:

$$CH_3CH_2COOH(aq) \rightleftharpoons CH_3CH_2COO^-(aq) + H^+(aq)$$

The sodium ion is a spectator ion (see Chapter 1) and hence not shown in the equilibrium reaction.

The strength of an acid can be described by its dissociation constant, K_a or K_b, or its negative logarithm ($pK_a = -\log_{10}K_a$ and $pK_b = -\log_{10}K_b$).

$$K_a = \frac{[\text{conjugate base}] \times [H^+(aq)]}{[\text{conjugate acid and } pK_a]} = \frac{-\log_{10}[\text{conjugate base}] \times [H^+(aq)]}{[\text{conjugate acid}]}$$

These two dissociation constants are related via the ionic product of water, K_w: $K_a \times K_b = K_w$ and $pK_a + pK_b = pK_w$.

Since $pH = -\log_{10}[H^+(aq)]$, the pK_a expression can be mathematically transformed into the Henderson–Hasselbalch equation (see Chapter 18 for its derivation):

$$pH = \frac{pK_a + \log_{10}[\text{conjugate base}]}{[\text{conjugate acid}]}$$

An equivalent expression is:

$$pOH = \frac{pK_b + \log_{10}[\text{conjugate base}]}{[\text{conjugate acid}]}$$

The Henderson–Hasselbalch equation allows chemists to calculate the pH of a buffer solution with known acid–base composition (in terms of concentrations), or the concentrations of the conjugate acid and base in a solution with a known pH.

Worked examples

Calculate the pH of a standard solution containing 0.20 mol dm^{-3} HF(aq) and 0.40 mol dm^{-3} KF(aq). The K_a for hydrofluoric acid is 6.8×10^{-4}.

$pH = pK_a + \log_{10}[\text{salt}]/[\text{acid}]$

$pH = -\log_{10}(6.8 \times 10^{-4}) + \log_{10}(0.4)/(0.2) = 3.47$

Calculate the mass in grams of ammonium chloride, NH_4Cl, that must be added to 0.030 mol of ammonia, NH_3, to prepare 1.00 dm^3 of buffer solution with a pH of 8.4. The K_a of ammonia is 6.8×10^{-4} and the molar mass of NH_4Cl is 53.50 g mol^{-1}.

$pOH = 14.0 - 8.4 = 5.6$

$pOH = pK_b + \log_{10} [salt]/[weak\ base]$

$5.6 = -\log_{10} (1.8 \times 10^{-5}) + \log_{10} [NH_4Cl]/[NH_3]$

$0.86 = \log_{10} [NH_4Cl]/0.030$

Amount of $NH_4Cl = 0.22\,mol\,dm^{-3} \times 1.00\,dm^3 = 0.22\,mol$

Mass of $NH_4Cl = 0.22\,mol \times 53.50\,g\,mol^{-1} = 11.77\,g = 12\,g$

Determine the ratio of potassium methanoate, HCOOK, to methanoic acid, HCOOH, to produce a buffer of pH 4.00. The pK_a of HCOOH is 3.75.

$pH = pK_a + \log_{10}[HCOOK]/[HCOOH]$

$4.0 = 3.75 + \log_{10}[HCOOK]/[HCOOH]$

$\log_{10}[HCOOK]/[HCOOH] = 0.25$

$[HCOOK]/[HCOOH] = 1.78$

The ratio of HCOOK to HCOOH = 1.78 to 1.

Determine the amount (in mol) of HCl that must be added to 0.80 mol of CH_3COOK to prepare 1.00 dm^3 of buffer solution with a pH of 4.50. The K_a of ethanoic acid, CH_3COOH, is 1.8×10^{-5}.

Equation:	HCl	+ CH_3COONa →	CH_3COOH +	NaCl
Initial amount (mol):	y	0.80	0	0
Change in amount (mol):	$-y$	$-y$	$+y$	$+y$
Final amount (mol):	0	y	y	$-y$

$pH = pK_a + \log_{10}[CH_3COONa]/[CH_3COOH]$

$4.50 = -\log_{10}(1.8 \times 10^{-5}) + \log_{10}(0.80-y/y)$

$\log_{10}(0.80-y/y) = -0.245$

$y = 0.51$; $[HCl(aq)] = 0.51\,mol\,dm^{-3}$, hence n $HCl(aq) = 0.51\,mol$

0.025 mol of HCl is added to 1.00 dm^3 of a buffer solution containing 0.35 $mol\,dm^{-3}$ NH_3 and 0.50 $mol\,dm^{-3}$ NH_4Cl. Calculate the pH of the solution after the addition of HCl. Assume that the volume of the solution does not change. The pK_b of ammonia is 4.75.

Equation:	HCl	+ NH_3 →	NH_4Cl
Initial amount (mol):	0.025	0.35	0.50
Change in amount (mol):	−0.025	−0.025	+0.025
Final amount (mol):	0	0.325	0.525

$pOH = pK_b + \log_{10}[NH_4Cl]/[NH_3]$

$pOH = -\log_{10}(1.8 \times 10^{-5}) + \log_{10}(0.525/0.325) = 4.95$; $pH = 9.05$

Determine the pH of a buffer solution prepared by dissolving 2.00 g of benzoic acid and 5.00 g of potassium benzoate in 250 cm^3 of water. K_a benzoic acid = 6.3×10^{-5}.

Molar masses: M_r $(C_6H_5COOH) = 122$, M_r $(C_6H_5COO^-K^+) = 160$,

Amount of acid $C_6H_5COOH = 2.00\,g/122\,g\,mol^{-1} = 0.0164\,mol$; molarity $= 0.0656\,mol\,dm^{-3}$

Amount of salt $C_6H_5COO^-K^+ = 5.0/160\,g\,mol^{-1} = 0.0313\,mol$; molarity $= 0.125\,mol\,dm^{-3}$

$[H^+(aq)] = K_a\,[acid(aq)]/[salt(aq)]$

$[H^+(aq)] = 6.3 \times 10^{-5} \times 0.0656 / 0.125 = 3.31 \times 10^{-5}\,mol\,dm^{-3}$

$pH = -\log_{10}[H^+(aq)] = -\log_{10}(3.31 \times 10^{-5}) = 4.48$

Hydrogen carbonate and carbonate buffers (in the body)

Blood is a complex liquid that contains cells suspended in plasma, which contains various ions and molecules dissolved in water. These dissolved solutes in blood contain buffer systems which regulate the body pH to a constant value of 7.4.

The main buffer system in blood consists of carbonic acid, $H_2CO_3(aq)$, hydrogen carbonate ions, $HCO_3^-(aq)$, and carbon dioxide, $CO_2(aq)$. Carbon dioxide is produced by the respiration (oxidation) of glucose in all body tissues. The carbon dioxide released by respiration diffuses out of the body cells into the blood and is transported in the plasma to the lungs, where it is exhaled (breathed out).

The following equilibria are responsible for the buffering action of the carbonic acid–hydrogen carbonate ion buffer. The equilibrium constants for reactions 1, 2 and 3 are K_1, K_2 and K_3, respectively:

1 $H_2CO_3(aq) + H_2O(l) \rightleftharpoons H_3O^+(aq) + HCO_3^-(aq); K_1$
2 $CO_2(aq) + H_2O(l) \rightleftharpoons H_2CO_3(aq); K_2$
3 $CO_2(g) \rightleftharpoons CO_2(aq); K_3$
4 $CO_2(g) + 2H_2O(l) \rightleftharpoons H_3O^+(aq) + HCO_3^-(aq); K_1 \times K_2 \times K_3 = K_4$

The overall reaction represented by equation 4 indicates that the concentration of oxonium ions, $H_3O^+(aq)$, and the pH of blood depend only on the concentration of hydrogen carbonate ions dissolved in blood and on the partial pressure of gaseous carbon dioxide, $CO_2(g)$, in the air spaces in the lungs.

Worked example

0.10 mol of solid sodium hydrogen carbonate and 0.20 mol of solid sodium carbonate are dissolved in the same beaker of water, transferred to a volumetric flask and made to 250.0 cm^3. The K_a for the hydrogen carbonate ion, HCO_3^-, is 4.7×10^{-11}. Determine the pH of the resulting buffer.

$$pH = pK_a + \log_{10}([A^-]/[HA]) = -\log_{10}(4.7 \times 10^{-11}) + \log_{10}(0.2/0.1) = 10.6$$

Buffer pH range

The ability of acid–base buffers to resist pH changes when strong acids or bases are added is limited. Its buffering ability depends on the concentrations and ratios of the conjugate acid and base in the aqueous solution of the buffer.

At pH = pK_a, an acidic acid–base buffer is able to neutralize the greatest amounts of acids or bases before any significant pH change occurs (which is known as breaking the buffer; see Chapter 18). The Henderson–Hasselbalch equation predicts that the ratio between the concentrations of a conjugate acid–base pair increases or decreases by a factor of 10 when the pH of the buffer solution changes by one unit. Therefore an acidic acid–base buffer can be used from pH = pK_a − 1 to pH = pK_a + 1.

For example, a dihydrogen phosphate buffer (pK_a = 2.12) works efficiently between pH = 3.12 and pH = 1.12. However, outside this pH range the concentration of one of the buffer components becomes low and the buffer loses its ability to maintain a constant pH in the solution.

25.5 Antiviral medications *– antiviral medications have been developed for some viral infections while others are being researched*

■ Viruses and bacteria

Viruses

Viruses are acellular parasites that replicate inside living cells. They consist of a nucleic acid (DNA, usually double stranded, or RNA, usually single stranded) surrounded by a protein capsid and in some cases a membrane-like envelope. The capsid consists of multiple protein units (capsomeres) arranged in a regular helical or polyhedral structure. X-ray diffraction is used to study the crystalline structure of viruses.

Viruses can be divided into four types, depending on the make-up of their nucleic acid:

1 single-stranded DNA
2 double-stranded DNA
3 single-stranded RNA
4 double-stranded RNA.

The RNA viruses can be divided again into two groups, depending on how their RNA is reproduced and used in the host cell. In ordinary RNA viruses, nucleic acid replication occurs entirely in the cytoplasm, but in RNA retroviruses their RNA is used as a template for making viral DNA, using a viral enzyme called reverse transcriptase. The RNA in human immunodeficiency virus (HIV) is single stranded. Once in the host cell it acts as the template for the formation of an RNA/DNA hybrid. The RNA strand is degraded and the surviving DNA strand then acts as a template for double-stranded DNA.

Once a virus or its nucleic acid enters a cell the host cell's enzymes and ribosomes are 'hijacked' to manufacture many new viral proteins and enzymes that self-assemble into viruses. The viruses will then exit from the cell through the cell membrane, leaving behind a dead or damaged cell. In some viral infections involving retroviruses, the viral DNA becomes integrated into the host's DNA and may not kill the host or cause any obvious illness.

Familiar animal diseases caused by viruses include warts, herpes, viral meningitis, ebola, measles, influenza (flu), rabies, hepatitis B, SARS, avian influenza ('bird flu') and AIDS (caused by HIV). Plants and bacteria are also prone to viral infection. The structures of the flu and herpes viruses are shown in Figure 25.81. The best defence against viruses has been immunization which involves injecting deactivated virus (capsid) which will allow the immune system to mount a response to a future infection. The viral proteins in the capsid are known as antigens and are recognized by the antibodies of the immune system.

7 Find out about smallpox and the role of vaccination in eradicating it.

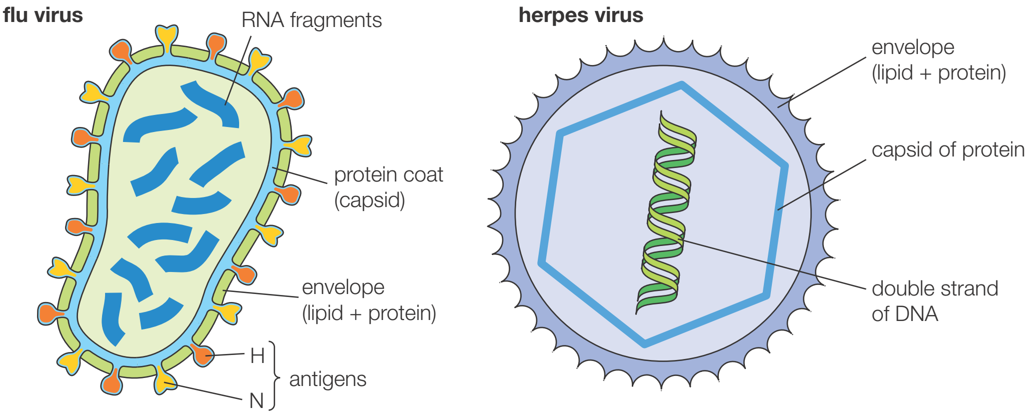
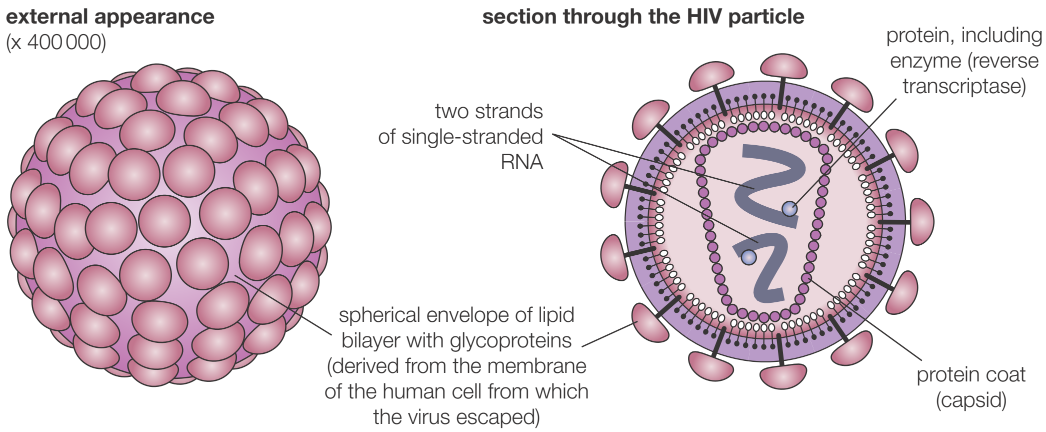


Figure 25.81 The structures of the flu and herpes viruses

The most familiar **retrovirus** is HIV (Figure 25.82), which is responsible for the condition known as acquired immune deficiency syndrome (AIDS). It reproduces inside a group of specialized white cells in the blood known as T4 lymphocytes or T4 helper cells, important regulators of the immune system, which is concerned with resisting and fighting infectious bacteria, fungi and viruses.

Figure 25.82 The external appearance and interior structure of HIV


external appearance
(x 400 000)
section through the HIV particle
protein, including enzyme (reverse transcriptase)
two strands of single-stranded RNA
spherical envelope of lipid bilayer with glycoproteins (derived from the membrane of the human cell from which the virus escaped)
protein coat (capsid)

Retroviruses like HIV are unusual because their genetic information is in the form of two single-stranded RNA molecules. The key step in infection is the interaction between a binding protein (a glycoprotein, gp120) on the HIV capsule and the CD4 receptor (also a glycoprotein) on the surface of the lymphocyte (a type of white blood cell). Once inside the T4 helper cell a viral enzyme called **reverse transcriptase** (hence the name retrovirus) produces a DNA copy of the RNA. This is then inserted into the DNA of the T4 lymphocyte (Figure 25.83).

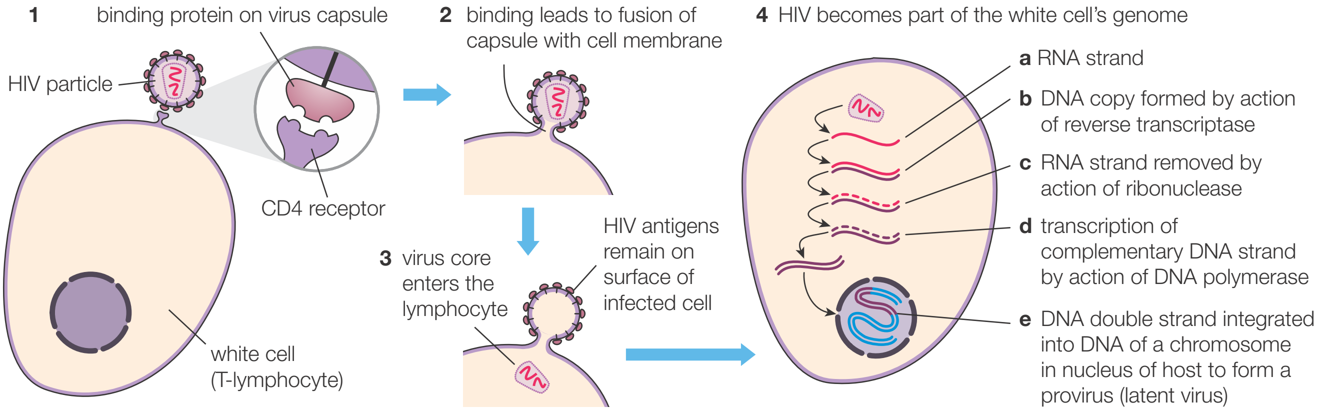

1 binding protein on virus capsule
HIV particle
CD4 receptor
white cell (T-lymphocyte)
2 binding leads to fusion of capsule with cell membrane
3 virus core enters the lymphocyte
HIV antigens remain on surface of infected cell
4 HIV becomes part of the white cell's genome
a RNA strand
b DNA copy formed by action of reverse transcriptase
c RNA strand removed by action of ribonuclease
d transcription of complementary DNA strand by action of DNA polymerase
e DNA double strand integrated into DNA of a chromosome in nucleus of host to form a provirus (latent virus)

Figure 25.83 HIV infection of a white blood cell

Bacteria

Bacteria are single-celled organisms that lack membrane-bound organelles and a true nucleus. There is no nuclear membrane and the only organelles present in the cytoplasm are ribosomes. Bacteria reproduce asexually via binary fission ('splitting in two'). They can also engage in conjugation which involves the lateral transfer of circular pieces of DNA. These are known as plasmids and are often responsible for antibiotic resistance. The cell can make multiple copies and transfer copies during conjugation.

Bacteria (Figure 25.84) often have a protective outer membrane and an inner plasma membrane that encloses the cytoplasm and the nucleoid – the area of cytoplasm that contains a single circular DNA molecule. Between the inner and outer membranes (the latter is not always present) is a thin but strong layer of sugar polymers cross-linked by amino acids (peptidoglycans). Many bacteria have rotating flagella to propel themselves through their surroundings and pili so they can adhere to the surfaces of animal cells or each other during conjugation.

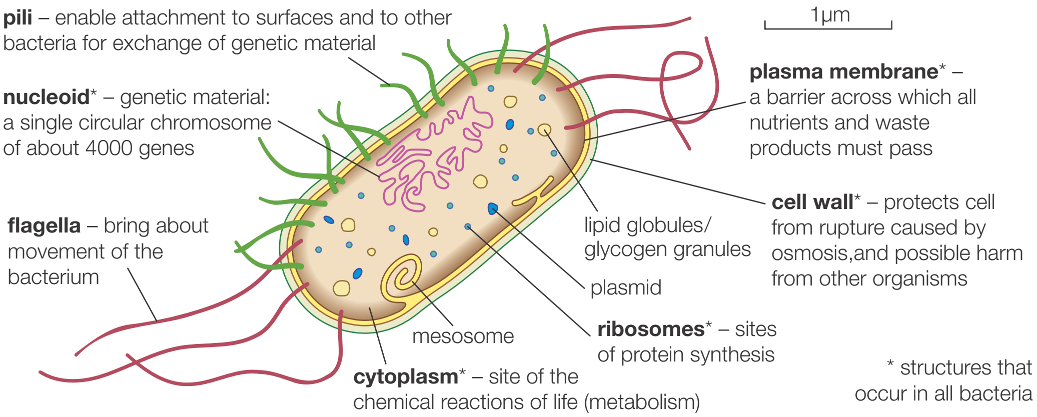

1μm
pili – enable attachment to surfaces and to other bacteria for exchange of genetic material
nucleoid* – genetic material: a single circular chromosome of about 4000 genes
flagella – bring about movement of the bacterium
plasma membrane* – a barrier across which all nutrients and waste products must pass
cell wall* – protects cell from rupture caused by osmosis,and possible harm from other organisms
lipid globules/ glycogen granules
plasmid
mesosome
ribosomes* – sites of protein synthesis
cytoplasm* – site of the chemical reactions of life (metabolism)
* structures that occur in all bacteria

Figure 25.84 The structure of *Escherichia coli* – a 'typical' bacterium

Description of how viruses differ from bacteria

Table 25.4 shows some of the main differences between bacteria and viruses.

Table 25.4 Differences and similarities between bacteria and viruses

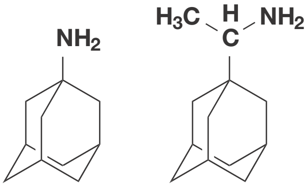
	Bacteria	Viruses
Ribosomes	Present	Absent
Number of cells	Unicellular; one cell (but can form cooperating colonies)	No cells (acellular)
Internal structure	DNA floating freely in cytoplasm; has cell wall and cell membrane	DNA or RNA enclosed inside a coat (capsid) of protein or glycoproteins
Cell wall composition	Peptidoglycan/lipopolysaccharide	No cell wall; protein coat (capsid) present instead
Treatment	Antibiotics	Vaccines prevent the spread of infection and antiviral drugs help to slow replication but cannot stop it completely
Enzymes	Yes	Yes in some, for example reverse transcriptase in retroviruses
Nucleus	No – the nuclear material (nucleoid) is not surrounded by a nuclear membrane	No
Virulence	Yes	Yes
Infection	Localized	Systemic
Reproduction	Binary fission – a form of asexual reproduction	Invades a host cell and takes over the cell causing it to make copies of the viral genome (DNA or RNA); destroys or damages the host cell, releasing new viruses
Size	Larger (1000 nm)	Smaller (20–400 nm)

How antiviral drugs work

Outside the cell, viruses are effectively in a state of suspended animation and so cannot be targeted by drugs. However, inside the cell they are incorporated into the host, hence it is difficult to kill the virus without killing the cell. For this reason it is only relatively recently that drugs capable of controlling viral diseases have been developed.

Antiviral drugs can theoretically act at a number of points in the viral lifecycle. The first step in viral infection is the adsorption of the virus on to the cell surface, often onto specific receptor proteins. Unfortunately, there are no antiviral drugs currently available to prevent the adsorption of viruses on to cell membranes.

The second possible point of attack for antiviral drugs is penetration, usually via engulfment of the virus particle by a small portion of the cell membrane during a process called exocytosis. Amantadine (Symmetrel) (Figure 25.85) appears to prevent the release



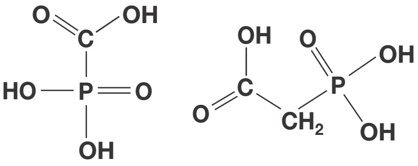
■ **Figure 25.85** Structures of amantadine and rimantadine

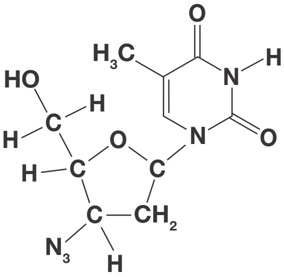
and uncoating of viruses from within the endosome (the virus surrounded by cell membrane from exocytosis) inside the cell and has been used to treat rubella (German measles) and influenza (flu). Rimantadine is a related molecule used to treat the common cold. Both drugs have become less effective due to the appearance of viral strains resistant to both drugs.

The third step of the viral lifecycle involves the shedding of the protective protein capsid, which is an enzyme-controlled process. The anti-malarial drug chloroquine is the only drug found to inhibit the uncoating of a virus (Newcastle disease virus).

The most promising point of attack for antiviral drugs is during the synthesis of viral proteins and viral nucleic acids. Phosphonoformic and phosphonoacetic acids (Figure 25.86) are two antiviral agents that inhibit the action of the viral DNA polymerase, the enzyme responsible for the production of viral DNA.

■ **Figure 25.86** Structures of phosphonoformic (foscarnet) and phosphonoacetic acid (fosfonet)





■ **Figure 25.87** Structure of azidothymidine (AZT)

Another approach is to use what is called the anti-metabolite concept, which usually involves the synthesis of modified chemical bases used in the synthesis of nucleic acids. An example of such a drug is zidovudine (Retrovir or AZT) (Figure 25.87) which is one of the few drugs used with some success in the clinical treatment of AIDS.

Once AZT is incorporated into growing DNA strands it causes chain termination and the enzyme, reverse transcriptase, is no longer able to extend the DNA chain. Two similar agents that block DNA replication in HIV are dideoxyinosine (ddI, DDI or Videx) and dideoxycytidine (ddC or Hivid).

The final stage of the influenza A and B virus replication cycle can also be targeted by antivirals. Two such drugs, oseltamivir (Tamiflu; Figure 25.88) and zanamivir (Relenza), prevent the release of viruses from the cells by inhibiting the active sites of certain viral enzymes known as neuraminidases. These enzymes trigger the process of budding, which allows viruses to depart through the outer cell membrane of the host cell. The inhibition of neuraminidases prevents viruses from departing from the cell and slows their spread around the body. Tamiflu may be given as a preventative measure either during a flu outbreak or following close contact with an infected individual.

■ **Figure 25.88** Oseltamivir (Tamiflu)

Both oseltamivir and zanamivir target the same enzymes, and their molecules have many similarities (Figure 25.89). Both molecules contain a six-membered ring with three chiral carbon atoms (marked with asterisks in Figure 25.89). Both drugs engage in a variety of favourable interactions, including hydrogen bonding and ionic interactions with the active sites of neuraminidases.

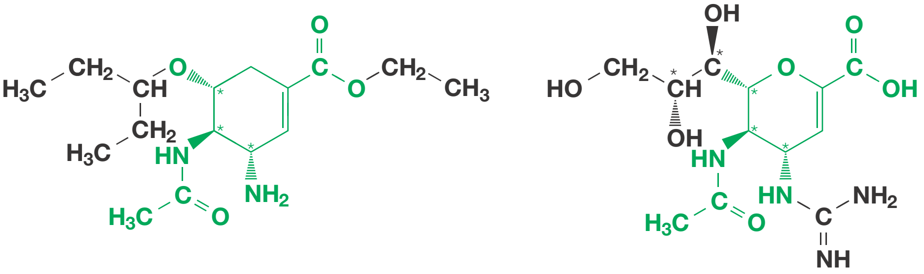


Figure 25.89 The structures of oseltamivir and zanamivir (the chiral carbon atoms are marked with *; common structural features are shown in green)

However, the side-chains in oseltamivir and zanamivir contain different functional groups, which affect the pharmacological properties of these drugs. In particular the presence of an ester group makes oseltamivir inactive in its original form.

In the body the ester group is rapidly metabolized (hydrolysed) into a carboxyl group, producing an active metabolite with enhanced antiviral activity. The zanamivir molecule already has a carboxyl group so it is active in its original form. Zanamivir has the disadvantage of having to be administered as a dry powder, rather like an asthma inhaler. However, it is reported that no flu virus currently has resistance to zanamivir.

The different ways in which antiviral medications work

In general there are four types of actions for antiviral drugs:

- Preventing the genetic material from being injected through the cell membrane: viruses have to interact and bind with specific receptors (proteins or glycoproteins) on the cell membrane of a cell and release its genetic material. To block entry, antiviral molecules can be synthesized that are structurally similar to the virus-associated protein so they bind strongly to the receptor or even bind to the viral capsid (protein coat).
- Inhibiting the replication of the virus: the drug may mimic nucleotides, the monomers of DNA or RNA, so that they are incorporated into the enzyme-controlled synthesis of DNA or RNA, which is then inhibited.
- Inhibiting the action of reverse transcriptase: reverse transcriptase is present in retroviruses, for example HIV. It converts the viral RNA of HIV into DNA, which enters the nucleus of the host cell. The new DNA is integrated (under enzyme control) with the DNA of the host cell and will control the production of new viral RNA and protein by the host cell.
- Preventing new viruses from leaving the cell: new DNA and viral proteins self-assemble into new viruses (viral particles). Following rupture of host cell membranes (or budding), these viruses leave the host cell; this then results in new infections in other cells of the body. Drugs may be developed that prevent the exit of the mature viruses.

8 Find out about enfuvirtide, the first of a new class of drugs approved for the treatment of HIV.

Nature of Science

Research into viral entry methods

Recent research in the biological community has improved our understanding of how viruses invade human and bacterial cells. Using a combination of imaging techniques, researchers have determined that the T4 bacteriophage infects bacterial cells by piercing the cell's outer membrane, digesting the bacterial cell wall and injecting virus DNA into the cell. These findings explain how viruses invade cells and offer a new way to deliver genes and drugs directly into cells.

The researchers combined X-ray crystallography, which provides a three-dimensional picture of the virus's proteins, with cryo-electron microscopy images to determine how T4 proteins rearrange themselves during cell infection. Now that scientists have established the relationships among T4 component proteins they are analysing other processes that occur during a viral infection. These studies provide hope that engineered viruses might be used to seek out and destroy specific cells or deliver a 'payload' that might include new genes or medicinal drugs.

The development of antivirals over recent decades for treating life-threatening or debilitating diseases is the result of the scientific collaboration and exchange of information on a global scale. The availability of protein, DNA and RNA sequences and crystal structures of viral enzymes, cultivation of viruses in the laboratory and extensive medical data accessible via shared databases has greatly expanded biochemical knowledge of the interactions between viruses and host cells at the molecular level.

Greater understanding of the structure, function and lifecycle of viruses leads to the development of new drugs that target viral infections at all stages of the viral lifecycle. The progress in antiviral therapy has already changed the way of treatment of many viral infections and will probably have the same effect on modern medicine as the discovery of antibiotics in the 20th century.

However, the high cost of developing antiviral drugs has limited the number of viral diseases of sufficient market size to a relatively short list. The fact that antiviral drugs are likely to be highly specific for one single infectious agent means that accurate diagnosis of an infection needs to be made before therapy.

Acquired immune deficiency syndrome (AIDS)

Drug treatment for AIDS

Very few antiviral compounds have been found to be effective against HIV. The first clinically available drugs are the nucleoside-based inhibitors of the viral enzyme reverse transcriptase, such as AZT. More recently a cell fusion inhibitor and integrase inhibitor have been made available to clinics.

HIV has been shown to undergo several mutations, resulting in new strains becoming increasingly resistant to drugs, many of which are significantly toxic. Using two or more drugs against different targets is more successful than using a single drug against a single target since it is more difficult for the virus to gain resistance against a drug 'cocktail'.

One approach in the development of drugs and vaccines to combat HIV is centred on a protein called HIV-1 protease, an essential enzyme needed for the production of the virus. Inhibition of this enzyme leads to the production of incomplete and non-infectious HIV particles.

There have been two approaches to drug development in this area. The first involves production of peptide-based inhibitors of HIV protease that mimic the transition state of the enzyme's substrate. The second approach involves synthesizing small non-peptide molecules that are able to bind to the active site of the protease. Many of these molecules have been designed by computer-based molecular modelling using information derived from X-ray diffraction studies of crystallized HIV-1 protease (Figure 25.90).

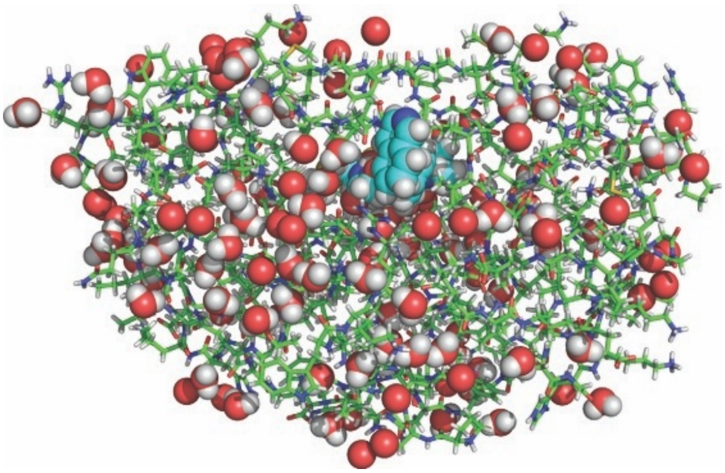


Figure 25.90 Computer model of the structure of HIV-1 protease and associated water molecules (determined by neutron diffraction). The blue represents a bound drug candidate, KNI-272

Societal and cultural factors

Although treatments for AIDS and HIV can slow the course of the disease, there is currently no vaccine or cure. Antiretroviral treatment prolongs the lives of people with HIV, but these drugs are expensive and routine access to antiretroviral medication is not available in all countries. Ninety-five per cent of HIV-positive people live in developing countries. Less than eight per cent of those needing retroviral therapy receive it. Due to the difficulty of treating HIV infection, preventing infection is a key aim in controlling the AIDS epidemic, with health organizations promoting safe-sex programs in attempts to slow the spread of the virus. Condoms, if used correctly without oil-based lubricants, is the single most effective available technology to reduce the sexual transmission of HIV.

AIDS stigma exists around the world in a variety of forms, including discrimination, rejection and avoidance of HIV-infected people, compulsory HIV testing without prior consent or protection of confidentiality, violence against HIV-infected individuals and the quarantine of HIV-infected individuals. The fear of violence prevents many people from seeking HIV testing, returning for their results or securing treatment. In many developed countries, there is an association between AIDS and homosexuality or bisexuality, and this association is correlated with anti-homosexual attitudes (homophobia). There is some evidence that homosexuality is inherited and influenced by several genes on the X chromosome. Epigenetic inheritance may also be involved. Epigenetics involves genetic control by factors other than an individual's DNA sequence.

The AIDS epidemic

Although news about HIV/AIDS is sometimes replaced by more recent disease outbreaks such as H1N1 and ebola, the virus remains a global threat. In 2009, the Joint United Nations Programme on HIV/AIDS and the World Health Organization estimated that 33.4 million people around the world are living with HIV. The organizations' report, *AIDS Epidemic Update*, also estimated 2.7 million new HIV infections and 2 million deaths from AIDS in 2008, indicating that advances in treatment, preventative methods and education are still urgently needed. Treatment for HIV/AIDS has slowly evolved since drugs such as AZT were first made available in 1989. Since the late 1980s, more than 30 possible vaccines have been tested; none has succeeded. About 1000 children die every day from HIV/AIDS. This is three times the number of children who die from cancer.

Many HIV cases and AIDS-related deaths have occurred in Africa, the origin of the HIV virus (Figure 25.91). As a result life expectancy has fallen sharply in some African countries, which has had a significant social and economic impact. Southern and sub-Saharan African countries are those most affected. The recent small increases in life expectancy are due to internationally supported programmes in healthcare and education.

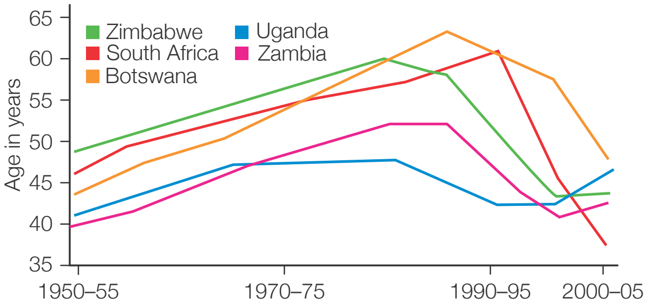


Figure 25.91 AIDS in Africa: life expectancy in selected African countries

25.6 Environmental impact of some medications

– the synthesis, isolation and administration of medications can have an effect on the environment

Introduction

Water is one of the substances required to sustain life and may be the source of many illnesses to humans. Over the years, surface and groundwater sources have become contaminated due to increased industrial and agricultural activity. However, there is also increasing demand for water free of colour, turbidity, taste, odour, nitrate, harmful metal ions, and a wide variety of organic chemicals such as pesticides and chlorinated solvents.

As a population increases, the demand for water grows accordingly, especially when this is accompanied by improved standards of living. In many situations in areas short of water, recycling of treated waste water will be required to avoid affecting industrial development.

Perhaps the most challenging environmental field at present is the treatment and disposal of industrial and hazardous wastes. Because of the great variety of wastes produced from established industries and the introduction of wastes from new processes, a knowledge of chemistry is essential to finding a solution for most of the problems. Some problems may be solved with a knowledge of inorganic chemistry; others may require a knowledge of organic, physical or colloidal chemistry biochemistry or even nuclear chemistry.

Medical waste and the environment

For many years the impact of medical waste substances on the environment has been largely ignored as chemists and biologists focused on well-known contaminants and pollutants generated by the agricultural (e.g. nitrates) and industrial (e.g. heavy metals and hydrocarbons) sectors.

Pharmacologically active compounds (PACs) used in medicine and biochemical studies have traditionally not been treated as potentially harmful and have been released to the environment. PACS detected in sewage, surface water and groundwater include analgesics, antibiotics, anti-epileptics, caffeine, anti-rheumatics, chemotherapeutics (organic and inorganic), steroid hormones and X-ray contrast media, for example barium sulfate. PACs also include illegal drugs, such as cocaine and ecstasy. Granular activated carbon may be used to adsorb and remove hormones and some pharmaceuticals from water, but many of these compounds are not eliminated during the processing of waste water and the production of drinking water.

Environmental xenobiotics are artificial bioactive compounds that are found as pollutants in the natural environment (usually in soil and water). The prefix *xeno-* means foreign. Together with industrial products, environmental xenobiotics include PACs as well as hospital disinfectants such as phenol, bisphenol (used to make certain plastics and epoxy resins), polychlorinated biphenyls (PCBs), azo dyes, aromatic hydrocarbons and steroids and steroid-based hormones. Concentrations of these chemicals are often monitored on a long-term basis. A variety of analytical techniques are used to detect and measure the concentrations of PACs and environmental xenobiotics, including liquid chromatography/mass spectrometry (LC-MS) (see Chapter 21).

However, prolonged exposure to PACs may cause significant changes in the metabolism and behaviour of various organisms, including humans. Of particular concern is the uncontrolled release of antibiotics to the environment which may lead to the selection of antibiotic-resistant bacteria. Other synthetic oestrogenic compounds such as PCBs, bisphenol and phthalates can act as endocrine disruptors, increasing the risk of breast cancer and reproductive disorders in humans, and causing reduced sperm counts.

Another type of environmental pollution is caused by radioactive materials used in medical treatment (radiotherapy) and radio-diagnostics. There are international guidelines for radioactive waste disposal from departments of nuclear medicine.

Radioactive waste from nuclear medicine procedures can be dealt with either by simply storing the wastes safely until radioactive decay reduces the activity to a relatively low level or by disposal of low-activity waste (with monitoring) into the sewerage system. Controlled disposal is defined as disposal with permission from the regulatory authority and appropriate monitoring. The waste must be diluted at the discharge point (from the hospital into the sewerage system) to a relatively low concentration to protect the local community.

Certain radioisotopes can undergo bioaccumulation and biomagnification, increasing the risk of radiation exposure for predators at the peak of a food chain. This has been demonstrated for caesium-137 in river ecosystems where there is a fourfold increase in each trophic (feeding) level.

The synthesis, production, storage and distribution of pharmaceutical drugs also contributes to environmental pollution though the release of greenhouse gases (such as carbon dioxide), ozone-depleting substances (such as chlorinated solvents) and toxic materials, including left-over solvent residues, excess acid or base, and biologically active by-products of organic synthesis. Inhaled anaesthetics from operating theatres in hospitals are potent greenhouse gases. These harmful effects can be greatly reduced by the introduction of sustainable industrial processes or green chemistry.

Antibiotic resistance

The widespread use of penicillins and other classes of beta-lactam antibiotics, such as cephalosporins, in the last half of the 20th century led to the selection and increase in the frequency of antibiotic resistance (Figure 25.92) in many strains of pathogenic (harmful) bacteria.

The large numbers of bacterial cells, combined with the short generation times, facilitate the development of mutants. In a typical bacterial population of 10^{11} bacterial cells (for example, in an infected patient) there can easily be 1000 mutants. If a mutant confers a

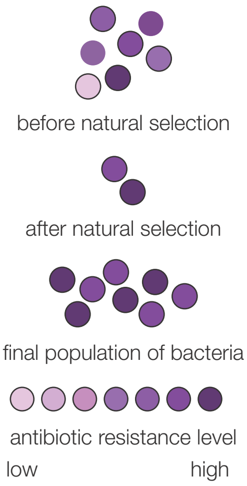


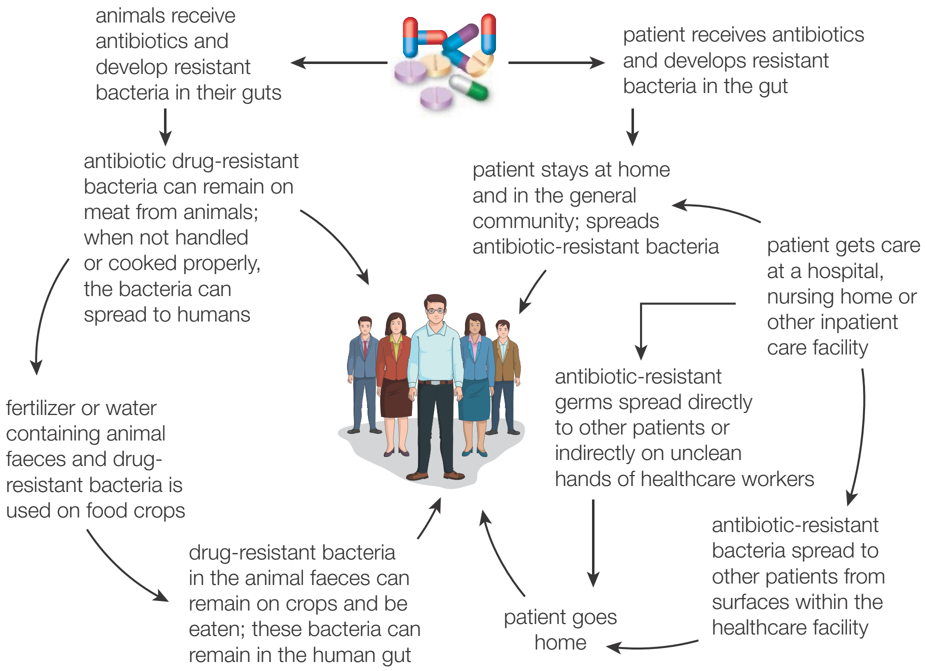
Figure 25.92 Mechanism of antibiotic resistance

selective advantage upon the bacterium (e.g. the ability to survive in the presence of an antibiotic) then that resistant bacterium will be selected and continue to grow while the neighbouring bacteria will die. This can happen in a matter of days in patients being treated with antibiotics.

As a result the efficiency of traditional antibiotics against common diseases has significantly decreased, so medical chemists need to synthesize and test new synthetic and semi-synthetic antibiotics in order to treat severe bacterial infections. However, it becomes progressively more difficult as bacteria constantly evolve and become resistant to increasing numbers of antibiotics and in some strains become multi-drug resistant.

Antibiotic resistance in bacteria is caused by several factors (Figure 25.93), including the over-prescription of antibiotics for mild bacterial infections, non-compliance of patients in finishing a course of treatment (that is, not taking the prescribed number of tablets at the required times), the use of antibiotics in farm animals to prevent infection, and the release of antibiotic waste by hospitals and the pharmaceutical industry into the water supply.

Figure 25.93 The spread of antibiotic resistance


animals receive antibiotics and develop resistant bacteria in their guts
patient receives antibiotics and develops resistant bacteria in the gut
antibiotic drug-resistant bacteria can remain on meat from animals; when not handled or cooked properly, the bacteria can spread to humans
patient stays at home and in the general community; spreads antibiotic-resistant bacteria
patient gets care at a hospital, nursing home or other inpatient care facility
antibiotic-resistant germs spread directly to other patients or indirectly on unclean hands of healthcare workers
fertilizer or water containing animal faeces and drug-resistant bacteria is used on food crops
antibiotic-resistant bacteria spread to other patients from surfaces within the healthcare facility
drug-resistant bacteria in the animal faeces can remain on crops and be eaten; these bacteria can remain in the human gut
patient goes home

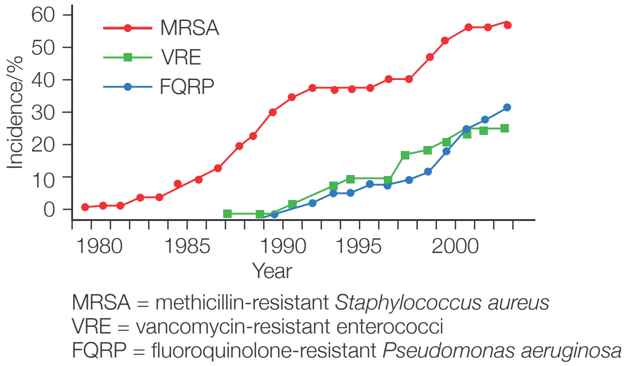


Figure 25.94 Rise of antibiotic resistance in various common bacterial infections

In all cases, exposure to low levels of antibiotics allows some mutated bacteria to survive and reproduce, eventually developing the ability to tolerate higher and higher concentrations of the antibiotic. Bacteria may be resistant because either they have no mechanism to transport the drug into the cell or they do not contain or rely on the antibiotic's target process or protein.

Such bacteria pass (via a process known as conjugation) their genetic resistance (on small strands of DNA called plasmids) to new generations of bacteria, gradually replacing non-resistant strains. This simple process of evolution by natural section can take place both in individual patients and in the environment. In the latter case, exposure to antibiotics increases the antibiotic resistance (Figure 25.94) of the whole bacterial population.

Over the past 30 years the use of antibiotics in agriculture (farming) has nearly doubled and now contributes to over half of the global consumption of meat, milk and eggs. Most of these antibiotics are given to healthy animals to prevent infectious diseases and promote growth of animals, such as cows, sheep, goats and chickens. Although this practice allows increased output and reduced prices in agricultural production, it is also the major source of antibiotic waste in the environment. Since antibiotics are never completely metabolized in animal organisms, a significant percentage of each antibiotic is excreted in unchanged form in the urine and released into the groundwater or absorbed by other organisms. Some of these antibiotics are eventually consumed by humans in meat, dairy products and water, further accelerating the development of resistant bacteria.

Restrictions on the use of antibiotics

Since 2006 the use of antibiotics as growth promoters in animal feed for farm animals has been banned in the European Union (EU); they have also been banned in some other countries, for example in South Korea in 2011. However, these measures had no immediate effect on bacterial resistance in humans while the rates of death and bacterial disease in animals increased significantly.

Excessive antibiotic intake has been shown to weaken the development of immune systems of children and make them more susceptible to infections. Antibiotics may also be promoting growth and obesity in the human population. Some people develop severe allergic reactions to antibiotics, including penicillins.

Many scientists believe antibiotic therapy should be restricted to the most severe cases of infections with multiplying bacteria while non-medical use of antibacterial drugs should be banned completely. At the same time, the amount of antibiotic waste from hospitals and the pharmaceutical industry must be reduced to a minimum and thoroughly processed before being released into the environment.

In addition, new antibiotic drugs must be produced and used under strict control to prevent the development of antibiotic resistance. To be effective, these measures need to be taken by all countries and coordinated at the international level. The discovery of new antibiotics has slowed down significantly. Developing new antibiotics has become too expensive and unattractive for pharmaceutical companies and no new classes of antibiotics have appeared since 2003 (lipopeptides).

Nuclear waste

Many medical procedures involve the use of radionuclides – unstable isotopes of certain elements that undergo spontaneous radioactive decay (see section 25.8) and release beta or gamma radiation. Some of these isotopes are administered to patients as water-soluble salts or radiopharmaceutical drugs where one or more atoms are radioactive (see section 25.8), whereas other radionuclides, such as cobalt-60, are used in medical equipment as sources of ionizing radiation. During medical procedures radionuclides and ionizing radiation come into contact with various materials that also become mildly radioactive. These materials together with left-over or spent radionuclides produce nuclear waste, which must be disposed of according to specific procedures.

Most radionuclides used in departments of nuclear medicine in hospitals and medical research centres have very low activity and short half-life times. The waste containing small amount of radionuclides is known as low-level waste and typically consists of contaminated syringes, clothes, swabs, paper, filters and protective clothing, for example gloves. Such waste has limited environmental impact, does not require shielding and is usually suitable for shallow land burial or incineration. Some types of low-level waste, such as concentrated solutions of medical radionuclides, must be stored for several days or weeks in lead-shielded containers until most of the radioactive nuclides have decayed and the radiation level has dropped to what is deemed a ‘safe’ level.

Medical equipment for radiotherapy may contain large quantities of radioactive isotopes such as cobalt-60 and caesium-137. These radionuclides remain active for many years and produce very high levels of ionizing radiation. This intermediate-level waste requires shielding during handling, processing and storage, but does not generate significant heat. Shielding can be barriers of lead, steel or concrete.

High-level nuclear waste is produced in nuclear reactors and contains a complex mixture of nuclear fission products (see section 24.3) with unused nuclear fuel (usually uranium). Many radionuclides in high-level nuclear waste have very long half-lives ranging from hundreds to billions of years. Due to nuclear reactions, high-level nuclear waste releases heat (as well as neutrons and gamma radiation) and must be constantly cooled and shielded with water for several years. When the radioactivity level decreases, high-level nuclear waste can be reprocessed and partly recycled. The remaining waste is fused with borosilicate glass (vitrification) or immobilized in certain minerals, producing water-resistant and chemically stable solid materials. These materials are encased in steel cylinders, covered with concrete, and buried deep underground in geologically stable locations. The containers must be continually cooled to avoid thermal stress and to prevent possible changes in the glass structure. Depending on how much heat is generated, stored containers can be cooled by natural or forced air convection

The treatment, transportation (by train, lorry and ship) and disposal of nuclear waste presents serious risks due to possible release of radionuclides to the environment. In high doses ionizing radiation is harmful to all living organisms, causing extensive cellular damage and genetic damage to DNA. Low doses of radiation increase the number of mutations and the risk of developing cancer, birth defects and reproductive disorders. In addition, ionizing radiation weakens the immune system by triggering apoptosis (programmed cell death) in lymphocytes (white blood cells that digest bacteria) and rapidly dividing bone marrow cells.

The effects of ionizing radiation can be cumulative. For example, radioactive materials discarded together with antibiotic waste can increase the mutation rate in bacteria and accelerate the appearance of drug-resistant mutant strains.

Waste products from the pharmaceutical industry

Many pharmaceutical drugs are produced on an industrial scale using a wide range of technological processes and synthetic approaches often using catalysts. Some of these processes traditionally involve the use of toxic chemicals that have to be recycled or disposed of after the synthesis is complete. Organic solvents used in the pharmaceutical industry constitute a significant proportion of chemical waste.

Most organic solvents, especially chlorinated solvents, are toxic and systemic poisons to living organisms, primarily affecting the respiratory system, liver, nervous system, kidneys and reproductive organs. Most of these compounds have an anaesthetic or narcotic effect, causing people to feel intoxicated if overexposed. Some solvents such as benzene (C_6H_6), methylbezne ($C_6H_5CH_3$) and chloroform ($CHCI_3$) (also used as an anaesthetic) increase the risk of cancer in humans and other animals. In addition, many solvents are highly flammable while the molecules of their vapours contribute to the greenhouse effect and hence global warming (see Section 24.5).

Chlorinated solvents such as carbon tetrachloride (CCl_4, also known as tetrachloromethane; used in fire extinguishers), chloroform ($CHCl_3$), dichloromethane (CH_2Cl_2) and trichloromethane ($CHCl_3$) are environmental hazards. Due to the relatively low bond enthalpies of the C–Cl bonds, these compounds act as ozone-depleting agents (due to formation of chlorine radical in the presence of ultraviolet radiation) and contribute to the formation of photochemical smog in large industrial cities. Some chlorinated solvents and chlorine-based pesticides have limited biodegradability and may accumulate in the groundwater causing long-term damage to local ecosystems.

The disposal of chlorinated solvents is an expensive and complex process. Chlorine-containing compounds cannot be incinerated together with common organic waste because their incomplete combustion could produce highly toxic phosgene (carbonyl chloride, $COCl_2$) and solvents must be oxidized separately at very high temperatures or recycled by distillation.

Research and rare diseases

Do pharmaceutical companies have a responsibility to do research on rare diseases that will not provide them with significant financial profit?

A rare disease is a disease that affects only a small percentage of a population yet they may affect between 10 and 20 million people in the USA alone. Pharmaceutical drugs are often 'orphaned' or are not produced and sold on the market, even when a compound is thought to be useful for the treatment of a rare disease. Due to the relatively small number of patients and the lack of money they would provide in sales revenue, pharmaceutical companies do not have any incentive to develop drugs and treatments for rare diseases.

Therefore the pharmaceutical company may have the ability to supply orphan drugs where, that is, there is a demand, but this demand is not significant enough to manufacture the needed product. This is also partially due to the lack of sponsorship, and no one to 'adopt' the orphan drug, to conduct the necessary testing (in the USA) to obtain the necessary approval from the Food and Drug Administration (FDA).

Pharmaceutical companies are faced by restrictions as to where time and money can actually be invested. They have limited resources available to them for research and development (R&D), so research must be prioritized. The drugs or treatments that will bring in the largest amount of money will be of most importance, while developing treatments for rare diseases will fall very low on the priority list.

Bringing a drug into the market can be very costly and time consuming. The estimated cost of bringing a single drug through screening processes and FDA procedures to obtain approval is some US$350–500 million, and it can take up to ten years to bring a single drug into the market.

Beyond the costs to itself, listed drug companies have an additional responsibility to their shareholders. In order to fulfil a shareholder's investment, a company must continue to grow and bring in money. An additional issue faced by the pharmaceutical industry is that many of the compounds found to treat rare diseases cannot be protected by a patent. A patent is the exclusive right to make, use or sell a product. Therefore, competitors are not able to sell a product that is the same as a product already protected by a patent.

Without this type of protection a company would be exposed to immediate competition with generic drugs. Generic drugs are less expensive than brand-name drugs, even though they are chemically identical and meet the FDA standards for safety, purity and effectiveness. The introduction of a generic drug into the market drives market prices down due to competition and therefore reduces the amount of revenue a non-generic could have potentially brought in. The high costs of bringing a drug into the market, added to the concerns about the amount of sales revenue a drug can actually provide, the financial responsibility to shareholders and the lack of patent protection all outweigh a pharmaceutical company's wishes or ethical obligation to provide drugs to treat people with rare diseases.

The US government realized the need to create incentives for pharmaceutical companies to develop drugs for rare diseases. It is the government's responsibility to ensure the 'base-line well-being of its citizens' and the government is already active in health issues and formulating health-related policies. The US government used to be involved in developing new drugs and treatment and new drugs were often developed in federally operated laboratories, non-profit medical research centres as well as the pharmaceutical industry. However, by the 1960s, most new drugs were manufactured for profit and the focus of federal laboratories and research universities concentrated on basic research and clinical studies.

Green chemistry

A quarter of the world's human population have to survive on less than US$1 a day. Millions of people have no healthcare. The world's population is expected to increase by about another 3 billion over the next 50 years. Even in developed nations, poverty, education and healthcare could be improved. To help tackle this situation, the world's economy needs to grow; in particular, the economies of developing nations need to expand. However, economic growth is often linked to global warming and environmental pollution problems, such as acid rain and water pollution.

The challenge is to develop in a way that meets the needs of the present generation without compromising the ability of future generations to meet their own needs – in other words, without causing environmental damage and wasting limited resources. This type of development is called 'sustainable development' and it will be increasingly critical as the population of the world continues to rise.

One of the ways in which the chemical industry is working towards sustainable development is by using 'green chemistry'. One of the basic ideas of green chemistry is to prevent pollution and the production of hazardous materials instead of producing materials and then cleaning them up. This means that green chemistry is safe, conserves raw materials and energy, and is more cost effective than conventional methods.

There are three main ways to make chemical processes 'greener'. Companies can redesign production methods to use different, less hazardous starting materials; use milder reaction conditions, better catalysts and less hazardous solvents; and use production methods with fewer steps and higher atom economy.

The concept of percentage yield is useful, but from a green chemistry and sustainable development perspective it is not the only issue. This is because yield is calculated by considering only one reactant and one product. One of the key principles of green chemistry is that processes should be designed so that the maximum amount of all the raw materials ends up in the product and a minimum amount of waste is produced. A reaction can have a high percentage yield but also generate large amount of waste product. This kind of reaction has a low atom economy. Both the yield and the atom economy have to be taken into account when designing a green chemical process.

Green chemistry is essentially a way of thinking rather than a new branch of chemistry. It is about using a set of principles that seek to reduce the environmental impact of chemical processes and products. It involves pulling together tools, techniques and technologies that can help chemists and chemical engineers in research, development and production to develop more eco-friendly and efficient products and processes, which may also have significant financial benefits. Green chemistry aims to improve the way that chemicals are both produced and used in chemical processes in order to reduce any impact on humans and the environment. It is not just about industrial production.

Green solvents

The main idea behind the use of green solvents is to minimize the impact they have on the environment. Four factors are involved in the movement towards the use of green solvents:

- substitution of hazardous solvents with those that have better environmental health and safety properties, for example glycerol;
- the use of solvents that are produced by renewable resources;
- substituting organic solvents with those that are environmentally harmless;
- the use of ionic liquids that show low vapour pressure (resulting in less emissions).

Traditionally solvents are disposed of in two ways – either recycled through a distillation process or treated in a waste incineration plant.

Supercritical fluids

If a volatile liquid, such as water, is heated in an open container, it gradually evaporates as it changes to a vapour (a gas near its boiling point). If the heating is carried out in a sealed container, the molecules at the surface of the liquid evaporate, increasing the pressure and density of the vapour. This is known as the critical point.

Eventually, as the temperature increases, a point is reached where the vapour and liquid have the same density, and the surface between the two disappears. Fluids above their critical temperature and pressure are known as supercritical fluids.

Supercritical fluids combine some of the properties of both liquids and gases. For example, they are effective solvents for a range of substances, but they have relatively low viscosities and high diffusion rates (a property normally associated with polar liquids, such as water). In recent years, supercritical fluids have found a range of uses in chemistry and are now used in industry.

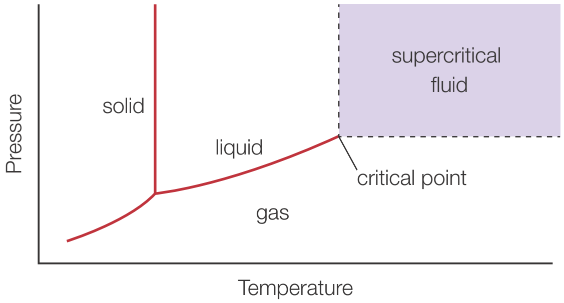
Supercritical water has very different properties from normal liquid water. It is relatively non-polar and dissolves hydrocarbons and other non-polar substance. It is a strong oxidizing agent and can be used to oxidize pollutants. However, it is corrosive, which means reactions must be carried out in specialized and expensive equipment.

A more commonly used supercritical fluid is liquid carbon dioxide. Liquid carbon dioxide does not exist at atmospheric pressure but the gas can be liquefied at pressures above 5 atmospheres. One of the first uses of supercritical carbon dioxide was in the extraction of molecules that give flavour to foods. Decaffeinated coffee was originally made by extracting the caffeine with the organic solvent dichloromethane, CH_2Cl_2. However, this also removed some of the molecules that give coffee its distinctive flavour. In addition it is a toxic organic solvent. It has been replaced by the 'green' solvent, supercritical carbon dioxide.

The use of supercritical carbon dioxide allows the removal of almost 100 per cent of the caffeine, while leaving most of the flavour. The supercritical carbon dioxide is circulated through the beans, and then into a second vessel at lower pressure which contains water. The caffeine dissolves in the water, and the carbon dioxide is repressurized and recycled. There are no problems with the contamination of the coffee beans with the solvent, and there is no harmful organic liquid to dispose of. The extracted caffeine is sold to manufacturers of carbonated (soft) drinks.

Figure 25.95 shows the phase diagram for carbon dioxide. The critical point marks the conditions above which a liquid does not exist. The system indicated by the shaded condition, bounded by the dashed lines, is a supercritical fluid. Carbon dioxide is an especially useful supercritical fluid since it has a conveniently low critical point (31 °C and 73 atmospheres) and is cheap, easily removed after the process and completely non-toxic.

Figure 25.95 Phase diagram for carbon dioxide



Green synthesis

Ibuprofen is a popular over-the-counter drug that has analgesic and anti-inflammatory properties. It was originally developed in the 1960s by the Boots Company of England. The Boots synthesis of ibuprofen involved a six-step process, known as the Brown process, that generated a large quantity of unwanted waste (Figure 25.96) due to reactants of the process not being incorporated into the product because of poor atom utilization.

The synthesis of ibuprofen begins with the compound isobutyl benzene (2-methylpropylbenzene).

Step 1 Friedel–Crafts acylation that uses the Lewis acid aluminium chloride and generates aluminium trichloride hydrate as a waste by-product

Step 2 Darzens reaction with ethylchloroethanoate that results in an epoxy ester compound

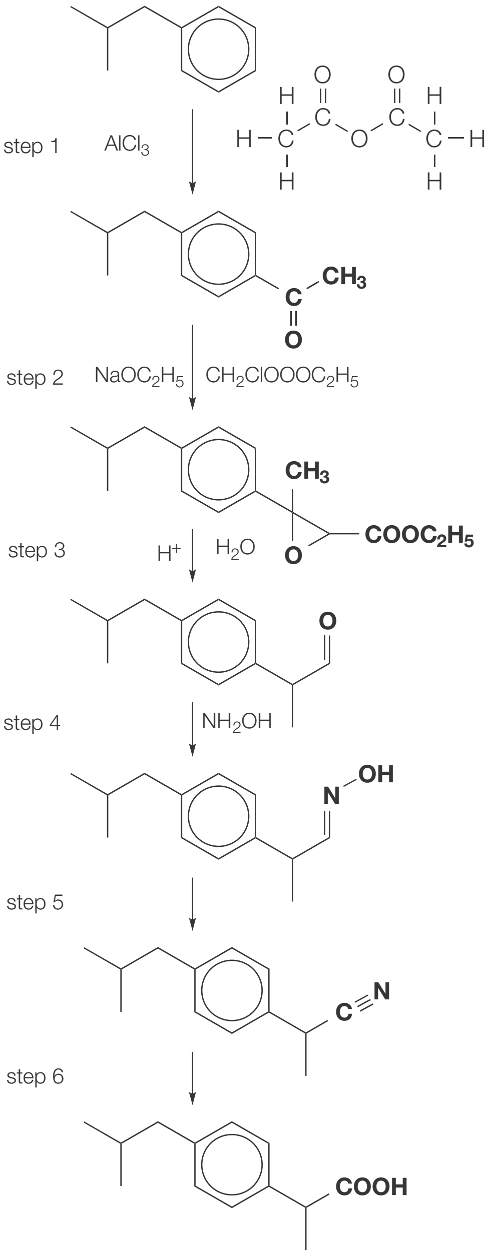
Step 3 decarboxylation and hydrolyzation forming an aldehyde

Step 4 reaction with hydroxylamine giving an oxime

Step 5 the oxime is converted to a nitrile

Step 6 hydrolyzation of the nitrile resulting in the final product 2,4-isobutylphenyl-propanoic acid.

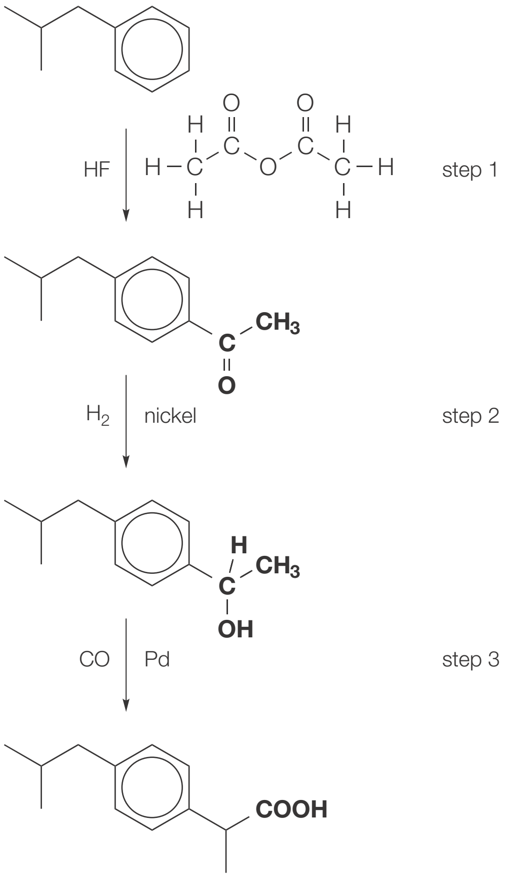
This process has a 40 per cent atom economy which translates into 60 per cent waste products. The solution to this problem was developed in 1991 and involves a green synthesis of ibuprofen that only required three steps (Figure 25.97). This process also incorporated most of the reactants into the final product, reducing or eliminating most of the waste by-products. The switch from the Brown process to the BHC process has resulted in a more efficient synthesis of ibuprofen.



■ **Figure 25.96** The Brown process for synthesis of ibuprofen

The green synthesis of ibuprofen also begins with the compound isobutyl benzene:

Step 1 Friedel-Crafts acylation using hydrogen fluoride as the catalyst that can be recovered and reused
Step 2 involves hydrogenation with nickel, which is recovered and reused, to produce an alcohol
Step 3 the alcohol undergoes carbonylation with the catalyst palladium, also recovered and reused, to produce ibuprofen or 2,4-isobutylphenylpropanoic acid.



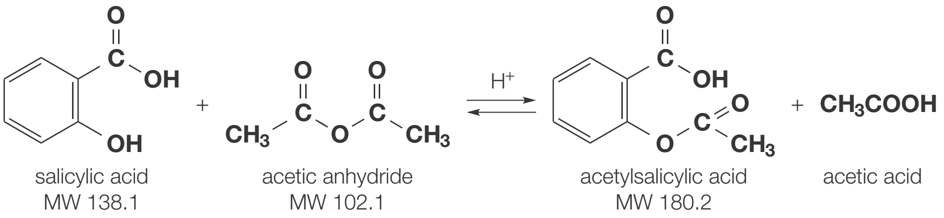
■ **Figure 25.97** This process has a 99% atom economy, which includes the recovered ethanoic acid that was generated in step 1. This means that only 1% is waste by-products

Atom economy

One of the most important concepts of chemistry is atom economy (Figure 25.98), which expresses the efficiency of a synthesis reaction as the ratio between the molar mass of the important target molecule and the sum of all the molar masses of the reactants (reagents) in the reaction. Atom economy also must consider substances such as solvents, separation agents and drying agents that are used in the process but are not directly part of the chemical reaction.

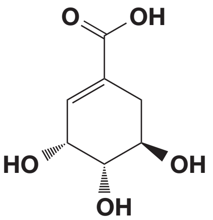
Using a green chemistry approach, chemists attempt to either reduce the amount of or eliminate completely as many of these substances as possible. Those that cannot be eliminated are reused or recycled when possible. Atom economy describes the conversion efficiency of a chemical process in terms of all atoms involved (desired products produced). In an ideal chemical process, the amount of starting materials or reactants equals the amount of all products generated and no atom is wasted.

■ **Figure 25.98** Atom economy for the synthesis of aspirin



$$\% \text{ atom economy} = \frac{180.2}{138.1 + 102.1} \times 100\% = 75.0\%$$

Another important aspect of green chemistry is the use of biotechnology and bioengineering in organic synthesis. Enzyme-catalysed biochemical reactions are highly selective, efficient and proceed in aqueous solution under relatively mild conditions of temperature and pH. Many pharmaceutical drugs or synthetic intermediates (primary precursors) can be produced from



■ **Figure 25.99** Structure of shikimic acid

renewable materials by genetically modified organisms. One such intermediate, shikimic acid (Figure 25.99), is a precursor to the antiviral drug for treating flu known as oseltamivir (Tamiflu).

For many years shikimic acid was extracted from the Chinese star anise, a spice plant, using solvent extraction with hot water and column chromatography. Roughly 80–90 per cent of the world's star anise is grown in south-western China, primarily in Guangxi and Yunnan provinces. An estimated 66 per cent of China's star anise harvest is used to make Tamiflu.

Outbreaks of 'bird flu' have increased the demand for Tamiflu and led to a worldwide shortage of this drug due to a limited supply of star anise. Modern biosynthetic technologies allow shikimic acid to be produced on an industrial scale by genetically modified *E. coli* bacteria, which effectively prevents any shortages of Tamiflu in the future.

The industrial use of natural products leads to various ecological and social issues, such as the extinction of plant species and the increasing of food prices. At the same time some non-hazardous substances branded as 'green' or 'environmentally friendly' still require toxic chemicals or large amounts of energy for their production. Therefore the criteria used to assess the 'greenness' of a substance or process must include all direct and indirect environmental implications.

Standards and practices in the pharmaceutical industry vary greatly around the world. Widespread adoption of green technological processes in developed countries has significantly reduced the emissions of many hazardous chemicals, such as chlorinated solvents and greenhouse gases. Although green technologies often involve expensive equipment and recycling facilities, they reduce the costs of environmental remediation, waste management and energy consumption, making green chemistry a commercially attractive and sustainable alternative to traditional organic synthesis.

An important issue in the chemical industry, including the pharmaceutical industry, is the E factor. This is the actual amount of waste formed, including solvent losses, acids and bases used in work-up and in principle waste from energy production.

It can be derived from the amount of raw materials purchased/amount of product sold; that is, from the mass balance E = [raw materials − product]/product. In the pharmaceutical industry product tonnage varies from 10 to 103 tonnes and the E factor from 25 to >100.

There are a number of origins for this waste: stoichiometric Brönsted acids and bases (e.g. aromatic nitration with HNO_3/H_2SO_4), stoichiometric Lewis acids (e.g. Friedel–Crafts acylation, for example $AlCl_3$), stoichiometric oxidizing and reducing agents (for example $KMnO_4$ and $LiAlH_4$) and solvent losses (via air emissions and aqueous effluent).

Green chemistry came from law enacted in the USA, and because laws result from political compromise and agreement among interested parties in order to get broad support, the ethical principles that underlie green chemistry reflect ethical beliefs regarding the environment that a large portion of the public share.

25.7 Taxol – a chiral auxiliary case study (AHL) – *chiral auxiliaries allow the production of individual enantiomers of chiral molecules*

■ Introduction

Plants have always been a rich source of lead compounds for the development of drugs, for example morphine, cocaine, digitalis (used to treat heart failure), quinine (used to treat malaria) and nicotine. Many of these lead compounds are drugs in themselves, for example morphine extracted from opium, and others have been the basis for semi-synthetic drugs, for example heroin.

Plants are and will continue to remain a promising source of new drugs. A clinically useful drug that has been isolated from plants is the anti-cancer drug Taxol (paclitaxel). Plants contain a 'library' of rich, complex and highly varied chemical structures that are difficult to synthesize in the laboratory. Many plant secondary metabolites remain undiscovered.

The discovery, isolation and development of Taxol (paclitaxel) is a good example of the time and resources required to extract a new drug or drug lead from natural material. It shows many of the difficulties and challenges and the multi-disciplinary nature of this process.

Taxol

Taxol (paclitaxel) is a potent anti-cancer natural product that is active against leukemia and tumours in the breast, ovary, brain and lungs. Taxol was isolated from the bark of the Pacific yew tree in 1971. In the initial extraction 12 kg of air-dried stem and bark were extracted with 95 per cent ethanol and the solution was concentrated.

The isolation of active compounds was followed by biological testing by measuring the inhibition of solid tumour cells. Activity was recorded as a *T*/C value, where *T*/C is defined as:

$$\frac{T}{\mathrm{C}} = \frac{\text{mean tumour mass of treated animal}}{\text{mean tumour mass of control animals}} \times 100$$

The smaller the *T*/C value relative to the dose used, the more active the sample. The concentrate was partitioned between water and a solvent containing chloroform (trichloromethane) and methanol (4 : 1). The organic layer yielded 146 g of solids active against solid tumours. These solids were further fractionated using two different solvent systems and the most active tubes were passed on to the next stage. Finally, trituration with benzene gave 0.5 g of a single pure compound, namely Taxol, an overall yield of 0.004 per cent. Taxol is a white crystalline powder. It is highly lipophilic and hence highly insoluble in water.

Trituration is used to purify crude chemical compounds containing solid impurities. A solvent is chosen in which the desired product is insoluble and the undesired by-products are very soluble (or vice versa). The crude material is washed with the solvent and filtered away, leaving the purified product in solid form and any impurities in solution.

Figure 25.100 shows the principle of multi-step liquid–liquid extraction of Taxol or other active compounds from biological material. The aqueous and solvent layers are both tested with tumour cells. After each step all the active fractions are combined, inactive fractions discarded and the process repeated using different solvents or combinations of solvents. Each extraction step produces an extract with progressively increasing anti-cancer activity.

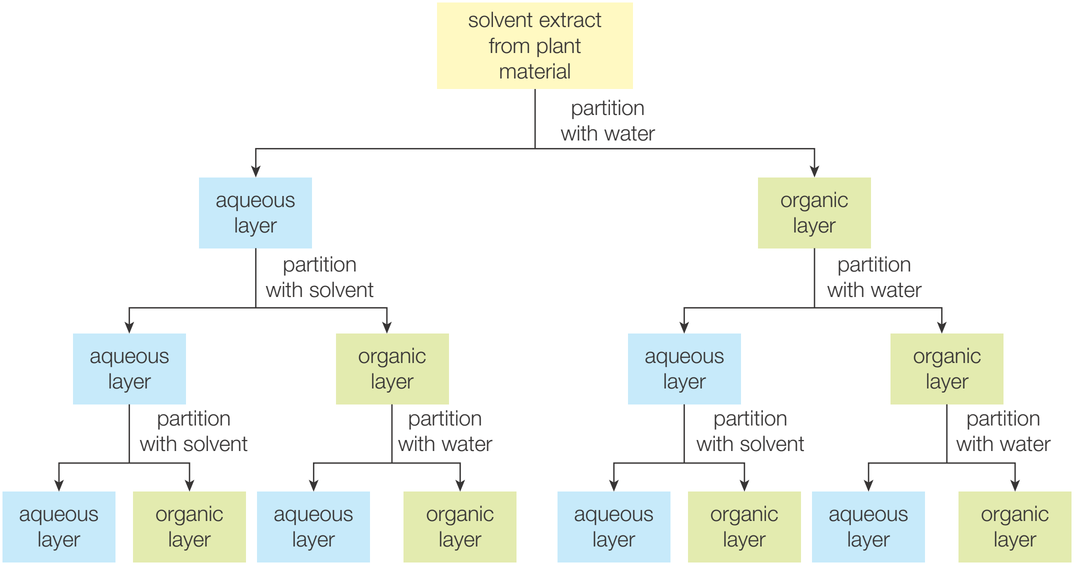


Figure 25.100 Multi-step liquid–liquid extraction

The structure of Taxol was determined by X-ray crystallography and chemical degradation and analysis of the fragments in 1971. Subsequent research showed that it prevented cell replication by binding to the polymerized form of tubulin, a protein present in structures called microtubules involved in chromosome movement during cell division.

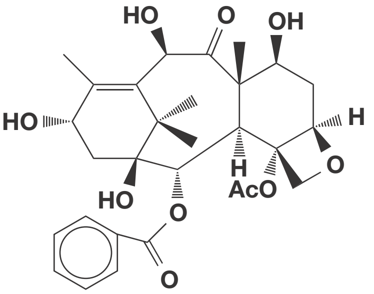
Clinical trials began in 1983, but Taxol was in short supply since it is present in only minute amounts in nature. The Pacific yew tree (Figure 25.101) is an environmentally protected species and one of the slowest growing trees in the world. Removal of the bark kills the tree and six 100-year-old trees are needed to treat just one patient.

Figure 25.101 Pacific yew (*Taxus brevifolia*)

Research efforts were then directed towards developing a total synthesis of Taxol from commercially available precursors. This was achieved after over ten years' efforts by many research groups worldwide who developed different syntheses.

In addition, a more efficient and sustainable semi-synthesis of Taxol was developed from a naturally occurring related compound called 10-deactylbaccatin III (Figure 25.102), a precursor that is easily coupled with a side-chain to give Taxol. This compound was discovered in the needles or leaves of the related European yew tree, *Taxus baccata*. This source is renewable and sufficient quantities of Taxol were prepared for clinical trials. The latest approach to Taxol production involves the cultivating of Pacific yew cells in tanks. The European yew tree has been investigated for the presence of other anti-cancer agents and screening has revealed the presence of a compound related to Taxol, which promises to be an even more potent anti-cancer agent than Taxol. Biotechnologists have obtained Taxol from cultivated callus cell of *Taxus brevifolia*. This area of research hopes to cultivate the Taxol-producing cells and speed up their production. On a large scale Taxol could be produced in tanks to supply the huge demand for Taxol and a related compound called Taxotere.

Taxol, whose structure is shown in Figure 25.103, is a relatively complex molecule that can be divided into two main parts: the side-chain and the Taxol skeleton. Many structural features of the side-chain and the Taxol skeleton are required for maintaining its anti-tumour activity.



10-Deacetylbaccatin III

Figure 25.102 Structure of 10-deacetylbaccatin III (Ac = ethanoyl (acetyl))

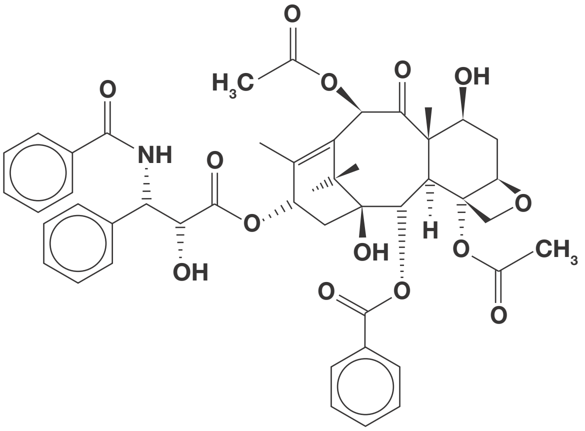


Figure 25.103 Structure of Taxol (paclitaxel)

The total synthesis of Taxol is highly challenging to synthetic organic chemists because of the presence of the following structural features: a double bond in a six-membered ring that is a bridgehead within the molecule, the central eight-membered ring and the unstable four-membered oxetane ring. In addition, the large number of oxygen atoms (in different oxidation states) coupled with nine chiral centres (implying 2^9 stereoisomers) posed further problems. There are 11 chiral (stereogenic) centres if the 2 in the side-chain are included.

Taxotere(docetaxel) has been synthesized on an industrial scale and is currently being administered in clinical trials. The structure of Taxotere (Figure 25.104) differs from Taxol in the presence of a tertiary butoxycarbonyl group instead of a benzoyl group on the Taxol side-chain and a hydroxy group instead of the acetyl group on the Taxol skeleton. The advantages of Taxotere over Taxol are its easier administration due to its greater solubility in water and its more potent anti-cancer properties against a variety of tumours, especially breast cancer.

The steps that make up the synthesis of a complicated natural product like Taxol are planned by a mental process called retrosynthetic analysis (see Chapter 20). In this approach, the 'target' molecule, in this example Taxol, is repeatedly disconnected or fragmented into progressively simpler 'building blocks' which can then be further simplified to readily available starting materials. Retrosynthesis is a 'paper exercise' and is not actually carried out as an experimental procedure.

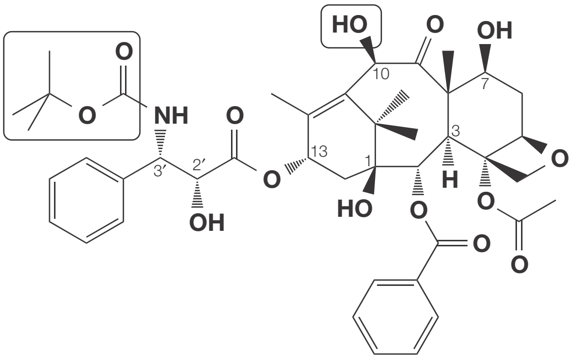
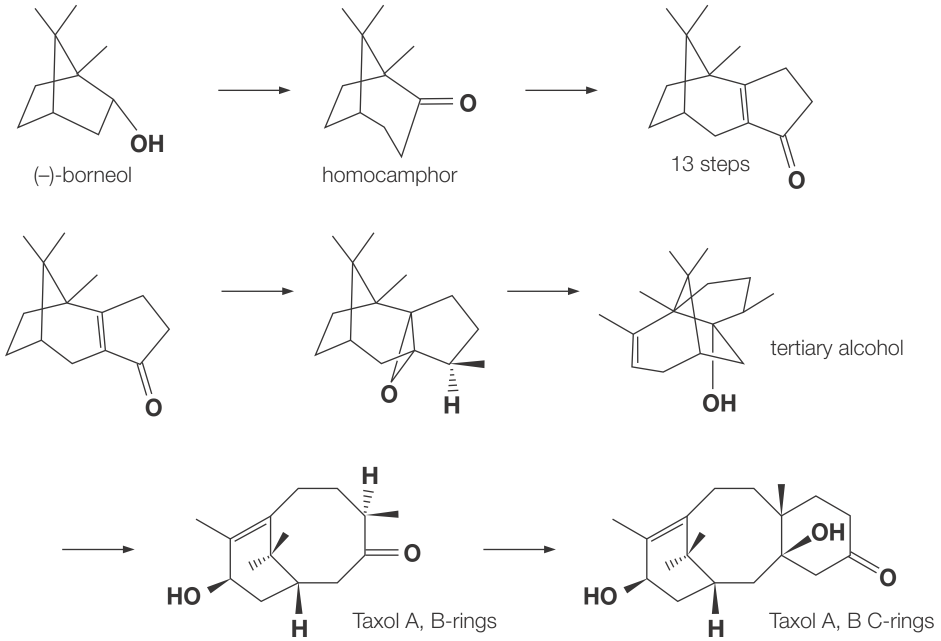


Figure 25.104 Structure of Taxotere

A number of total syntheses of Taxol have been carried out and part of one of these syntheses is very briefly outlined below. Robert Holton of Florida State University used one of the enantiomers of a cyclic alcohol called borneol as his starting compound. This is a natural product obtained from *Dryobalanops aromatica*, a member of the teak family. This was converted in a series of 13 synthetic steps to an unsaturated ketone. Further reactions were carried out, including a rearrangement to create two of the rings present in Taxol – a so-called linear route. The third ring was then introduced later (Figure 25.105).

Figure 25.105 A summary of Holton's Taxol synthesis



Another approach, as exemplified by K. C. Nicolaou of the Scripps Research Institute, is to utilize a so-called convergent route where the three carbocyclic rings of Taxol are made separately and then joined together.

Although a number of different total syntheses have been carried out for Taxol, all are very complex, use chiral auxiliaries and are not easy to perform. This, coupled with their low yields, means that they are, at the present time, not economically feasible for the industrial manufacture of Taxol. Their main use is for designing analogues of Taxol in an effort to develop a range of active anti-cancer compounds based upon the Taxol skeleton. This is arguably more important, for stimulating and testing new, generally applicable synthetic reagents, transformations and strategies.

Chiral auxiliaries

To synthesize Taxol or Taxotere from their precursor molecules, the side-chains of these drugs need to be synthesized in the laboratory. Since these chains contain two chiral centres their synthesis from achiral molecules would lead to a mixture of several stereoisomers, which would need separation.

Hence both side-chains are synthesized using a chiral auxiliary (see Section 25.1), which converts an achiral molecule S into the preferred enantiomer, A or B. This avoids the need to separate enantiomers from a racemic mixture. In other words it converts the problem into one based on diastereo-control and the easier separation of diasteroisomers. A chiral auxiliary (X) works by bonding itself to the non-chiral molecule S to produce the stereochemical intermediate structure required to make the reaction favour the desired 'stereochemical' direction (i.e. the desired enantiomer). Once the new intermediate stereoisomer molecule X–A or X–B is formed, the chiral auxiliary molecule X can be cleaved and recycled, leaving the desired enantiomer required, A or B.

A chiral auxiliary is a control element temporarily incorporated into the structure of the substrate in order to direct the stereochemistry at new stereogenic centre(s) formed in a reaction. The auxiliary is removed (either immediately during work-up or in a separate subsequent step) and may be recovered for re-use. The sequence is shown in diagrammatic form in Figure 25.106.

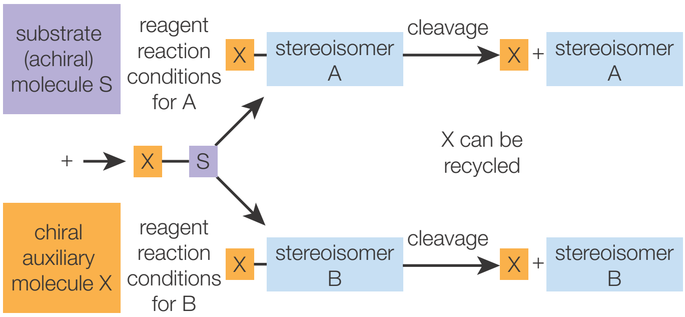

substrate (achiral) molecule S
reagent reaction conditions for A
X
stereoisomer A
cleavage
X
+
stereoisomer A
+
X
S
X can be recycled
chiral auxiliary molecule X
reagent reaction conditions for B
X
stereoisomer B
cleavage
X
+
stereoisomer B

Figure 25.106 The action of a chiral auxiliary

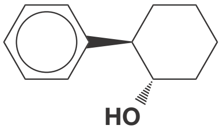


Figure 25.107 The structure of the chiral auxiliary *trans*-2-phenylcyclohexanol

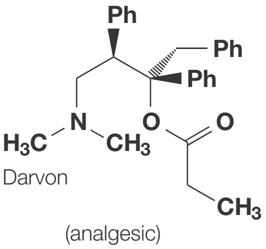
The chiral auxiliary used in the synthesis of Taxol and Taxotere is a single stereoisomer of *trans*-2-phenylcyclohexanol (Figure 25.107). It is a large molecule with two chiral carbon centres, which favour the formation of specific diasteroisomers in the subsequent steps of the synthesis. It is also known as Whitesell's chiral auxiliary and its chirality is ultimately derived from a plant-derived terpene.

The use of chiral auxiliaries allows specific stereoisomers to be synthesized in a stereoselective reaction, but small quantities of the other stereoisomer are always formed along with the target molecule. Since the configuration of the chiral centre(s) in the chiral auxiliary is fixed, the unwanted stereoisomers will be diastereoisomers which will have specific chemical properties and physical properties, for example solubility and melting point. Unwanted diastereoisomers can be separated and removed from the reaction mixture by fractional crystallization, solvent extraction or high-pressure liquid chromatography (HPLC). The identity and purity of chiral compounds can be determined using a polarimeter and analysing a sample of the chiral substance, either as a pure liquid or in solution.

Optical activity

Enantiomers have the property of rotating the plane of polarized light as it passes through a sample of the molecules. Such molecules are said to be optically active. The most common cause of optical activity is the presence of one or more chiral (asymmetric) carbon atoms.

Optical activity is an integral property of many natural products and biomolecules. Medicinal chemistry is often involved with their isolation, study and synthesis. Hence the medicinal chemist will usually need to measure an optical rotation at some stage during drug development.



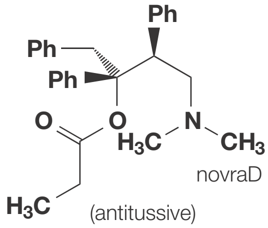


Figure 25.108 Darvon and novraD are enantiomers of each other with different biological properties

For example, for the commercial compound propoxyphene (Figure 25.108), one enantiomer is an analgesic whereas the other shows antitussive (anti-coughing) properties. Even the names of the two enantiomers are mirror images.

New drugs are often prepared by a stereoselective reaction and usually contain only a single enantiomer or stereoisomer of the active compound. The levels of the other enantiomer or stereoisomers are usually kept at a minimum because often only one isomer has a therapeutic effect, and sometimes the other enantiomer is toxic or otherwise detrimental.

The instrument used to measure optical rotation is called a polarimeter. Plane polarized light is obtained by passing light from a sodium lamp through a polarizer. This light is then passed through the sample, either as the pure liquid or in solution, which causes the polarization plane to be rotated to an extent depending on the optical activity of the sample.

To describe the outcome of these measurements we need to use the following terms:

- A dextrorotatory compound rotates the plane of polarized light in a clockwise direction and is designated as d or (+).
- A levorotatory compound rotates the plane of polarized light in an anticlockwise direction and is designated as l or (−).

There is no relationship between (+)/(−) and the R/S system of denoting absolute configurations (Chapter 20).

Optical rotation is the rotation of the plane of polarized light by an angle *a*. If it is clockwise it is designated (+), whereas if it is anticlockwise the designation is (−). Each optically active isomer (enantiomer) has a unique rotation angle and hence a pure isomer of an unknown compound can be identified by its rotation angle.

Specific rotation $[\alpha]$ is the optical rotation of a non-racemic mixture of a chiral substance in solution at a specified concentration (g/100 mL) in a cell of specified length (dm).

When quoting the specific rotation of a sample of an enantiomer, it is necessary to give the concentration of the solvent, as well as the temperature of the solution at which the measurement was recorded, because these factors affect the value obtained.

For example, the specific rotations are +123 and −123 degrees for the (+) and (−) enantiomers respectively of limonene (a hydrocarbon present in lemon peel) at 25 °C for a concentration of 1 g/100 mL).

A racemic mixture of two enantiomers (1 : 1 mixture by amount) will be optically inactive as the opposing effects of the two enantiomers will cancel each other out. Other proportions of enantiomers will produce rotation angles from +*X*° to −*X*°, where +*X* and −*X* are the rotation angles of the pure enantiomers.

The optical purity of a mixture of enantiomers can be calculated from the following expression: (observed specific rotation/specific rotation of pure enantiomer) × 100.

For example, a sample of (+) limonene contaminated with (−) limonene has an observed specific rotation of +106° but the literature value of the pure enantiomer is +123°. Hence the optical purity = (+106/+123) × 100 = 86%.

Nature of Science

Vinblastine and vincristine

Vinblastine and vincristine are alkaloids (weak bases) found in the Madagascar periwinkle, *Catharanthus roseus* (formerly known as *Vinca rosea*). The natives of Madagascar traditionally used *V. rosea* to treat diabetes. It has also been used for centuries throughout the world to treat all kinds of ailments from wasp stings, in India, to eye infections in the Caribbean.

When researchers began to analyse the plant in the 1950s they discovered it contained over 70 alkaloids. Some were found to lower blood sugar levels and others to act as hemostatics (to stop bleeding), but the most interesting were vinblastine and vincristine, which were found to lower the number of white cells in blood. A high number of white cells in the blood indicates leukemia and so a new naturally occurring anti-cancer drug had been discovered.

Vinblastine and vincristine work by preventing mitosis (cell division) by binding to the protein tubulins that form microtubules, thereby preventing the cell from making the spindle it needs to be able to divide. The spindle consists of microtubules that pull the chromosomes to the poles of the dividing cell. This is different from the action of Taxol, which interferes with cell division by keeping the spindle from being broken down.

Vinblastine is mainly useful for treating Hodgkin's disease (a form of lymphoma), advanced testicular cancer and advanced breast cancer. Vincristine is mainly used to treat acute leukemia and other lymphomas.

The periwinkle plant produces these compounds in very low concentrations, ranging from less than 0.0005 per cent to 0.3 by mass. As a result, the market price for vinblastine and vincristine extracted from the plant are relatively high. The supply of these compounds is limited by the availability of the plant. Chemists are seeking total syntheses for vinblastine and vincristine as well as cell cultures of the periwinkle.

There is an unequal availability and distribution of certain drugs and medicines around the globe

More than 10 million children die each year, most from preventable causes and almost all in poor countries. Six countries account for 50 per cent of worldwide deaths in children younger than 5 years; these include three countries in South Asia. These deaths are preventable, but in many poor countries basic healthcare remains financially and often physically out of reach for millions of people.

A small number of infectious diseases, including diarrhoea, are treatable with inexpensive generic drugs but still account for 70–90 per cent of all childhood illness and deaths in the developing world. Existing access to essential medicines for many of the world's poor is very inadequate. A large percentage of the most vulnerable populations live hours or days from the nearest doctor or hospital.

The health situation of the developing world poor is characterized by a high incidence of communicable diseases. Communicable diseases spread from one person to another or from an animal to a person. The majority of deaths still occur largely due to conditions that are preventable or could be stopped by early intervention. In the poorest parts of the world, illnesses associated with poverty remain major contributors to the burden of disease. The most common illnesses causing death include neonatal disorders, respiratory infections, diarrhoeal diseases and nutritional deficiency disorders.

Diarrhoea kills more children than HIV/AIDS, malaria and tuberculosis combined, yet compared to these diseases it receives little financing and is not prioritized by governments in donor countries. A report by WaterAid shows that between 2004 and 2006 only $1.5 billion was spent globally on improving sanitation. In the same period, $10.8 billion was spent on interventions for HIV/AIDS (responsible for 315 000 child deaths in that period) and $3.5 billion on interventions for malaria (responsible for 840 000 child deaths).

25.8 Nuclear medicine (AHL) *– nuclear radiation, whilst dangerous owing to its ability to damage cells and cause mutations, can also be used to both diagnose and cure diseases*

Introduction

Nuclear chemistry deals with the study of nuclear particles, nuclear forces and nuclear reactions. This branch of chemistry began with the discovery of radioactivity (see Chapter 2). A great deal of knowledge was accumulated during the next few decades and this led to the discovery of artificial radioactivity. The release of huge amounts of energy through nuclear reactions became possible with the development of artificial radioactivity. Nuclear chemistry is important not only in providing alternatives to fossil fuels but also in providing radioisotopes for use in nuclear medicine.

Nuclear medicine is the use of radioactive materials (usually in the form of soluble complexes or salts) in medicine for diagnostic imaging and therapy (usually cancer therapy). In therapy, the aim is to localize the radioactive material in cancerous (malignant) cells so they are killed by the ionizing radiation. Since the normal cells surrounding the diseased tissue must not be killed, short-range ionizing radiation (alpha and beta radiation or emission) is used.

Ionizing radiation

Ionizing or nuclear radiation is usually composed of three types of radiation, which differ in their nature and properties (Table 25.5).

Table 25.5 The properties of ionizing radiation

	Alpha particle	Beta (minus) particle	Gamma ray
Relative charge and relative mass	+2 and 4	−1 and 0	0 and 0
Nature	Helium nucleus or helium ion	High-speed electron	Electromagnetic wave with very short wavelength $<10^{-10}$ m
Representation	${}^{4}_{2}He$	${}^{0}_{-1}e$	${}^{0}_{0}\gamma$
Action of magnetic field	Deflected towards cathode (negative)	Deflected towards anode (positive)	Not deflected
Velocity	1/10 speed of light	Up to 9/10 speed of light	Speed of light ($3.00 \times 10^{8}\,m\,s^{-1}$)
Ionizing power	Very high: nearly 100 × that of beta particles	Low: nearly 100 × less than that of gamma rays	Low
Penetrating power	Low; easily stopped by air	100 × that of alpha particles	10 × that of beta particles
Kinetic energy	High	Considerably less than alpha particles	Zero
Nature of product	Product obtained by the loss of one alpha particle has an atomic number less 2 units and a mass number less 4 units	Product obtained by the loss of one beta particle has an atomic number greater by 1 unit, without any change in mass number	There is no change in atomic number or mass number

The emission of a beta (minus) particle is accompanied by the emission of another particle, the anti-neutrino, but this is not required for the current IB Chemistry syllabus.

Another type of radioactive decay is positron emission. A positron is a particle that has the same mass as an electron but an opposite charge. It is the anti-particle of the electron. The positron is represented as $^{0}_{1}e$. The isotope carbon-11 undergoes decay by positron emission:

$$^{11}_{6}C \rightarrow ^{11}_{5}B + ^{0}_{1}e$$

Positron emission causes the atomic number to decrease from 6 to 5. The emission of a positron has the effect of converting a proton to a neutron, thereby decreasing the atomic number of the nucleus by 1:

$$^{1}_{1}p \rightarrow ^{1}_{0}n + ^{0}_{+1}e$$

The positron has a very short life because it is annihilated and converted to energy (gamma rays).

$$^{0}_{1}e + ^{0}_{-1}e \rightarrow 2^{0}_{0}\gamma$$

This process is used in positron emission tomography (PET) (see Chapter 2): compounds containing radionuclides that decay by positron emission are injected into a patient. These compounds are chosen to allow researchers to monitor blood flow and oxygen and glucose metabolic rates. The changes in how glucose is metabolized by the brain may suggest cancer, epilepsy, Parkinson's disease or schizophrenia.

Table 25.6 summarizes the nuclear symbols used to represent the various nuclear and subatomic particles commonly found in nuclear reactions.

Table 25.6 Nuclear symbols

Particle	Symbol
Neutron	$^{1}_{0}n$
Proton	$^{1}_{1}H$ or $^{1}_{1}p$
Electron	$^{0}_{-1}e$
Alpha particle	$^{4}_{2}He$ or $^{4}_{2}\alpha$
Beta particle	$^{0}_{-1}e$, $^{0}_{-1}\beta$ or β^{-}
Positron	$^{0}_{+1}e$, $^{0}_{1}\beta$ or β^{+}

Positron emission topography

PET provides images of blood flow or biochemical functions, depending upon the type of molecule that is radioactively tagged. PET can show images of glucose metabolism in the brain, or rapid changes in biochemical activity in various areas of the body. However, there are usually relatively few PET centres in most countries because they must be located near a particle accelerator device that produces the short-lived radioisotopes used in the technique. Cost and availability are clearly important factors in determining the use of PET by patients.

Nuclear equations

1 Nuclear reactions are written like a chemical reaction: the reactants in a nuclear reaction are written on the left-hand side and the products on the right-hand side with an arrow in between them.

2 The mass number and atomic number of each element are written in a nuclear reaction. They are inserted as superscripts and subscripts respectively on the symbol of the element. For example, $^{27}_{13}Al$ represents a nuclide of aluminium with a mass number of 27 and an atomic number 13. A nuclide is a nucleus of a particular mass number.

3 Mass number and atomic number are conserved. In a nuclear reaction the sum of the mass numbers and the sum of the atomic numbers are balanced on the two sides of the reaction (recall that in a chemical reaction the total numbers of atoms of the various elements are balanced on the two sides).

4 The energy involved in the nuclear reactions may be indicated in the product as +Q or –Q for reactions accompanied by release or absorption of energy, respectively.
For example a nitrogen-14 nucleus reacts with a helium-4 nucleus to form an oxygen-17 nucleus and a proton. Energy is absorbed:

$$^{14}_{7}N + ^{4}_{2}He \rightarrow ^{17}_{8}O + ^{1}_{1}H + -Q$$

Types of nuclear reactions

There are two types of nuclear reactions: natural nuclear reactions and artificial nuclear reactions. In natural nuclear reactions, a nucleus of a single atom (element) undergoes a spontaneous change. Alpha and beta decays are examples of natural nuclear reactions.

In artificial nuclear reactions, two nuclei of different elements (atoms) are brought to interact artificially. This is done by bombarding a relatively large nucleus (non-radioactive) with a lighter nucleus, e.g. protons, alpha particles or helium atoms. For example, the artificial radionuclide phosphorus-15 is prepared by bombarding aluminium-27 with alpha particles:

$$^{27}_{13}\text{Al} + {}^{4}_{2}\text{He} \rightarrow {}^{30}_{15}\text{P} + {}^{1}_{0}\text{n}$$

Many of the radionuclides used in nuclear medicine are directly or indirectly prepared by these types of reactions, such as iodine-131 and technetium-99m (molybdenum-90 reactor produced). Two other important nuclear reactions are nuclear fission and nuclear fusion (see Chapter 24).

Production of artificial isotopes

Most of the radioactive nuclides used in nuclear medicine do not occur naturally. Some are produced in nuclear fission reactors at nuclear power stations from natural, often stable isotopes; others are produced in a cyclotron (a particle accelerator). Usually a sample of the natural isotope is bombarded (irradiated) with neutrons obtained as a by-product in the reactors (see Chapter 24).

For example, the radioisotope sodium-24 can be produced from non-radioactive sodium-23 by neutron bombardment:

$$^{23}_{11}\text{Na} + {}^{1}_{0}\text{n} \rightarrow {}^{24}_{11}\text{Na} + {}^{0}_{0}\gamma$$

Not all the nuclides will be changed and it is not possible to separate the two isotopes on a chemical basis. Therefore it is not possible to have a pure sample of the radioactive sodium. Sodium-23 is said to be the carrier for the radioisotope sodium-24.

Neutrons are often used as projectiles in many transmutation processes. Since neutrons are electrically neutral and not strongly repelled by the target nuclei, high kinetic energy neutrons are not necessary. They are sufficiently accelerated by heat, hence called thermal neutrons, which have just enough energy to bind to the target nuclei.

Neutron bombardments are often used to produce radioisotopes such as cobalt-60 from cobalt-59 or iron-59 from iron-58:

$$^{59}_{27}\text{Co} + {}^{1}_{0}\text{n} \rightarrow {}^{60}_{27}\text{Co} + \text{Y} \qquad {}^{58}_{26}\text{Fe} + {}^{1}_{0}\text{n} \rightarrow {}^{59}_{26}\text{Fe}$$

9 Complete the following nuclear equations and identify the other products.

a $^{13}_{6}\text{C} + {}^{1}_{0}\text{n} \rightarrow {}^{4}_{2}\text{He} +$ _____

b $^{14}_{7}\text{N} + {}^{4}_{2}\text{He} \rightarrow {}^{1}_{0}\text{n} +$ _____

c $^{253}_{99}\text{Es} + {}^{4}_{2}\text{He} \rightarrow {}^{1}_{0}\text{n} +$ _____

d $^{53}_{24}\text{Cr} + {}^{4}_{2}\text{He} \rightarrow {}^{1}_{0}\text{n} +$ _____

e $^{250}_{98}\text{Cf} + {}^{11}_{5}\text{B} \rightarrow$ _____ $+ 4\,{}^{1}_{0}\text{n}$

Natural radioactivity

Radioactive elements (radioisotopes) are generally heavy elements which are unstable; the origin of that instability is the nucleus. This unstable nucleus undergoes a spontaneous breakdown with the emission of either an alpha particle or a beta particle. This may also be accompanied by gamma radiation.

The decay is a random process; that is, each and every atom has an equal chance for radioactive decay at any time. The number of atoms that decay per second is directly proportional to the number of remaining undecayed radioactive atoms present at any time. The decay is independent of temperature, pressure and whether the radioisotope is an element or part of a compound.

A daughter nuclide is different from its parent nuclide, not only in its radioactive properties but also in other chemical and physical properties. Such a process is termed transmutation, meaning one element has been converted into another element.

Types of radioactive decay

The three most common types of ionizing radiation released when a radioisotope decays are alpha, beta (negative and positive) and gamma radiation.

During alpha decay (Figure 25.109) an alpha particle, composed of two protons and two neutrons, is released. When a nucleus decays by alpha particle emission, proton number or atomic number (Z) decreases by two and its mass number or nucleon number (A) decreases by four.

During beta (minus) decay (Figure 25.110) the mass or nucleon number (A) remains unchanged, but the atomic number (Z) increases by one. During this process, a neutron splits into a proton and an electron (and an anti-neutrino). The proton number increases and the energetic electron leaves the nucleus as a beta particle.

In a special type of beta decay, known as beta positive decay (Figure 25.111), a positron is emitted. The positron is the anti-particle of the electron. The positron has a positive charge +*e* and the mass of a positron is equal to that of the electron. (A neutrino is also released, but not shown in the nuclear equation.)

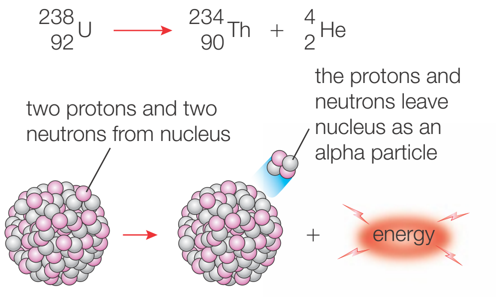

$^{238}_{92}U \rightarrow\ ^{234}_{90}Th + \ ^{4}_{2}He$
two protons and two neutrons from nucleus
the protons and neutrons leave nucleus as an alpha particle
+
energy

Figure 25.109 Alpha decay of uranium-238

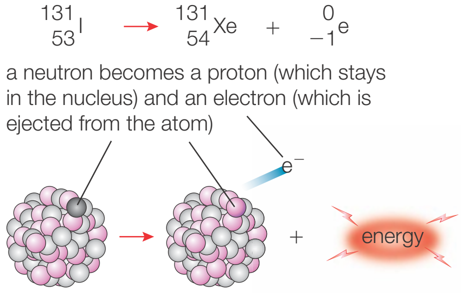

$^{131}_{53}I \rightarrow\ ^{131}_{54}Xe + \ ^{0}_{-1}e$
a neutron becomes a proton (which stays in the nucleus) and an electron (which is ejected from the atom)
e^-
+
energy

Figure 25.110 Beta (minus) decay of iodine-131

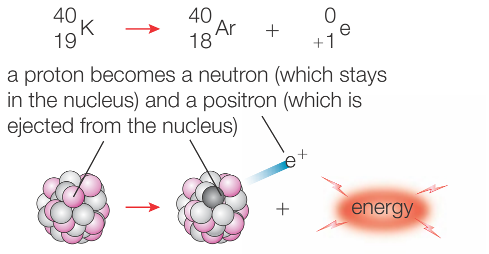
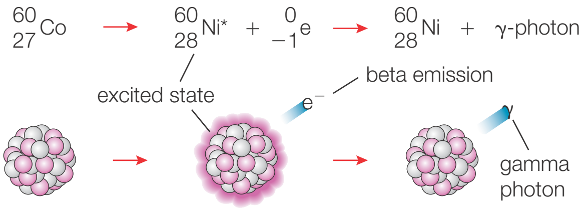

$^{40}_{19}K \rightarrow\ ^{40}_{18}Ar + \ ^{0}_{+1}e$
a proton becomes a neutron (which stays in the nucleus) and a positron (which is ejected from the nucleus)
e^+
+
energy

Figure 25.111 Beta (positive) decay of potassium-40

In the nucleus, protons and neutrons are bound together by a very strong nuclear binding energy, which is effective only over a very short distance (within ~10^{-15} m). Both nuclear particles also appear to exist in a set of quantized energy shells, analogous to electronic shells in atoms.

The emission of gamma rays (Figure 25.112) has no effect on the nucleon number or the proton number of the nucleus. Gamma rays are usually emitted at the same time as alpha and beta particles. For some radioactive nuclides, the emission of alpha and beta particles from the nucleus leaves the protons and neutron in an excited arrangement. The protons and neutrons undergo transitions within the nuclear energy levels and release energy as gamma photons.

Figure 25.112 Gamma and beta emission from cobalt-60


$^{60}_{27}Co \rightarrow\ ^{60}_{28}Ni^* + \ ^{0}_{-1}e \rightarrow\ ^{60}_{28}Ni + \gamma$-photon
excited state
beta emission
e^-
γ
gamma photon

Balancing nuclear equations involving alpha and beta particles

Deduce what product is formed when radium-226 undergoes alpha decay.

This is best solved by writing a nuclear equation for the process.

The periodic table shows that radium has an atomic number of 88. The nuclide notation for radium-226 is therefore $^{226}_{88}\mathrm{Ra}$. An alpha particle is a helium-4 nucleus, and so its nuclide notation is $^{4}_{2}\mathrm{He}$. The alpha particle is a product of the nuclear reaction, and so the nuclear equation is:

$$^{226}_{88}\mathrm{Ra} \rightarrow {}^{A}_{Z}\mathrm{X} + {}^{4}_{2}\mathrm{He}$$

where A is the mass number of the product (daughter) nucleus and Z is the atomic (proton) number. Mass numbers and atomic numbers must balance, so:

$$226 + A + 4; 88 = Z + 2; \text{hence } A = 222 \text{ and } Z = 86.$$

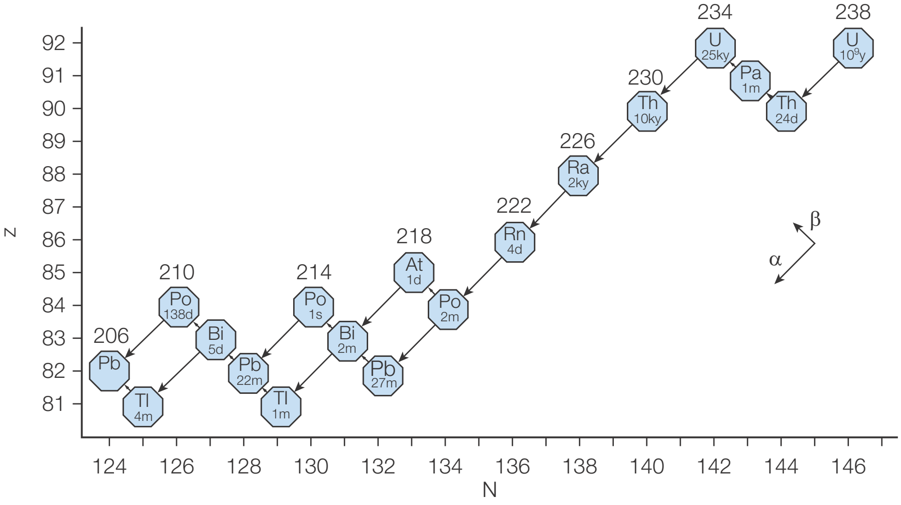
From the periodic table, the element with $Z = 86$ is radon (Rn). The product is therefore $^{222}_{86}\mathrm{Rn}$, and the nuclear equation is:

$$^{226}_{88}\mathrm{Ra} \rightarrow {}^{222}_{86}\mathrm{Rn} + {}^{4}_{2}\mathrm{He}$$

Decay series

When the product of a radioactive disintegration is itself unstable, then a further radioactive disintegration will occur. Thus a whole series of spontaneous radioactive disintegration can occur (see the uranum-235 decay series in Figure 25.113), each step involving the loss of either an alpha or a beta particle or very occasionally both, until eventually a stable isotope is formed.

Figure 25.113 The uranium decay series



10 Balance the following nuclear equations:

a $^{239}_{94}\mathrm{Ra} \rightarrow {}^{4}_{2}\mathrm{He} +$ ____

b $^{40}_{19}\mathrm{K} \rightarrow {}^{0}_{-1}\mathrm{e} +$ ____

c $^{99}_{43}\mathrm{Tc} \rightarrow {}^{99}_{44}\mathrm{Ru} +$ ____

d $^{218}_{84}\mathrm{Po} \rightarrow {}^{214}_{82}\mathrm{Pb} +$ ____

11 Deduce the final product from the decay of uranium-235. One alpha particle and two beta particles are emitted.

12 Write balanced nuclear equations for the alpha decay of uranium-238 and the beta decay of lead-210.

13 Write a nuclear equation showing the emission of gamma radiation from an excited xenon-131 nucleus.

14 Deduce which element undergoes alpha decay to form lead-208.

The kinetics of radioactive decay

Nuclear activity is the rate of nuclear decay. The SI unit of nuclear activity is called the becquerel (Bq), where 1 Bq = 1 event/s or 1 disintegration/s.

With time, the number of parent nuclei decreases because of radioactive decay. This decrease obeys the law of radioactive decay. At the initial instant of time, $t = 0$, let there be Δt nuclei of the same element that will remain undecayed by time t. Since we are dealing with spontaneous changes, it is natural to assume over a longer interval of time that a greater number of nuclei will decay. If we have N undecayed nuclei present at time t, and $N - \Delta N$ undecayed nuclei existing at time $t + \Delta t$ then the change in the number of undecayed nuclei, that is, the number of nuclei decaying in time Δt, will be proportional to N; that is:

$$\Delta N = N\Delta t \text{ or } \Delta N = -\lambda N\Delta t$$

where λ is a positive proportionality factor called the decay constant, which it has a definite value for each nuclide. The minus sign on the right-hand side of the equation indicates that ΔN decreases with time. Thus it follows that the decay constant is the fractional decrease in the number of nuclei decaying per unit time:

or using calculus notation: $\dfrac{dN}{dt} = \lambda N$

In other words, the decay constant represents the proportion of nuclei decaying per unit time, or the decay rate. The decay constant is independent of the surrounding conditions (such as pressure and temperature) and is only determined by the internal properties of the nucleus. It has dimensions of $|\lambda| = T^{-1}$.

In order to find the time dependence for radioactive decay we can show that the number of atoms of the original nuclei remaining after time t is:

$N = N_0\exp(-\lambda t)$ (this equation is given in Table 1 of the *IB Chemistry data booklet*)

where N_0 is the initial number of radioactive nuclei existing at $t = 0$ and N is the number of radioactive nuclei present at time t. A plot of $\ln(N/N_0)$ as a function of time shows the decrease is exponential. The decay constant λ can be found from the slope of the curve.

In practice the stability of radioactive nuclei against decay and the decay rate are most often estimated in terms of the half-life, $t\frac{1}{2}$, rather than the decay constant λ.

The half-life is defined as the time at which half of the original nuclei have decayed. The half-life is the time after which one-half the original number of nuclei remains unchanged.

By this definition and on the basis of the exponential decay law, $t_{\frac{1}{2}}$ and λ are related; cancelling N_0 and taking a logarithm, we obtain:

$t_{\frac{1}{2}} = \dfrac{\ln 2}{\lambda} = \dfrac{0.693}{\lambda}$ (this equation is given in Table 1 of the *IB Chemistry data booklet*)

The half-life represents a 50 per cent chance that a nuclide will decay during that period. The activity of a radioactive substance is normally given in the form of its half-life, from which the rate constant, k, can be calculated. For example,

$${}^{14}_{6}\text{C} \rightarrow {}^{14}_{7}\text{N} + {}^{0}_{-1}\beta$$

$$t_{\frac{1}{2}} = (0.693/k) = 5730 \text{ years}; k = (0.693/5730 \text{ years}) = 1.21 \times 10^{-4} \text{ year}^{-1}$$

The data booklet also has the following equation that can be used in radioactivity calculations:

$$N_t = N_0(0.5)^{t/k}$$

This equation is a way of enabling direct calculation of N from time and half-life, t, except it is using the rate constant, k, instead of the half-life in the expression used earlier in Section 1 of the *IB Chemistry data booklet*.

Worked example

The half-life of radium is equal to 1590 years. Determine the number of nuclei in 1 g of radium and find its decay constant.

The number of radium atoms per gram is equal to Avogadro's constant, N_A, divided by the molar mass:

$= 2.67 \times 10^{24}\,\text{g}^{-1}$

Then the activity of 1 g of radium will be:

$= 3.7 \times 10^{10}\,\text{s}^{-1}$

That is, the number of decays per second in 1 g of radium is 37 000 million.

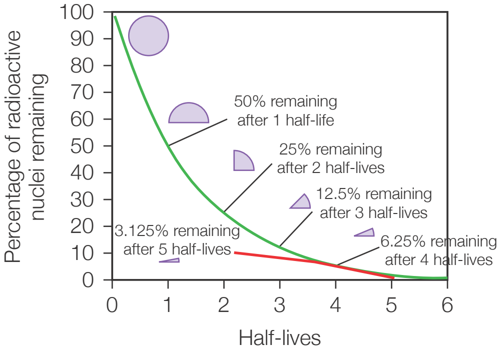
Since the decay constant is a constant, then the half-life of a particular radioisotope is independent of the amount of radioisotope. Whatever the amount of radioisotope present at a particular time, it will always decay to half of that amount at the end of one half-life. This can be illustrated in the form of Table 25.7, where *x* represents the amount of radioisotope at the start (that is, when time = 0).

Table 25.7 Decay of a radioisotope

Number of half-life periods	Mass initially present	Mass of substance that has undergone decay	Mass of remaining radioactive substance
0	*x*	0	*x*
1	*x*	*x*/2	*x*/2
2	*x*/2	*x*/4	*x*/4
3	*x*/4	*x*/8	*x*/8
4	*x*/8	*x*/16	*x*/16

Because radioactive decay is a first-order process the activity of a radionuclide decreases exponentially with time. Each radionuclide has a specific and constant half-life which can vary from less than a second to many millions of years. The concept of radioactive decay and half-life is shown in Figure 25.114.

Figure 25.114 The half-life concept – the time it takes for one-half of a radioactive sample to decay



The half-life is a measure of the radioactivity of the element since the shorter the half-life of an element, the greater is the number of the decaying atoms and hence the greater its radioactivity. The half-lives of different radioisotopes vary widely, from a few seconds to millions of years.

Calculating the percentage and amount of radioactive material decayed and remaining after a certain period of time using the nuclear half-life equation

Shown below are a wide range of examples of calculations involving the half-life concept and a range of different mathematical expressions related to radioactive decay.

Worked examples

Technetium-99m is used for brain scans. If a laboratory receives a shipment of 200 g of this isotope and after 24 hours only 12.5 g of this isotope remains, what is the half-life of technetium-99m?

Total time of decay = number of half-lives × number of years/half-life

Number of years/half-life = total time of decay/number of half-lives

Fraction of sample remaining = final mass sample/initial mass sample = $\frac{12.5\,\text{g}}{200\,\text{g}} = 0.0625 = \frac{1}{16}$

$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$ so four half-lives have passed

Number of hours/half-life = $\frac{24}{4}$ = 6 hours

The half-life of uranium-232 is 70 years. How many half-lives will it take for 10 g of it to be reduced to 1.25 g?

Number of years/half-life = total time of decay/number of half-lives

Fraction of sample remaining = final mass sample/initial mass sample = $\frac{1.25\,\text{g}}{10\,\text{g}} = 0.125 = \frac{1}{8}$

After one half-life has passed = $\frac{1}{2}$; after two half-lives $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$; after three half-lives $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$; so three half-lives have passed

The disintegration constant of ^{55}Fe is 0.2665 years^{-1}. Calculate the half-life of the radioisotope.

$\lambda = 0.2665$ years^{-1}

$t_{1/2} = 0.693/\lambda$

$t_{1/2} = \frac{0.693}{0.2665} = 2.6$ years

The half-life of ^{45}Ca is 165 days. Calculate the disintegration constant of the radioisotope.

$t_{1/2} = 165$ days

$\lambda = \frac{0.693}{t_{1/2}}$

$\lambda = \frac{0.693}{165} = 0.0042$ days^{-1}

The half-life of technetium-99 m is 6 hours. Calculate the percentage remaining after 4 hours.

$\lambda = \frac{0.693}{t_{1/2}}$

$\lambda = \frac{0.693}{6} = 0.116$ hours^{-1}

Percentage remaining = $100 \times e^{-kt} = 100 \times e^{-0.1155\times4} = 63.00\%$

0.250 kg of radium-226 emits alpha particles at a measured rate of $9.00 \times 10^{12}\,\text{s}^{-1}$. What is the half-life of radium-226? (The Avogadro constant is $6.02 \times 10^{23}\,\text{mol}^{-1}$.)

$t_{1/2} = 0.693/\lambda$

$dN/dt = -\lambda N$

226 g of radium-226 contains 6.02×10^{23} atoms, hence

$N = \frac{250}{260} \times 6.02 \times 10^{23} = 6.65 \times 10^{24}$ atoms

$-9.00 \times 10^{12}\,\text{s}^{-1} = -\lambda \times 6.65 \times 10^{23}$

$\lambda = 9.00 \times 10^{12}\,\text{s}^{-1}/6.65 \times 10^{23} = 1.35 \times 10^{-11}\,\text{s}^{-1}$

$t_{1/2} = \frac{0.693}{1.35} \times 10^{-11}\,\text{s}^{-1} = 5.1 \times 10^{10}\,\text{s} \approx 1620$ years

The half-life of radium is 1.62×10^3 years. Calculate what fraction of radium would remain after 1.00×10^3 years.

$\lambda = \ln 2/t_{1/2} = 0.693/1.62 \times 10^3$

$\lambda t = (0.693/1.62 \times 10^3)\ 1.00 \times 10^3 = 0.617 \ln 2$

$N = N_0 e^{-kt}$

$N/N_0 = e^{-kt} = e^{-0.617}\ln 2 = 0.652$

Radiation hazards

The radiation hazards to humans arise from exposure of the body to external nuclear radiation and ingestion (eating) or inhalation (breathing in) of radioactive materials. The effect of the nuclear radiation depends on the nature of the radiation, the part of the body irradiated (exposed to radiation) and the dose received.

The hazard from alpha particles is small, unless the source enters the body, since they cannot penetrate the outer layers of unbroken skin. Beta particles are more penetrating in general, but most of their energy is absorbed by surface tissues. Gamma rays are the most hazardous since they penetrate deeply into the body. Alpha and beta emitters can be harmful if ingested; for example, there was increased risk of thyroid cancer in children exposed to fallout from the Chernobyl disaster.

Intense nuclear radiation can cause immediate damage to tissue and, depending on the dose, is accompanied by radiation burns (redness of the skin followed by blistering and sores which are slow to heal), radiation sickness, loss of hair and, in extreme cases, death.

Damage to body cells is due to the formation of a range of ions that upset cellular reactions or even kill them. The most susceptible parts of the body are the reproductive organs, blood-forming organs such as the liver, and to some extent the eyes. Delayed effects such as cancer (especially leukemia) and eye cataracts may appear many years later. Hereditary defects (mutations) may also occur in children due to damage to the DNA of sex cells.

The absorbed dose D is the energy absorbed (in joules per kilogram) by unit mass of the irradiated material. The unit of D is termed the gray (Gy), where $1\,\text{Gy} = 1\,\text{J}\,\text{kg}^{-1}$ although the unit rad (*r*adiation *a*bsorbed *d*ose) is more commonly used, where $1\text{ rad} = 10^{-2}\,\text{J}\,\text{kg}^{-1}$ and $1\,\text{Gy} = 100\,\text{rad}$).

Equal doses of nuclear or ionizing radiation provide the same amount of energy in a given absorbing tissue, but they do not have the same biological effects on the human body. To allow for this the dose equivalent (H) is used as a measure of the effect that a certain dose of radiation has on a person. It takes into account the type of radiation as well as the amount of energy absorbed and is obtained by multiplying D by a number termed the relative biological effectiveness. The unit of H is the sievert (Sv).

dose equivalent = absorbed dose × relative biological effectiveness

For beta particles, X-rays and gamma rays the relative biological effectiveness is about 1; for alpha particles and fast neutrons it is about 20.

The traditional unit of radiation dose that takes into account the type of radiation and its relative biological effectiveness (RBE) value is called the rem (Table 25.8), where $1\text{ rem} = 1\text{ rad} \times \text{RBE}$; $1\text{ rem} = 0.01\text{ Sv}$.

Table 25.8 Effects of short-term exposures to radiation

Dose (in rem)	Clinical effects
0–25	No detectable effect
25–100	Temporary decrease in white blood cell counts
50+	Temporary sterility in men
100–200	Strong decrease in white blood cell counts
	Mild radiation sickness: vomiting, diarrhoea and tiredness in a few hours
	Reduction in infectious resistance
	Possible bone growth retardation in children
300+	Permanent sterility in men and women
500+	Serious radiation sickness: marrow and intestine destruction
	50% death of exposed population within a month
1000+	Acute illness; death within days

When ionizing radiation passes through living tissue, electrons are removed from water molecules, forming highly reactive H_2O^+ ions. An H_2O^+ ion can react with another water molecule to form an oxonium ion, H_3O^+, and a neutral $^{\bullet}OH$ molecule.

$$H_2O^+ + H_2O \rightarrow H_3O^+ + {}^{\bullet}OH$$

The $^{\bullet}OH$ species is a highly reactive hydroxyl free radical. In cells and tissues, this hydroxyl free radical reacts with a variety of biological molecules damaging them and releasing more free radicals. Hence, the formation of a single free radical can initiate a large number of reactions that may disrupt the normal activities of cells.

15 Find out about the poisoning of Alexander Litvinenko in London with polonium-210.

Nature of Science

Protection from radiation

It is generally assumed that even very small doses of ionizing or nuclear radiation can potentially be harmful (this is known as the linear no-threshold hypothesis). This implies that a person must be protected from ionizing radiation at all dose levels. This includes the nuclear medicine patient, members of their family, doctors and nurses working with nuclear medicines, and the general public. People should be protected by optimizing protection, justifying the exposure to the radioisotope and limiting the dose.

This is the system of radiological protection as defined by the International Commission on Radiological Protection (ICRP), a non-governmental international organization.

The recommended system of radiation protection by the ICRP is based upon three principles. The benefit of a practice must offset the radiation detriment (justification). Exposures and likelihood of exposure should be kept as low as reasonably achievable, economic and social factors being taken into account (optimization). Dose limits should be set to ensure that no person faces an unacceptable risk in normal circumstances.

No use of ionizing or nuclear radiation is justified if there is no benefit. All applications must be medically justified. This applies to even the smallest exposures that are potentially harmful, and the risk must be offset by a benefit. When radiation is to be used then the exposure should be optimized to minimize any possibility of harm.

This means radiation exposure should be limited as much as possible, keeping in mind the risk/benefit relation of radiation and its applications. For example, it is unreasonable to refuse an X-ray after a bone fracture because statistically the X-ray exposure may shorten your life expectancy by just one day and the benefits of the X-ray with its diagnostic value by far outweigh the risk associated with the radiation exposure.

Nuclear diagnosis

Radioactive nuclides or radioisotopes are used in nuclear medicine. The radioactive materials used are radiopharmaceuticals which are a pharmaceutical (medicine) and a radionuclide. The pharmaceutical used will depend on the organ to be studied, for example iodine-131, ^{131}I, is used to study the thyroid gland.

In nuclear diagnosis, radioisotopes are injected into the bloodstream and their passage through an organ followed and measured with a detection instrument. Alternatively, specific organs of the body can be radiolabelled with a specific radioisotope. Biochemical processes involving a series of chemical reactions can be followed using a radioactive isotope of an element. Table 25.9 shows some of the medical uses for a range of radioactive tracers.

Table 25.9 Medical uses for selected radioisotopes as tracers

Organ or tissue	Tracers	Uses
General body composition	^{24}Na and ^{42}K	Used to measure volumes of body fluids, and estimate quantities of salts
Blood	^{32}P, ^{125}I and ^{132}I	Used to measure volumes of blood and the different components of blood (plasma and red blood cells)
Bone	^{45}Ca, ^{47}Ca and ^{99m}Tc	Used to investigate absorption of calcium ions, location of bone disease and how bone is metabolizing minerals
Cancerous tumours	^{32}P, ^{133}Xe, ^{99m}Tc and ^{131}I	Used to detect, locate and diagnose tumours
Thyroid	^{123}I and ^{99m}Tc	Tests of thyroid function and morphology
Kidney	^{99m}Tc agents (with gamma camera)	Kidney function tests

The quantity of tracer used must be as small as possible to minimize exposure to harmful ionizing radiation. Exposure time is reduced if the radioactive substance that is labelled with a tracer is quickly eliminated from the body, when the radioisotope has a short half-life.

The effective lifetime of the tracer must be matched to the time scale of any process being studied. Often, a short half-life is useful, as in monitoring blood flow, which can be studied in a short period of time. The tracer should be easy to detect and its position within the body identified accurately. The preferred type of radiation is gamma because gamma rays travel easily through biological tissue and cause little ionization. Gamma radiation is long-range radiation that can be monitored from outside the patient's body. Low beta radiation is also useful. Most the isotopes in Table 25.9 are a mixture of beta, positron and gamma emitters, but technetium-99m is a pure gamma emitter.

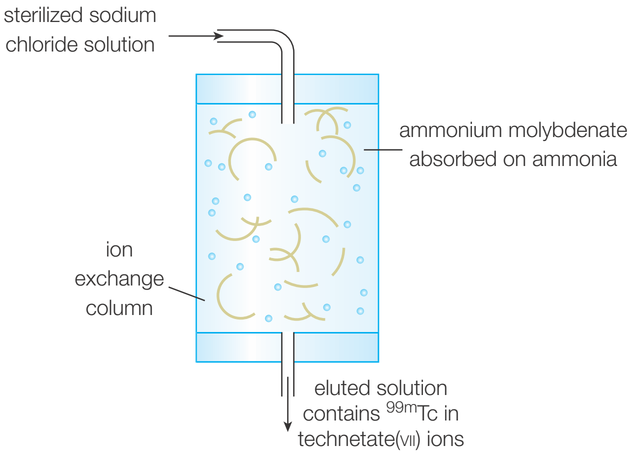
Technetium-99m – the most common radioisotope used in nuclear medicine

One of the most useful tracers in nuclear medicine is an isotope of the element technetium $^{99m}_{43}Tc$ (half-life six hours). The m indicates that it is a metastable nuclide, meaning that the nucleons in its nucleus are at an energy level higher than in a stable technetium nucleus. The excited nuclei return to the ground state with a half-life of six hours, emitting gamma rays with an energy that makes them easy to detect. The decay produces $^{99}_{43}Tc$, which is a naturally occurring radioactive isotope of half-life 216 000 years, and so is relatively stable, emitting very little radiation.

If the radionuclide has a long half-life then the patient would receive a dangerously high radiation dose. Technetium-99m has a half-life of six hours, which limits the radiation dose. However, there is the issue of the time required to transport the radionuclide from the production site (nuclear reactor) to the hospital. One approach to this problem is to use a long-lived 'parent' radionuclide, for the transportation stage, which decays to a radioactive material with a short half-life for use with patients. The two substances must be easily separated.

Technetium-99m is produced from molybdenum-99 which decays to technetium-99m by beta emission with a half-life of 67 hours. The molybdenum-99 is transported to the hospital in a generator (Figure 25.115) which consists of a saline (salt) reservoir, an alumina column which contains adsorbed molybdenum (as ammonium molybdenate, $(NH_4)_2MoO_4$) and a filter which sterilizes material passing through it. The column is shielded with lead to protect the user from gamma radiation.

Figure 25.115 Technetium generator


sterilized sodium chloride solution
ammonium molybdenate absorbed on ammonia
ion exchange column
eluted solution contains ^{99m}Tc in technetate(VII) ions

Saline (salt) solution is passed through the column containing the adsorbed molybdenum. The technetium is soluble in the salt solution, but molybdenum is insoluble. This elution produces a solution of sodium technetate(VII), $NaTcO_4$. This can be diluted and divided into a number of patients' doses.

Technetium is a transition metal and exists in several stable oxidation states (+3, +4 and +7) and readily forms complex ions with various ligands which can be administered by injection and delivered to specific organs or tissues.

Lutetium-177 and yttrium-90 – common isotopes used for radiotherapy

Lutetium-177 is a radioisotope with a half-life of 6.7 days and is a beta/gamma emitter. Lutetium-177 is often used in conjunction with a peptide (small protein) to selectively target neuro-endocrine tumours.

Yttrium-90 has a half-life of 2.67 days and is a beta emitter. Yttrium-90 therapy is often used in the treatment of liver cancer. Large numbers of small plastic or glass beads containing yttrium-90 are injected directly into the arteries that supply blood to the liver.

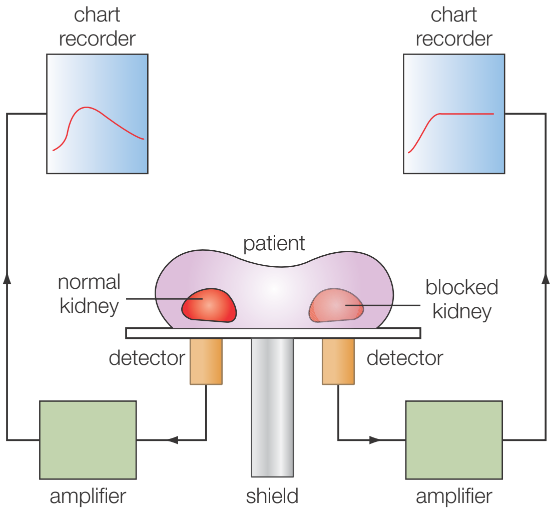
Radioisotopes used in diagnostic nuclear medicine should ideally be gamma emitters, since alpha radiation is easily absorbed and would cause unnecessary exposure to highly ionizing radiation. Also for *in vivo* measurements inside the body it would not be detected outside the body. Beta radiation *in vivo* would be partly absorbed and would cause unnecessary exposure to ionizing radiation. Beta transport is typically only a few millimetres so is also not detected externally.

Examples of nuclear diagnostic tests

Kidney function (*in vivo*)

The patient is usually given a large glass of water about 30 minutes before an intravenous injection of a fluid termed hippuran labelled with iodine-123 (half-life 13 hours). A radiation detector connected to a chart connector is held against the kidney to record the activity in the kidney over 15 minutes. The reading from a normal kidney rises and then decreases as the radioactive iodine passes through (Figure 25.116). However, the reading from a blocked kidney would rise but not decrease. Many kidney studies now use gamma cameras for imaging.

Figure 25.116 Kidney test with iodine-123

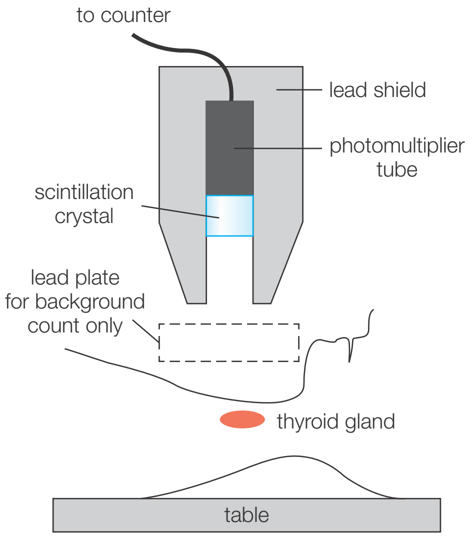


Thyroid tests

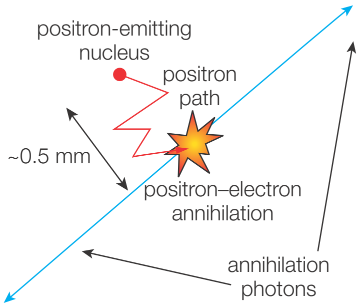
The thyroid gland absorbs iodine to maintain its function of producing a hormone (thyroxine). The rate of uptake is measured by giving the patient a solution containing a solution containing sodium iodide labelled with radioactive iodine (Figure 25.117). The activity of the patient's thyroid and the activity of a 'phantom' solution prepared at the same time is measured 24 hours later. The percentage uptake by the patient is calculated from the following expression:

$$\frac{\text{corrected count rate of the thyroid}}{\text{corrected count rate of the phantom}} \times 100\%$$

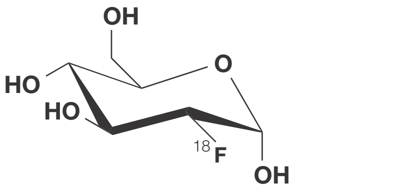
A normal thyroid has a percentage uptake of 20 – 50 per cent after 24 hours.



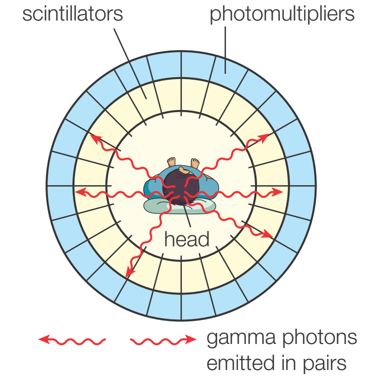
■ **Figure 25.117** Thyroid test



■ **Figure 25.118** General principle of PET imaging



■ **Figure 25.119** FDG with a fluorine-18 radiotracer



■ **Figure 25.120** Arrangement of PET scanners for a PET scan

Blood flow through the brain (*in vivo*)

In this diagnostic test, the head is positioned against two counters, each against one hemisphere. No more than 0.5 cm^3 of sodium technetate(VII), $NaTcO_4$, labelled with ^{99m}Tc is injected intravenously and the counters detect the movement of the sample through the brain's blood vessels a few seconds later. By comparing the traces with each other and with traces known to be normal, information on blood through the brain's hemispheres is obtained. Another approach is to use a Tc-99m labelled cerebral blood flow agent and gamma camera imaging.

Positron emission tomography

PET offers a number of unique advantages compared to other medical imaging techniques. PET (see Chapter 2) measures the two annihilation photons that are produced after positron emission (Figure 25.118) from a radionuclide tagged tracer molecule, which is chosen to mark a specific function in the body. Hence PET provides molecular imaging of biological function instead of anatomy. The detection of both annihilation photons yields increased sensitivity and accuracy over all other forms of medical imaging.

Only small amounts of imaging substrate need to be injected (tracer principle) because of the high sensitivity of PET. In addition, positron-emitting isotopes that are used in medical imaging (C-11, N-13, O-15, F-18) are relatively short-lived, which enables optimal use of imaging photons while keeping the radiation dose low. Furthermore, many of these isotopes can be incorporated into biological substrates (glucose, water, ammonia, carbon dioxide and oxygen) and pharmaceuticals without altering their biological activity.

The most common substance used in PET is 2-fluoro-2-deoxyglucose (FDG) containing a radiotracer, fluorine-18 (Figure 25.119). Fluorine-18 is produced by irradiation of the stable oxygen-18 isotope with high-energy protons (18 MeV) from a cyclotron:

$$^{18}_{8}O + ^{1}_{1}p \rightarrow ^{18}_{9}F + ^{1}_{0}n$$

The target is water.

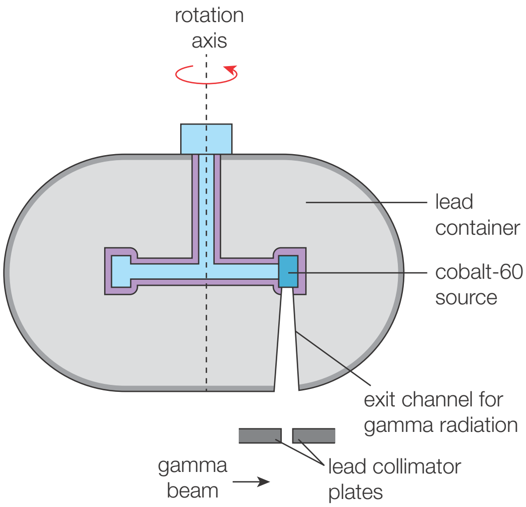
When FDG is injected into the circulation it is distributed around the blood and the body in a similar manner to glucose. FDG is taken up by body cells that are high users of glucose such as brain and cancer cells. The presence of fluorine in the molecule stops the body from metabolizing it as a normal glucose molecule. Once the fluorine has decayed to oxygen-18 the normal metabolic process proceeds. Fluorine-18 has a half-life of 100 minutes.

Positrons (the anti-particles of electrons) are emitted by fluorine-18 nuclei and undergo an interaction ('collision') with electrons, producing pairs of high-energy photons (gamma rays) moving in opposite directions:

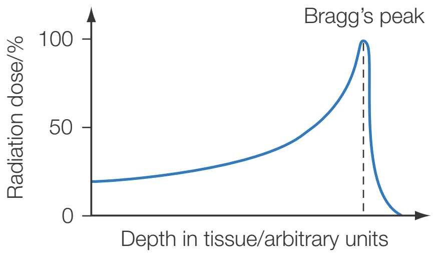
$$^{18}_{9}F \rightarrow ^{18}_{8}O + ^{0}_{+1}e$$

$$^{0}_{+1}e + ^{0}_{-1}e \rightarrow 2\gamma$$

These pairs of photons can be detected by a gamma camera (Figure 25.120) and the raw data processed and displayed in a computer producing a two- or three-dimensional image of the brain or other part of the body. The intensity of the gamma radiation detected is proportional to the concentration of FDG, which is dependent on the metabolic activity of cellular tissues. Any unusual variation in such activity may suggest a pathological problem, such as cancer, brain degeneration or damage or heart problems.



■ **Figure 25.121** Gamma therapy



■ **Figure 25.122** Absorption of protons (with high kinetic energy) by cells in body tissue

Ionizing radiation therapy

Ionizing radiation in a narrow beam can be directed at a tumour from different directions, either by moving the patient or by moving the radioactive source. This movement is necessary to ensure healthy tissue in the path of the beam is exposed much less than the target tissue. X-rays are used for tumours near the surface, but gamma radiation from cobalt-60 was traditionally used for deeper tumours (Figure 25.121). The radioisotope has a half-life of five years and is enclosed in a thick lead container in which the source is rotated to the inner end of the exit channel when it is to be used. External gamma therapy from cobalt-60 sources is increasingly used only for superficial skin conditions and is being replaced by linear particle accelerator (LINAC) therapy.

Proton beam therapy or proton therapy is a recent development of nuclear medicine where protons are used to irradiate cancerous tissue. This is a very expensive technique that uses a linear accelerator to impart high kinetic energies to protons in a narrow beam which is then directed at a tumour. The protons damage the DNA of cells, ultimately causing them to die or be unable to divide. Cancer cells are particularly vulnerable to DNA damage because of their high rate of cell division.

In contrast to other types of nuclear or ionizing radiation the absorption of protons reaches a maximum within a narrow range, deep inside the patient's body. This phenomenon is known as **Bragg's peak** effect (Figure 25.122) and allows the proton beam to be focused on the tumour with minimal damage to surrounding healthy tissue. The Bragg peak effect occurs because all the protons in the beam have the same speed and hence the same kinetic energy.

To treat tumours at greater depths, the proton accelerator must produce a beam of protons with higher kinetic energy. Tumours closer to the surface of the body are treated with protons with lower kinetic energy. In most treatments, protons of different kinetic energies with Bragg peaks at different depths are applied to treat the entire tumour.

Radiotherapy

Radiotherapy, also called radiation therapy, is the treatment of cancer and other diseases with ionizing radiation. Ionizing radiation injures or destroys cells in the area being treated by damaging their DNA, making it impossible for these cells to continue to grow and divide.

Although ionizing radiation damages both cancer cells and normal cells, the latter are able to repair themselves and function normally. Radiotherapy may be used to treat localized solid tumours, such as cancers of the skin, tongue, larynx, brain, breast or uterine cervix. It can also be used to treat leukemia and lymphoma (cancers of the blood-forming cells and lymphatic system, respectively).

One type of radiation therapy commonly used involves X-rays. Depending on the amount of energy they possess, X-rays can be used to destroy cancer cells on the surface or deeper in the body. The higher the energy of the X-rays, the deeper the X-rays can go into the target tissue. Linear accelerators are machines that produce X-rays of increasingly greater energy. The use of machines to focus radiation such as X-rays on a cancer site is called external beam radiotherapy.

Gamma rays are another form of electromagnetic radiation used in radiotherapy. Gamma rays are produced spontaneously as certain elements (such as radium, uranium and cobalt-60) release ionizing radiation as they decay. Each element decays at a specific rate and releases energy in the form of gamma rays and other particles. X-rays and gamma rays have the same effect on cancer cells.

Another investigational approach is particle beam radiation therapy. This type of therapy differs from photon radiotherapy in that it involves the use of fast-moving subatomic particles to treat localized cancers. A very sophisticated machine is needed to produce and accelerate the particles required for this procedure. Some particles (neutrons, protons and heavy ions) deposit more energy along the path they take through tissue than do X-rays or gamma rays, thus causing more damage to the cells they interact with. This type of radiation is often referred to as high linear energy transfer (LET) radiation.

Radioactive sources can be packed into metal tubes, metal needles and plaques. These kinds of sources can be used in three different ways. Surface applicators are arrays of sources placed near or in contact with the patient's skin. They are used in treating cancer of the eye. Interstitial implants use radioactive wire, pellets or arrays of needle-like radioactive sources implanted directly into the target organ. They can be used for breast cancer. This technique is known as internal radiotherapy or brachytherapy. With intercavitary methods, the sources are sealed into special containers and inserted into the body cavity. They can be used for the treatment of cancer of the cervix.

Radiation therapy may be used alone or in combination with chemotherapy or surgery. Like all forms of cancer treatment, radiation therapy can have side effects. Possible side effects of treatment with radiation include temporary or permanent loss of hair in the area being treated, skin irritation, temporary change in skin colour in the treated area, and tiredness. Other side effects are largely dependent on the area of the body that is being treated.

Techniques in nuclear medicine

Targeted alpha therapy

Cancer is a major health problem for older people in developed countries. Unfortunately, radiotherapy and chemotherapy, even though often effective, have low success rates against cancers that have formed metastases and have spread around the body.

Targeted alpha therapy is a new and promising approach to cancer treatment. It combines alpha particle emitting radioisotopes such as lead-212 (Figure 25.123), with tumour-selective carrier molecules, such as monoclonal antibodies or peptides (small proteins). Often a chelating or complexing agent is used. In addition to lead-212, bismuth-213 and actinium-225 can be used for targeted alpha therapy.

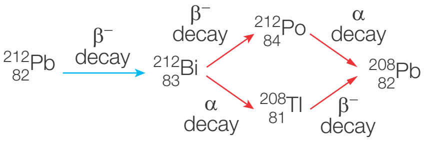


Figure 25.123 The decay chain of lead-212

A monoclonal antibody is obtained from an immune system B cell (grown in culture) that produces a single antibody. Monoclonal antibodies with specific properties can be produced in large quantities and may be used to develop antibody drugs,

Peptides and monoclonal antibodies have the ability to selectively target tumour cells even if they are spread throughout the body. They recognize the targeted cancer cells through antigens (proteins specific to cancer cells) that are located on the outer surface of the cell membrane and can bind selectively to these cells (compare with the lock and key model of enzyme activity). In targeted alpha therapy (Figure 25.124) these carrier molecules transport the radioisotopes to the cancer cells to kill them. This is called the 'magic bullet' approach.

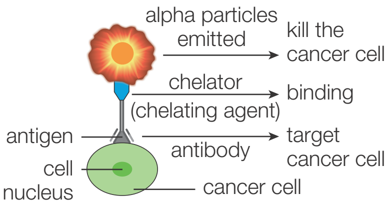


Figure 25.124 Targeted alpha therapy

Radioisotopes that emit alpha particles are very suitable for selectively destroying cancer cells. Alpha particles have a high kinetic energy (due to their large mass) and a very short path length in human tissue corresponding to less than the diameter of ten cells. Hence, alpha emitters allow the specific targeting and killing of individual cancer cells, while avoiding damaging surrounding healthy tissue.

The use of alpha emitters is most useful in treating cancers where individual cancer cells or small clusters of cancer cells need to be targeted, for example leukemia, treatment of micrometastases (small groups of cancer cells that have spread from the original tumour) or surgical removal of a large tumour.

Radium-223 is another widely used radioisotope in alpha targeted therapy. It is used to treat prostate cancers that have spread to the bones. Radium is very similar to calcium. Like calcium it is taken up by active bone cells. This makes it a good way of targeting bone cancer cells. Cancer cells are more active than normal bone cells and so are more likely to absorb and assimilate the radium-223.

Boron neutron capture therapy

High intensity boron neutron beams are used in boron neutron capture therapy, which uses the ability of stable boron-10 nuclei to absorb neutrons. After capturing a low-energy (thermal) neutron, the nucleus of the boron-10 atom is transmuted into an unstable boron-11 atom, which immediately undergoes alpha decay:

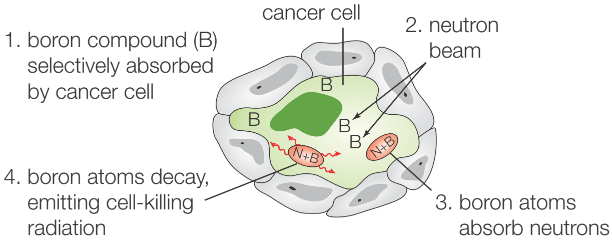
$$^{10}_{5}B + ^{1}_{0}n \rightarrow [^{11}_{5}B] \rightarrow ^{7}_{3}Li + ^{4}_{2}He$$

The recoiling lithium-7 ions and alpha particles have high kinetic energies and cause extensive cellular damage, but within a very limited range, 0.005–0.01 mm, which is approximately the diameter of a cell. Hence tumours can be destroyed by boron neutron capture therapy (Figure 25.125) provided they accumulate sufficient boron-10. The boron-10 isotope is administered to the patient by intravenous injection of an organo-boron compound.

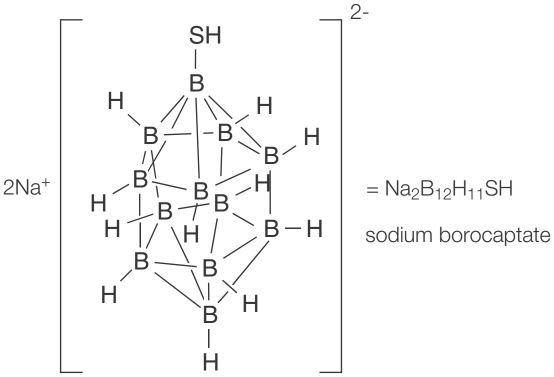
A wide variety of boron delivery agents have been synthesized, but only two of these currently are being used in clinical trials. The first, which has been used primarily in Japan, is a polyhedral borane anion, sodium borocaptate BSH ($Na_2B_{12}H_{11}SH$) (Figure 25.126), and the second is a dihydroxyboryl derivative of the amino acid phenylalanine, referred to as boronophenylalanine or BPA (Figure 25.127). The latter has been used in clinical trials in the USA, Europe, Japan, Argentina and Taiwan. Following administration of either BPA or BSH by intravenous infusion, the tumour site is irradiated.

Boron neutron capture therapy can be used to treat cancers that are normally treated with radiotherapy, such as lymphomas and skin cancers, as well as cancers of the brain, breast, lung, head and neck, bone, prostate, pancreas and cervix. In addition, when surgical removal of a tumour is planned, boron neutron capture therapy may also be used to help reduce the size of the tumour and to reduce the associated normal tissue loss. Boron neutron capture therapy is less demanding for the patient than conventional radiotherapy as it can be given several times over a period of 2–4 days; in contrast conventional radiotherapy needs to be given up to 30 times over a period of 6 weeks.

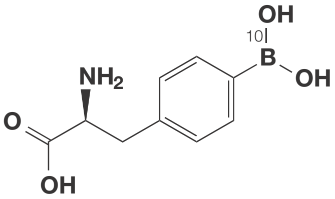
Unlike some forms of ionizing radiation, such as X-rays, alpha particles do not require oxygen to enhance their biological effectiveness. A rapidly expanding tumour frequently outgrows its blood supply, so that some regions receive less oxygen than normal tissues do. As a result of this oxygen depletion, the tumour can become more resistant to the effects of conventional photon or electron (beta particle) radiation therapy. Tumour sensitivity to alpha particles is retained, however, even when the tumour has limited oxygen supply.


cancer cell
1. boron compound (B) selectively absorbed by cancer cell
2. neutron beam
3. boron atoms absorb neutrons
4. boron atoms decay, emitting cell-killing radiation
B
N+B

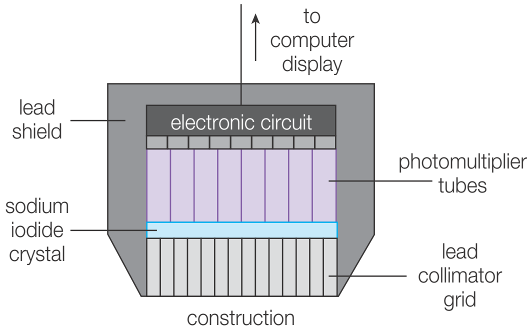
■ **Figure 25.125** Boron neutron capture therapy



■ **Figure 25.126** Structure of sodium borocaptate



■ **Figure 25.127** Structure of borophenylalanine



■ **Figure 25.128** Gamma camera

Gamma camera

A gamma camera (Figure 25.128) is used to detect gamma radiation. It consists of a series of photomultiplier tubes. If a gamma ray (emitted from a radioactive substance in a patient's body) interacts with an ionic crystal such as sodium iodide, the gamma photon is converted into a photon of visible light. The crystal is said to scintillate, and because of this effect the gamma camera scan is termed a scintigram. The number of photons of visible light emitted depends upon the energy of the gamma rays. Scintillation was used by Rutherford in his discovery of the nucleus (Chapter 2).

■ **Figure 25.129** Gamma knife

The photons then strike a photocathode, a material that ejects electrons when bombarded by photons. This is termed the photoelectric effect. The number of electrons is then multiplied using a sequence of charged plates called dynodes.

The array of photomultiplier tubes is connected to a recoding and display system. An image is formed which matches the distribution of gamma rays emitted from the patient with the radioactive gamma emitter.

Gamma knife

The gamma knife (Figure 25.129) is used for brain surgery but is non-invasive (it is performed without cutting the skin or muscles and the skull does not need to be opened), though a frame needs to be attached to the skull using four screws (fitted under local anaesthetic).

The gamma knife has a large number of cobalt-60 sources, which emit gamma rays and have a half-life of 5.26 years. The sources are positioned in a hemisphere inside the unit. The patient's head, held in the frame, is held inside a helmet with a large number of holes to precisely target the radiation. When treatment starts, the patient's head is moved inside the unit.

Magnetic resonance imaging (MRI)

MRI depends on the magnetic properties of some nuclei, most notably the protons in the hydrogen atoms of water, and was developed from its parent technique NMR spectroscopy, which is widely used in chemical analysis. It relies on the use of magnetic field gradients to provide spatial information, and these field gradients can be manipulated electronically, giving the technique great versatility.

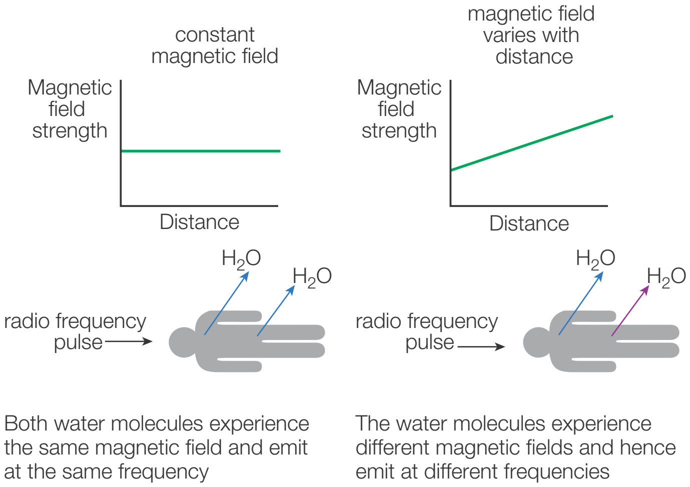
Modern MRI scanners use superconducting magnets to create powerful magnetic fields (up to a million times stronger than the relatively weak magnetic field of the Earth). The MRI scanner also produces electromagnetic radiation of low energy and frequency and long wavelength (radio waves). The photons have insufficient energy to break bonds in molecules.

When a patient is placed inside the magnetic field, hydrogen atoms (protons) of water molecules in the body absorb radio waves to generate an NMR signal, which decays by a process called relaxation. The relaxation time varies with the type of tissue. The MRI scanner detects the NMR signal produced by the hydrogen nuclei as an oscillating voltage induced in a coil placed around the patient.

The technique relies on using a magnetic field gradient (Figure 25.130). The two hydrogen atoms (protons) in two water molecules at different places in the human body would resonate at the same frequency if placed in the same applied magnetic field.

An MRI scanner therefore places the patient in a magnetic field that varies linearly between the poles of the magnet. Two identical hydrogen atoms now resonate at different frequencies depending on their position in the magnetic field – that is depending on their position in the body. A hydrogen atom in a region where the magnetic field is stronger resonates at a higher radiowave frequency than in a region where the magnetic field is weaker.

The patient is irradiated with a pulse of a range of radio frequencies and the absorption of the radiation measured. This allows the measurement of the concentration of ^{1}H nuclei in water molecules at a corresponding position. This is repeated with a succession of a range of pulses at different frequency ranges to measure the number of ^{1}H nuclei at the different positions in the scanner. In this way, a 'map' of the ^{1}H nuclei throughout the body can be built up. These maps can be combined by the computer software to give a three-dimensional or tomographic image with a resolution better than 1 mm.


constant magnetic field
Magnetic field strength
Distance
H$_2$O
H$_2$O
radio frequency pulse
Both water molecules experience the same magnetic field and emit at the same frequency
magnetic field varies with distance
Magnetic field strength
Distance
H$_2$O
H$_2$O
radio frequency pulse
The water molecules experience different magnetic fields and hence emit at different frequencies

■ **Figure 25.130** MRI scanners rely on a magnetic field gradient to give different frequency signals from ^{1}H nuclei in different positions in the body

MRI is a non-invasive technique and produces more spatially detailed images of the human body than nuclear medicine or PET scanning techniques. The protons in water, lipids, carbohydrates and proteins have different chemical environments in both healthy and diseased tissue (which often has a lower concentration of water), which can be easily distinguished by ^1H NMR chemical shifts.

Because the concentrations of water and organic compounds in various tissues are different, MRI provides highly detailed images of soft tissues such as the brain, heart, muscles and body fluids. There are many ways of increasing the contrast in MRI images, for example by making use of the different relaxation times in different kinds of tissues. The technique does not use ionizing or nuclear radiation so can be used repeatedly without increasing the risk of cancer to the patient. The only drawbacks of MRI are the high cost of the equipment and the interaction of magnetic fields with metal body implants such as prosthetics (for example, hip replacements) and heart pacemakers.

Multinuclear MRI

As well as proton or ^1H NMR, modern MRI instruments can detect other nuclei, including carbon-13, sodium-23, nitrogen-14 and phosphorus-31. Each nuclide or element has a characteristic resonance frequency, so a series of NMR spectra can be recorded examining each nuclide in turn.

Multinuclear MRI studies are particularly useful for the imaging of organs that have insufficient contrast in ^1H NMR. For example, images of the air spaces (alveoli) in the lungs can be obtained by ^3He or ^{129}Xe NMR where a noble gas (helium or xenon respectively) is inhaled by the patient during the MRI scan. Another nucleus, naturally occurring ^{31}P, can provide important information about the metabolism of phosphorylated compounds such as ATP.

ToK Link

MRI and NMR

There is no reference to the term 'nuclear' in MRI. Are names simply labels or do they influence our other ways of knowing? How does public perception influence scientific progress and implementation?

Magnetic resonance imaging (MRI) is a form of nuclear resonance spectroscopy applied to the living human body. The term 'nuclear' refers to the fact that the technique is studying the spin of hydrogen nuclei (protons). However, the term 'nuclear' to the general public means nuclear power stations and nuclear bombs. The use of the term 'nuclear' is therefore linked with ionizing or nuclear radiation. Hence the term MRI is used rather than NMR. If MRI was given the label NMR, then many patients who required an MRI scan may not have the procedure due to the negative connotations associated with the word 'nuclear'. MRI does not use ionizing radiation which, even in small doses, will create a radiation hazard. Radio waves and a powerful magnetic field are used and both are believed to be safe. Names are simply 'labels' and interchangeable, but they influence emotion as a way of knowing.

One way the public can influence scientific progress is via emotion or ethics. Two current examples of ethical principles and perhaps emotion hindering scientific progress are stem cell research and the development of genetically modified organisms (GMO).

25.9 Drug detection and analysis (AHL) *– a variety of analytical techniques is used for detection, identification, isolation and analysis of medicines and drugs*

■ Analytical techniques

A variety of analytical techniques is used for the separation, detection and analysis of pharmaceutical drugs. These techniques include chromatographic techniques, electrophoresis (including a form known as capillary electrophoresis), infrared spectroscopy, ultraviolet–visible spectroscopy, nuclear magnetic resonance (especially high resolution), mass spectrometry, X-ray crystallography, titrations and electrochemical methods.

Illegal drugs are often detected by mass spectrometry, gas chromatography and combined techniques, for example gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS). Urine is easier to test than blood, largely because it has higher drug concentrations and fewer proteins to complicate extraction.

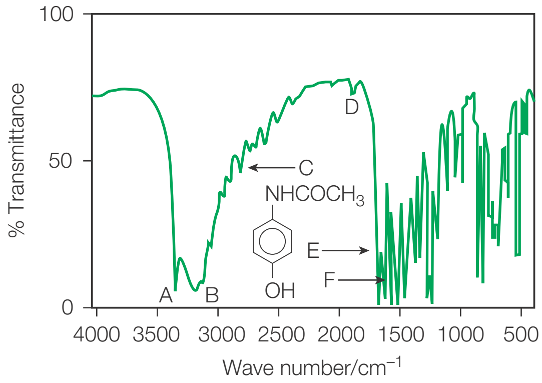


Figure 25.131 Infrared spectrum of paracetamol

Spectroscopic identification of drugs

Many pharmaceutical drugs, such as aspirin, paracetamol and omeprazole, are relatively simple and stable organic molecules containing various functional groups. The presence or absence of these functional groups can be determined by mass spectrometry and spectroscopic techniques. Information about the detailed structure of a drug molecule is obtained from mass spectrometry and high-resolution NMR.

Figure 25.131 shows the infrared spectrum of paracetamol and Table 25.10 summarizes an interpretation of some of the absorption bands.

Table 25.10 Interpretation of selected peaks in the IR spectrum of paracetamol

Wave number/cm^{-1}	Assignment	Comment
A: 3360	N–H amide stretch	This band can be seen quite clearly, although it is on top of the broad O–H stretch
B: 3000–3500	Phenol–OH stretch	Very broad due to strong hydrogen bonding and hence obscures (hides) other bonds in this region
C: approximately 3000	C–H stretching	Not clear due to underlying –OH absorption
D:1840–1940	Benzene ring	Molecular vibrations
E: 1650	>C=O amide stretch	
F: 1608	Aromatic C=C stretch	

Figure 25.132 shows the high-resolution ^{1}H NMR spectrum of paracetamol and Table 25.11 summarizes an interpretation of some of the peak assignments.

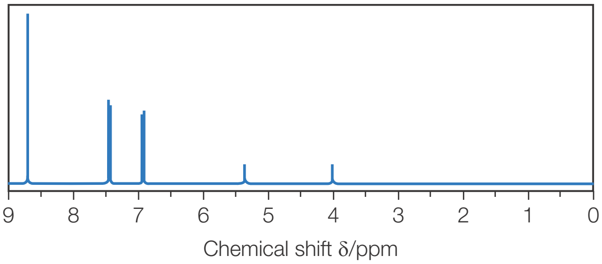


Figure 25.132 High-resolution ^{1}HNMR spectrum of paracetamol

Table 25.11 Interpretation of peaks in the ^{1}H NMR spectrum of paracetamol

Peak number (from right to left)	Comment
1	Signal due to 3 × ^{1}H on the *sp*^{3} –COCHH_3 group which contains an electronegative oxygen atom (integration = 3). There are no ^{1}H on the neighbouring atoms so the signal is not split by coupling
2	Signal due to 2 × ^{1}H on the aromatic, *sp*^{2} CH groups (integration = 2). The signal is split into a 1 : 1 doublet by coupling to the ^{1}H on the neighbouring CH group
3	Signal due to 2 × ^{1}H on the aromatic, *sp*^{2} CH groups (integration = 2). The signal is split into a 1 : 1 doublet by coupling to the ^{1}H on the neighbouring CH group
4	Signal due to ^{1}H on –OH group (integration = 1). This ^{1}H rapidly exchanges with ^{1}H atoms in the solvent, causing the peak to be broad
5	Signal due to ^{1}H on –NH group (integration = 1). This ^{1}H rapidly exchanges with ^{1}H atoms in the solvent, causing the peak to be broad

Figure 25.133 shows the mass spectrum of paracetamol following its analysis by mass spectrometry.

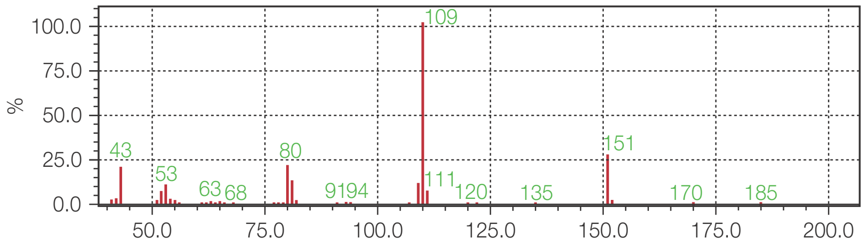


Figure 25.133 Mass spectrum of paracetamol

The peak at 151 is the molecular ion and the peak at 43 is due to the removal of a C_2H_3O group which exists as a resonance stabilized ion. Not all of the peaks can be accounted for by fragmentations. Some of the peaks are due to the presence of ions formed by rearrangements (Figure 25.134).

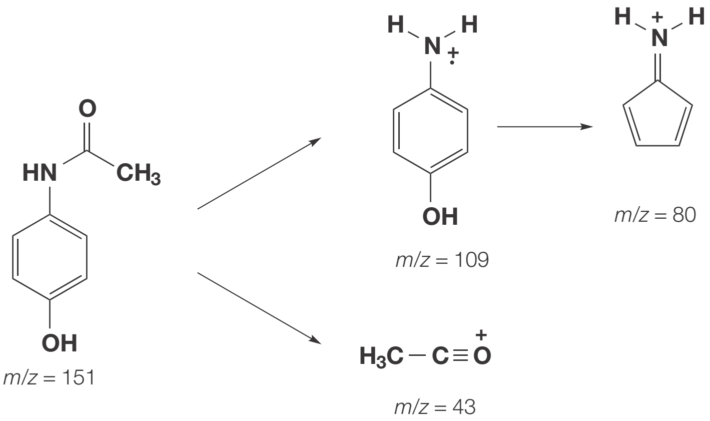


Figure 25.134 Mass spectral fragmentation of paracetamol

Identifying unknown organic compounds

A common task for medicinal chemists is the identification of a pharmaceutical or illegal drug from various analytical and chemical data. If information about the drug, such as molar mass, elemental composition from elemental analysis and retention time, or retention factor from chromatography has been determined, the molecule may identified by comparison with a library of known compounds. Computers attached to analytical instruments, such as GC-MS, may have a library of over 220 000 spectra that can be used to identify an unknown chemical in the sample mixture. The library compares the mass spectrum from a sample component and compares it to mass spectra in the library. It reports a list of likely identifications along with the statistical probability of the match.

Confirmation of the identity of the drug can be obtained from its mass spectrum, its infrared spectrum and high-resolution ^1H NMR. All of this information together with the preliminary data can be used to establish the identity of the molecule. It is helpful for the sample of the drug to be free of impurities before being subjected to a detailed chemical and analytical analysis.

Extraction and purification of organic compounds

When the last reaction for a multi-step drug synthesis has finished, the desired product is described as being in a crude state. It is often contaminated by unused reactants, solvents or a reaction by-product. Hence there is a requirement to separate and purify the product from these impurities.

Some of the most commonly used separation techniques are listed in Table 25.12. The choice of technique is governed by the physical state of the product and its impurities.

Table 25.12 Some commonly used physical separation techniques

Technique	Used to separate
Distillation	A liquid from a dissolved non-volatile solid or liquids with very different boiling points
Fractional distillation	Liquids with similar boiling points
Steam distillation	Two liquids with very different boiling points
Solvent extraction	A solid or liquid from its solution
Recrystallization	A solid from other solid impurities
Suction filtration	An insoluble solid from a liquid
Chromatography	Various types of mixtures

16 State which technique(s) could be used to separate the following mixtures:

a Propan-1-ol (bp 97 °C) and 1-iodopropane (bp 103 °C).

b Phenylamine (amino benzene) (bp 184 °C) and tin(II) chloride (s).

c Pentane (bp 36 °C), hexane (bp 69 °C) and heptane (bp 99 °C).

d Benzoic acid (a solid) contaminated by charcoal powder.

Solvent extraction

Many natural and synthetic products used in pharmaceuticals have to be isolated from their mixtures with other compounds of differing solubilities. This is commonly done by liquid–liquid extraction, a physical process that involves the distribution or partitioning of a solute between two immiscible liquids.

In a typical experiment a mixture of organic compounds is shaken with water and an organic solvent (such as ethoxyethane or trichloromethane (chloroform)) and the resulting emulsion is allowed to settle. Since water and the organic solvents are almost immiscible (do not mix) they form two separate layers.

Polar compounds tend to be more soluble in polar solvents (such as water) – the 'like dissolves like' principle – and therefore remain in the aqueous layer, whereas non-polar substances dissolve in the organic layer. Each layer can be run into a different beaker using a separation funnel (Figure 25.135).

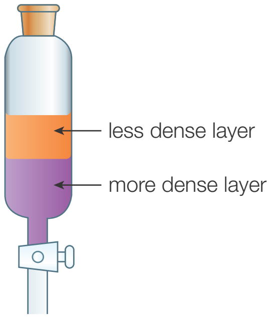


Figure 25.135 Use of a separating funnel

The organic solvent and water can be evaporated from the separated layers, leaving the components of the original mixture. For a complex mixture the separation process can be repeated many times using the same or different solvents. In the case of the anti-cancer drug Taxol (see section 25.7), the isolation of the target compound required several hundred extractions and took over two years to complete. This process can now be automated.

The partition of a solute between two immiscible liquids can be described as a heterogeneous equilibrium (Chapters 7 and 17) between different states of the same compound. For example, when molecular bromine, Br_2, is partitioned between water ('aq') and an organic solvent ('organic') the following equilibrium takes place (at constant temperature):

$$Br_2(aq) \rightarrow Br_2(organic)$$

The equilibrium constant of this equilibrium is known as the **partition coefficient, P_c**:

$$P_c = \frac{[Br_2(organic)]}{[Br_2(aq)]}$$

The partition coefficient depends on the chemical nature of the participating chemical species and the temperature of the reaction mixture. Non-polar molecules and molecules with low polarity will have high values of partition coefficients (due to their low solubility in water), but polar compounds will have low values of partition coefficients (due to their high solubility in water).

Solvent extraction is often used to separate an organic product which is contaminated by an aqueous solution. Such extracts will always contain a small amount of water and this is removed by 'drying' it with an anhydrous ionic salt, such as magnesium sulfate. Usually, drying takes a few hours, after which time the ionic salt, which is now partially hydrated, is filtered off.

Finally, volatile solvent must be removed from the dried extracts. This is easily achieved by distillation. The resulting product is then further purified by fractional distillation (for liquids) or recrystallization (for solids).

Worked examples

A solution contains 1.00 g of iodine dissolved in 20.00 cm^3 of potassium iodide solution. The solution is shaken with 20.00 cm^3 of tetrachloromethane (CCl_4), an organic solvent. Determine how much iodine will be transferred into the tetrachloromethane. The partition coefficient between tetrachloromethane and water at 25 °C is 85.00.

P_c = concentration of iodine in tetrachloromethane/concentration of iodine in water = 85.00

Assume that x g of iodine will enter the tetrachloromethane. This will leave $(1 - x)$ g of iodine in the water. Hence:

$x\,\text{g}/20.00\,\text{cm}^3/(1.00 - x)\,\text{g}/20.00\,\text{cm}^3 = 85.00$

So, $x = 85.00(1.00 - x)$ and $x = 85.00/86.00 = 0.988$

Hence 0.988 g of iodine will be present in the tetrachloromethane and 0.012 g left in the water.

Use the value of the partition coefficient to determine how much iodine would have been removed from the solution if it were shaken with (i) 40.00 cm^3 of tetrachloromethane; (ii) 10.00 cm^3 of tetrachloromethane and then another 10.00 cm^3 of tetrachloromethane. Compare the effectiveness of the two approaches with using 20.00 cm^3 of the solvent.

$x\,\text{g}/40.00\,\text{cm}^3/(1 - x)\,\text{g}/20.00\,\text{cm}^3 = 85.00$; so $x = 170(1 - x)$ and $x = 0.994$

$x\,\text{g}/10.00\,\text{cm}^3/(1 - x)\,\text{g}/20.00\,\text{cm}^3 = 85.00$; $x = 0.977$. This leaves 0.023 g of iodine in 20.00 cm^3 of water. With the second 10.00 cm^3 of solvent, if x g of iodine enters the solvent, $(0.023 - x)$ g will be left in the water. Therefore:

$x\,\text{g}/10.00\,\text{cm}^3/(0.023 - x)\,\text{g}/20.00\,\text{cm}^3 = 85.00$; $2x = 85(0.023 - x)$ and $x = 0.0225$.

By using two separate 10.00 cm^3 portions, a total of 0.977 g + 0.0225 g = 0.995 g has been separated. This is more efficient than using one 20.00 cm^3 portion, which removed only 0.988 g. It is also more efficient than using one portion of 40.00 cm^3.

Polarity of drug molecules

The pharmacological properties of a drug depend, in part, on its polarity. Polar (hydrophilic) drug molecules tend to stay in the plasma of the blood, which is an aqueous solution. Non-polar (hydrophobic or lipophilic) drug molecules accumulate in lipid (adipose) tissues.

The polarity of a drug influences its solubility, absorption, distribution, metabolism and excretion. Highly polar drugs do not easily cross the cell membranes of the small intestine. They can be injected but they cannot be used against intracellular targets since they will not cross cell membranes. If non-polar drugs are administered orally they are likely to be absorbed in fat globules in the intestines and will be poorly absorbed. If they are injected, they are poorly soluble in blood.

In medicinal chemistry the hydrophobicity of a molecule is normally measured by its $\log_{10} P_c$ value, where P_c is known as the partition coefficient. P_c can be measured by measuring the relative solubility of a compound in an octan-1-ol–water mixture, where

$$P_c = \frac{\text{concentration of compound in octan-1-ol}}{\text{concentration of compound in water}}$$

The more hydrophobic the compound is, the greater the proportion of it will dissolve in the organic layer, and the higher the value of P_c or $\log_{10} P_c$. It is also possible to calculate $\log_{10} P_c$ values for a given molecular structure using suitable software programs.

Many drugs exist as an equilibrium between an ionized and un-ionized form, but $\log_{10} P_c$ measures the relative distribution of the un-ionized species between water and octan-1-ol. The polarity of drugs can be changed easily by chemically modifying the functional groups. For example, the $\log_{10} P_c$ values for morphine and diamorphine (heroin) are 0.9 and 1.6, which explains the greater solubility of diamorphine to cross the blood–brain barrier and produce a stronger analgesic effect in the brain.

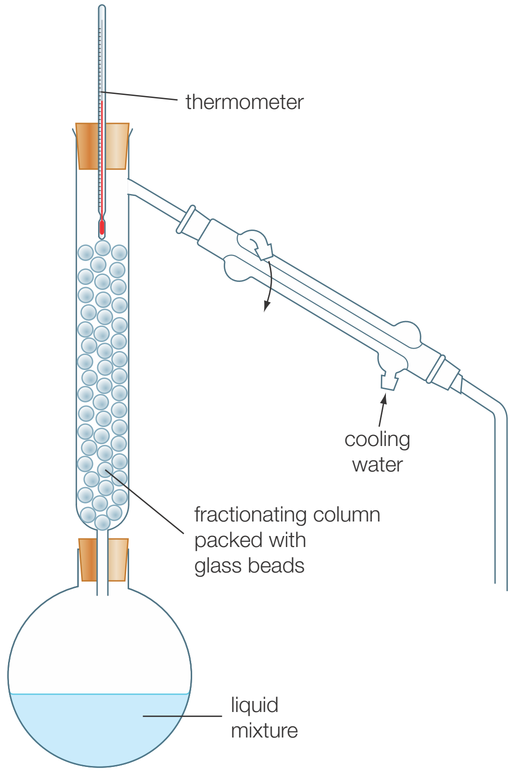


Figure 25.136 Fractional distillation

With most drugs, activity in the body (*in vivo* activity) tends to increase as the $\log_{10}P_c$ value increases. In other words, activity increases with increasing hydrophobicity (lipophilicity). This is usually an indication that increasing hydrophobicity allows easier movement of the drug molecule through cell membranes in order to reach target sites.

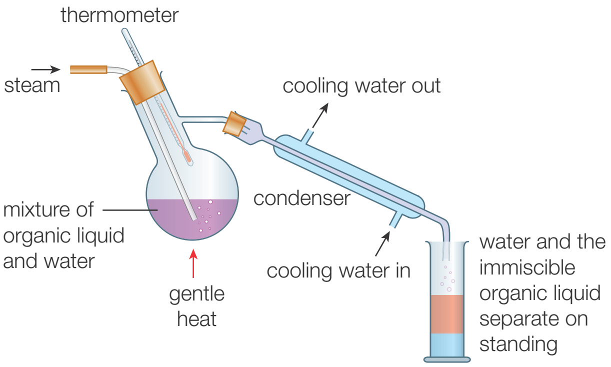
Fractional distillation

Fractional distillation (Figure 25.136) is used to separate liquids that have similar boiling points. The fractionating column is a long tube packed with small glass beads which provide a large surface area for condensation. Vapour rising from the flask warms up the bottom of the column and condenses back into liquid. As it falls, it meets fresh hot vapour moving upwards. This process is repeated continually up the column and each time the vapour becomes richer in the more volatile component – that is, the one with the lower boiling point. If the column is long enough, the vapour which escapes at the top is the pure, more volatile component, and this may be confirmed by checking its boiling point on the thermometer.

Steam distillation

An organic liquid which has a high boiling point and very low solubility in water, for example eucalyptus oil, may be removed from a mixture by steam distillation (Figure 25.137). As steam is passed through the hot mixture, a mixture of the two components distils out of the flask and the condensate is collected as shown in Figure 25.137. On standing, the condensate forms two layers; one is the organic product, the other is aqueous. These are separated and any product in the aqueous layer is removed by solvent extraction.

Figure 25.137 Steam distillation


thermometer
steam
cooling water out
mixture of organic liquid and water
condenser
gentle heat
cooling water in
water and the immiscible organic liquid separate on standing

Liquid–vapour equilibria and Raoult's law

Any liquid left in a close container at a fixed temperature will establish an equilibrium with its vapour:

$$\text{Liquid} \rightleftharpoons \text{vapour}$$

The pressure exerted at that temperature is known as the saturated vapour pressure (svp).

In a system consisting of a mixture of two miscible liquids at equilibrium, Raoult's law states that each component will exert a vapour pressure proportional to its mole fraction. Raoult's law can be expressed mathematically: $P(\text{A}) = P^{\circ}(\text{A}) - x(\text{A})$, where $P(\text{A})$ is the vapour pressure of A over the mixture of the two liquids (at a given temperature), $P^{\circ}(\text{A})$ is the vapour pressure over a pure sample of A at the same temperature and $x(\text{A})$ is the mole fraction of A, which is the ratio of the amount of A to the sum of all the amounts of the component liquids in the mixture.

Worked example

At a given temperature, you have a mixture of benzene (vapour pressure of pure benzene = 745 mmHg) and methylbenzene (vapour pressure of pure methylbenzene = 290 mmHg). The mole fraction of benzene in the solution is 0.590. Assuming ideal behaviour, calculate the mole fraction of methylbenzene in the vapour above the solution.

Since it is a binary mixture, the mole fraction of benzene in the solution is

$x_{total} = 1 - x_{benzene}$

The total vapour pressure, P_{total}, can be calculated with Raoult's law (ideal behaviour):

$P_{total} = x_{benzene} \times P^{o}_{benzene} + x_{total}\ P^{o}_{total}$

$= 0.590 \times 745\,mmHg + (1 - 0.590) \times 290\,mmHg = 558.45\,mm\,Hg$

The partial vapour pressure of methylbenzene over the solution follows also from Raoult's law:

$P_{total} = x_{total} \times P^{o}_{total} = (1 - 0.590) \times 290\,mmHg = 118.90\,mm\,Hg$

From Dalton's law of partial pressures follows the mole fraction of methylbenzene in the vapour, y_{total}:

$Y_{total} = \frac{P_{total}}{P_{total}} = \frac{118.90}{558.45} = 0.2129 = 0.213$

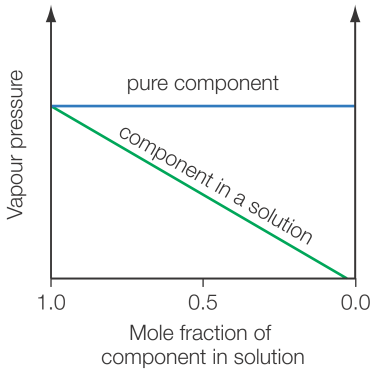


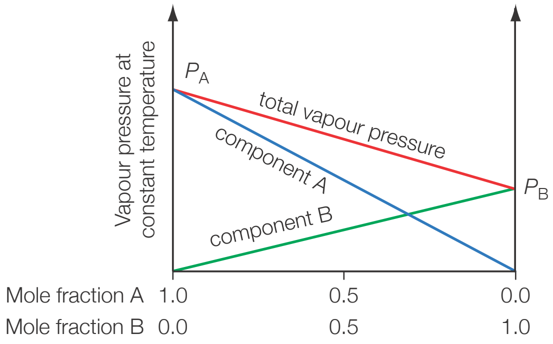
Figure 25.138 The relationship between vapour pressure and mole fraction for a component in an ideal mixture of liquids

This means that a graph of mole fraction against vapour pressure for each of the liquids will be of the type shown in Figure 25.138 – that is, a straight line passing through the origin.

The total vapour pressure of the liquid mixture is found by adding the vapour pressure reached by the partial pressures of the two components (two liquids) and it will vary with composition as shown in Figure 25.139.

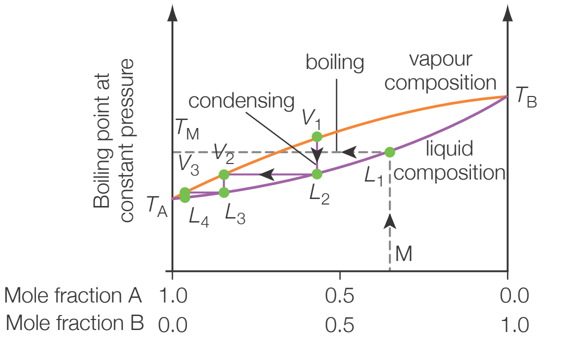
The boiling point of a liquid is the temperature at which its vapour pressure reached the external pressure. Since the component (liquid) with the lower saturated vapour pressure will have a higher boiling point, a boiling-point composition diagram of the type shown in Figure 25.140 can be drawn.

Note that this graph has two curves since the liquid and vapour in equilibrium at any temperature will have different compositions. The vapour will always contain a higher proportion of the more volatile component – that is, the substance with the lower boiling point. In Figure 25.140 a liquid mixture of composition M when heated will boil at temperature T_M. The vapour in equilibrium with liquid of composition L_1 will have composition V_1 that is much



where P_A is the vapour pressure of pure A
P_B is the vapour pressure of pure B

Figure 25.139 The relationship between mole fractions and partial and total vapour pressures for an ideal mixture of two liquids



where T_A is the boiling point of pure A
T_B is the boiling point of pure B

Figure 25.140 The boiling-point composition diagram for the ideal liquid mixture represented in Figure 25.139

richer in the lower boiling component A. This difference between the equilibrium liquid and vapour compositions allows a mixture to be separated by distillation. If the vapour V_1 is cooled it will condense to a liquid of the same composition, L_2. Boiling L_2 will produce a vapour V_2 which is even richer in A. Meanwhile the proportion of A in the remaining liquid will decrease and its composition will follow the liquid curve up until it is almost pure B. Hence any mixture of liquid which obeys Raoult's law may be separated by repeated distillation (i.e. boiling and condensing). This is known as fractional distillation and is carried out using a fractionating column designed to allow many such steps to take place over its length. Each distillation step is known as a theoretical plate. If the column is efficient enough – that is, contains enough theoretical plates – the thermometer at the top will indicate the boiling point of the pure, more volatile (lower boiling point) component.

Liquid mixtures which obey Raoult's law exactly are rare, but often the deviations are sufficiently small that they may be sufficiently separated by fractional distillation. Raoult's law is obeyed by a pair of ideal liquids. When ideal liquids are mixed there is no enthalpy change and no change in total volume.

However, when ethanol and water are mixed, there is a slight increase in the total volume. This occurs because the very strong intermolecular hydrogen bonding in pure water is disrupted by ethanol molecules which can form weaker hydrogen bonds. This reduction in intermolecular attractions means that as well as an increase in volume, each component liquid vaporizes more easily and therefore contributes more than expected to the total vapour pressure. This is known as positive deviation from Raoult's law.

Worked example

A solution is prepared by mixing 5.81 g of propanone (molar mass of propanone = 58.1 g mol^{-1}) and 11.9 g trichloromethane (molar mass of trichloromethane = 119.4 g mol^{-1}). At 35 °C, this solution has a total vapour pressure of 260 mmHg. Deduce whether this is an ideal solution. The vapour pressures of pure propanone and pure trichloromethane at 35 °C are 345 and 293 mmHg, respectively.

$P_{\text{total}} = X_A\, P^o{}_A + X_B\, P^o{}_B$

$N_{\text{propanone}}$ = 5.81 g/(58.1 g mol^{1}) = 0.100 mol propanone

$N_{\text{trichloromethane}}$ = 11.9 g/(119 g mol^{1}) = 0.100 mol trichloromethane

$X_{\text{propanone}} = X_{\text{trichloromethane}} = 0.500$

Then

P_{total} = (0.500) × (345 mmHg) + (0.500) × (293 mmHg) = 319 mmHg

Since the observed pressure is 260 mmHg (lower than expected), the solution does not behave ideally.

The dipole–dipole interaction between propanone and trichloromethane molecules lowers their tendency to escape from solution.

Recrystallization

Recrystallization is used to separate an organic solid product from other solid impurities. In this technique, a solvent is chosen in which:

- the organic product readily dissolves when hot but hardly does so when cold;
- the impurities must either not dissolve or remain dissolved at low temperatures.

Often, the selection of this solvent is simply a case of 'trial and error'. A solid product can sometimes be purified by recrystallization from water. The impure solid is dissolved in the minimum quantity of hot water. Impurities which are insoluble in water can be removed by quickly filtering through a fluted filter paper held in a hot funnel.

As the filtrate cools, any soluble impurities remain solution while the pure product crystallizes out. The crystals are filtered off using a vacuum filtration apparatus (Figure 25.141). Apart from giving a rapid filtration, vacuum filtration has the added advantage that the flow of air past the crystals helps them to dry out. The final drying of an organic solid is carried out in either a desiccator at low pressure or an oven. The method chosen will depend on the properties of the solvent and the solid, for example the latter might decompose in an oven.

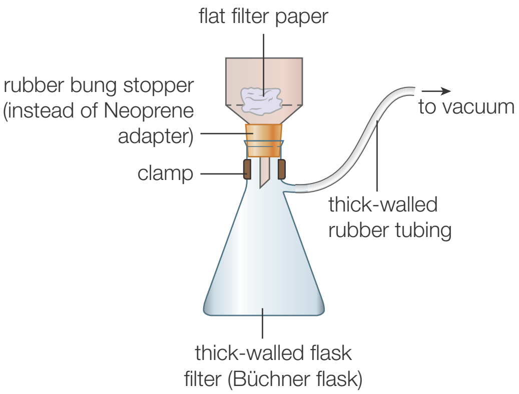

flat filter paper
rubber bung stopper (instead of Neoprene adapter)
clamp
to vacuum
thick-walled rubber tubing
thick-walled flask filter (Büchner flask)

Figure 25.141 Suction filtration apparatus

Drug detection in sports and forensic studies

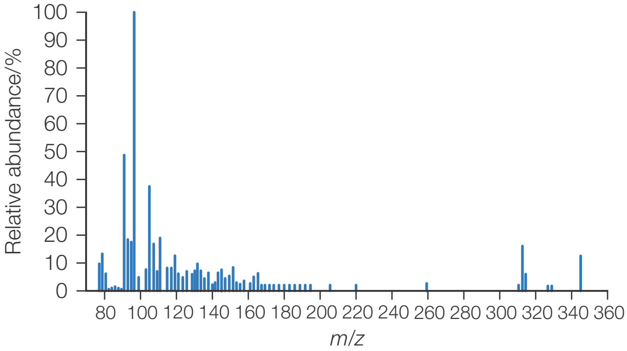
Anabolic steroids

The misuse of performance-enhancing substances in sports is a serious international problem. The most common of these substances, anabolic steroids, accelerate the synthesis of proteins and cellular growth, especially in the skeletal muscles and bone tissues. Anabolic steroids are drugs whose molecules are structurally related to the cyclic steroid ring system and have similar effects to testosterone in the body. They are used medically to induce puberty and to treat chronic wasting conditions such as cancer and AIDS.

Anabolic steroids are banned by most sports and athletics organizations including the International Olympic Committee. Athletes and sports players are regularly required to provide a urine and blood samples for laboratory analyses in which steroids and their metabolites can be detected by a combination of gas chromatography (GC) or high-performance liquid chromatography (HPLC) with mass spectrometry (MS) (Chapters 11 and 21).

Steroids are predominantly non-polar compounds, so they can be extracted from biological materials with organic solvents (or solid-phase extraction techniques) and concentrated for further studies. Each steroid produces a characteristic mass spectrum (Figure 25.142) that can be compared with a library of known compounds. Modern GC-MS and HPLC-MS instruments can detect anabolic steroids and their metabolites at concentrations at low as 1 ng cm^{-3} (3 × 10^{-9} mol dm^{-3}) giving positive results for many weeks or even months after the use of these drugs was stopped.

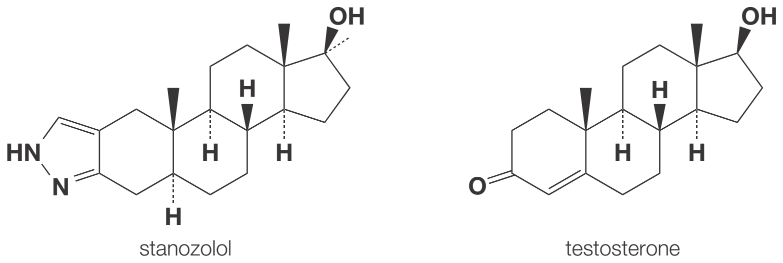
Figure 25.142 Mass spectrum of 3-hydroxystanozolol



Competitive sports and steroids

Stanozolol is an anabolic steroid related to the natural hormone testosterone (Figure 25.143). Taking stanozolol increases protein synthesis in cells and so allows athletes to build up muscle and improve their performance. However, taking anabolic steroids for long periods of time can have serious side effects, including the risk of heart attacks.

Figure 25.143 Structures of stanozolol and testosterone, the male sex hormone



The rewards of fame, glory and money in commercial sport and competitions mean that some athletes may be tempted to gain a competitive advantage by taking anabolic steroids and other performance-enhancing drugs. This corruption of attitude has not been restricted solely to individuals, or individuals and their coaches. Historical examples of state or governing-body sponsored malpractice have been well-documented – the East German athletics team of the 1950s and 1960s is a case in point, with the pressure on individual athletes to conform being immense.

Detecting and measuring the amount and concentration of a banned drug in an athlete's body is a challenging task for analytical chemists. In the body, drugs chemically break down to form other compounds known as metabolites – and so the concentration of these compounds must also be measured. In some cases only the metabolites can be detected in the athlete's body. Sampling for steroids is carried out on athletes' urine samples

Urine contains a large number of different substances and detection of very low concentrations is needed – less than 10 ng cm^{-3} (<0.01 ppm) by mass. Analysis for stanozolol involves extraction from the urine on to a solid surface, hydrolysis by enzymes, extraction into a non-polar solvent and chemical reaction to form stable compounds. The analysis is carried out by GC-MS. The peaks in the mass spectrum and the gas chromatogram are characteristic of stanozolol.

One of the most well-known uses of drugs at the Olympics occurred in 1988 when Ben Johnson won the Olympic men's 100 metre sprint title. He was stripped of his Olympic gold medal and world record after testing positive for banned steroids. He returned to athletics in 1991 but in 1993 he again tested positive for steroids and was banned from competition for life. The recent high-profile instances of doping in a range of sports, particularly road cycling, show how intense the pressure to perform at the highest levels of human fitness and endurance can create a culture of acceptance of illicit practices. The World Anti-Doping Agency was formed in 1999 as a response to the rapidly increasing instances of the use of anabolic steroids in sports. The Agency's mission is to rationalize and unify the anti-doping policies and regulations of sport governing bodies and governments so that there can be a consistency of practice and detection. The development of the World Anti-Doping Code and its broad acceptance worldwide has been a major achievement. Integral to this code is the specification that drug-testing laboratories should be accredited to meet international standards.

Ethanol

Ethanol, the alcohol in alcoholic drinks, causes mild intoxication at around 30–50 milligrams (mg) per 100 cm^3 of blood, resulting in a sense of euphoria (great happiness). In people who have not developed tolerance to ethanol, silly behaviour is observed. Once the concentration of ethanol has reached 100 mg per 100 cm^3, most people suffer neurological problems, resulting in slurred speech and staggering.

Aggressive and dangerous behaviour is also common, even in experienced drinkers. At concentrations of 200 mg per 100 cm^3 blood, vision and movement are difficult and at 400 mg per 100 cm^3 of blood, coma and death are likely. The usual cause of death is respiratory distress which is due to an inhibition of nerve impulses from the reticular formation in the brain to the intercostal muscles between the ribs.

The actual mode of action of ethanol is not fully understood: at low concentrations it is a stimulant, but at high concentrations it becomes a depressant. However, most of the problems attributed to ethanol (C_2H_5OH) are probably due to the action of ethanal (CH_3CHO) formed from ethanol by the enzyme alcohol dehydrogenase in the liver. This compound interacts with neurotransmitters to produce chemicals with psychotropic activity. Ethanal is then metabolized to ethanoate ions (CH_3COO^-) and eventually to water and carbon dioxide. The average 70 kg human male can oxidize about 10 grams of ethanol per hour.

The effects of alcohol on society are considerable. Studies have shown that there are strong correlations between heavy drinking, strokes, cirrhosis of the liver, suicide rates and all cancers, especially stomach cancer.

Long-term consumption of large quantities of alcohol can also lead to brain damage, due to its dehydrating action. Mothers who drink during pregnancy run the risk of giving birth to babies with fetal alcohol syndrome where the baby is born underweight and with brain damage.

Alcoholism is medically defined as a disease; it is often progressive and frequently fatal. It often appears to run in families and has recently been established to have a strong genetic component in some people. It seems to be related to the levels of specific enzymes inside the body.

Alcohol rapidly passes through the stomach wall and into the bloodstream. The majority of the absorbed alcohol is decomposed into carbon dioxide and water. The remainder leaves the body via sweat, in the breath or in the urine. The highest concentration of alcohol in the blood is reached about an hour after drinking ceases, although this depends on a number of factors (including the amount of food in the stomach, which tends to slow the absorption of alcohol from the small intestine).

A sample of breath taken from deep in the lungs has an alcohol concentration related to the alcohol concentration in the blood, because a rapid equilibrium is set up between the blood and air in the lungs. The concentration of the alcohol in the breath is 1/2300 of that in the blood. A measurement of the concentration of alcohol in the breath can therefore be used to calculate the concentration of alcohol in the blood.

The 'breathalyser' (Chapter 13) was introduced in the UK in 1967 to test whether a driver has exceeded the legal limit of alcohol in his or her blood (currently the value in many countries around the world is 50 or (as in the UK) 80 milligrams per 100 cm^3 of blood; this last value corresponds to 35 micrograms per 100 cm^3 of air). Drivers suspected of excessive drinking are asked to blow through a previously sealed tube until they have provided a volume of one litre (1 dm^3). The tube is packed with potassium dichromate(VI) crystals and sulfuric acid on a silica (silicon dioxide) support. Ethanol in the breath is oxidized to ethanal and then to ethanoic acid:

$$\underset{\text{ethanol}}{CH_3CH_2OH} \rightarrow \underset{\text{ethanal}}{CH_3CHO} + 2H^+ + 2e^-$$

$$\underset{\text{ethanal}}{CH_3CHO} + H_2O \rightarrow \underset{\text{ethanoic acid}}{CH_3COOH} + 2H^+ + 2e^-$$

The orange crystals of potassium dichromate(VI) are reduced to compounds containing green chromium(III) ions:

$$Cr_2O_7^{2-} + 14H^+ + 6e^- \rightarrow 2Cr^{3+} + 7H_2O$$

The concentration of alcohol in the driver's breath can be determined by seeing what length of the crystals in the tube have changed colour.

In 1980 the British police started to use an electronic instrument, known as an Alcolmeter, in place of the breathalyser. This instrument, like the breathalyser, makes use of the oxidation of ethanol to ethanal and ethanoic acid, but the oxidation process forms one of the electrodes in an electrochemical cell and the voltage of the cell is measured.

The cell consists of a permeable membrane containing phosphoric(V) acid or sodium hydroxide in sintered glass – a form of porous glass – sandwiched between catalytic coatings of silver or gold. When the driver breathes into the instrument the ethanol in their breath is catalytically oxidized to ethanoic acid at the metal electrode on one side of the membrane:

$$CH_3CH_2OH + H_2O \rightarrow CH_3COOH + 4H^+ + 4e^-$$

Oxygen is kept in contact with the metal on the other side to act as a reference electrode (c.f. the standard hydrogen electrode). The oxygen is reduced to water:

$$O_2 + 4H^+ + 4e^- \rightarrow 2H_2O$$

The instrument is calibrated by passing air with known ethanol concentrations through and measuring the resulting cell voltage.

At the police station many drunk driver suspects are given a second type of breath test, where the ethanol concentration is measured by absorption of infrared radiation (Chapter 21). The driver is asked to breathe continuously into a breath testing instrument, known as the Intoximeter. The breath passes into a cell where it is irradiated with infrared radiation (Figure 25.144). The amount of radiation absorbed at a particular wavelength or frequency (specifically 2950 cm^{-1} to detect C–H bonds) is proportional to the number of ethanol molecules present (Figure 25.145).

Figure 25.144 Components of the infrared analyser unit of the Lion Intoximeter 3000

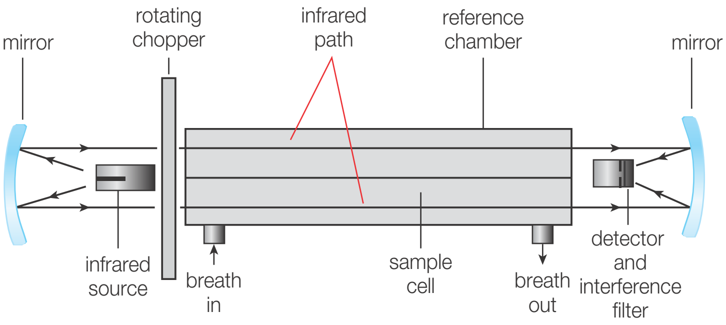
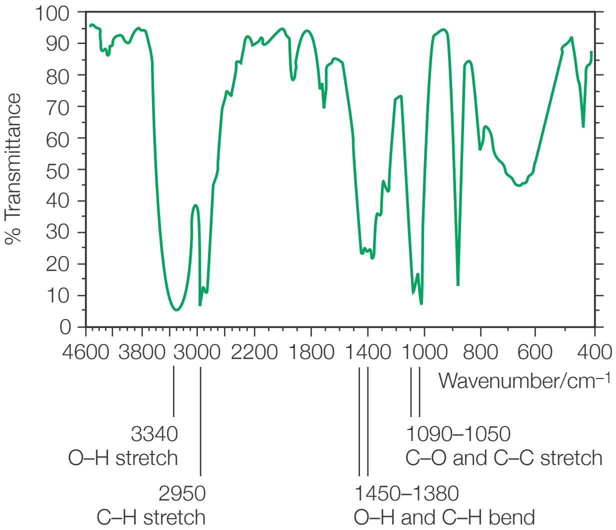


Figure 25.145 Infrared spectrum of ethanol



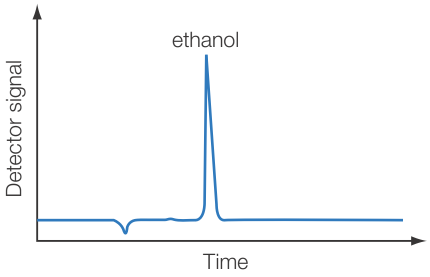


Figure 24.146 The reading from a chromatogram showing the presence of alcohol in a person's breath. The area under the peak gives the amount of alcohol in the sample

Alternatives to breath testing are also available and involve taking samples of blood and urine and subjecting them to GC. The samples are injected into a gas–liquid chromatograph (GLC) which separates out the different components. Their presence in the sample is displayed as a series of peaks on a chart recorder (Figure 24.146). The area under the peak due to alcohol is then used to calculate the concentration of alcohol. Not only can alcohol be detected and measured, but other drugs can be detected and measured at the same time.

Alcohol abusers frequently use alcohol in combination with other drugs. There is often a synergistic effect: the combination of the two drugs may be more harmful than either drug alone.

For example, when alcohol is used with cocaine, a substance called cocaethylene is formed, which may extend and enhance the cocaine 'high'.

However, cocaethylene is far more toxic than cocaine and alcohol used separately and causes severe vasoconstriction (narrowing of blood vessels leading to a rise in blood pressure) and cardiac arrhythmia (irregular heart beat).

Hyperacidity is a consequence of excess production of gastric juice and hence hydrochloric acid. This is often caused by damage to the mucous membranes of the stomach wall by alcohol or aspirin, or a synergistic combination of both drugs. If untreated a gastric ulcer is likely to develop.

Nature of Science

Analytical chemistry and drug testing

Recent advances in the instrumentation used in analytical chemistry have dramatically improved the sensitivity and accuracy of molecular drug analysis in medicine, forensic science (crime investigation) and the pharmaceutical industry. Modern analytical techniques can detect and measure trace amounts of illegal substances in the urine and blood of the human body and distinguish between stereoisomers of biologically active compounds. They can also be used to confirm the identity and purity (degree of contamination) of pharmaceutical drugs.

Safety and efficacy of pharmaceuticals are two fundamental issues of importance in drug therapy. The safety of a drug is determined by its pharmacological–toxicological profile as well as the adverse effects caused by the impurities. The impurities in drugs often possess unwanted pharmacological–toxicological effects by which any benefit from their administration may be outweighed.

These technological changes improve the quality of our lives and protect individuals and society from the harmful effects of drug abuse. At the same time an increasing number of people are now legally required to provide samples of their blood or urine for random and routine drug tests. This limits their personal freedom and affects the ethical choices of individuals.

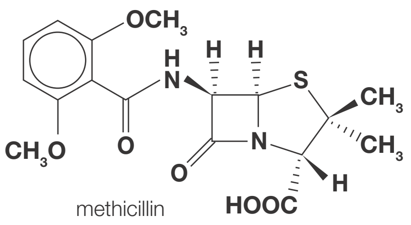
Supporters of routine or random drug testing claim that employers have a moral right to a fair day's work in exchange for a fair day's pay. They also have a right to inquire into any issue that seriously interferes with an employee giving a fair day's work. Drugs can significantly impair a person's work performance, lowering productivity. Employees who use drugs have double the rate of absenteeism, have higher job turnover rates, and incur more medical benefits compared to those who do not use drugs.

It can be argued that the government has a moral duty to protect the health and safety of its citizens. Drug abuse in the workplace constitutes a serious hazard to others. According to one survey, employees who use drugs have three times the accident rates of non-users. And, in some cases, for example nuclear plants and air-traffic control, such accidents could involve the loss of many lives.

Critics of drug-testing programs argue that employees have a basic right to privacy. Employers cannot intrude on this privacy without serious cause and in a manner that is reasonable. Routine and random drug testing, they claim, clearly violates an employee's right to privacy. First, these programs, by their nature, subject employees to humiliation and invade their privacy routinely and randomly, not because there is reasonable suspicion of drug abuse. Second, drug testing is not an effective means for screening out employees whose on-the-job performance is being impaired by drugs. The results of drug testing only indicate that traces of a drug are present in a person's body and not whether a drug is affecting a person at work.

Examination questions – a selection

1 Methicillin, a semi-synthetic penicillin, was developed to counter various strains of penicillin-resistant bacteria, but now some bacteria have even developed a resistance to methicillin. They are called MRSA (methicillin-resistant *Staphylococcus aureus*).



methicillin

a Define the term *antibiotic*. [1]

b State two reasons for chemically modifying the side-chain in penicillins and other antibiotics. [2]

c A large number of penicillins exist: some are narrow spectrum and some are broad spectrum. Explain the difference between broad-spectrum and narrow-spectrum antibiotics. [1]

d Describe the mode of action of penicillin in preventing the growth of bacteria. [2]

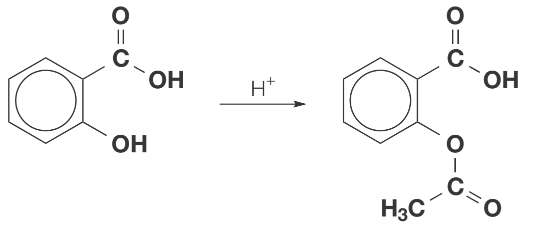
e Discuss one effect of over-prescription of penicillin to humans. [1]

f Give one reason why a course of antibiotics, such as methicillin, must be completed. [1]

g Explain the importance of the beta-lactam ring in the action of penicillins. [3]

h Antibiotics are present as pollutants in water. State the names of three other evironmental xenobiotics that are non-radioactive. [3]

Q2 The bark of the willow tree contains 2-hydroxybenzoic acid which when extracted can be used as an antipyretic and to treat fever. It is very irritating to the stomach and is now used as 2-(acetoxy)benzoic acid, or aspirin.



2-hydroxybenzoic acid 2-(acetoxy)benzoic acid

a Identify the type of chemical reaction used to convert 2-hydroxybenzoic acid to aspirin. [1]

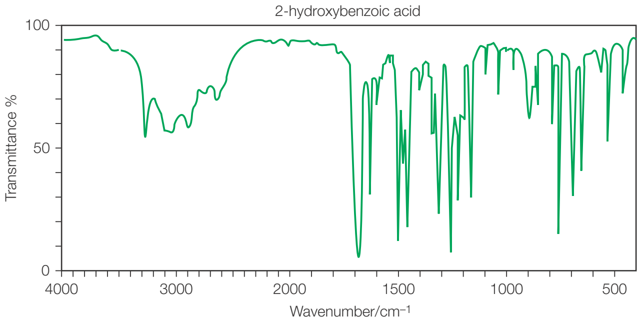
b The infrared spectra for 2-hydroxybenzoic acid and 2-(acetoxy)benzoic acid are shown below. Both molecules have many structural features in common and thus show similar peaks in their spectra. State and explain one similarity and one difference between the two spectra with respect to the bonds present. [2]

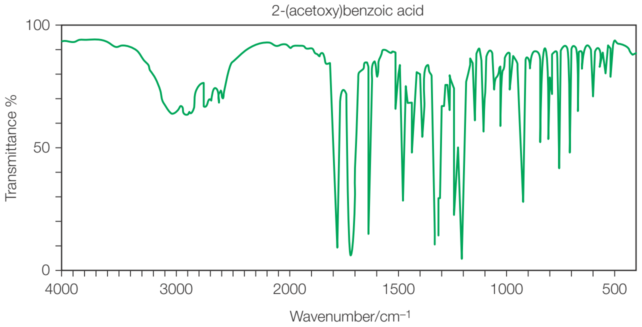
c State the number of peaks in the ^{1}H NMR spectrum (ignore the peaks due to the hydrogen atoms on the benzene ring and the reference sample) and describe the splitting pattern for each of the peaks. [2]

d A very soluble form of aspirin is prepared by reacting it with a strong base, such as potassium hydroxide. Explain why this process can increase the bioavailability of the drug. [3]

e The use of aspirin can have beneficial effects for the user, but can also produce some unwanted side effects. State one beneficial effect (other than its analgesic activity) and one unwanted side effect. [2]

f Explain how aspirin produces an analgesic effect at the molecular level. [1]





Q3 The structures of some analgesics are shown on page 38 of the *IB Chemistry data booklet*. Refer to this table when answering parts (b) and (c) of this question.

a Explain the differences in the method of action of mild analgesics and strong analgesics. [2]

b State the name of the nitrogen-containing functional group in paracetamol and heroin. [2]

c Naturally occurring morphine can be converted into synthetic heroin by reaction with ethanoic acid. Identify the group in the morphine molecule that reacts with ethanoic acid, the name of the type of reaction and the other product of the reaction. [3]

d Morphine, codeine and heroin are classified as strong analgesics.

i Name two functional groups common to morphine, codeine and heroin. [2]

ii A hospital patient has been prescribed morphine after surgery. State the main effect and a major side effect of this drug. [2]

e For two comparable populations, the LD_{50} values (expressed as mg per kg body mass) for morphine and heroin are 20 and 4, respectively.

i Explain what is meant by the term LD_{50}. [2]

ii Identify which of the two substances is more toxic with respect to these populations. [1]

f Explain what is meant by the terms TD_{50} and ED_{50}. [2]

g Explain why diamorphine (heroin) is often administered as an ionic salt, diamorphine hydrochloride. [2]

Q4 Mylanta is a commercial antacid which often contains aluminium and magnesium hydroxides in equal proportions by mass.

a Write an equation for the reaction of hydrochloric acid with one of the above antacids. [2]

b Identify which antacid neutralizes the greater amount of hydrochloric acid if 0.2 mol of each antacid is used to neutralize the hydrochloric acid present in the stomach. [1]

c Give one reason why potassium hydroxide is not used instead of these antacids. [1]

d Explain how heartburn is caused. [1]

e Explain why dimethicone and alginates are added to some antacids. [2]

f State one drawback to using calcium carbonate as an antacid. [1]

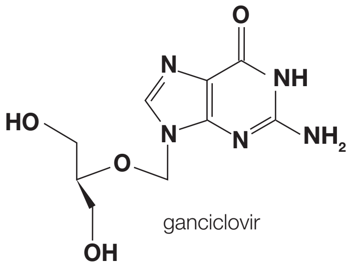
g The pH of gastric juice is 2.00. Calculate how many milligrams of HCl there are in 1.000 cm^3 of gastric juice. [2]

h 4.28 g of ammonium chloride was added to 250 cm^3 of 0.50 $mol\,dm^{-3}$ ammonia solution. Calculate the pH of the resulting solution. The pK_a for the ammonium ion is 9.25. [2]

i Calculate the mass of sodium ethanoate needed to be added to 500 cm^3 of 0.10 $mol\,dm^{-3}$ ethanic acid to produce a buffer of 4.5. The pK_a for ethanoic acid is 4.76. [2]

j Outline how the drugs rantidine (Zantac) and omeprazole regulate acid secretion. [2]

Q5 Ganciclovir is an antiviral medication used to treat or prevent infection by cytomegalovirus infections. It prevents replication of viral DNA. It is usually added to the eye or taken orally. One of its side effects is anemia. Ganciclovir and foscarnet show synergistic inhibition of cytomegalovirus.



ganciclovir

a Identify four functional groups in Ganciclovir and outline why the molecule is water soluble. [3]

b Many drugs are taken orally. State two other ways in which medicinal drugs are taken by a patient. State which method has the most rapid effect. [2]

c What term is given to a preparation which is pharmacologically inert but which may have a medical effect based solely on the power of suggestion? [1]

d State what is meant by the term side effect. [1]

e State what is meant by the term synergistic effect. [1]

f State what is meant by the term therapeutic window. Explain why it is important for cytomegalovirus patients prescribed with ganciclovir. [2]

g Give one reason why viral infections can be difficult to treat. [1]

h State two differences in the structure and replication of viruses. [2]

Q6 Acidified sodium dichromate(VI) is commonly used in roadside tests for ethanol in the breath of drivers of motor vehicles. The ethanol, C_2H_5OH, present in the breath reacts in a redox reaction to form ethanoic acid, CH_3COOH.

a State the function of sodium dichromate(VI) during this redox reaction and give the colour change that takes place. [2]

b Identify two other methods used in the police station for the detection of ethanol in a person's breath or blood that are considered to be more accurate. [2]

c State one harmful effect of aspirin that is more likely to occur if it is taken with ethanol. [1]

d Explain how anabolic steroids can be dectected in athletes using chromatrography and mass spectrometry. [4]

Acknowledgements for Option chapters

The publisher would like to thank the following for permission to reproduce copyright photographs.

Chapter 22/Option A

p.2 © Marcel – Fotolia; **p.3** *t* David Talbot, *b* © Stephen Chung/Alamy; **p.6** © Galit Seligmann/Alamy; **p.7** David Talbot; **p.9** *l*, *r* David Talbot; **p.15** Chris Talbot; **p.20** © The Natural History Museum/Alamy; **p.33** David Talbot; **p.34** Chris Talbot; **p.35** From Festschrift zur Jahrhundertfeier des Joanneums. Herausgegeben vom Verband ehemaliger Grazer Techniker 1911, Courtesy of Technische Universitaet Graz'; **p.36** David Talbot; **p.39** Chris Talbot; **p.40** Chris Talbot; **p.44** David Talbot; **p.46** David Talbot; **p.49** David Talbot; **p.50** *t* © Sablin – Fotolia, *b* Chris Talbot; **p.51** Chris Talbot; **p.52** © IBM Research; **p.53** © IBMResearch/Science Photo Library; **p.57** © NASA; **p.63** © LCDR Eric Johnson/NOAA, https://creativecommons.org/licenses/by/2.0; **p.69** © Phil Degginger/Alamy; **p.81** Richard Harwood; **p.82** *tl* Chris Talbot, *tr* © Charles D. Winters/Science Photo Library, *b* David Talbot; **p.83** David Talbot; **p.84** Chris Talbot.

Chapter 23/Option B

p.1 © Steve Gschmeissner/Science Photo Library; **p.2** © Dr Torsten Wittmann/Science Photo Library; **p.3** © Laguna Design/Science Photo Library; **p.6** © Georgette Douwma/Science Photo Library; **p.21** *l*, *r* Richard Harwood; **p.25** Courtesy of Dr J. Cooper, UCL, UK; **p.27** © Laguna Design/Science Photo Library; **p.28** © Pascal Goetgheluck/Science Photo Library; **p.31** courtesy of Dr. J. Cooper, UCL, UK; **p.32** *t* Richard Harwood; **p.33** *b* © R.A. Longuehaye/Science Photo Library; **p.37** courtesy of Dr. J. Cooper, UCL, UK; **p.49** © Cordelia Molloy/Science Photo Library; **p.54** Richard Harwood; **p.56** © Steve Gschmeissner/Science Photo Library; **p.57** David Talbot; **p.60** *bl* © Tammy Kalber & Simon Richardson, UCL CABI/Science Photo Library, *br* © David Mccarthy/Science Photo Library; **p.62** © Alfred Pasieka/Science Photo Library; **p.69** Richard Harwood; **p.71** © Dr. Lloyd M. Beidler/Science Photo Library; **p.74** © Medimage/Science Photo Library; **p.76** Cesar Reyes; **p.77** David Talbot; **p.85** © Biophoto Associates/Science Photo Library; **p.89** © Martin Bond/Science Photo Library; **p.92** *t* © Sunnys – Fotolia, *b* © Image Created By Reto Stöckli, Nazmi El Saleous and Marit Jentoft-Nilsen, NASA GSFC; **p.96** David Talbot; **p.98** © AFP/Getty Images; **p.99** Richard Harwood; **p.107** Public Domain/Http://En.Wikipedia.Org/Wiki/Brewery#/Media/File:The_Brewer_Designed_And_Engraved_In_The_Sixteenth._Century_By_J_Amman; **p.116** © A. Barrington Brown/Science Photo Library; **p.117** © Science Photo Library; **p.124** *t* © Biophoto Associates/Science Photo Library, *b* © Power And Syred/Science Photo Library; **p.125** © Pasieka/Science Photo Library; **p.129** © Pasieka/Science Photo Library; **p.130** Richard Harwood; **p.131** Chris Talbot; **p.132** © /Science Photo Library; **p.134** Richard Harwood; **p.136** Richard Harwood; **p.137** *t* David Talbot , *b* © Dr Jon Hare; p.**140** *l*, *r* Chris Talbot; **p.141** © Laguna Design/Science Photo Library; **p.142** *l*, *r* David Talbot; **p.145** *t* © GIPhotoStock/Science Photo Library, *b* Richard Harwood; **p.150** Richard Harwood; **p.154** Chris Talbot; **p.165** Chris Talbot; **p.167** Courtesy of Dr J. Cooper, UCL, UK.

Chapter 24/Option C

p.8 Chris Talbot; **p.14** all Cesar Reyes; **p.23** © The Nature Conservancy /http://www.nature.org/greenliving/carboncalculator/index.htm; **p.39** © ITER Organization; **p.44** original image courtesy of Dr. Jon Nield, http://Macromol.Sbcs.Qmul.Ac.Uk; for further information see the publication Pagliano C. Et Al., (2014) Biochim Biophys Acta 1837:1454–62; **p.49** *l* fotolia_77510924, *r* © Fotos 593 – Fotolia; **p.57** David Talbot; **p.58** *t* David Talbot, *b* Philippe Sainte-Laudy; **p.59** David Talbot; **p.64** © George Doyle/Stockbyte/Getty Images; **p.68** *l* ©

Johann Hinrichs/Alamy, *r* © Kevin Maskell/Alamy; **p.70** © Action Press/Rex; **p.85** © Aap/ Lukas Coch/ Press Association Images; **p.89** Chris Talbot; **p.94** © Alain Herzog/EPFL; **p.95** © Wenzhong Shen, Xiaoliang Yuan, Maojun Zheng and Li Ma/Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China.

Chapter 25/Option D

p.1 Chris Talbot; **p.8** © Leonard Mccombe/The LIFE Picture Collection/Getty Images; **p.15** © Greir – Fotolia; **p.20** David Talbot; **p.22** © Octavio Campos Salles/Alamy; **p.24** Chris Talbot; **p.26** David Talbot; **p.32** *l*, *r* Cesar Reyes; **p.34** Cesar Reyes; **p.36** Chris Talbot; **p.39** © Discovod – Fotolia; p.41 Chris Talbot; **p.42** Chris Talbot; **p.44** © / Science Photo Library; **p.52** Cesar Reyes; **p.67** © Pat & Tom Leeson/Science Photo Library; **p.87** © Astier – Chru Lille/Science Photo Library.

Examination questions

Examination questions have been reproduced with kind permission from the International Baccalaureate Organization.