18 Quantitative structure–activity relationships (QSAR)

In chapters 13 and 14 we studied the various strategies which can be used in the design of drugs. Several of these strategies involved a change in shape such that the new drug had a better 'fit' for its target binding site. Other strategies involved a change in functional groups or substituents such that the drug's pharmacokinetics or binding site interactions were improved. These latter strategies often involved the synthesis of analogues containing a range of substituents on aromatic or heteroaromatic rings or accessible functional groups. The number of possible analogues that could be made is infinite, if we were to try and synthesize analogues with every substituent and combination of substituents possible. Therefore, it is clearly advantageous if a rational approach can be followed in deciding which substituents to use. The quantitative structure–activity relationship (QSAR) approach has proved extremely useful in tackling this problem.

The QSAR approach attempts to identify and quantify the physicochemical properties of a drug and to see whether any of these properties has an effect on the drug's biological activity. If such a relationship holds true, an equation can be drawn up which quantifies the relationship and allows the medicinal chemist to say with some confidence that the property (or properties) has an important role in the pharmacokinetics or mechanism of action of the drug. It also allows the medicinal chemist some level of prediction. By quantifying physicochemical properties, it should be possible to calculate in advance what the biological activity of a novel analogue might be. There are two advantages to this. First, it allows the medicinal chemist to target efforts on analogues which should have improved activity and thus cut down the number of analogues that have to be made. Second, if an analogue is discovered which does not fit the equation, it implies that some other feature is important and provides a lead for further development.

What are these physicochemical features that we have mentioned?

Essentially, they could be any structural, physical, or chemical property of a drug. Clearly, any drug will have a large number of such properties and it would be a Herculean task to quantify and relate them all to biological activity at the same time. A simple, more practical approach is to consider one or two physicochemical properties of the drug and to vary these while attempting to keep other properties constant. This is not as simple as it sounds, since it is not always possible to vary one property without affecting another. Nevertheless, there have been numerous examples where the approach has worked.

It is important that the QSAR method is used properly, and in relevant situations. First of all, the compounds studied must be related structurally, act at the same target, and have the same mechanism of action. Secondly, it is crucial that the correct testing procedures are used. *In vitro* tests carried out on isolated enzymes are relevant for a QSAR study since the activities measured for different inhibitors are related directly to how each compound binds to the active site. *In vivo* tests carried out to measure the physiological effects of enzyme inhibitors are not valid, however, since both pharmacodynamic and pharmacokinetic factors come into play. This makes it impossible to derive a sensible QSAR equation.

18.1 Graphs and equations

In the simplest situation, a range of compounds is synthesized in order to vary one physicochemical property (e.g. log P) and to test how this affects the biological activity (log $1/C$) (we will come to the meaning of log $1/C$ and log P in due course). A graph is then drawn to plot the biological activity on the y-axis versus the physicochemical feature on the x-axis (Fig. 18.1).

It is then necessary to draw the best possible line through the data points on the graph. This is done by a

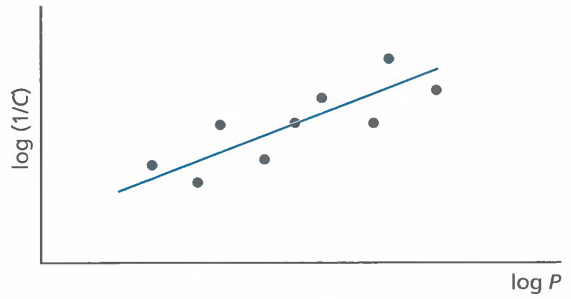


FIGURE 18.1 Biological activity versus log *P*.

procedure known as '**linear regression analysis by the least squares method**'. This is quite a mouthful and can produce a glazed expression on any chemist who is not mathematically orientated. In fact, the principle is quite straightforward. If we draw a line through a set of data points, most of the points will be scattered on either side of the line. The best line will be the one closest to the data points. To measure how close the data points are, vertical lines are drawn from each point (Fig. 18.2). These verticals are measured and then squared in order to eliminate the negative values. The squares are then added up to give a total (the sum of the squares). The best line through the points will be the line where this total is a minimum. The equation of the straight line will be $y = k_1x + k_2$ where k_1 and k_2 are constants. By varying k_1 and k_2, different equations are obtained until the best line is obtained. This whole process can be speedily done using relevant software.

The next stage in the process is to see whether the relationship is meaningful. As any good politician knows, you can make figures mean anything and so statistical evidence has to be obtained to support the QSAR equation and to quantify the goodness of fit. Otherwise, we may have obtained a straight line through points which are so random that it means nothing. The **regression or correlation coefficient** (r) is a measure of how well the equation explains the variance in activity observed in terms of the physicochemical parameters present in the equation. An explanation of how r is derived is given in Appendix 3. For a perfect fit, $r = 1$, in which case the observed activities would be the same as those calculated by the equation. Such perfection is impossible with biological data and so r values greater than 0.9 are considered acceptable. The regression coefficient is often quoted as r^2, in which case values over 0.8 are considered a good fit. If r^2 is multiplied by 100 it indicates the percentage variation in biological activity that is accounted for by the physicochemical parameters used in the equation. Thus, an r^2 value of 0.85 signifies that 85% of the variation in biological activity is accounted for by the parameters used. There are dangers in putting too much reliance on r, as the value obtained takes no account of the number of compounds (n) involved in the study and it is possible to obtain higher values of r by increasing the number of compounds tested.

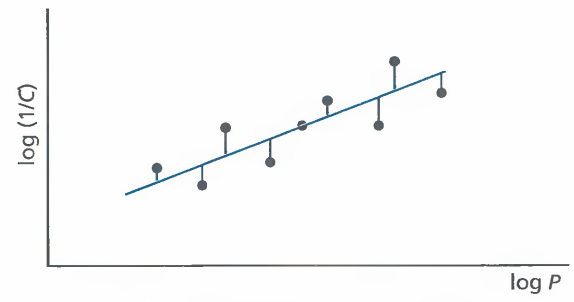


FIGURE 18.2 Proximity of data points to line of best fit.

Therefore, another statistical measure for the goodness of fit should be quoted alongside r. This is the standard error of estimate or the standard deviation (s). Ideally, s should be zero, but this would assume there were no experimental errors in the experimental data or the physicochemical parameters. In reality, s should be small, but not smaller than the standard deviation of the experimental data. It is therefore necessary to know the latter to assess whether the value of s is acceptably low. Appendix 3 shows how s is obtained and demonstrates that the number of compounds (n) in the study influences the value of s.

Statistical tests called Fishers F-tests are often quoted. These tests are used to assess the significance of the coefficients k for each parameter in the QSAR equation. Normally p values (derived from the F-test) should be less than or equal to 0.05 if the parameter is significant. If this is not the case, the parameter should not be included in the QSAR equation.

18.2 Physicochemical properties

Many physical, structural, and chemical properties have been studied by the QSAR approach, but the most common are hydrophobic, electronic, and steric properties. This is because it is possible to quantify these effects. Hydrophobic properties can be easily quantified for complete molecules or for individual substituents. On the other hand, it is more difficult to quantify electronic and steric properties for complete molecules, and this is only really feasible for individual substituents.

Consequently, QSAR studies on a variety of totally different structures are relatively rare and are limited to studies on hydrophobicity. It is more common to find QSAR studies being carried out on compounds of the same general structure, where substituents on aromatic

rings or accessible functional groups are varied. The QSAR study then considers how the hydrophobic, electronic, and steric properties of the substituents affect biological activity. The three most studied physicochemical properties are now considered in some detail.

18.2.1 Hydrophobicity

The hydrophobic character of a drug is crucial to how easily it crosses cell membranes (section 11.2) and may also be important in receptor interactions. Changing substituents on a drug may well have significant effects on its hydrophobic character and hence its biological activity. Therefore, it is important to have a means of predicting this quantitatively.

The partition coefficient (*P*)

The hydrophobic character of a drug can be measured experimentally by testing the drug's relative distribution in an *n*-octanol/water mixture. Hydrophobic molecules will prefer to dissolve in the *n*-octanol layer of this two-phase system, whereas hydrophilic molecules will prefer the aqueous layer. The relative distribution is known as the partition coefficient (*P*) and is obtained from the following equation:

$$P = \frac{\text{Concentration of drug in octanol}}{\text{Concentration of drug in aqueous solution}}$$

Hydrophobic compounds have a high *P* value, whereas hydrophilic compounds have a low *P* value.

Varying substituents on the lead compound will produce a series of analogues having different hydrophobicities and therefore different *P* values. By plotting these *P* values against the biological activity of these drugs, it is possible to see if there is any relationship between the two properties. The biological activity is normally expressed as $1/C$, where C is the concentration of drug required to achieve a defined level of biological activity. The reciprocal of the concentration ($1/C$) is used, since more active drugs will achieve a defined biological activity at lower concentration.

The graph is drawn by plotting log ($1/C$) versus log *P*. In studies where the range of the log *P* values is restricted to a small range (e.g. log P = 1–4), a straight-line graph is obtained (Fig. 18.1) showing that there is a relationship between hydrophobicity and biological activity. Such a line would have the following equation:

$$\log\left(\frac{1}{C}\right) = k_1 \log P + k_2$$

For example, the binding of drugs to serum albumin is determined by their hydrophobicity, and a study of 42 compounds resulted in the following equation:

$$\log\left(\frac{1}{C}\right) = 0.75 \log P + 2.30 \quad (n = 42, r = 0.960, s = 0.159)$$

The equation shows that serum albumin binding increases as log *P* increases. In other words, hydrophobic drugs bind more strongly to serum albumin than hydrophilic drugs. Knowing how strongly a drug binds to serum albumin can be important in estimating effective dose levels for that drug. When bound to serum albumin, the drug cannot bind to its receptor and so the dose levels for the drug should be based on the amount of unbound drug present in the circulation. The equation above allows us to calculate how strongly drugs of similar structure will bind to serum albumin and gives an indication of how 'available' they will be for receptor interactions. The r value of 0.96 is close to 1, which shows that the line resulting from the equation is a good fit. The value of r^2 is 92%, which indicates that 92% of the variation in serum albumin binding can be accounted for by the different hydrophobicities of the drugs tested. This means that 8% of the variation is unaccounted for, partly as a result of the experimental errors involved in the measurements.

Despite such factors as serum albumin binding, it is generally found that increasing the hydrophobicity of a lead compound results in an increase in biological activity. This reflects the fact that drugs have to cross hydrophobic barriers such as cell membranes in order to reach their target. Even if no barriers are to be crossed (e.g. in *in vitro* studies), the drug has to interact with a target system such as an enzyme or receptor where the binding site is more hydrophobic than the surface. Therefore, increasing hydrophobicity aids the drug in crossing hydrophobic barriers or in binding to its target site.

This might imply that increasing log *P* should increase the biological activity *ad infinitum*. In fact, this does not happen. There are several reasons for this. For example, the drug may become so hydrophobic that it is poorly soluble in the aqueous phase. Alternatively, it may be 'trapped' in fat depots and never reach the intended site. Finally, hydrophobic drugs are often more susceptible to metabolism and subsequent elimination.

A straight-line relationship between log *P* and biological activity is observed in many QSAR studies because the range of log *P* values studied is often relatively narrow. For example, the study carried out on serum albumin binding was restricted to compounds having log *P* values in the range 0.78 to 3.82. If these studies were to be extended to include compounds with very high log *P* values then we would see a different picture. The graph would be parabolic, as shown in Fig. 18.3. Here, the biological activity increases as log *P* increases until a maximum value is obtained. The value of log *P* at the maximum (log P^0) represents the optimum partition coefficient for biological activity. Beyond that point, an increase in log *P* results in a decrease in biological activity.

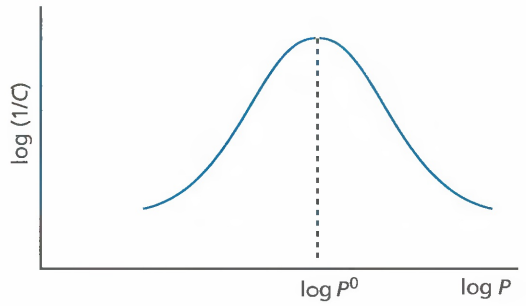


FIGURE 18.3 Parabolic curve of log (1/*C*) versus log *P*.

If the partition coefficient is the only factor influencing biological activity, the parabolic curve can be expressed by the equation:

$$\log\left(\frac{1}{C}\right) = -k_1 (\log P)^2 + k_2 \log P + k_3$$

Note that the $(\log P)^2$ term has a minus sign in front of it. When P is small, the $(\log P)^2$ term is very small and the equation is dominated by the log *P* term. This represents the first part of the graph where activity increases with increasing *P*. When *P* is large, the $(\log P)^2$ term is more significant and eventually 'overwhelms' the log *P* term. This represents the last part of the graph where activity drops with increasing *P*. k_1, k_2, and k_3 are constants and can be determined by a suitable software program.

There are relatively few drugs where activity is related to the log *P* factor alone. Such drugs tend to operate in cell membranes where hydrophobicity is the dominant feature controlling their action. The best example of drugs which operate in cell membranes are the general anaesthetics. Although they also bind to $GABA_A$ receptors, general anaesthetics are thought to function by entering the central nervous system and 'dissolving' into cell membranes where they affect membrane structure and nerve function. In such a scenario, there are no specific drug-receptor interactions and the mechanism of the drug is controlled purely by its ability to enter cell membranes (i.e. its hydrophobic character). The general anaesthetic activity of a range of ethers was found to fit the following parabolic equation:

$$\log\left(\frac{1}{C}\right) = -0.22 (\log P)^2 + 1.04 \log P + 2.16$$

According to this equation, anaesthetic activity increases with increasing hydrophobicity (*P*), as determined by the log *P* factor. The negative $(\log P)^2$ factor shows that the relationship is parabolic and that there is an optimum value for log *P* ($\log P^0$) beyond which increasing hydrophobicity causes a decrease in anaesthetic activity. With this equation, it is now possible to predict the anaesthetic activity of other compounds, given their partition coefficients. However, there are limitations. The equation is derived purely for anaesthetic ethers and is not applicable to other structural types of anaesthetics. This is generally true in QSAR studies. The procedure works best if it is applied to a series of compounds which have the same general structure.

However, QSAR studies *have* been carried out on other structural types of general anaesthetics and in each case a parabolic curve has been obtained. Although, the constants for each equation are different, it is significant that the optimum hydrophobicity (represented by log P^0) for anaesthetic activity is close to 2.3, regardless of the class of anaesthetic being studied. This finding suggests that all general anaesthetics are operating in a similar fashion, controlled by the hydrophobicity of the structure.

Since different anaesthetics have similar log P^0 values, the log *P* value of any compound can give some idea of its potential potency as an anaesthetic. For example, the log *P* values of the gaseous anaesthetics **ether, chloroform,** and **halothane** are 0.98, 1.97, and 2.3 respectively. Their anaesthetic activity increases in the same order.

Since general anaesthetics have a simple mechanism of action based on the efficiency with which they enter the central nervous system, it implies that log *P* values should give an indication of how easily any compound can enter the central nervous system. In other words, compounds having a log *P* value close to 2 should be capable of entering the central nervous system efficiently. This is generally found to be true. For example, the most potent barbiturates for sedative and hypnotic activity are found to have log *P* values close to 2.

As a rule of thumb, drugs which are to be targeted for the central nervous system should have a log *P* value of approximately 2. Conversely, drugs which are designed to act elsewhere in the body should have log *P* values significantly different from 2 in order to avoid possible central nervous system side effects (e.g. drowsiness) (Box 18.1).

The substituent hydrophobicity constant (π)

We have seen how the hydrophobicity of a compound can be quantified using the partition coefficient *P*. In order to get *P*, we have to measure it experimentally and that means that we have to synthesize the compounds. It would be much better if we could calculate *P* theoretically and decide in advance whether the compound is worth synthesizing. QSAR would then allow us to target the most promising looking structures. For example, if we were planning to synthesize a range of barbiturate structures, we could calculate log *P* values for them all and concentrate on the structures which had log *P* values closest to the optimum log P^0 value for barbiturates.

Partition coefficients can be calculated by knowing the contribution that various substituents make to hydrophobicity. This contribution is known as the substituent

BOX 18.1 Altering log *P* to remove central nervous system side effects

The cardiotonic agent (I) was found to produce 'bright visions' in some patients, which implied that it was entering the central nervous system. This was supported by the fact that the log *P* value of the drug was 2.59. In order to prevent the drug entering the central nervous system, the 4-OMe group was replaced by a 4-S(O)Me group. This particular group is approximately the same size as the methoxy group, but more hydrophilic. The log *P* value of the new drug (**sulmazole**) was found to be 1.17. The drug was now too hydrophilic to enter the central nervous system and was free of central nervous system side-effects.

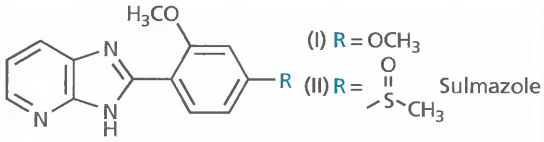


FIGURE 1 Cardiotonic agents.

hydrophobicity constant (π) and is a measure of how hydrophobic a substituent is, relative to hydrogen. The value can be obtained as follows. Partition coefficients are measured experimentally for a standard compound such as benzene with and without a substituent (X). The hydrophobicity constant (π_X) for the substituent (X) is then obtained using the following equation:

$$\pi_X = \log P_X - \log P_H$$

where P_H is the partition coefficient for the standard compound, and P_X is the partition coefficient for the standard compound with the substituent.

A positive value of π indicates that the substituent is more hydrophobic than hydrogen; a negative value indicates that the substituent is less hydrophobic. The π values for a range of substituents are shown in Table 18.1. These π values are characteristic for the substituent and can be used to calculate how the partition coefficient of a drug would be affected if these substituents were present. The *P* value for the lead compound would have to be measured experimentally, but once

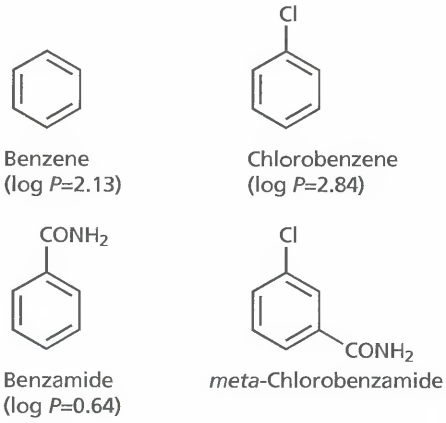


FIGURE 18.4 Values for log *P*.

that is known, the *P* value for analogues can be calculated quite simply.

As an example, consider the log *P* values for benzene (log *P* = 2.13), chlorobenzene (log *P* = 2.84), and benzamide (log *P* = 0.64) (Fig. 18.4). Benzene is the parent compound, and the substituent constants for Cl and $CONH_2$ are 0.71 and −1.49 respectively. Having obtained these values, it is now possible to calculate the theoretical log *P* value for *meta*-chlorobenzamide:

$$\begin{aligned}\log P_{(\text{chlorobenzamide})} &= \log P_{(\text{benzene})} + \pi_{Cl} + \pi_{CONH_2} \\ &= 2.13 + 0.71 + (-1.49) \\ &= 1.35\end{aligned}$$

The observed log *P* value for this compound is 1.51.

It should be noted that π values for aromatic substituents are different from those used for aliphatic substituents. Furthermore, neither of these sets of π values are in fact true constants and they are accurate only for the structures from which they were derived. They can be used as good approximations when studying other structures, but it is possible that the values will have to be adjusted in order to get accurate results.

In order to distinguish calculated log *P* values from experimental ones, the former are referred to as Clog ***P*** values. There are also software programs which will calculate Clog *P* values for a given structure.

P versus π

QSAR equations relating biological activity to the partition coefficient *P* have already been described, but

TABLE 18.1 Values for π for a range of substituents.

Group	CH_3	*t*-Bu	OH	OCH_3	CF_3	Cl	Br	F
π (aliphatic substituents)	0.50	1.68	−1.16	0.47	1.07	0.39	0.60	−0.17
π (aromatic substituents)	0.52	1.68	−0.67	−0.02	1.16	0.71	0.86	0.14

there is no reason why the substituent hydrophobicity constant π cannot be used in place of *P* if only the substituents are being varied. The equation obtained would be just as relevant as a study of how hydrophobicity affects biological activity. That is not to say that *P* and π are exactly equivalent—different equations would be obtained with different constants. Apart from that, the two factors have different emphases. The partition coefficient *P* is a measure of the drug's overall hydrophobicity and is therefore an important measure of how efficiently a drug is transported to its target and bound to its binding site. The π factor measures the hydrophobicity of a specific region on the drug's skeleton and if it is present in the QSAR equation, it could emphasize any important hydrophobic interactions involving that region of the molecule with the binding site.

Most QSAR equations have a contribution from *P* or from π, but there are examples of drugs for which they have only a slight contribution. For example, a study on antimalarial drugs showed very little relationship between antimalarial activity and hydrophobic character. This finding supports the theory that these drugs act in red blood cells, since previous research has shown that the ease with which drugs enter red blood cells is not related to their hydrophobicity.

18.2.2 Electronic effects

The electronic effects of various substituents will clearly have an effect on a drug's ionization or polarity. This in turn may have an effect on how easily a drug can pass through cell membranes or how strongly it can interact with a binding site. It is therefore useful to measure the electronic effect of a substituent.

As far as substituents on an aromatic ring are concerned, the measure used is known as the **Hammett substituent constant** (σ). This is a measure of the electron-withdrawing or electron-donating ability of a substituent and has been determined by measuring the dissociation of a series of substituted benzoic acids compared to the dissociation of benzoic acid itself.

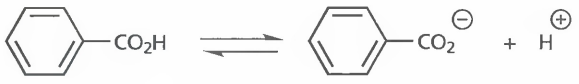


FIGURE 18.5 Ionization of benzoic acid in water.

Benzoic acid is a weak acid and only partially ionizes in water (Fig. 18.5). An equilibrium is set up between the ionized and non-ionized forms, where the relative proportion of these species is known as the **equilibrium** or **dissociation constant** K_H (the subscript H signifies that there are no substituents on the aromatic ring).

$$K_H = \frac{[PhCO_2^-]}{[PhCO_2H]}$$

When a substituent is present on the aromatic ring, this equilibrium is affected. Electron-withdrawing groups, such as a nitro group, result in the aromatic ring having a stronger electron-withdrawing and stabilizing influence on the carboxylate anion. The equilibrium will therefore shift more to the ionized form such that the substituted benzoic acid is a stronger acid and has a larger K_X value (X represents the substituent on the aromatic ring) (Fig. 18.6).

If the substituent X is an electron-donating group such as an alkyl group, then the aromatic ring is less able to stabilize the carboxylate ion. The equilibrium shifts to the left and a weaker acid is obtained with a smaller K_X value (Fig. 18.6).

The Hammett substituent constant (σ_X) for a particular substituent (X) is defined by the following equation:

$$\sigma_X = \log \frac{K_X}{K_H} = \log K_X - \log K_H$$

Benzoic acids containing electron-withdrawing substituents will have larger K_X values than benzoic acid itself (K_H) and therefore the value of σ_X for an electron-withdrawing substituent will be positive. Substituents such as Cl, CN, or CF_3 have positive σ values.

Benzoic acids containing electron-donating substituents will have smaller K_X values than benzoic acid itself

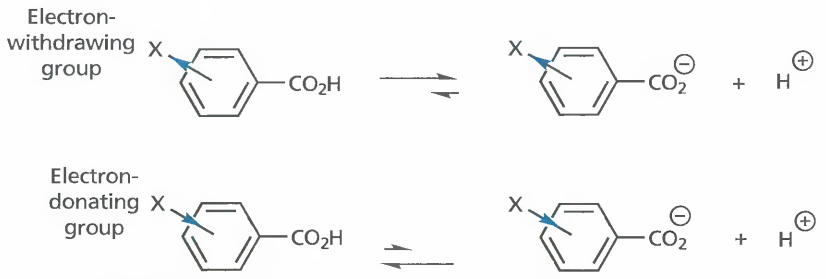


FIGURE 18.6 Position of equilibrium dependent on substituent group X.

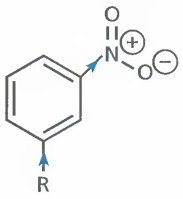
and hence the value of σ_x for an electron-donating substituent will be negative. Substituents such as Me, Et, and *t*-Bu have negative values of σ. The Hammett substituent constant for H is zero.

The Hammett constant takes into account both resonance and inductive effects. Therefore, the value of σ for a particular substituent will depend on whether the substituent is *meta* or *para*. This is indicated by the subscript *m* or *p* after the σ symbol. For example, the nitro substituent has $\sigma_p = 0.78$ and $\sigma_m = 0.71$. In the *meta* position, the electron-withdrawing power is due to the inductive influence of the substituent, whereas at the *para* position inductive and resonance both play a part and so the σ_p value is greater (Fig. 18.7).

For the hydroxyl group $\sigma_m = 0.12$ and $\sigma_p = -0.37$. At the *meta* position, the influence is inductive and electron-withdrawing. At the *para* position, the electron-donating influence due to resonance is more significant than the electron-withdrawing influence due to induction (Fig. 18.8).

Most QSAR studies start off by considering σ, and if there is more than one substituent, the σ values are summed ($\Sigma\sigma$). However, as more compounds are synthesized, it is possible to fine-tune the QSAR equation. As mentioned above, σ is a measure of a substituent's inductive and resonance electronic effects. With more detailed studies, the inductive and resonance effects can be considered separately. Tables of constants are available which quantify a substituent's inductive effect (*F*) and its resonance effect (*R*). In some cases, it might be found that a substituent's effect on activity is due to *F* rather than *R*, and *vice versa*. It might also be found that a substituent has a more significant effect at a particular position on the ring and this can also be included in the equation.

meta Nitro group—electronic influence on R is inductive



para Nitro group—electronic influence on R is due to inductive and resonance effects

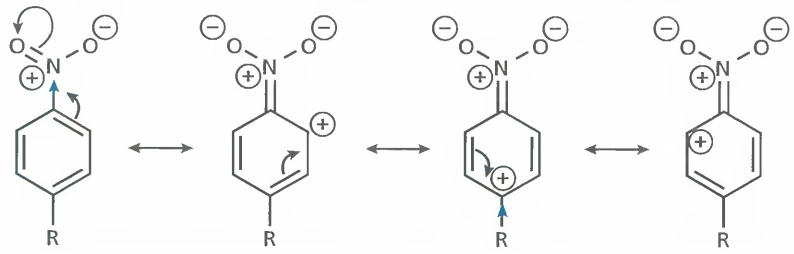
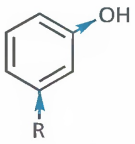


FIGURE 18.7 Substituent effects of a nitro group at the *meta* and *para* positions.

meta Hydroxyl group—electronic influence on R is inductive



para Hydroxyl group—electronic influence on R dominated by resonance effects

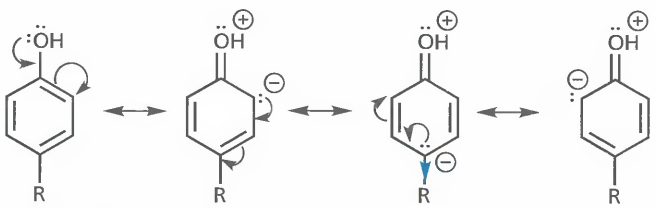


FIGURE 18.8 Substituent effects of a phenol at the *meta* and *para* positions.

There are limitations to the electronic constants described so far. For example, Hammett substituent constants cannot be measured for *ortho* substituents since such substituents have an important steric, as well as electronic, effect.

There are very few drugs whose activities are solely influenced by a substituent's electronic effect, as hydrophobicity usually has to be considered as well. Those that do are generally operating by a mechanism whereby they do not have to cross any cell membranes (Box 18.2). Alternatively, *in vitro* studies on isolated enzymes may result in QSAR equations lacking the hydrophobicity factor, since there are no cell membranes to be considered.

The constants σ, *R*, and *F* can only be used for aromatic substituents and are therefore only suitable for drugs containing aromatic rings. However, a series of aliphatic electronic substituent constants are available. These were obtained by measuring the rates of hydrolysis for a series of aliphatic esters (Fig. 18.9). Methyl ethanoate is the parent ester and it is found that the rate of hydrolysis is affected by the substituent X. The extent to which the rate of hydrolysis is affected is a measure of the substituent's electronic effect at the site of reaction (i.e. the ester group). The electronic effect is purely inductive and is given the symbol σ_I. Electron donating groups reduce the rate of hydrolysis and therefore have negative values. For example, σ_I values for methyl, ethyl, and propyl are −0.04, −0.07, and −0.36 respectively. Electron withdrawing groups increase the rate of hydrolysis and have positive values. The σ_I values for NMe_3^+ and CN are 0.93 and 0.53 respectively.

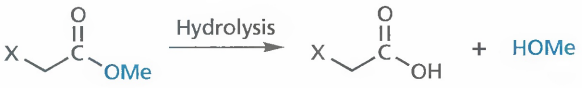


FIGURE 18.9 Hydrolysis of an aliphatic ester.

It should be noted that the inductive effect is not the only factor affecting the rate of hydrolysis. The substituent may also have a steric effect. For example, a bulky substituent may shield the ester from attack and lower the rate of hydrolysis. It is therefore necessary to separate out these two effects. This can be done by measuring hydrolysis rates under basic conditions and also under acidic conditions. Under basic conditions, steric and electronic factors are important, whereas under acidic conditions only steric factors are important. By comparing the rates, values for the electronic effect (σ_I), and the steric effect (E_S) (see below) can be determined.

BOX 18.2

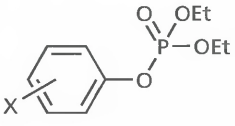
Insecticidal activity of diethyl phenyl phosphates

The insecticidal activity of diethyl phenyl phosphates is one of the few examples where activity is related to electronic factors alone:

$$\log\left(\frac{1}{C}\right) = 2.282\sigma - 0.348. \quad (r^2\ 0.952,\ r\ 0.976,\ s\ 0.286)$$

The equation reveals that substituents with a positive value for σ (i.e. electron-withdrawing groups) will increase activity. The fact that a hydrophobic parameter is not present is a good indication that the drugs do not have to pass into or through a cell membrane to have activity. In fact, these drugs are known to act on an enzyme called **acetylcholinesterase**, which is situated on the outside of cell membranes (section 22.14).

The value of *r* is close to 1, which demonstrates that the line is a good fit, and the value of r^2 demonstrates that 95% of the data are accounted for by the σ parameter.



Diethyl phenyl phosphates

18.2.3 Steric factors

The bulk, size, and shape of a drug will influence how easily it can approach and interact with a binding site. A bulky substituent may act like a shield and hinder the ideal interaction between a drug and its binding site. Alternatively, a bulky substituent may help to orientate a drug properly for maximum binding and increase activity. Steric properties are more difficult to quantify than hydrophobic or electronic properties. Several methods have been tried, of which three are described here. It is highly unlikely that a drug's biological activity will be affected by steric factors alone, but these factors are frequently found in Hansch equations (section 18.3.).

18.2.3.1 Taft's steric factor (E_s)

Attempts have been made to quantify the steric features of substituents by using Taft's steric factor (E_s). The value for E_s can be obtained by comparing the rates of hydrolysis of substituted aliphatic esters against a standard ester under acidic conditions. Thus,

$$E_s = \log k_x - \log k_o$$

where k_x represents the rate of hydrolysis of an aliphatic ester bearing the substituent X, and k_o represents the rate of hydrolysis of the reference ester.

TABLE 18.2 Values of E_s for various substituents.

Substituent	H	F	Me	Et	*n*-Pr	*n*-Bu	*i*-Pr	*i*-Bu	Cyclopentyl
E_S	1.24	0.78	0	−0.07	−0.36	−0.39	−0.47	−0.93	−0.51

The substituents that can be studied by this method are restricted to those that interact sterically with the tetrahedral transition state of the reaction and not by resonance or internal hydrogen bonding. For example, unsaturated substituents that are conjugated to the ester cannot be measured by this procedure. Examples of E_s values are shown in Table 18.2. Note that the reference ester is X=Me. Substituents such as H and F which are smaller than a methyl group result in a faster rate of hydrolysis ($k_x > k_o$), making E_s positive. Substituents which are larger than methyl reduce the rate of hydrolysis ($k_x < k_o$), making E_s negative. A disadvantage of E_s values is that they are a measure of an *intramolecular* steric effect, whereas drugs interact with target binding sites in an *intermolecular* manner. For example, consider the E_s values for *i*-Pr, *n*-Pr, and *n*-Bu. The E_s value for the branched isopropyl group is significantly greater than that for the linear *n*-propyl group since the bulk of the substituent is closer to the reaction centre. Extending the alkyl chain from *n*-propyl to *n*-butyl has little effect on E_s. The larger *n*-butyl group is extended away from the reaction centre and so has little additional steric effect on the rate of hydrolysis. As a result, the E_s value for the *n*-butyl group undervalues the steric effect which this group might have if it was present on a drug that was approaching a binding site.

18.2.3.2 Molar refractivity

Another measure of the steric factor is provided by a parameter known as molar refractivity (*MR*). This is a measure of the volume occupied by an atom or a group of atoms. The *MR* is obtained from the following equation:

$$MR = \frac{(n^2 - 1)}{(n^2 + 2)} \times \frac{MW}{d}$$

where n is the index of refraction, MW is the molecular weight, and d is the density. The term MW/d defines a volume, and the $(n^2 - 1)/(n^2 + 2)$ term provides a correction factor by defining how easily the substituent can be polarized. This is particularly significant if the substituent has π electrons or lone pairs of electrons.

18.2.3.3 Verloop steric parameter

Another approach to measuring the steric factor involves a computer program called **Sterimol** which calculates steric substituent values (Verloop steric parameters) from standard bond angles, van der Waals radii, bond lengths, and possible conformations for the substituent. Unlike E_s, the Verloop steric parameters can be measured for any substituent. For example, the Verloop steric parameters for a carboxylic acid group are demonstrated in Fig. 18.10. L is the length of the substituent and B_1–B_4 are the radii of the group in different dimensions.

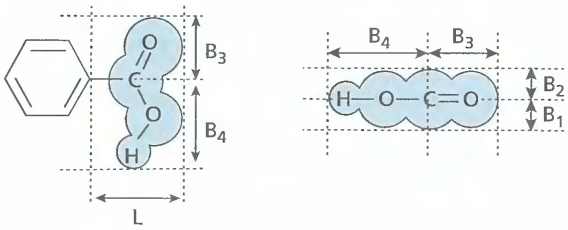


FIGURE 18.10 Verloop parameters for a carboxylic acid group.

18.2.4 Other physicochemical parameters

The physicochemical properties most commonly studied by the QSAR approach have been described above, but other properties have been studied including dipole moments, hydrogen bonding, conformations, and interatomic distances. Difficulties in quantifying these properties limit the use of these parameters, however. Several QSAR formulae have been developed based on the highest occupied and/or the lowest unoccupied molecular orbitals of the test compounds. The calculation of these orbitals can be carried out using semi-empirical quantum mechanical methods (section 17.7.3). Indicator variables for different substituents can also be used. These are described in section 18.7.

18.3 Hansch equation

In section 18.2 we looked at the physicochemical properties commonly used in QSAR studies and how it is possible to quantify them. In a situation where biological activity is related to only one such property, a simple equation can be drawn up. The biological activity of most drugs, however, is related to a combination of physicochemical properties. In such cases, simple equations involving only one parameter are relevant only if the other parameters are kept constant. In reality, this is not easy to achieve and equations which relate biological activity to a number of different parameters are

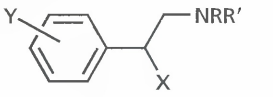
more common (Box 18.3). These equations are known as Hansch equations, and they usually relate biological activity to the most commonly used physicochemical properties (log *P* or π, σ, and a steric factor). If the range of hydrophobicity values is limited to a small range then the equation will be linear, as follows:

$$\log\left(\frac{1}{C}\right) = k_1 \log P + k_2\sigma + k_3 E_s + k_4$$

If the log *P* values are spread over a large range, then the equation will be parabolic for the same reasons described in section 18.2.1.

$$\log\left(\frac{1}{C}\right) = -k_1(\log P)^2 + k_2 \log P + k_3\sigma + k_4 E_s + k_5$$

The constants k_1–k_5 are determined by computer software in order to get the best fitting equation. Not all the parameters will necessarily be significant. For example, the adrenergic blocking activity of β-halo-arylamines (Fig. 18.11) was related to π and σ and did not include a steric factor. This equation tells us that biological activity increases if the substituents have a positive π value and a



β-Halo-arylamines

$$\log\left(\frac{1}{C}\right) = 1.22\pi - 1.59\sigma + 7.89$$

$(n = 22, r^2 = 0.841, s = 0.238)$

FIGURE 18.11 QSAR equation for β-halo-arylamines.

BOX 18.3 Hansch equation for a series of antimalarial compounds

A series of 102 phenanthrene aminocarbinols was tested for antimalarial activity. In the structure shown, X represents up to 4 substituents on the left hand ring while Y represents up to 4 substituents on the right hand ring. Experimental log *P* values for the structures were not available and equations were derived which compared the activity with some or all of the following terms:

- π_{sum} the π constants for *all* the substituents in the molecule (i.e. all the X and Y substituents, as well as the amino substituents R and R″). This term was used in place of log *P* to represent the overall hydrophobicity for the molecule.
- σ_{sum} the σ constants for *all* the substituents in the molecule
- $\Sigma\pi_X$ the sum of the π constants for all the substituents X in the left-hand ring
- $\Sigma\pi_Y$ the sum of the π constants for all the substituents Y in the right-hand ring
- $\Sigma\pi_{X+Y}$ the sum of the π constants for all the substituents X and Y in both the left and right-hand rings
- $\Sigma\sigma_{X+Y}$ the sum of the σ constants for all the substituents X and Y in both the left and right-hand rings
- $\Sigma\sigma_X$ the sum of the σ constants for all the substituents X in the left-hand ring
- $\Sigma\sigma_Y$ the sum of the σ constants for all the substituents Y in the right-hand ring

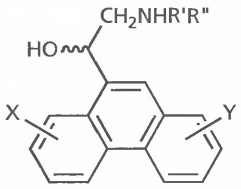
Equations such as equations 1–3 were derived which matched activity against one of the above terms, but none of them had an acceptable value of r^2.

A variety of other equations were derived which included two of the above terms but these were not satisfactory either. Finally, an equation was derived which contained 6 terms and proved satisfactory:

$$\log\left(\frac{1}{C}\right) = -0.015(\pi_{sum})^2 + 0.14\pi_{sum} + 0.27\Sigma\pi_X + 0.40\Sigma\pi_Y + 0.65\Sigma\sigma_X + 0.88\Sigma\sigma_Y + 2.34$$

$(n = 102,\ r = 0.913, r^2 = 0.834, s = 0.258)$

The equation shows that antimalarial activity increases very slightly as the overall hydrophobicity of the molecule (π_{sum}) increases (the constant 0.14 is low). The $(\pi_{sum})^2$ term shows that there is an optimum overall hydrophobicity for activity and this is found to be 4.44. Activity increases if hydrophobic substituents are present on ring X and in particular on ring Y. This could be taken to imply that some form of hydrophobic interaction is involved near both rings. Electron withdrawing substituents on both rings are also beneficial to activity, more so on ring Y than ring X. The r^2 value is 0.834, which is above the minimum acceptable value of 0.8.



Phenanthrene aminocarbinols

1) $\log\left(\frac{1}{C}\right) = 0.557\Sigma\pi x{+}y + 2.699$ $(n = 102, r = 0.768, r^2 = 0.590, s = 0.395)$

2) $\log\left(\frac{1}{C}\right) = 0.017\pi_{sum} + 3.324$ $(n = 102, r = 0.069, r^2 = 0.005, s = 0.616)$

3) $\log\left(\frac{1}{C}\right) = 1.218\sigma_{sum} + 2.721$ $(n = 102, r = 0.814, r^2 = 0.663, s = 0.359)$

negative σ value. In other words, the substituents should be hydrophobic and electron donating.

When carrying out a Hansch analysis, it is important to choose the substituents carefully to ensure that the change in biological activity can be attributed to a particular parameter. There are plenty of traps for the unwary. Take, for example, drugs which contain an amine group. One of the studies most frequently carried out on amines is to synthesize analogues containing a homologous series of alkyl substituents on the nitrogen atom (i.e. Me, Et, *n*-Pr, *n*-Bu). If activity increases with the chain length of the substituent, is it due to increasing hydrophobicity, increasing size, or both? If we look at the π and *MR* values of these substituents, we find that both sets of values increase in a similar fashion across the series and we would not be able to distinguish between them (Table 18.3). In this example, a series of substituents would have to be chosen where π and *MR* are not correlated. The substituents H, Me, OMe, $NHCOCH_2$, I, and CN would be more suitable.

18.4 Craig plot

Although tables of π and σ factors are readily available for a large range of substituents, it is often easier to visualize the relative properties of different substituents by considering a plot where the y-axis is the value of the σ factor and the x-axis is the value of the π factor. Such a plot is known as a Craig plot. The example shown in Fig. 18.12 is the Craig plot for the σ and π factors of *para*-aromatic substituents. There are several advantages to the use of such a Craig plot.

- The plot shows clearly that there is no overall relationship between π and σ. The various substituents are scattered around all four quadrants of the plot.
- It is possible to tell at a glance which substituents have positive π and σ parameters, which substituents have negative π and σ parameters, and which substituents have one positive and one negative parameter.

TABLE 18.3 Values for π and MR for a series of substituents.

Substituent	H	Me	Et	*n*-Pr	*n*-Bu	OMe	$NHCONH_2$	I	CN
π	0.00	0.56	1.02	1.50	2.13	−0.02	−1.30	1.12	−0.57
MR	0.10	0.56	1.03	1.55	1.96	0.79	1.37	1.39	0.63

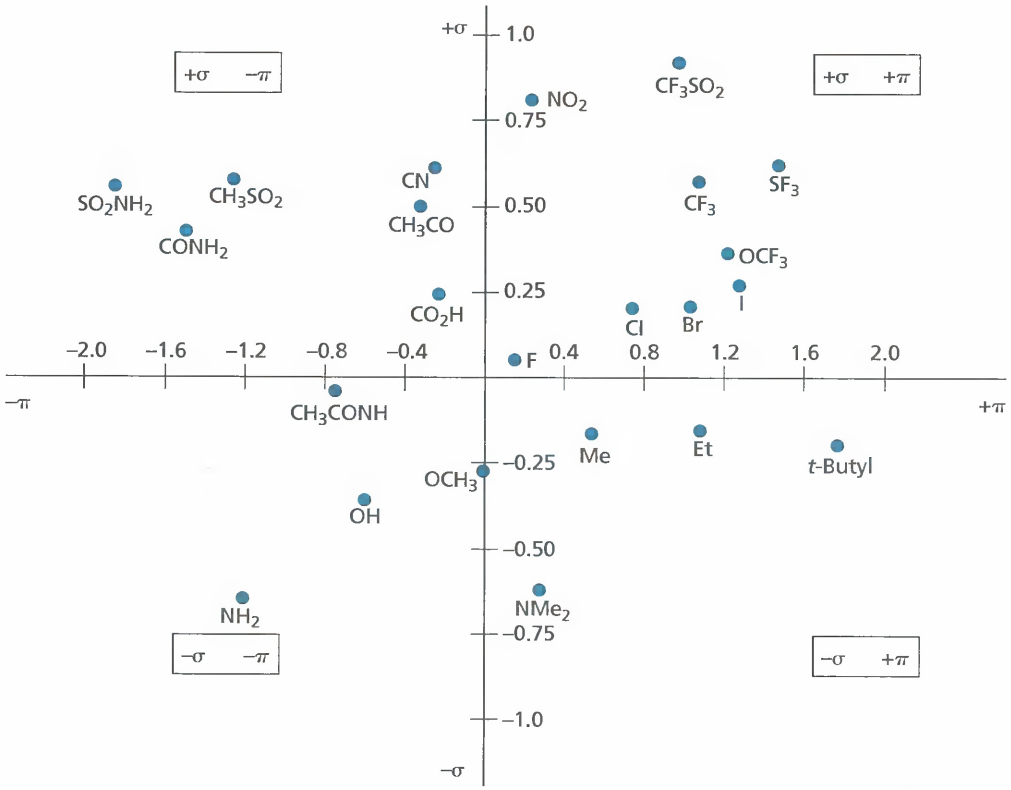


FIGURE 18.12 Craig plot comparing the values of σ and π for various substituents.

- It is easy to see which substituents have similar π values. For example, the ethyl, bromo, trifluoromethyl, and trifluoromethylsulfonyl groups are all approximately on the same vertical line on the plot. In theory, these groups could be interchangeable on drugs where the principal factor affecting biological activity is the π factor. Similarly, groups which form a horizontal line can be identified as being isoelectronic or having similar σ values (e.g. CO_2H, Cl, Br, I).
- The Craig plot is useful in planning which substituents should be used in a QSAR study. In order to derive the most accurate equation involving π and σ, analogues should be synthesized with substituents from each quadrant. For example, halogen substituents are useful representatives of substituents with increased hydrophobicity and electron-withdrawing properties (positive π and positive σ), whereas an OH substituent has more hydrophilic and electron-donating properties (negative π and negative σ). Alkyl groups are examples of substituents with positive π and negative σ values, whereas acyl groups have negative π and positive σ values.
- Once the Hansch equation has been derived, it will show whether π or σ should be negative or positive in order to get good biological activity. Further developments would then concentrate on substituents from the relevant quadrant. For example, if the equation shows that positive π and positive σ values are necessary, then further substituents should only be taken from the top right quadrant.

Craig plots can also be drawn up to compare other sets of physicochemical parameters, such as hydrophobicity and *MR*.

18.5 Topliss scheme

In certain situations, it might not be feasible to make the large range of structures required for a Hansch equation. For example, the synthetic route involved might be so difficult that only a few structures can be made in a limited time. In these circumstances, it would be useful to test compounds for biological activity as they are synthesized and to use these results to determine the next analogue to be synthesized.

A Topliss scheme is a 'flow diagram' which allows such a procedure to be followed. There are two Topliss schemes, one for aromatic substituents (Fig. 18.13) and one for aliphatic side-chain substituents (Fig. 18.14). The schemes were drawn up by considering the hydrophobicity and electronic factors of various substituents, and are designed such that the optimum substituent can be found as efficiently as possible. They are not meant to be a replacement for a full Hansch analysis, however. Such an analysis would be carried out in due course, once a suitable number of structures have been synthesized.

The Topliss scheme for aromatic substituents (Fig. 18.13) assumes that the lead compound has been tested for biological activity and contains a monosubstituted aromatic ring. The first analogue in the scheme is the 4-chloro derivative, as this derivative is usually easy to synthesize. The chloro substituent is more hydrophobic and electron-withdrawing than hydrogen and therefore, π and σ are positive.

Once the chloro analogue has been synthesized, the biological activity is measured. There are three possibilities. The analogue will have less activity (L), equal activity

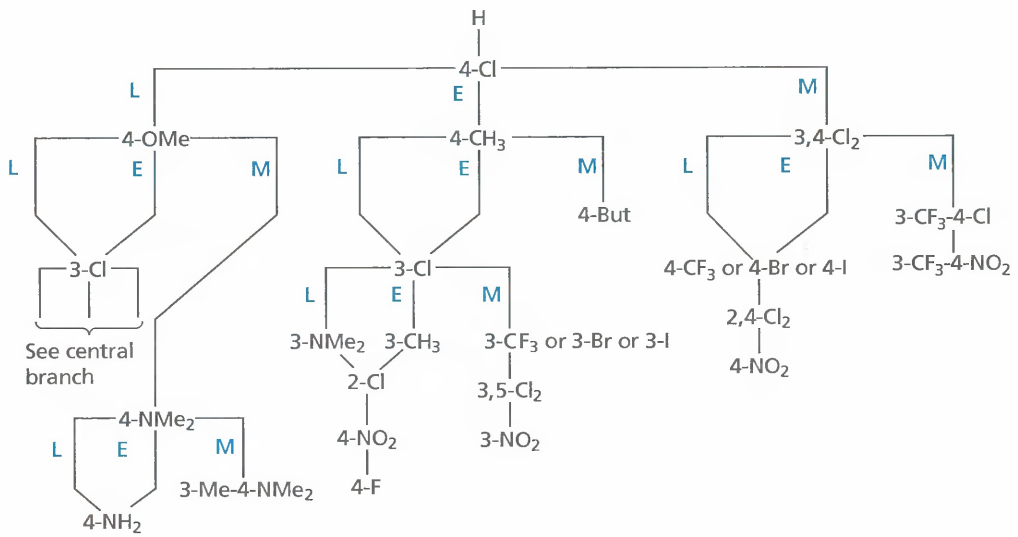


FIGURE 18.13 Topliss scheme for aromatic substituents.

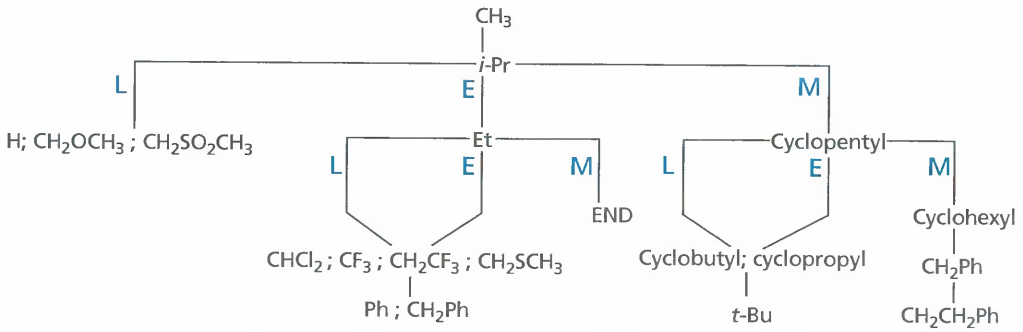


FIGURE 18.14 Topliss scheme for aliphatic side chain substituents.

(E), or more activity (M). The type of activity observed will determine which branch of the Topliss scheme is followed next.

If the biological activity increases, the (M) branch is followed and the next analogue to be synthesized is the 3,4-dichloro-substituted analogue. If, on the other hand, the activity stays the same, then the (E) branch is followed and the 4-methyl analogue is synthesized. Finally, if activity drops, the (L) branch is followed and the next analogue is the 4-methoxy analogue. Biological results from the second analogue now determine the next branch to be followed in the scheme.

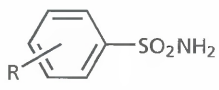
What is the rationale behind this?

Let us consider the situation where the 4-chloro derivative increases in biological activity. The chloro substituent has positive π and σ values, which implies that one or both of these properties are important to biological activity. If both are important, then adding a second chloro group should increase biological activity yet further. If it does, substituents are varied to increase the π and σ values even further. If it does not, then an unfavourable steric interaction or excessive hydrophobicity is indicated. Further modifications then test the relative importance of π and steric factors.

Now consider the situation where the 4-chloro analogue drops in activity. This suggests either that negative π and/or σ values are important to activity or that a *para* substituent is sterically unfavourable. It is assumed that an unfavourable σ effect is the most likely reason for the reduced activity and so the next substituent is one with a negative σ factor (i.e. 4-OMe). If activity improves, further changes are suggested to test the relative importance of the σ and π factors. On the other hand, if the 4-OMe group does not improve activity, it is assumed that an unfavourable steric factor is at work and the next substituent is a 3-chloro group. Modifications of this group would then be carried out in the same way as shown in the centre branch of Fig. 18.13.

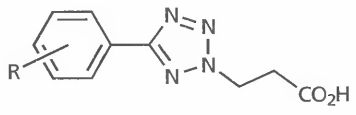
The last scenario is where the activity of the 4-chloro analogue is little changed from the lead compound. This could arise from the drug requiring a positive π value and a negative σ value. As both values for the chloro group are positive, the beneficial effect of the positive π value might be cancelled out by the detrimental effects of a positive σ value. The next substituent to try in that case is the 4-methyl group which has the necessary positive π value and negative σ value. If this still has no beneficial effect, then it is assumed that there is an unfavourable steric interaction at the *para* position and the 3-chloro substituent is chosen next. Further changes continue to vary the relative values of the π and σ factors.

The validity of the Topliss scheme was tested by looking at structure–activity results for various drugs which had been reported in the literature. For example, the biological activities of 19 substituted benzenesulfonamides (Fig. 18.15) have been reported. The second most active compound was the nitro-substituted analogue, which would have been the fifth compound synthesized if the Topliss scheme had been followed.



Order of synthesis	R	Biological activity	High potency
1	H	–	
2	4-Cl	More	
3	3,4-Cl_2	Less	
4	4-Br	Equal	
5	4-NO_2	More	*

FIGURE 18.15 The order of benzenesulfonamide synthesis as directed by the Topliss scheme.



Order of synthesis	R	Biological activity	High potency
1	H	–	
2	4-Cl	Less	
3	4-OMe	Less	
4	3-Cl	More	*
5	3-CF_3	Less	
6	3-Br	More	*
7	3-I	Less	
8	3,5-Cl_2	More	*

FIGURE 18.16 The order of synthesis for substituted aryltetrazolylakonoic acids as directed by the Topliss scheme.

Another example comes from the anti-inflammatory activities of substituted aryltetrazolylalkanoic acids (Fig. 18.16), of which 28 were synthesized. Using the Topliss scheme, 3 out of the 4 most active structures would have been prepared from the first 8 compounds synthesized.

The Topliss scheme for aliphatic side-chains (Fig. 18.14) was set up following a similar rationale to the aromatic scheme, and is used in the same way for side-groups attached to a carbonyl, amino, amide, or similar functional group. The scheme attempts to differentiate only between the hydrophobic and electronic effects of substituents, and not their steric properties. Thus, the substituents involved have been chosen to try to minimize any steric differences. It is assumed that the lead compound has a methyl group. The first analogue suggested is the isopropyl analogue. This has an increased π value and in most cases would be expected to increase activity. It has been found from experience that the hydrophobicity of most lead compounds is less than optimum.

Let us concentrate first of all on the situation where activity increases. Following this branch, a cyclopentyl group is now used. A cyclic structure is used since it has a larger π value, but keeps any increase in steric factor to a minimum. If activity rises again, more hydrophobic substituents are tried. If activity does not rise, then there could be two explanations. Either the optimum hydrophobicity has been passed or there is an electronic effect (σ_I) at work. Further substituents are then used to determine which is the correct explanation.

Let us now look at the situation where the activity of the isopropyl analogue stays much the same. The most likely explanation is that the methyl group and the isopropyl group are on either side of the hydrophobic optimum. Therefore, an ethyl group is used next, since it has an intermediate π value. If this does not lead to an improvement, it is possible that there is an unfavourable electronic effect. The groups used have been electron-donating, and so electron-withdrawing groups with similar π values are now suggested.

Finally, we shall look at the case where activity drops for the isopropyl group. In this case, hydrophobic and/or electron-donating groups could be bad for activity and the groups suggested are suitable choices for further development.

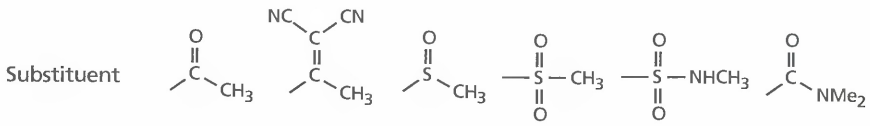
18.6 Bioisosteres

Tables of substituent constants are available for various physicochemical properties. A knowledge of these constants allows the medicinal chemist to identify substituents which may be potential bioisosteres. Thus, the substituents CN, NO_2, and COMe have similar hydrophobic, electronic, and steric factors, and might be interchangeable. Such interchangeability was observed in the development of **cimetidine** and cimetidine analogues (sections 25.2.6 and 25.2.8). The important thing to note is that groups can be bioisosteric in some situations, but not others. Consider for example the table shown in Fig. 18.17.

This table shows physicochemical parameters for six different substituents. If the most important physicochemical parameter for biological activity is σ_p, then the $COCH_3$ group (0.50) would be a reasonable bioisostere for the $SOCH_3$ group (0.49). If, on the other hand, the dominant parameter is π, then a more suitable bioisostere for $SOCH_3$ (−1.58) would be SO_2CH_3 (−1.63).

18.7 Free–Wilson approach

In the Free–Wilson approach to QSAR, the biological activity of a parent structure is measured then compared with the activities of a range of substituted analogues. An equation is then derived which relates biological activity



Substituent						
π	−0.55	0.40	−1.58	−1.63	−1.82	−1.51
σ_p	0.50	0.84	0.49	0.72	0.57	0.36
σ_m	0.38	0.66	0.52	0.60	0.46	0.35
MR	11.2	21.5	13.7	13.5	16.9	19.2

FIGURE 18.17 Physicochemical parameters for six substituents.

to the presence or otherwise of particular substituents (X_1–X_n).

$$\text{Activity} = k_1X_1 + k_2X_2 + k_3X_3 + \ldots\ldots\ldots\ldots + k_nX_n + Z.$$

In this equation, X_n is defined as an **indicator variable** and is given the value 1 or 0, depending on whether the substituent (n) is present or not. The contribution that each substituent makes to the activity is determined by the value of k_n. Z is a constant representing the average activity of the structures studied.

Since the approach considers the overall effect of a substituent to biological activity rather than its various physicochemical properties, there is no need for physicochemical constants and tables, and the method only requires experimental measurements of biological activity. This is particularly useful when trying to quantify the effect of unusual substituents that are not listed in the tables, or when quantifying specific molecular features which cannot be tabulated.

The disadvantage in the approach is the large number of analogues which have to be synthesized and tested to make the equation meaningful. For example, each of the terms k_nX_n refers to a specific substituent at a specific position in the parent structure. Therefore, analogues would not only have to have different substituents, but have them at different positions of the skeleton as well.

Another disadvantage is the difficulty in rationalizing the results and explaining why a substituent at a particular position is good or bad for activity. Finally, the effects of different substituents may not be additive. There may be intramolecular interactions which affect activity.

Nevertheless the idea of using indicator variables can be useful in certain situations and they can also be used as part of a Hansch equation. An example of this can be seen in the later case study (section 18.9).

18.8 Planning a QSAR study

When starting a QSAR study it is important to decide which physicochemical parameters are going to be studied and to plan the analogues such that the parameters under study are suitably varied. For example, it would be pointless to synthesize analogues where the hydrophobicity and steric volume of the substituents are correlated, if these two parameters are to go into the equation.

It is also important to make enough structures to make the results statistically meaningful. As a rule of thumb, five structures should be made for every parameter studied. Typically, the initial QSAR study would involve the two parameters π and σ, and possibly E_s. Craig plots could be used in order to choose suitable substituents.

Certain substituents are worth avoiding in the initial study, as they may have properties other than those being studied. For example, it is best to avoid substituents that might ionize (CO_2H, NH_2, SO_3H), and groups that might easily be metabolized (e.g. esters or nitro groups).

If there are two or more substituents, then the initial equation usually considers the total π and σ contribution.

As more analogues are made, it is often possible to consider the hydrophobic and electronic effect of substituents at specific positions of the molecule. Furthermore, the electronic parameter σ can be split into its inductive and resonance components (*F* and *R*). Such detailed equations may show up a particular localized requirement for activity. For example, a hydrophobic substituent may be favoured in one part of the skeleton, while an electron-withdrawing substituent is favoured at another. This in turn gives clues about the binding interactions involved between drug and receptor.

18.9 Case Study

An example of how a QSAR equation can become more specific as a study develops is demonstrated from work carried out on the anti-allergic activity of a series of pyranenamines (Fig. 18.18). In this study, substituents were varied on the aromatic ring, and the remainder of the molecule was kept constant. Nineteen compounds were synthesized and the first QSAR equation was

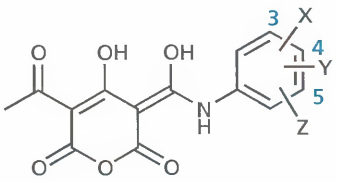


FIGURE 18.18 Structure of pyranenamines.

obtained by considering π and σ:

$$\log\left(\frac{1}{C}\right) = -0.14\Sigma\pi - 1.35(\Sigma\sigma)^2 - 0.72$$
$$(n\ 19,\ r^2\ 0.48,\ s\ 0.47,\ F_{2,16}\ 7.3)$$

where Σπ and Σσ are the total π and σ values for all the substituents present.

The negative coefficient for the π term shows that activity is inversely proportional to hydrophobicity, which is quite unusual. The $(\Sigma\sigma)^2$ term is also quite unusual. It was chosen because there was no simple relationship between activity and σ. In fact, it was observed that activity decreased if the substituent was electron-withdrawing *or* electron-donating. Activity was best with neutral substituents. To take account of this, the $(\Sigma\sigma)^2$ term was introduced. As the coefficient in the equation is negative, activity is lowered if σ is anything other than zero.

A further range of compounds was synthesized with hydrophilic substituents to test this equation, making a total of 61 structures. This resulted in the following inconsistencies.

- The activities for the substituents 3-NHCOMe, 3-NHCOEt, and 3-NHCOPr were all similar, but according to the equation these activities should have dropped as the alkyl group got larger, due to increasing hydrophobicity.
- Activity was greater than expected if there was a substituent such as OH, SH, NH_2, or NHCOR at position 3, 4, or 5.
- The substituent $NHSO_2R$ was bad for activity.
- The substituents 3,5-$(CF_3)_2$ and 3,5-$(NHCOMe)_2$ had much greater activity than expected.
- An acyloxy group at the 4-position resulted in an activity five times greater than predicted by the equation.

These results implied that the initial equation was too simple and that properties other than π and σ were important to activity. At this stage, the following theories were proposed to explain the above results.

- The similar activities for 3-NHCOMe, 3-NHCOEt, and 3-NHCOPr could be due to a steric factor. The substituents had increasing hydrophobicity, which is bad for activity, but they were also increasing in size and it was proposed that this was good for activity. The most likely explanation is that the size of the substituent is forcing the drug into the correct orientation for optimum receptor interaction.
- The substituents that unexpectedly increased activity when they were at positions 3, 4, or 5 are all capable of hydrogen bonding. This suggests an important hydrogen bonding interaction with the receptor. For some reason, the $NHSO_2R$ group is an exception, which implies there is some other unfavourable steric or electronic factor peculiar to this group.
- The increased activity for 4-acyloxy groups was explained by suggesting that these analogues are acting as prodrugs. The acyloxy group is less polar than the hydroxyl group and so these analogues would be expected to cross cell membranes and reach the receptor more efficiently than analogues bearing a free hydroxyl group. At the receptor, the ester group could be hydrolysed to reveal the hydroxyl group, which would then take part in hydrogen bonding with the receptor.
- The structures having substituents 3,5-$(CF_3)_2$ and 3,5-$(NHCOMe)_2$ are the only di-substituted structures where a substituent at position 5 has an electron-withdrawing effect, so this feature was also introduced into the next equation.

The revised QSAR equation was as follows:

$$\log\left(\frac{1}{C}\right) = -0.30\Sigma\pi - 1.5(\Sigma\sigma)^2 + 2.0(F\text{-}5) + 0.39(345\text{-}HBD) - 0.63(NHSO_2) + 0.78(M\text{-}V) + 0.72(4\text{-}OCO) - 0.75$$
$$(n\ 61,\ r^2\ 0.77,\ s\ 0.40,\ F_{7,53}\ 25.1)$$

The π and σ parameters are still present, but a number of new parameters have now been introduced.

- The *F*-5 term represents the inductive effect of a substituent at position 5. The coefficient is positive and large, showing that an electron-withdrawing group substantially increases activity. However, only 2 compounds of the 61 synthesized had a 5-substituent, so there might be quite an error in this result.
- The *M-V* term represents the volume of any *meta* substituent. The coefficient is positive, indicating that substituents with a large volume at the *meta* position increase activity.
- The advantage of having hydrogen bonding substituents at position 3, 4, or 5 is accounted for by including a hydrogen bonding term (*345-HBD*). The value of this term depends on the number of hydrogen bonding substituents present. If one such group is present, the *345-HBD* term is 1. If two such groups are present, the parameter is 2. Therefore, for each hydrogen bonding substituent present at positions 3, 4, or 5, log (1/*C*)

increases by 0.39. This sort of term is known as an **indicator variable**, which is the basis of the **Free–Wilson** approach described earlier. There is no tabulated value one can use for a hydrogen bonding substituent and so the contribution that this term makes to the biological activity is determined by the value of k, and whether the relevant group is present or not. Indicator variables were also used for the following terms.

- The $NHSO_2$ term was introduced because this group was bad for activity despite being capable of hydrogen bonding. The negative coefficient indicates the drop in activity. A figure of 1 is used for any $NHSO_2R$ substituent present, resulting in a drop of activity by 0.63.
- The *4-OCO* term is 1 if an acyloxy group is present at position 4, and so log (1/*C*) is increased by 0.72 if this is the case.

A further 37 structures were synthesized to test steric and *F*-5 parameters as well as exploring further groups capable of hydrogen bonding. Since hydrophilic substituents were good for activity, a range of very hydrophilic substituents were also tested to see if there was an optimum value for hydrophilicity. The results obtained highlighted one more anomaly, in that two hydrogen bonding groups *ortho* to each other were bad for activity. This was attributed to the groups hydrogen bonding with each other rather than to the receptor. A revised equation was obtained as follows:

$$\log\left(\frac{1}{C}\right) = -0.034(\Sigma\pi)^2 - 0.33(\Sigma\pi) + 4.3(F\text{-}5) + 1.3(R\text{-}5) - 1.7(\Sigma\sigma)^2 + 0.73(345\text{-}HBD) - 0.86(HB\text{-}INTRA) - 0.69(NHSO_2) + 0.72(4\text{-}OCO) - 0.59$$

$$(n\ 98,\ r^2\ 0.75,\ s\ 0.48,\ F_{9,88}\ 28.7)$$

The main points of interest from this equation are as follows.

- Increasing the hydrophilicity of substituents allowed the identification of an optimum value for hydrophobicity ($\Sigma\pi = -5$) and introduced the $(\Sigma\pi)^2$ parameter into the equation. The value of -5 is remarkably low and indicates that the receptor site is hydrophilic.
- As far as electronic effects are concerned, it is revealed that the resonance effects of substituents at the 5-position also have an influence on activity.
- The unfavourable situation where two hydrogen bonding groups are *ortho* to each other is represented by the *HB-INTRA* parameter. This parameter is given the value 1 if such an interaction is possible and the negative constant (-0.86) shows that such interactions decrease activity.
- It is interesting to note that the steric parameter is no longer significant and has disappeared from the equation.

The compound having the greatest activity has two $NHCOCH(OH)CH_2OH$ substituents at the 3- and 5-positions and is 1000 times more active than the original lead compound. The substituents are very polar and are not ones that would normally be used. They satisfy all the requirements determined by the QSAR study. They are highly polar groups which can take part in hydrogen bonding. They are *meta* with respect to each other, rather than *ortho*, to avoid undesirable intramolecular hydrogen bonding. One of the groups is at the 5-position and has a favourable *F*-5 parameter. Together the two groups have a negligible $(\Sigma\sigma)^2$ value. Such an analogue would certainly not have been obtained by trial and error and this example demonstrates the strengths of the QSAR approach.

All the evidence from this study suggests that the aromatic ring of this series of compounds is fitting into a hydrophilic pocket in the receptor which contains polar groups capable of hydrogen bonding.

It is further proposed that a positively charged residue such as arginine, lysine, or histidine might be present in the pocket which could interact with an electronegative substituent at position 5 of the aromatic ring (Fig. 18.19).

This example demonstrates that QSAR studies and computers are powerful tools in medicinal chemistry. However, it also shows that the QSAR approach is a long way from replacing the human factor. One cannot put a series of facts and figures into a computer and expect it to magically produce an instant explanation of how a drug works. The medicinal chemist still has to interpret results, propose theories, and test those theories by incorporating the correct parameters into the QSAR equation. Imagination and experience still count for a great deal.

KEY POINTS

- QSAR relates the physicochemical properties of a series of drugs to their biological activity by means of a mathematical equation.
- The commonly studied physicochemical properties are hydrophobicity, electronic factors and steric factors.
- The partition coefficient is a measure of a drug's overall hydrophobicity. Values of log *P* are used in QSAR equations, with larger values indicating greater hydrophobicity.
- The substituent hydrophobicity constant is a measure of the hydrophobic character of individual substituents. The value is different for aliphatic and aromatic substituents and is only directly relevant to the class of structures from which the values were derived. Positive values represent substituents more hydrophobic than hydrogen; negative values represent substituents more hydrophilic than hydrogen.

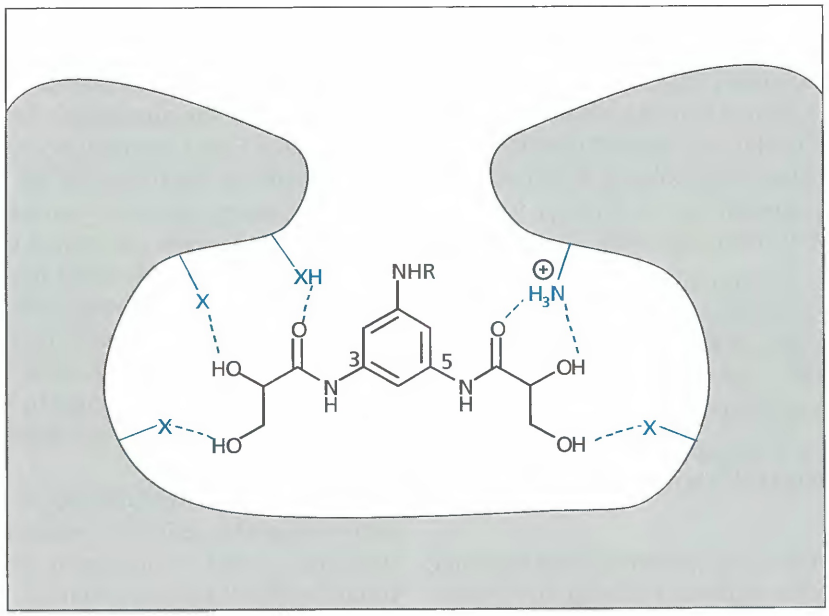


FIGURE 18.19 Hypothetical binding interactions between a pyranenamine and the target binding site.

- The Hammett substituent constant is a measure of how electron-withdrawing or electron-donating an aromatic substituent is. It is measured experimentally and is dependent on the relative position of the substituent on the ring. The value takes into account both inductive and resonance effects.
- The parameters *F* and *R* are constants quantifying the inductive and resonance effects of an aromatic substituent.
- The inductive effect of aliphatic substituents can be measured experimentally and tabulated.
- Steric factors can be measured experimentally or calculated using physical parameters or computer software.
- The Hansch equation is a mathematical equation that relates a variety of physicochemical parameters to biological activity for a series of related structures.
- The Craig plot is a visual comparison of two physicochemical properties for a variety of substituents. It facilitates the choice of substituents for a QSAR study such that the values of each property are not correlated.
- The Topliss scheme is used when structures can only be synthesized and tested one at a time. The scheme is a guide to which analogue should be synthesized next in order to get good activity. There are different schemes for aromatic and aliphatic substituents.
- Indicator variables are used when there are no tabulated or experimental values for a particular property or substituent. The Free–Wilson approach to QSAR only uses indicator variables, whereas the Hansch approach can use a mixture of indicator variables and physicochemical parameters.

18.10 3D QSAR

In recent years, a method known as 3D QSAR has been developed in which the three-dimensional properties of a molecule are considered as a whole rather than by considering individual substituents or moieties. This has proved remarkably useful in the design of new drugs. Moreover, the necessary software and hardware are readily affordable and relatively easy to use. The philosophy of 3D QSAR revolves around the assumption that the most important features about a molecule are its overall size and shape, and its electronic properties.

If these features can be defined, then it is possible to study how they affect biological properties. There are several approaches to 3D QSAR, but the method which has gained ascendancy was developed by the company Tripos and is known as **CoMFA (Comparative Molecular Field Analysis)**. CoMFA methodology is based on the assumption that drug-receptor interactions are non-covalent and that changes in biological activity correlate with the changes in the steric and/or electrostatic fields of the drug molecules.

18.10.1 Defining steric and electrostatic fields

The steric and electrostatic fields surrounding a molecule can be measured and defined using the grid and probe method described in section 17.7.5. This can be repeated

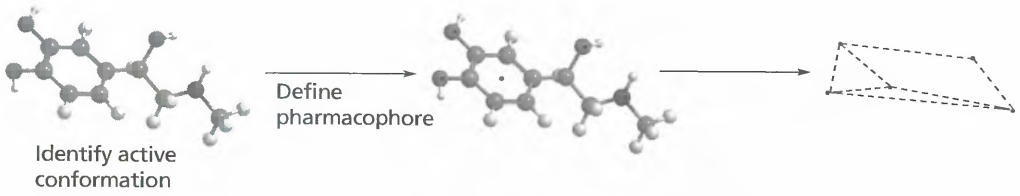


FIGURE 18.20 Identification of the active conformation and pharmacophore.

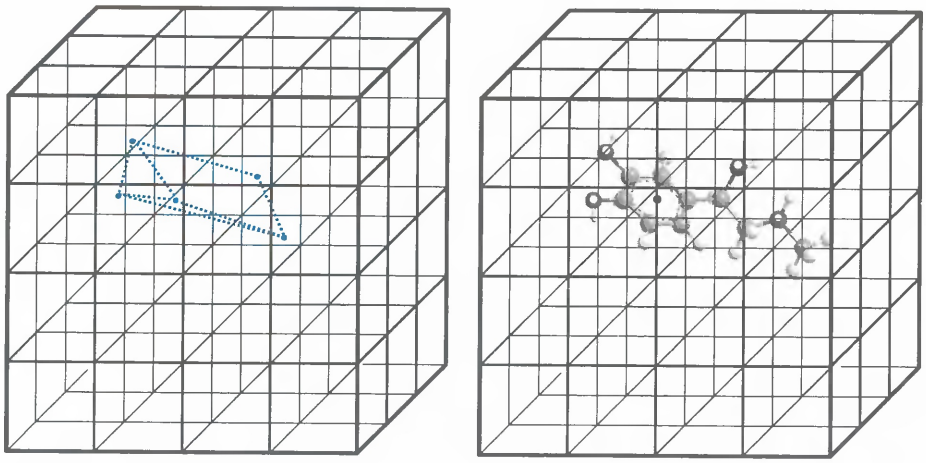


FIGURE 18.21 Positioning a pharmacophore and molecules into a lattice.

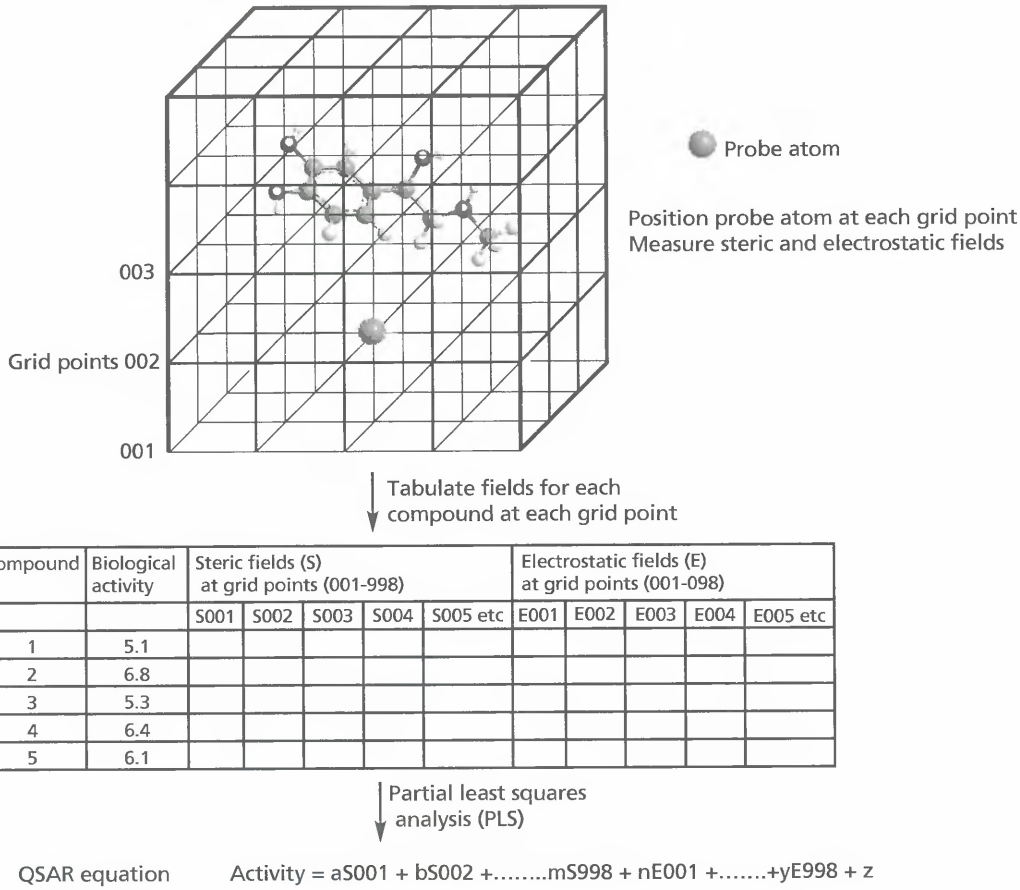
for all the molecules in the 3D QSAR study, but it is crucial that the molecules are all in their **active conformation**, and that they are all positioned within the grid in exactly the same way. In other words, they must all be correctly aligned. Identifying a **pharmacophore** (section 13.2) that is common to all the molecules can assist in this process (Fig. 18.20).

The pharmacophore is placed into the grid and its position is kept constant such that it acts as a reference point when positioning each molecule into the lattice. For each molecule studied, the active conformation and pharmacophore is identified and then the molecule is placed into the lattice such that its pharmacophore matches the reference pharmacophore (Fig. 18.21). Once a molecule has been placed into the lattice, the steric and electrostatic fields around it are measured as described in section 17.7.5.

18.10.2. Relating shape and electronic distribution to biological activity

Defining the steric and electrostatic fields of a series of molecules is relatively straightforward and is carried out automatically by the software program. The next stage is to relate these properties to the biological activity of the molecules. This is less straightforward and differs significantly from traditional QSAR. In traditional QSAR, there are relatively few variables involved. For example, if we consider hydrophobicity, π, σ and a size factor for each molecule, then we have four variables per molecule to compare against biological activity. With 100 molecules in the study, there are far more molecules than variables, and it is possible to come up with an equation relating variables to biological activity as previously described.

In 3D QSAR, the variables for each molecule are the calculated steric and electronic interactions at a couple of thousand lattice points. With 100 molecules in the study, the number of variables now far outweighs the number of structures, and it is not possible to relate these to biological potency by the standard multiple linear regression analysis described in section 18.1. A different statistical procedure has to be followed using a technique called **partial least squares** (PLS). Essentially, it is an analytical computing process which is repeated over and over again (iterated) to try to find the best formula relating biological property against the various variables. As part of the process, the number of variables is reduced as the software filters out those which are clearly unrelated to biological activity.



Compound	Biological activity	Steric fields (S) at grid points (001-998)					Electrostatic fields (E) at grid points (001-098)				
		S001	S002	S003	S004	S005 etc	E001	E002	E003	E004	E005 etc
1	5.1										
2	6.8										
3	5.3										
4	6.4										
5	6.1										

FIGURE 18.22 Measuring steric and electronic fields.

An important feature of the analysis is that a structure is deliberately left out as the computer strives to form some form of relationship. Once a formula has been defined, the formula is tested against the structure which was left out. This is called **cross-validation** and tests how well the formula predicts the biological property for the molecule which was left out. The results of this are fed back into another round of calculations, but now the structure which was left out is included in the calculations and a different structure is left out. This leads to a new improved formula which is once again tested against the compound that was left out, and so the process continues until cross validation has been carried out against all the structures.

At the end of the process, the final formula is obtained (Fig. 18.22). The predictability of this final equation is quantified by the **cross-validated correlation coefficient** r^2, which is usually referred to as q^2. In contrast to normal QSAR, where r^2 should be greater than 0.8, values of q^2 greater than 0.3 are considered significant. It is more useful, though, to give a graphical representation showing which regions around the molecule are important to biological activity on steric or electronic grounds. Therefore, a steric map shows a series of coloured contours indicating beneficial and detrimental steric interactions around a representative molecule from the set of molecules tested (Figure 18.23). A similar contour map is created to illustrate beneficial or detrimental electrostatic interactions.

An example of a 3D QSAR study is described in Case Study 6 and in the Case Study described in section 18.10.6.

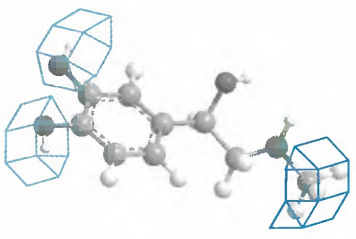


FIGURE 18.23 Definition of favourable and unfavourable interactions around a representative molecule.

18.10.3 Advantages of CoMFA over traditional QSAR

Some of the problems involved with a traditional QSAR study include the following:

- Only molecules of similar structure can be studied.
- The validity of the numerical descriptors is open to doubt. These descriptors are obtained by measuring reaction rates and equilibria constants in model reactions and are listed in tables. However, separating one property from another is not always possible in experimental measurement. For example, the Taft steric factor is not purely a measure of the steric factor because the measured reaction rates used to define it are also affected by electronic factors. Also, the *n*-octanol/water partition coefficients which are used to measure log *P* are known to be affected by the hydrogen bonding character of molecules.
- The tabulated descriptors may not include entries for unusual substituents.
- It is necessary to synthesize a range of molecules where substituents are varied in order to test a particular property (e.g. hydrophobicity). However, synthesizing such a range of compounds may not be straightforward or feasible.
- Traditional QSAR equations do not directly suggest new compounds to synthesize.

These problems are avoided with CoMFA, which has the following advantages:

- Favourable and unfavourable interactions are represented graphically by 3D contours around a representative molecule. A graphical picture such as this is easier to visualize than a mathematical formula.
- In CoMFA, the properties of the test molecules are calculated individually by computer program. There is no reliance on experimental or tabulated factors. There is no need to confine the study to molecules of similar structure. As long as one is confident that all the compounds in the study share the same pharmacophore and interact in the same way with the target, they can all be analysed in a CoMFA study.
- The graphical representation of beneficial and non-beneficial interactions allows medicinal chemists to design new structures. For example, if a contour map shows a favourable steric effect at one particular location, this implies that the target binding site has space for further extension at that location. This may lead to further favourable receptor–drug interactions.
- Both traditional and 3D QSAR can be used without needing to know the structure of the biological target.

18.10.4 Potential problems of CoMFA

There are several potential problems in using CoMFA:

- It is important to know the active conformation for each of the molecules in the study. Identifying the active conformation is easy for rigid structures such as steroids, but it is more difficult for flexible molecules that are capable of several bond rotations. Therefore, it is useful to have a conformationally restrained analogue which is biologically active and which can act as a guide to the likely active conformation. More flexible molecules can then be constructed on the computer with the conformation most closely matching that of the more rigid analogue. If the structure of the target binding site is known, this can be useful in deciding the likely active conformations of molecules.
- Each molecule in the study must be correctly positioned in the grid so that it is properly aligned with respect to all the other molecules. A common pharmacophore can be used to aid this process as described earlier. However, it may be difficult to identify the pharmacophore in some molecules. In that case, a pharmacophore mapping exercise could be carried out (section 17.11). This is likely to be successful if there are some rigid active compounds amongst the compounds being studied. An alternative method of alignment is to align the molecules based on their structural similarity. This can be done automatically using what is known as 'topomer' methodology.
- One has to be careful to ensure that all the compounds in the study interact with the target in similar ways. For example, a 3D QSAR study on all possible **acetylcholinesterase** inhibitors is doomed to failure. In the first place, the great diversity of structures involved makes it impossible to align these structures in an unbiased way or to generate a 3D pharmacophore. Second, the various inhibitors do not interact with the target enzyme in the same way. X-ray crystallographic studies of enzyme-inhibitor complexes show that the inhibitors **tacrine**, **edrophonium** and **decamethonium** all have different binding orientations in the active site.
- 3D QSAR provides a summary of how structural changes in a drug affect biological activity, but it is dangerous to assume too much. For example, a 3D QSAR model may show that increasing the bulk of the molecule at a particular location increases activity. This might suggest that there is an accessible hydrophobic pocket allowing extra binding interactions. On the other hand, it is possible that the extra steric bulk causes the molecule to bind in a different orientation from the other molecules in the analysis, and that this is the reason for the increased activity.

- It has been found that slightly different orientations of the grid from one study to another can produce different results for the same set of compounds.

18.10.5 Other 3D QSAR methods

CoMFA continues to be the most popular program for studies into 3D QSAR, but it does suffer a number of disadvantages as described above. Two of the more serious problems are the spurious results that can be obtained if compounds are not properly aligned, or if the orientation of the grid box is slightly different between studies. Users of the program can also choose different spacings between the grid points, and this can result in poor results if too course or too fine a grid is used. The method is also computationally expensive requiring a lot of calculations for each molecule in the study, and so there is a huge memory requirement requiring powerful computers.

Other 3D QSAR programs have been developed in an attempt to address some of these issues. Many have been designed in order to study large numbers of structures, as well as having some predictive power of biological activity. Each of these methods can work well in particular situations, and the following are some examples.

HINT

HINT is a program which can be used alongside CoMFA to measure a hydrophobic field. In traditional QSAR, the hydrophobic factor is very important and most equations include a log *P* value. Surprisingly, the calculation of a hydrophobic factor in CoMFA is not so crucial, and most successful studies have been based on steric and electrostatic factors alone.

CoMSIA

The **CoMSIA** (comparative molecular similarity index analysis) approach includes hydrogen bonding and hydrophobic fields as well as steric and electrostatic fields. The program has the advantage that it is less sensitive to small changes in the alignment of compounds or in the orientation of the grid. This is achieved by using modified equations to measure steric interactions. In CoMFA, the equations used to calculate the steric interactions between a probe atom and the molecule are such that the interaction energy increases dramatically when the probe atom gets too close to the surface of the molecule. A cut off maximum value for the interaction has to be introduced and this results in the inaccuracies that can arise from poor alignment of the molecules or different orientations of the grid. In CoMSIA, a 'smoother' equation is used to reduce this problem. Despite the advantages that CoMSIA offer over CoMFA, there are instances where CoMFA performs better, and so it is not possible to say that CoMSIA is better than CoMFA.

CoMASA

CoMASA (comparative molecular active site analysis) is a method of carrying out 3D QSAR studies more quickly and with less calculations. Instead of using a complete lattice, a smaller number of representative points are selected based on the molecules used in the study. In order to identify the representative points, the molecules must first be superimposed using programs that recognise similar molecular features. The positions of all the atoms of all the molecules are then identified. If two atoms from different molecules are close together in 3D space, a pseudoatom is created midway between them which will represent both. In this way, a series of points representing all the atoms of the molecules present is set up.

The electrostatic and steric fields are then measured by placing probe atoms at these representative points. This radically reduces the number of calculations that need to be carried out. For example, in a study on steroids, the CoMSIA program used 7200 lattice points, whereas CoMASA used only 92 points. The number of calculations required in CoMASA is greatly reduced and the program can be used on a standard PC.

Another advantage of the method is the way that the results are presented. In CoMFA, favourable or unfavourable interactions are represented by contours surrounding the molecule. With CoMASA, favourable or unfavourable interactions can be represented by spheres generated at the representative points used in the study. Since these points were generated from the molecules themselves, they are close in space to the specific atoms of the molecules, making it easier to identify molecular features that are beneficial or detrimental. Moreover, the results can be 'sharper' and identify a small volume in space that might be beneficial in steric terms.

3D QSAR methods based on intrinsic molecular properties

Some 3D QSAR programs use the intrinsic molecular properties of compounds, rather than using a probe to measure the property fields surrounding them. Four examples are SOMFA, HASL, CoMMA and MS-Whim.

SOMFA (self organizing molecular field analysis) is designed to give greater prominence to the most active and the least active compounds in the study. A 'mean centred activity' is measured for each molecule in the training set. This represents the difference between the actual activity of the molecule and the mean activity for the whole set. This is in turn identifies the most active and the least active structures.

Each molecule is then placed into a grid, and it is important that they are properly aligned. Grid points outside the van der Waals envelope of a molecule are given

a value of 0, as opposed to those inside which are given a value of 1. Electrostatic potential values at the grid points are calculated from the partial charges of the atoms in the molecule. The grid point values for a particular molecule are then multiplied by the mean centred activity for that molecule. This weights the grid points in order to give greater prominence to the most active and least active molecules, and acts as a form of descriptor filtering which tends to increase the quality of the correlation. The grids for each molecule are combined to give master grids for each property studied. Consequently, grid points with high positive values for a particular property are beneficial and can be used to highlight pharmacophores. Grid points with high negative values for a particular property reveal detrimental features. QSAR equations can be determined relating activity to the properties studied. The activity of a molecule which has not been used in the training set can be predicted by placing it into the grid and identifying which features match the important grid points, then adding up the values of the grid points.

HASL (hypothetical active site lattice) is a program which uses the structures and biological activities of a group of molecules to set up a lattice where the grid points are given binding values for different atoms interacting with a hypothetical binding site. The molecules in the study are first superimposed on each other and then they are used to create a lattice of regularly spaced grid points representing the positions in space occupied by the atoms of the various molecules. Calculations are then carried out to give each grid point on the lattice a binding value, based on the biological activity of the molecules used in the study. Each grid point represents the binding influence of a particular atom. When each molecule is studied, its atoms will correspond to particular grid points. The binding value for these grid points is calculated for the whole set of molecules in the study to ensure that the total matches the activity of each molecule concerned. Once binding values have been calculated from the test set of molecules, it can be used in a predictive manner. New molecules can be superimposed on the test set and placed into the lattice. The predicted biological activity will then be the sum of the binding values of the grid points corresponding to the atoms present.

CoMMA (comparative molecular moment analysis) is a method that uses descriptors to define the spatial and electrostatic properties of a molecule. Moments of inertia are the descriptors used to define shape, while the dipole and quadrupole moments are used to represent electrostatic properties. In total, 13 different descriptors are required to describe these features in each molecule. Three moments of inertia are required to represent shape. The magnitude of the dipole moment and the quadrupole moment define the charge distribution in the molecule. The remaining 8 descriptors define the relationship between shape and charge distribution. The 13 descriptors are inherent to the structure of the molecule and provide 3D information about it. They are also independent of where the molecule is positioned in space. This rules out the necessity to superimpose and align molecules. Since the descriptors are inherent features of the molecule as a whole, they can be viewed in much the same way as the descriptors used in conventional QSAR. Indeed, CoMMA has been described as being a 'half way house' between conventional QSAR and 3D QSAR. 3D QSAR equations can be drawn up relating the molecular moments to biological activity. A drawback of the method is the fact that the molecules are viewed as sets of molecular moments rather than atoms and so one loses the 'feel' for the molecule, as well as some of the fine detail that can be obtained from other methods of 3D QSAR.

The concept behind the **MS-Whim** (molecular surface weighted holistic invariant molecular) approach is similar to CoMMA and also avoids the need to align molecules. The geometry, size and physicochemical properties of each molecule are defined mathematically using a set of 12 indices based on the properties of the molecular surface. This generates 36 molecular descriptors. The QSAR analysis then relates these descriptors to biological activity. The method is faster than CoMFA and can be used to predict the activity of new molecules, but it does not give information of specific interactions that might occur between the molecule and its binding site.

Methods based on creating hypothetical pseudoreceptors

Some 3D QSAR programs are non-grid based and model hypothetical pseudoreceptors. This involves the creation of a Receptor Surface Model (RSM) or a Genetically Evolved Receptor Model (GERM). In the RSM method, a receptor surface is constructed based on the shapes and the properties of the most active compounds in the study. The pseudoreceptor is then used as a template to calculate various 3D energy descriptors.

In the GERM method, the molecules in the study are superimposed and probe atoms are then placed on the surface shell of the superimposed ligands. A genetic algorithm is then used to identify the type and location of the probe atoms having the best interactions with the ligands. These are then used to define the surface of the binding site and its binding characteristics. The **Quasar** program (quasi-atomistic SAR) is an extension of this approach. The fields that are calculated take into account a variety of interactions such as hydrophobic interactions, hydrogen bonding and salt bridges. The flexibilities of the receptor and the ligands are also taken into account.

The WeP (weighted probe interaction energies) method is a newer method. It assumes that certain regions of the binding site play more important roles to binding than others. The method is similar to Quasar but differs in one important feature. In the Wep method, it is the importance of each surface probe that is to be optimised while in other methods it is the atom type or property to be assigned to each probe. This is done by giving each probe a weighting factor based on its relative degree of contribution to the binding energy. The procedure first superimposes all the molecules. Probe atoms are then constructed around the surface of the superimposed molecules. Each probe atom is the size of a methyl group with a partial charge that is complementary to the average electrostatic potential of the aligned ligands. Interaction energies between the probes and the ligands are then calculated for van der Waals and electrostatic interactions. A genetic algorithm is then used to optimise the weightings of the probes. The chromosome used encodes the weightings of each probe. In other words, the weightings of each probe are randomly generated and the genetic process is carried out to find the best weightings for the different probes. This identifies the probes that have a significant weighting. These are then taken for a second round of optimisation and the surface probes with non-zero average weight values are labelled as 'surviving' probes, and are used to define the hypothetical pseudoreceptor. In some cases, this can mean that only 10% of the original probes survive. It has been noted that areas lacking probes correspond to regions that are solvent accessible.

GRIND (grid independent descriptors) makes use of a grid and probe atoms to calculate various interactions that might be possible between the grid points and a molecule. The measured interactions are then used to define areas around each molecule that could be potential binding regions for a hypothetical receptor. A filtering process takes place such that the most important binding regions are identified. This is done independently for each molecule and so there is no need to align the molecules for the study. The results that are obtained can then be used to define the hypothetical binding site.

KEY POINTS

- CoMFA is an example of a 3D QSAR program that measures steric and electrostatic fields round a series of structures and relates that to biological activity.
- A comparison of the steric and electrostatic fields for different molecules against their biological activity allows the definition of steric and electrostatic interactions that are favourable and unfavourable for activity. These can be displayed visually as contour lines.
- It is necessary to define the active conformation and pharmacophore for each molecule in a CoMFA study. Alignment of the molecules is crucial.
- Unlike conventional QSAR studies, molecules of different structural classes can be compared if they share the same pharmacophore.
- 3D QSAR does not depend on experimentally measured parameters.
- A variety of different 3D QSAR programs have been developed.

18.10.6 Case Study: inhibitors of tubulin polymerization

Colchicine (Fig. 18.24) is a lead compound for agents which act as inhibitors of tubulin polymerization (section 2.7.1) and which might be useful in the treatment of arthritis. Other lead compounds have been discovered which bind to tubulin at the same binding site and so a study was carried out to compare the various structural classes interacting in this way. In this 3D QSAR study, 104 such agents were tested, belonging to 4 distinct families of compounds (Fig. 18.25); 51 compounds were used as a 'training set' for the analysis itself and 53 were used as a 'testing set' to test the predictive value of the results. Both sets contained a mixture of structural classes having both low and high activity.

The first task was to work out how to align these different classes of molecule. Colchicine is the most rigid of the four and also has a high affinity for tubulin. Therefore, it was chosen as the template on to which the other structures would be aligned. The relevant pharmacophore in colchicine was identified as the two aromatic rings. Molecular modelling was now carried out on each of the remaining structures to generate various conformations. Each conformation was compared with colchicine to find the one that would allow the pharmacophores in each structure to be aligned. This was then identified as the active conformation.

Once the active conformations for each structure had been identified, they were fitted into the lattice of grids previously described such that each structure was

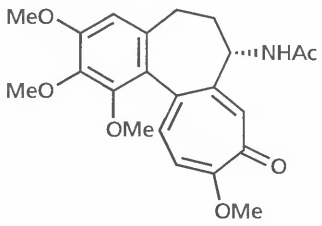


FIGURE 18.24 Colchicine.

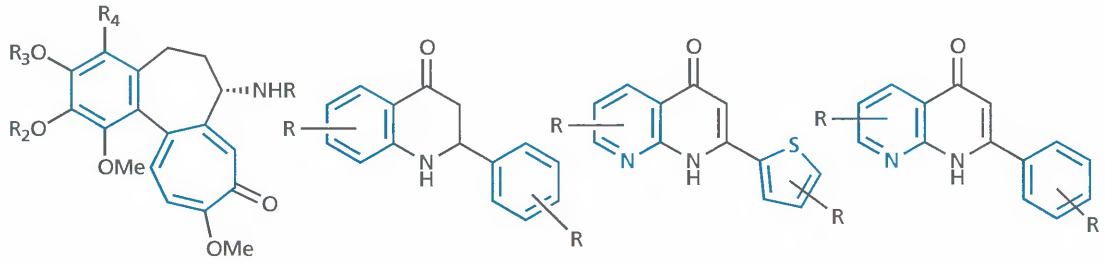


FIGURE 18.25 Structural classes used in the 3D QSAR study.

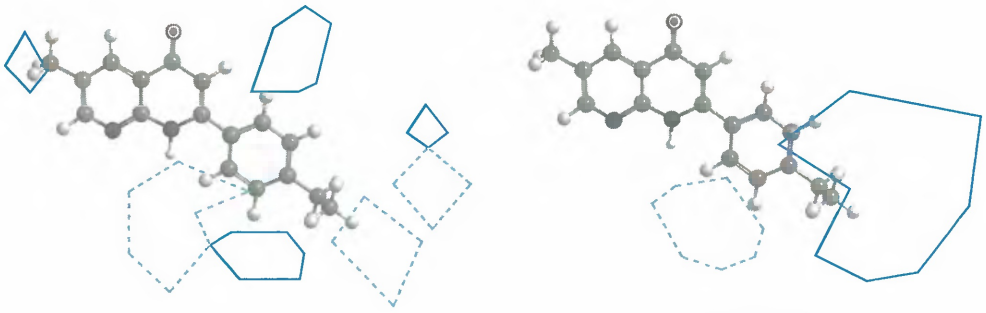


FIGURE 18.26 Results of the 3D QSAR analysis ($q^2 = 0.637$).

properly aligned. The steric and electrostatic fields round each molecule were calculated using a probe atom, then the 3D QSAR analysis carried out to relate the fields to the measured biological activity.

The results of the 3D QSAR analysis are summarized as contour lines round a representative molecule (Fig. 18.26). For the steric interactions, solid contours represent fields which are favourable for activity, and the dashed lines show fields which are unfavourable. For the electrostatic interactions, solid lines are regions where positive charged species improve affinity, and dashed lines indicate regions where negatively charged groups are favourable.

The results showed that introducing steric bulk round the aromatic ring is more crucial to activity than introducing steric bulk round the bicyclic system. Based on this evidence, the structure shown in Fig. 18.27 was synthesized. The predicted value of pIC_{50} for the compound was 5.62. The actual value was in close agreement at 6.04.

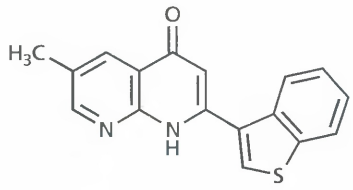


FIGURE 18.27 Novel agent designed on the basis of the 3D QSAR study.

($pIC_{50} = -\log [IC_{50}]$ where IC_{50} is the concentration of inhibitor required to produce 50% enzyme inhibition).

The steric fields of the Tripos' CoMFA analysis (Fig. 18.26) were subsequently placed into a model of the binding site. It was found that the bad steric regions were in the same regions as the peptide backbone, whereas the favourable steric areas were in empty spaces.

QUESTIONS

1. Using values from Table 18.1, calculate the log *P* value for structure (I) (log *P* for benzene = 2.13).
2. Several analogues of a drug are to be prepared for a QSAR study that will consider the effect of various aromatic substituents on biological activity. You are asked whether the substituents (SO_2NH_2, CF_3, CN, CH_3SO_2, SF_3, $CONH_2$, OCF_3, CO_2H, Br, I) are relevant to the study. What are your thoughts?

3. A lead compound has a monosubstituted aromatic ring present as part of its structure. An analogue was synthesized containing a *para*-chloro substituent that had approximately the same activity. It was decided to synthesize an analogue bearing a methyl group at the *para*-position. This showed increased activity. What analogue would you prepare next, and why?

4. The following QSAR equation was derived for the pesticide activity of structure (II). Explain what the various terms mean and whether the equation is a valid one. Identify what kind of substituents would be best for activity.

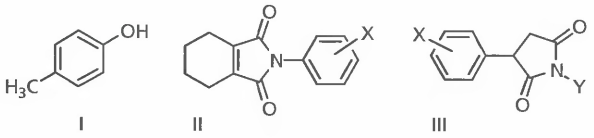
   $\log 1/C = 1.08\pi_x + 2.41F_x + 1.40\ R_x - 0.072\ MR_x + 5.25$ ($n = 16$, $r^2 = 0.840$, $s = 0.59$)

5. A QSAR equation for the anticonvulsant (III) was derived as follows:

   $\log 1/C = 0.92\ \pi_x - 0.34\ \pi_x^2 + 3.18$ (n=15, r^2=0.902, s=0.09, π_0=1.35)

   What conclusions can you draw from this equation? Would you expect activity to be greater if X=CF_3 rather than H or CH_3?

6. The following QSAR equation is related to the mutagenic activity of a series of nitrosoamines; $\log 1/C = 0.92\ \pi + 2.08\ \sigma - 3.26$ ($n = 12$, $r^2 = 0.794$, $s = 0.314$). What sort of substituent is likely to result in high mutagenic activity?



I II III

FURTHER READING

Craig, P. N. (1971) Interdependence between physical parameters and selection of substituent groups for correlation studies. *Journal of Medicinal Chemistry*, **14**, 680–684.

Cramer, R. D., *et al.* (1979) Application of quantitative structure–activity relationships in the development of the antiallergenic pyranenamines. *Journal of Medicinal Chemistry*, **22**, 714–725.

Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988) Comparative field analysis (CoMFA). *Journal of the American Chemical Society*, **110**, 5959–5967.

Cramer, R. D. (2003) Topomer CoMFA: A design methodology for rapid lead optimization. *Journal of Medicinal Chemistry*, **46**, 374–388.

Hansch, C., and Leo, A. (1995) *Exploring QSAR*. American Chemical Society, Washington, DC.

Kellogg, G. E. and Abraham, D. J. (2000) Hydrophobicity: Is $LogP_{o/w}$ more than the sum of its parts? *European Journal of Medicinal Chemistry*, **35**, 651–661.

Kotani, T., and Higashiura. (2004) Comparative molecular active site analysis (CoMASA). 1. An approach to rapid evaluation of 3D QSAR. *Journal of Medicinal Chemistry*, **47**, 2732–2742.

Kubini, H., Folkers, G., and Martin, Y. C. (eds.) (1998) *3D QSAR in drug design*. Kluwer/Escom, Dordrecht.

Martin, Y. C., and Dunn, W. J. (1973) Examination of the utility of the Topliss schemes by analog synthesis. *Journal of Medicinal Chemistry*, **16**, 578–579.

Sutherland, J. J., O'Brien, L. A., and Weaver, D. F. (2004) A comparison of methods for modeling quantitative structure-activity relationships. *Journal of Medicinal Chemistry*, **47**, 5541–5554.

Taft, R. W. (1956) Separation of polar, steric and resonance effects in reactivity. Chapter 13 in: *Steric effects in organic chemistry*, ed. M. S. Newman. John Wiley and Sons, New York.

Topliss, J. G. (1972) Utilization of operational schemes for analog synthesis in drug design. *Journal of Medicinal Chemistry*, **15**, 1006–1011.

Verloop, A., Hoogenstraaten, W., and Tipker, J. (1976) Development and application of new steric substituent parameters in drug design. *Medicinal Chemistry*, **11**, 165–207.

van de Waterbeemd, H., Testa, B., and Folkers, G. (eds.) (1997) *Computer–assisted lead finding and optimization*. Wiley–VCH, New York.

Zhang, S.–X., *et al.* (2000) Antitumor agents. 199. Three–dimensional quantitative structure–activity relationship study of the colchicine binding site ligands using comparative molecular field analysis. *Journal of Medicinal Chemistry*, **43**, 167–176.

Titles for general further reading are listed on p. 725.

CASE STUDY 5

Design of a thymidylate synthase inhibitor

In this case study, we shall look at an early example of how the strategies of *de novo* drug design (section 17.15) and structure-based drug design (section 13.3.11) were used together to develop an active compound that went forward for clinical trials. The research in question involved the design of inhibitors for the enzyme **thymidylate synthase** (section 21.3.2). This enzyme catalyses the methylation of **deoxyuridylate monophosphate (dUMP)** to **deoxythymidylate monophosphate (dTMP)** using **5,10-methylenetetrahydrofolate** as a coenzyme (Fig. CS5.1). Inhibitors of this enzyme have been shown to be antitumour agents which prevent the biosynthesis of one of the required building blocks for DNA. Traditional inhibitors have been modelled on dUMP or the enzyme cofactor 5,10-methylenetetrahydrofolate (Fig. CS5.2), which means that these inhibitors are structurally related to the natural substrate and cofactor. Unfortunately, this increases the possibility of side effects resulting from inhibition of other enzymes and receptors that use these molecules as natural ligands. Therefore, it was decided that *de novo* drug design would be used to design a novel structure that was unrelated to either of the natural substrates.

Before starting the *de novo* design, a good supply of the enzyme was required. Although human thymidylate synthase was not readily available in large quantities, it was possible to obtain good quantities of the bacterial version from *Escherichia coli*, by using recombinant DNA technology to clone the gene and then expressing it in fast-growing cells (section 6.4). The bacterial enzyme is not identical to the human version, but it is very similar,

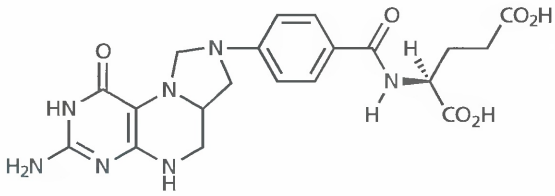


FIGURE CS5.2 5,10-Methylenetetrahydrofolate.

and so it was considered a reasonable analogue. The enzyme was crystallized along with the known inhibitors **5-fluorodeoxyuridylate** and **CB 3717** (Fig. CS5.3). These structures mimic the substrate and the coenzyme respectively, and bind to the sites normally occupied by these structures. The structure of the enzyme–inhibitor complex was then determined by X-ray crystallography, and downloaded on to the computer.

A study of the enzyme–inhibitor complex revealed where the inhibitors were bound and also the binding interactions involved. For CB 3717, the binding interactions around the pteridine portion of the inhibitor were identified as involving hydrogen bonding interactions to two amino acids (the carboxylate ion of Asp-169 and the main chain peptide link next to Ala-263). There was also a hydrogen bonding interaction to a water molecule, which acted as a hydrogen bonding bridge to Arg-21 (Fig. CS5.4). Using molecular modelling, the inhibitor was deleted from the binding site to allow further analysis of the empty binding site. Generating the empty binding site from the enzyme–ligand complex is better than studying the empty binding site from the pure enzyme,

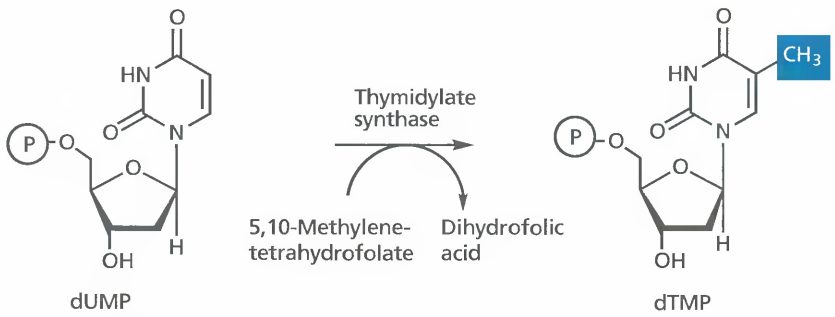


FIGURE CS5.1 Reaction catalysed by thymidylate synthase.

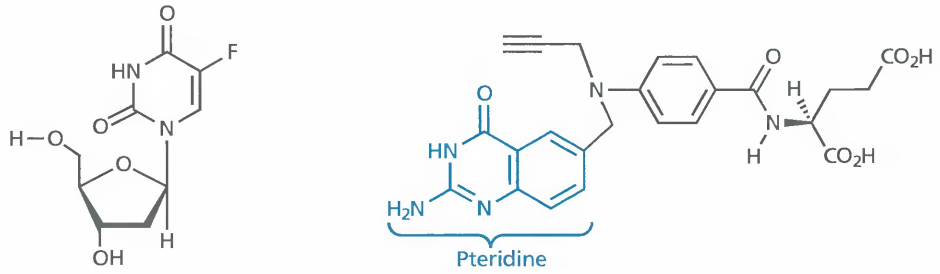


FIGURE CS5.3 Inhibitors of thymidylate synthase.

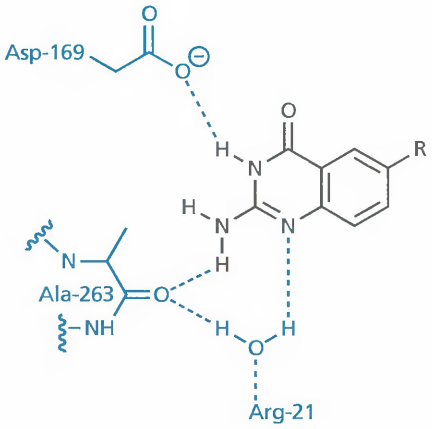


FIGURE CS5.4 Binding interactions holding a pteridine moiety in the active site.

because the latter does not take into account the induced fit that occurs on ligand binding.

A grid was set up within the binding site, and an aromatic CH probe was placed at each grid point to measure hydrophobic interactions and thus identify hydrophobic regions (see section 17.7.5). From this analysis, it was discovered that the pteridine portion of CB 3717 was positioned in a hydrophobic pocket despite the presence of the hydrogen bonding interactions which held it there. The boundaries of this hydrophobic region were determined and a naphthalene ring was found to be a suitable hydrophobic molecule to fit the pocket, yet still leave room for the addition of a functional group which would be capable of forming the important hydrogen bonds.

The functional group chosen was a cyclic amide which was fused to the naphthalene scaffold to create a naphthostyryl scaffold (Fig. CS5.5). Modelling suggested that the NH portion of the amide would bind to Asp-169 while the carbonyl group would bind to the water molecule identified above. A substituent was now added to the naphthostyryl scaffold in order to gain access to the space normally occupied by the benzene ring of the cofactor. A dialkylated amine was chosen as the linking unit and was placed at position 5 of the structure. There were several reasons for this. First of all, adding an amine at this position was easy to carry out synthetically. Second, the two substituents on the amine could be easily varied, which would allow fine tuning of the compound. Last, by using an amine, it would be possible to have a branching point which could have different substituents without adding

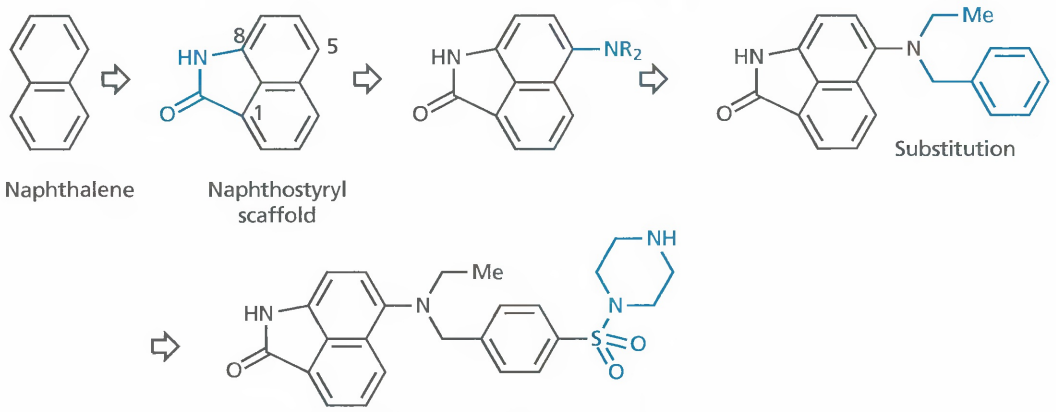


FIGURE CS5.5 Design of an enzyme inhibitor by *de novo* methods.

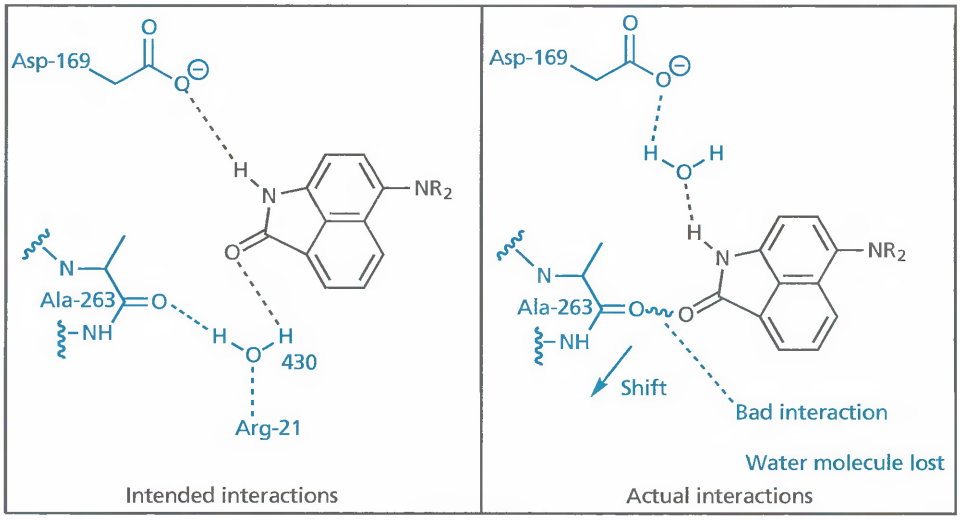


FIGURE CS5.6 Intended versus actual interactions of the inhibitor with the active site.

an asymmetric centre. If a carbon atom had been added instead, two different substituents would have inevitably led to an asymmetric centre with all the attendant complications that would entail (section 13.3.8).

Modelling demonstrated that a benzyl group was a suitable substituent for the amine in order to access the space normally occupied by the benzene ring of the cofactor. A (phenylsulfonyl)piperazine group was then added to the *para* position of the aromatic ring, in order to make the molecule more water soluble—a necessary property if the synthesized structures were to be bound to the enzyme and crystallized for further X-ray crystallographic studies. The positioning of this group was important since the modelling studies showed that it would protrude from the binding site and still make contact with surrounding water. This meant that the group did not have to be desolvated in order for the drug to bind—a process that would involve an energy penalty (section 1.3.6).

This structure was now synthesized and was found to inhibit both the bacterial and human versions of the enzyme with higher activity for the human enzyme. This represents the successful *de novo* design of a novel lead structure.

It was now time to move to structure-based drug design such that the binding interactions and activity of the lead compound could be optimized.

A crystal structure of the novel inhibitor bound to the bacterial enzyme was successfully obtained and studied to see whether the inhibitor had fitted the binding site as expected. In fact, it was found that the naphthalene ring of the inhibitor was wedged deeper into the pocket than expected, because of more favourable hydrophobic interactions. As a result, the cyclic amide failed to form the direct hydrogen bond interaction to Asp-169 which had been planned, and was hydrogen bonding to a bridging water molecule instead. The lactam carbonyl oxygen was also too close to Ala-263 and this had caused this residue to shift 1 Å from its usual position. This in turn had displaced the water molecule which had been the intended target for hydrogen bonding (Fig. CS5.6).

By studying the actual position of the structure in the binding site, it was possible to identify four areas where extra substituents could fill up empty space and perhaps improve binding. These are shown below in Fig CS5.7.

Various structures were proposed then overlaid on the lead compound (still docked within the binding site) to see whether they fitted the binding site. Only those which passed this test and were in stable conformations were synthesized and tested for activity (41 in total). The optimum substituent at each position was then identified.

- In region 1 (R^1), modelling showed that this substituent fitted into a hydrophobic pocket which became hydrophilic the deeper one got. This suggested that a hydrogen bonding substituent at the end of an alkyl chain might be worth trying (the extension strategy; section 13.3.2) and, indeed, a CH_2CH_2OH group led to an improvement in binding affinity. It was also found that a methyl group was better than the original ethyl group.

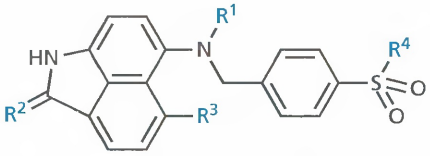


FIGURE CS5.7 Variable positions (in colour).

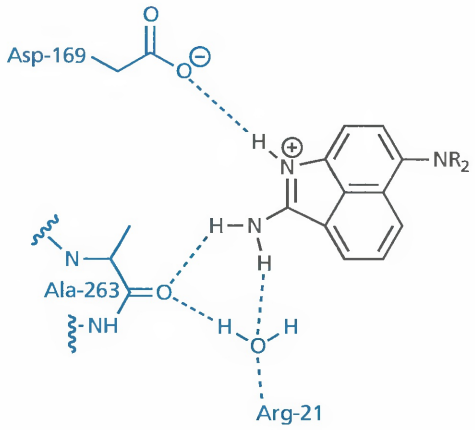


FIGURE CS5.8 Binding interactions of the modified inhibitor with the active site.

- In region 2 (R^2), the carbonyl oxygen was replaced by an amidine group that would be capable of hydrogen bonding to the carbonyl oxygen of Ala-263, rather than repelling it. An added advantage in using a basic amidine group was the fact that there was a good chance that it would become protonated, allowing a stronger ionic interaction with Asp-169, as well as a better hydrogen bonding interaction with Ala-263. When this structure was synthesized, it was found to have improved inhibition, and a crystal structure of the enzyme-inhibitor complex showed that the expected interactions were taking place (Fig. CS5.8). Moreover, Ala-263 had returned to its original position allowing the return of the bridging water molecule.
- In region 3 (R^3), there was room for a small group such as a chlorine atom or a methyl group and both of these substituents led to an increase in activity.
- Region 4 (R^4) was relatively unimportant for inhibitory activity, since groups at this position protrude out of the active site into the surrounding solvent and have only minimal contact with the enzyme. Nevertheless, the piperazine ring was replaced by a morpholine group because the latter had some advantages with respect to selectivity and pharmacological properties.

Having identified the optimum groups at each position, structures were synthesized combining some or all of these groups. The presence of the amidine resulted in the best improvement in activity and so the presence of this group was mandatory. Interestingly, adding all the optimum groups is not as beneficial as adding some of them.

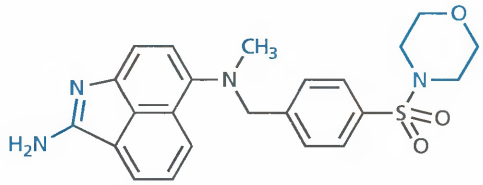


FIGURE CS5.9 Modified inhibitor put forward for clinical trials.

The modified structure (Fig. CS5.9) was synthesized and was found to be a potent inhibitor that was 500 times more active than the original amide. A crystal structure of the enzyme–inhibitor complex showed a much better fit, and the compound was put forward for clinical trials as an antitumour agent.

The case study illustrates many of the general principles behind *de novo* design that were described in section 17.15.1. For example, the designed molecule is fairly flexible and did not fill up all the space available in the binding site. As we have seen, a different binding mode took place from that predicted, and this might not have occurred if a more rigid and more closely fitting structure had been designed. The fact that binding did take place allowed structure-based drug design to be carried out on the information obtained.

The importance of synthetic feasibility was considered throughout the process and was one of the reason for introducing an amine substituent to the naphthostyryl ring system.

Finally, it is important to appreciate that binding studies using target proteins from different species may produce slightly different results. For example, the computer modelling studies described above were carried out on bacterial thymidylate synthase enzyme rather than the human version. Fortunately, the activities of the designed inhibitors were actually greater for the human enzyme than for the bacterial enzyme, and this was put down to the fact that the hydrophobic space available for the naphthalene ring was larger in the human enzyme than in the bacterial one. Fortunately most changes carried out had beneficial effects for both enzymes, with one exception; adding a methyl group at R^3 of the amidine led to an increase in activity for the bacterial enzyme, but not for the human enzyme.

FURTHER READING

Greer, J., Erickson, J. W., Baldwin, J. J., and Varney, M. D. (1994) Application of the three-dimensional structures of protein target molecules in structure-based drug design. *Journal of Medicinal Chemistry*, **37**, 1035–1054.

CASE STUDY 6

Design of a serotonin antagonist as a possible anxiolytic agent

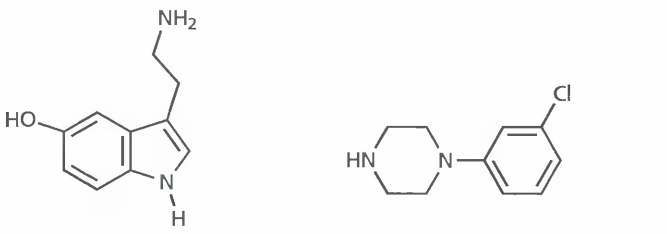
It was once claimed that drugs could be designed using molecular modelling alone (i.e *de novo* design) (section 17.15). However, to date, no clinically useful drug has been designed purely by molecular modelling. *De novo* design has certainly been effective in designing new lead compounds, but further drug optimization with the aid of structure-based drug design has always been necessary to develop a clinically useful drug (section 13.3). Molecular modelling should be seen as one of several powerful tools which the medicinal chemist has to hand. The following case study illustrates how computer modelling has been used as an aid in drug design alongside many of the tools and strategies described in chapters 12–18. It also demonstrates that the design process considers both the pharmacodynamic and pharmacokinetic properties of the compounds produced. In other words, it would be inefficient designing a molecule that interacts strongly with a target if it has no chance of reaching that target in a living system, or is likely to have toxic side effects. The project in question was carried out by SmithKline Beecham, and was aimed at finding novel anxiolytic and antidepressant agents.

The target

First, a suitable target had to be identified and it was decided to design an agent that would interact with a serotonin receptor. **Serotonin** (or 5-hydroxytryptamine, 5-HT) (Fig. CS6.1) is an important neurotransmitter in the central nervous system, and abnormalities resulting in altered levels of serotonin are thought to be involved in a variety of disorders such as anxiety, depression, and migraine. Agents that can influence serotonin levels are therefore useful in treating these diseases. Unfortunately, these agents have various problems and side effects associated with them (Case Study 7). The discovery of various types of serotonin receptor allows the possibility of designing agents which might be more selective in their action and have fewer side effects.

There are seven main types of serotonin receptor (5-HT_1 – 5-HT_7) and 14 subtypes of these. For example, the 5-HT_2 receptor consists of three subtypes, (5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C}), which are distributed differently around the body. The 5-HT_{2A} receptor is found in both the peripheral and central nervous systems (CNS), the 5-HT_{2B} receptor is found mostly in the peripheral nervous system, and the 5-HT_{2C} receptor is found only in the CNS. All the serotonin receptors are G-protein-coupled receptors (sections 4.7 and 5.1), with the exception of 5-HT_3 which is a ligand-gated ion channel. The 5-HT_1, 5-HT_2, and 5-HT_7 receptors have all been cited as targets for potential new mood-altering drugs.

The 5-HT_{2C} receptor was of particular interest to this project, as there is evidence that this receptor might be involved in anxiety. For example, ***meta*-chlorophenylpiperazine** (*m*CPP) is an agonist that shows some selectivity for the 5-HT_{2B} and 5-HT_{2C} receptors, and which has been shown to cause anxiety in animal and human studies. As the 5-HT_{2B} receptor is mainly in the peripheral nervous system, it is likely that these CNS effects are caused by the agonist acting on the 5-HT_{2C} receptor, which is only present in the CNS. At the start of this project, there were several known serotonin



Serotonin (5-HT) *meta*-Chlorophenylpiperazine (*m*CPP)

FIGURE CS6.1 Serotonin and *m*CPP.

antagonists for the 5-HT_2 receptor, which were used for treating schizophrenia, feeding disorders, perception, depression, and anxiety, but none showed any selectivity for the 5-HT_{2C} over the 5-HT_{2A} receptor. Therefore, it was not possible to state whether one or other of these receptor subtypes was responsible for the CNS properties and/or side effects of these drugs. It was argued that an antagonist showing selectivity for the 5-HT_{2C} over the 5-HT_{2A} receptor would give useful information regarding the functions of these receptor subtypes and possibly result in a compound with improved properties over established drugs.

Having identified a likely target, the aim was now to identify a suitable profile of activity for the intended drug. Clearly, the antagonist should be as selective as possible for the 5-HT_{2C} receptor in order to reduce the possibility of side effects. In particular, the challenge was to obtain selectivity relative to the 5-HT_{2A} receptor. Furthermore, the new drug should not affect metabolic enzymes (e.g. cytochrome P450 enzymes; section 11.4.2), again to reduce possible side effects and to avoid the possibility of drug-drug interactions (section 11.4.6). The ideal drug should also be non-sedating, have no interaction with alcohol, have a fast onset of action and a high response rate, and show no withdrawal effects.

Testing procedures

Both *in vitro* and *in vivo* tests were required for the study. The *in vitro* tests involved radioligand binding studies (section 8.9) carried out on human 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C} receptors, which had been cloned and expressed in fast growing cells. Testing for any inhibitory activity against cytochrome P450 enzymes was also done *in vitro*.

In vivo activity was measured by the ability of compounds to block hypoactivity in rats, brought on by administering *m*CPP. The effects of *m*CPP are thought to be due to it acting as an agonist at 5-HT_{2C} receptors as stated earlier.

Modelling studies also played an important part in the design strategies that were used in this study, and these were carried out on model binding sites for the receptors involved (*homology modelling; section 17.14.1*).

Lead compound to SB 200646

The lead compound for the project was a drug that had been produced by Lilly Pharmaceuticals. One of the problems with this compound was its insolubility in water, and so it was decided to replace the phenyl ring with a more polar pyridine ring (*ring variation; section 13.3.5*) (Fig. CS6.2). Various analogues were synthesized to find the best substitution positions for the urea group linking the ring systems, for both the pyridine ring and the indole ring. This revealed that 3- and 5- substitution respectively were ideal (*substituent variation; section 13.3.1*). Finally, the methyl and ethyl substituents on the indole ring were not essential and could be absent (*simplification; section 13.3.8*). The final compound (**SB 200646**) proved to be the first selective 5$HT_{2B/2C}$ antagonist, having modest 5HT_{2C} affinity *in vitro*, some oral activity *in vivo*, and a 50-fold selectivity over the closely related 5HT_{2A} receptor.

SB 200646 to SB 206553

The urea functional group is the most flexible region of SB 200646 and so the next tactic was to reduce this flexibility by locking one of the bonds into a ring (*rigidification; section 13.3.9*). The aim here was to reduce the total number of conformations, while retaining the active conformation. Four conformationally restrained structures were prepared (Fig. CS6.3), and it was found that the introduction of a six-membered ring decreased affinity, whereas the five-membered ring increased affinity.

The difference in affinity between the six- and five-membered rings was rationalized by a molecular modelling study, which showed a marked difference in geometry between the two ring systems (Fig. CS6.4). With a five-membered ring (A), the overall structure is roughly planar. This is not the case with the six-membered ring (B) where the overall structure is more twisted. Clearly, the

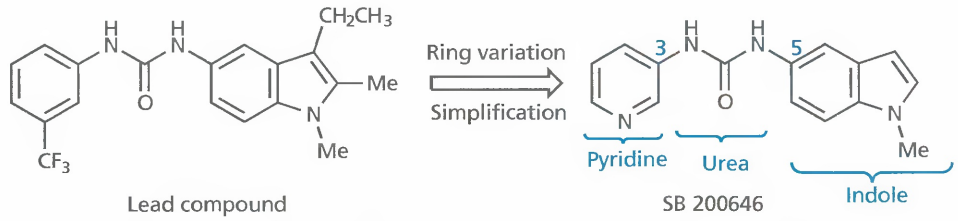


FIGURE CS6.2 Simplification and ring variation of the lead compound.

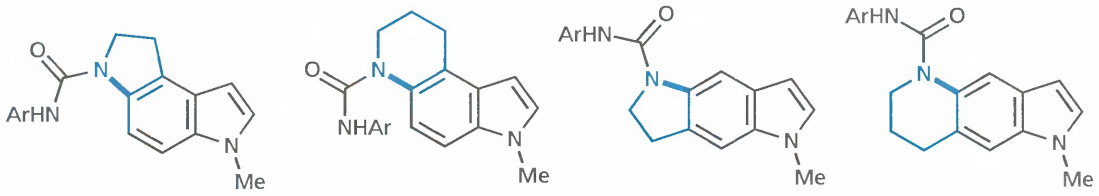


FIGURE CS6.3 The strategy of rigidification (locked bond shown in bold).

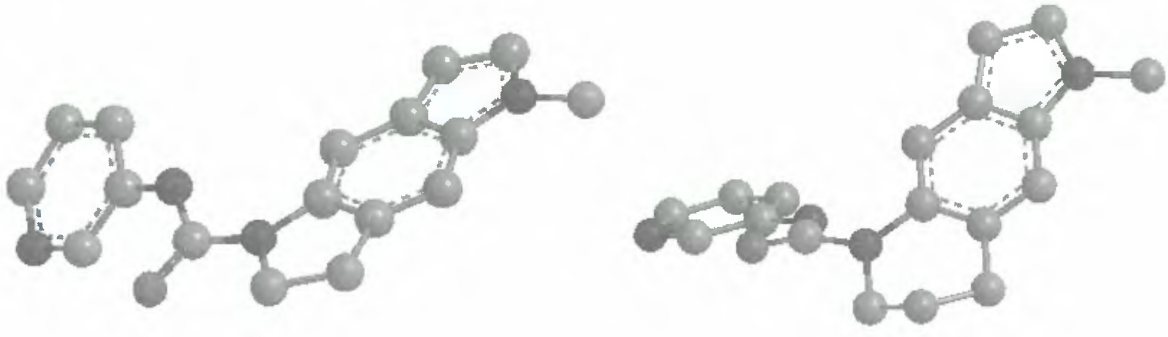


FIGURE CS6.4 Relative geometries of the rigidified analogues using (**a**) a five membered ring and (**b**) a six membered ring.

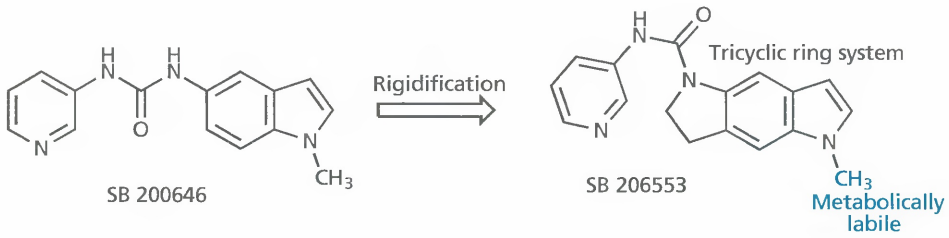


FIGURE CS6.5 Development of SB 206553.

planar molecule interacts with the receptor more strongly than the twisted molecule.

Therefore, rigidification was carried out using the five-membered ring, leading to **SB 206553** (Fig. CS6.5) and resulting in a 10-fold increase in *in vitro* affinity and a 160-fold selectivity over the 5-HT_{2A} receptor. The compound was also shown to relieve anxiety *in vivo*, with a 4-fold increase in potency in the rat hypolocomotion assay. Unfortunately, this compound was found to undergo metabolism in the body resulting in *N*-demethylation (section 11.4.2) and the production of a non-selective, active metabolite.

Analogues of SB 206553

Various strategies were tried to replace or stabilize the metabolically labile *N*-methyl group of SB 206553. First, a number of tricyclic analogues of SB 206553 were synthesized to explore the effects of fusing the five-membered ring at different positions of the aromatic ring (*group shift; section 14.2.6*) (Fig. CS6.6). Such a variation shifts the methyl group relative to the benzene ring, which might make it unrecognizable to metabolic enzymes, yet still allow the molecule to interact with its target binding site. Unfortunately, all of these compounds showed decreased activity and a drop in selectivity between the receptor subtypes.

Removing the *N*-methyl group from the indole ring to replace it with NH led to a loss of affinity, suggesting that the methyl group was important and might bind to a hydrophobic pocket in the binding site. In order to explore the size of this proposed pocket, the methyl group was replaced by larger alkyl groups (*varying substituents; section 13.3.1*) to see whether this would lead to increased hydrophobic interactions and an increased affinity. It was found that groups such as Et, Pr, and

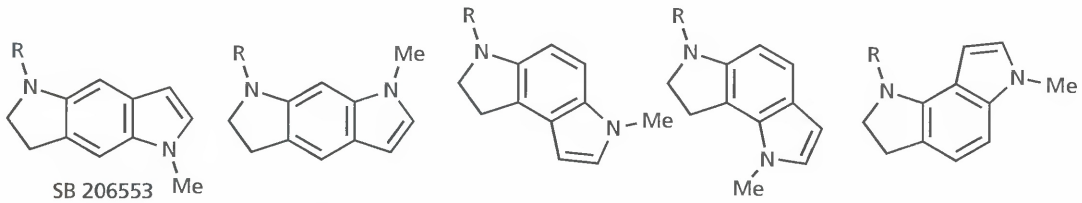


FIGURE CS6.6 Tricyclic analogues of SB 206553 (R = CONH(3-pyridyl).

^{i}Pr resulted in similar 5-HT_{2C} affinity and slightly increased selectivity, which suggests that the pocket for the 5-HT_{2C} receptor has slightly more volume than the pocket for the 5-HT_{2A} receptor. An *N*-benzyl group was bad for both affinity and selectivity, which suggests that the pocket is neither extensive nor long. Substitution elsewhere in the pyrrole ring was also bad for affinity (Fig CS6.7).

The improvements observed above were not substantial and so an isostere for the pyrrole ring was now sought, with various five-membered heterocycles being tried (Fig. CS6.8) (*ring variation; sections 13.3.5 and 14.2.7*, and *bioisosteres; sections 13.3.7 and 14.1.5*). A thiophene ring led to a loss in selectivity and this was attributed to the loss of the methyl group, which could no longer fit into the proposed hydrophobic pocket. A substituted furan ring was used instead, with a methyl substituent positioned at an analogous position. As this group is attached to carbon rather than nitrogen, it was expected that it would be more stable to metabolism. The resulting compound retained good *in vitro* affinity and selectivity, but had poor oral activity *in vivo*. However, *in vivo* activity was better if the compound was administered intravenously, suggesting that the compound was poorly absorbed from the gut or was rapidly metabolized in the liver.

Molecular modelling studies on SB 206553 and analogues

Having synthesized several analogues, molecular modelling was carried out to overlay both active and inactive compounds, using the urea group as the common feature (*overlays; section 17.9*). The space occupied by each molecule was then analysed and compared with its affinity for both the 5-HT_{2A} and 5-HT_{2C} receptors (3D-QSAR; section 18.10). It was then assumed that compounds that showed poor activity must have substituents in sterically disallowed areas of the relevant binding site. This allowed a map to be created (Fig. CS6.9) showing a volume in space (coloured light blue) which could be accessed by molecules having 5-HT_{2C} affinity, and which included a volume (coloured dark blue) that was disallowed for 5-HT_{2A} affinity. This crucial dark blue area was in the

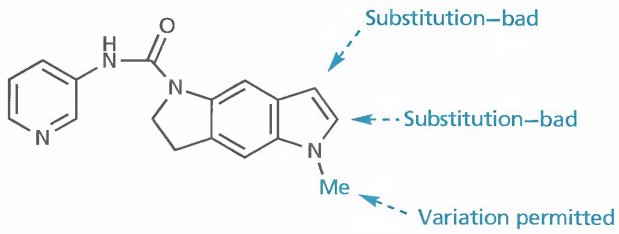


FIGURE CS6.7 Varying substituents on SB 206553.

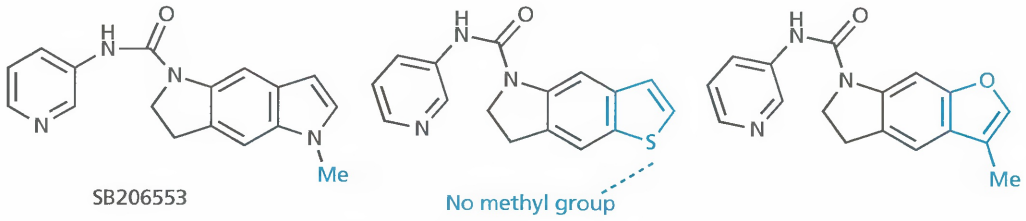


FIGURE CS6.8 Ring variation of SB 206553.

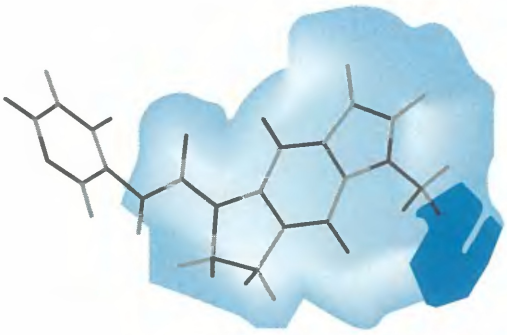


FIGURE CS6.9 5-HT_{2C} allowed volume (light blue) includes a 5-HT_{2A} disallowed volume (dark blue) which is crucial for selectivity (shown around the structure of SB 206553).

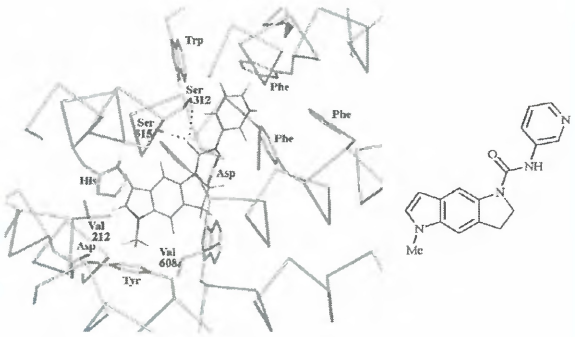


FIGURE CS6.10 Stable binding mode for SB 206553.

region of the *N*-methyl group of SB 206553. (The pyridine ring was not included in the calculations since it can freely rotate.)

A model 5-HT_{2C} receptor was now built using molecular modelling software. The amino acid sequence for the 5-HT_{2C} receptor was obtained and compared with other related serotonin receptors to identify the likely transmembrane (TM) regions. The receptor is a member of the 7-TM receptor family (section 4.7) and so a model was built by analogy with the crystal structure of **bacteriorhodopsin** to create a model binding site for the receptor (*homology modelling; section 17.14*). Such models are never perfect and a model binding site is unlikely to have *exactly* the same structure as the true binding site. Nevertheless, the model can still be useful in indicating what drug design strategies are worth pursuing (Case Study 5).

The next stage was to dock SB 206553 into the binding site (*docking; section 17.12*). Standard agonists and antagonists for serotonin receptors have a strongly basic nitrogen which binds to an aspartate residue. However, SB 206553 does not have this feature. Moreover, the molecule is quite rigid and does not fit into the normal agonist cavity. Therefore, other binding modes were investigated and because structure–activity relationship (SAR) studies (*section 13.1*) emphasized the importance of the carbonyl oxygen in the urea group, the receptor model was studied to see whether there were any amino acid residues that could act as possible hydrogen bond donors for this carbonyl oxygen. Two serine residues were identified that were positioned fairly close to the important aspartate residue, one of which (Ser-312) is unique to the 5-HT_2 class of receptors. SB 206553 does not bind to 5-HT_1 receptors, which lack Ser-312, and so it was presumed that this amino acid was responsible for the selectivity.

Therefore, SB 206553 was positioned into the pocket such that it interacted with Ser-312, and it was found that a second hydrogen bonding interaction involving the urea carbonyl group was possible to another serine residue (Ser-315). This binding interaction also allowed the two aromatic regions (the pyridine and tricyclic ring system) to fit into hydrophobic pockets, one of which was dominated by aromatic amino acids, and the other made up of both aliphatic and aromatic amino acids. It could not be determined, however, which ring fitted which pocket and so both possibilities were considered. Once docked, energy minimizations were carried out for both binding modes, and interaction energies were calculated. The more stable binding mode was the one where the pyridine ring was placed in the more aromatic of the hydrophobic pockets (Fig. CS6.10).

In this binding mode, the carbonyl oxygen of the urea interacts through hydrogen bonds with two serine residues, while the indole *N*-methyl group is placed in a hydrophobic pocket adjacent to two valine residues (Val-608 and Val-212) (Fig. CS6.11). The receptor–ligand model was now studied to identify methods of achieving better selectivity for the 5-HT_{2C} receptor over the 5-HT_{2A} receptor. These studies revealed a difference between the two model receptors in the pocket occupied by the *N*-methyl group. In the 5-HT_{2C} receptor, there are two valine residues in this pocket. In the 5-HT_{2A} receptor, these amino acids are replaced by leucine. Leucine is bulkier than valine, so it was proposed that the hydrophobic pocket is smaller in the 5-HT_{2A} receptor, making it more difficult for the *N*-methyl group of SB 206553 to fit and thus accounting for the observed selectivity.

SB 206553 to SB 221284

The next stage was to try and take advantage of the above findings. The metabolically labile *N*-methyl group had to be removed, but it was important to replace it with something that could fit the identified hydrophobic pocket in order to retain selectivity. It was decided to remove the pyrrole ring containing the *N*-methyl group and add

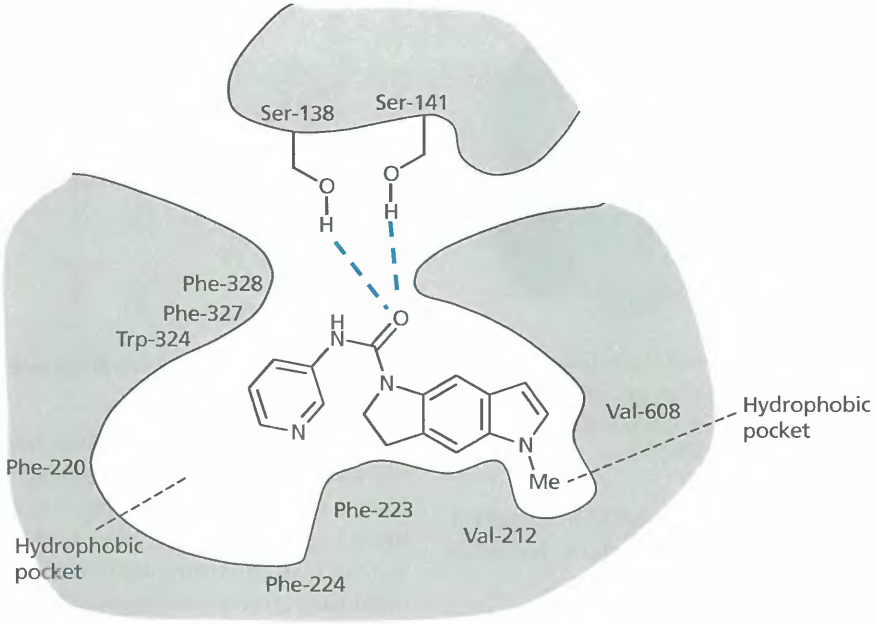


FIGURE CS6.11 Binding site showing two hydrophobic pockets.

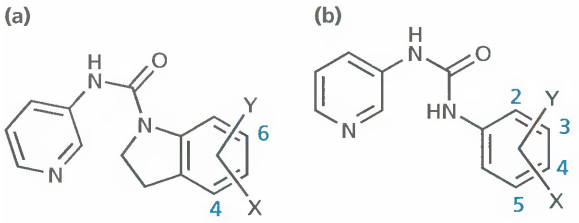


FIGURE CS6.12 Search for an isostere of the pyrrole ring. (**a**) Indolines and (**b**) phenylureas. Note the different numbering systems for structures **a** and **b**.

substituents to the remaining bicyclic indoline system (Fig. CS6.12a) to act as isosteres for the lost ring. These structures were fairly difficult to synthesize, so an initial study was carried out on the simpler phenyl urea system (Fig.CS6.12b). A large number of analogues (86) were prepared by **parallel synthesis** (chapter 16) in order to find the best substituents for that system.

Mono-, di-, and tri-substituted structures were prepared at various positions, and it was found that the preferred substitution pattern consisted of two substituents, *ortho* to each other at positions 3 and 4. Substitution at position 2 was bad for affinity, and it was rationalized that this was due to a steric effect between the substituent and the urea group, which caused the urea group to move out of the plane relative to the aromatic ring. This would result in a twist in the molecule and poorer binding as a consequence.

Having identified positions 3 and 4 as the ideal substitution points, the nature of the substituents was studied to see whether good affinity was related to particular substituents. At position 3, small hydrophobic, electron-withdrawing substituents were preferred (e.g. Cl or CF_3). In contrast, small hydrophobic electron-donating groups were preferred at position 4 (e.g. SMe or OMe). **QSAR studies** (chapter 18) showed a good correlation between the lipophilicity (π) of small substituents at position 4, and 5-HT_{2C} binding affinity (pK_i; section 8.9) (Fig. CS6.13).

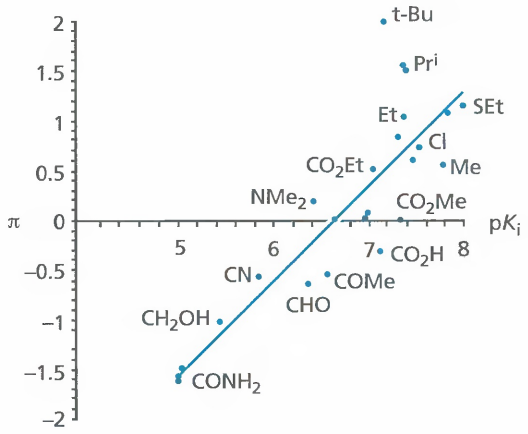


FIGURE CS6.13 QSAR plot relating lipophilicity of 4-substituents versus binding affinity.

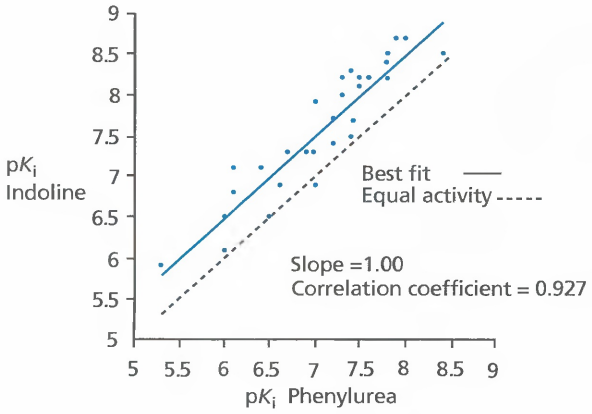


FIGURE CS6.14 Comparison of pK_i values for indolines and phenylureas. The dotted line indicates what would be expected if the indolines and phenylureas had equal activity.

Having established the preferred substituents in the phenyl urea series, it was important to check that the results were also valid for the indoline series, and so a selection of indolines was prepared and their affinities checked against the corresponding phenylureas. The graph (Fig. CS6.14) showed the same substituent effect for the indolines as for the phenylureas, and demonstrated that the indolines were more potent than the corresponding phenylureas.

In the indoline series (Fig. CS6.15), an electron-withdrawing group at position 6 was preferred (e.g. Cl or CF_3). The substituent at position 5 is the one expected to fit the hydrophobic pocket, which is so crucial to selectivity. Increasing the bulk of this substituent using different alkyl groups (*varying substituents*) increased the selectivity but also led to a fall in affinity. The best balance of affinity versus selectivity was achieved with a thioether or an ether group. The best selectivity was found with groups such as SEt , S(^{n}Pr), or O(^{i}Pr). However, *in vivo* activity was better with smaller groups such as SMe or OMe. Consequently, **SB 221284** (Fig. CS6.15) was chosen for further study and proved to be a potent inhibitor of 5-$HT_{2B/2C}$ receptors with good selectivity over the 5-HT_{2A} receptor. Unfortunately, the compound had a major drawback in that it inhibited cytochrome P450 enzymes.

Modelling studies on SB 221284

Ligand docking studies were now carried out with SB 221284 on the model 5-HT_{2C} receptor (Figs. CS6.16 and CS6.17). The resulting complex showed that the pyridine and indoline rings were in the same hydrophobic pockets as before. The hydrogen bonding interactions to the two serine residues were present, and the *S*-methyl group fitted into the hydrophobic pocket previously occupied by the *N*-methyl group. Moreover, it was noted that the CF_3 group helped to orientate the thioether correctly for this interaction. The CF_3 group, which was *ortho* to the thiomethyl group, restricted the latter's rotation (*conformational blocker*), thus favouring conformations where the thiomethyl group was directed towards its binding pocket. It was also found that both the pyridine ring and the SMe group were twisted out of the plane of the indoline ring and were almost perpendicular to it.

Attention now turned to the hydrophobic pocket occupied by the pyridine ring and it was observed that this pocket was quite deep, suggesting that further aromatic substituents could be added to increase the binding

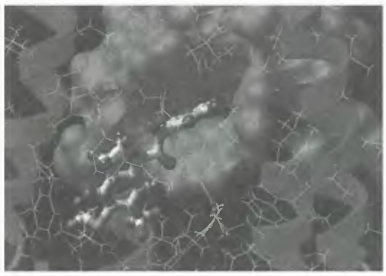


FIGURE CS6.16 Ligand docking studies on SB 221284.

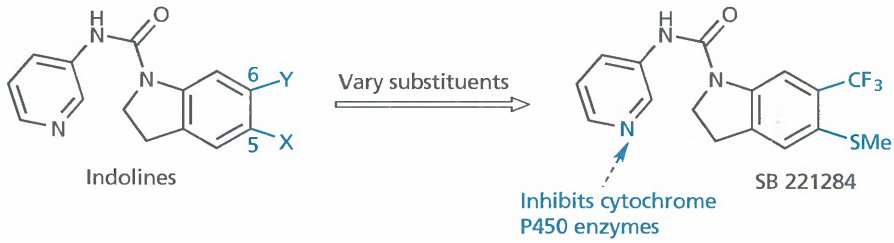


FIGURE CS6.15 Development of SB 221284.

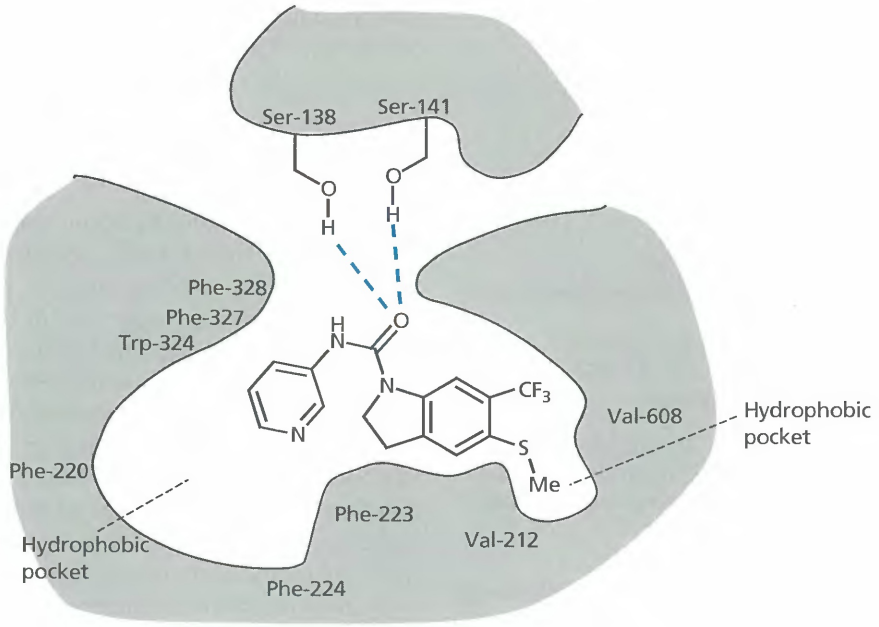


FIGURE CS6.17 Binding site interactions for SB 221284.

interactions. Adding an extra substituent to the pyridine ring might also be expected to have a separate beneficial effect. Previous work had shown that the pyridine nitrogen in this class of compounds was responsible for the cytochrome P450 inhibitory activity and that this activity could be prevented if substituents were placed close to the pyridine nitrogen to act as steric shields (section 14.2.1).

3D QSAR studies on analogues of SB 221284

As a check on the authenticity of the receptor model, 55 substituted analogues of SB 221284 were synthesized and a 3D QSAR study (section 18.10) was carried out. Each structure was docked into the model receptor and the receptor–ligand complex minimized. The ligands were then removed from the receptor and subjected to CoMFA analysis, leaving out eight structures to serve as a test set. Therefore, 47 compounds were analysed and a QSAR equation was derived, which was checked against the test set of eight structures. Affinity values for the 47 structures used in the analysis were calculated, and a graph of predicted affinity versus actual affinity showed a good relationship (Fig. CS6.18). Further analysis showed that the steric fields derived for the CoMFA analysis were of more importance to the equation than the electrostatic fields.

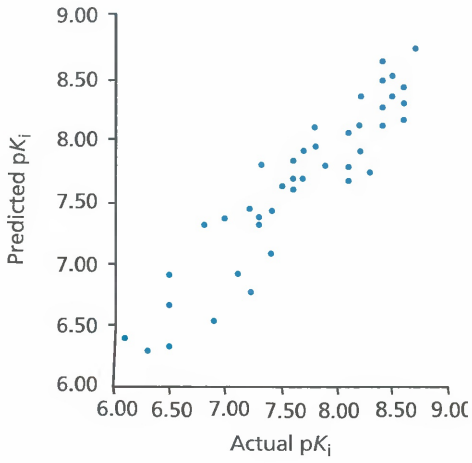


FIGURE CS6.18 3D QSAR study on SB 221284 analogues—predicted versus actual pK_i ($q^2 = 0.656$).

A graphical representation of the steric fields around the aromatic ring of the indoline system was produced which showed beneficial areas for affinity in dark blue, and detrimental areas in light blue (Fig. CS6.19). The large number of detrimental areas that were present suggested that the indoline ring and its substituents were in a tight pocket allowing little scope for further variation. These results were in agreement with what had been expected from the proposed receptor binding model.

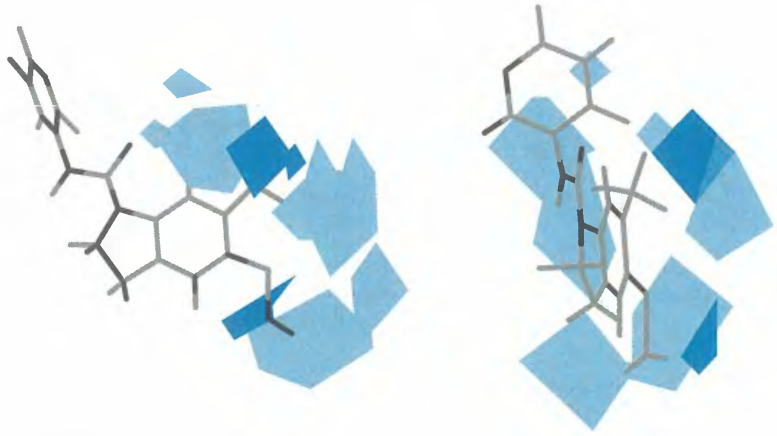


FIGURE CS6.19 3D QSAR study on SB 221284—steric fields (detrimental in light blue, beneficial in dark blue) around the aromatic ring of the indoline system. Diagram shows the molecule and the steric fields from two different viewpoints.

SB 221284 to SB 228357

Work was now carried out to determine whether adding further aromatic groups to the pyridine ring would have the expected beneficial effects on affinity, and also remove cytochrome P450 inhibition. For this work, compounds were prepared with a methoxy group at position 5 instead of a methylthio group (Fig. CS6.20). It had previously been shown that the methoxy and methylthio groups were essentially equivalent as far as affinity and selectivity were concerned.

Substitutions at various positions of the pyridine ring were carried out, with position 5 being the most promising. An aromatic substituent (structure I) resulted in slightly increased affinity, increased selectivity and also led to a 100-fold drop in cytochrome P450 inhibition. However, the level of cytochrome P450 inhibition remaining was still unacceptable and the structures also had poor oral activity. It was proposed that low oral activity might be due to the aromatic substituent being susceptible to metabolism since it is relatively electron rich, and so fluorine substituents were added to make the ring less electron rich. Unfortunately, this was detrimental to both affinity and selectivity.

A pyridine ring was then introduced in place of the aromatic ring in order to increase water solubility and metabolic stability (structure II). This resulted in a 10-fold increase in affinity, which suggested that an additional binding interaction might be taking place involving the pyridine nitrogen. Regrettably, affinity for the 5-HT_{2A} receptor increased even more significantly, resulting in a drop in selectivity.

Fortunately, selectivity was recovered by placing a methyl group at position 4' (structure III). This group forces the two heteroaromatic rings out of plane with each other, because of a steric clash with the *ortho* proton of the neighbouring

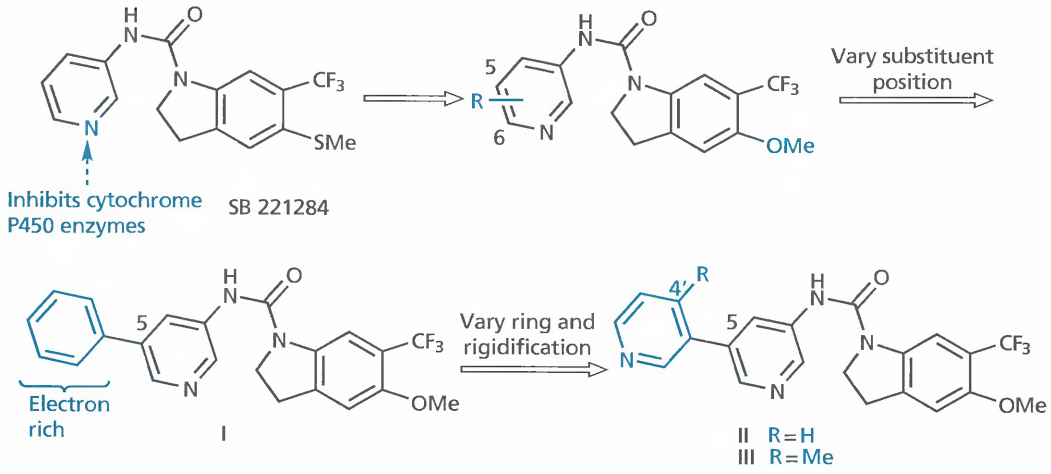


FIGURE CS6.20 The development process from SB221284 involving modifications of the pyridine ring.

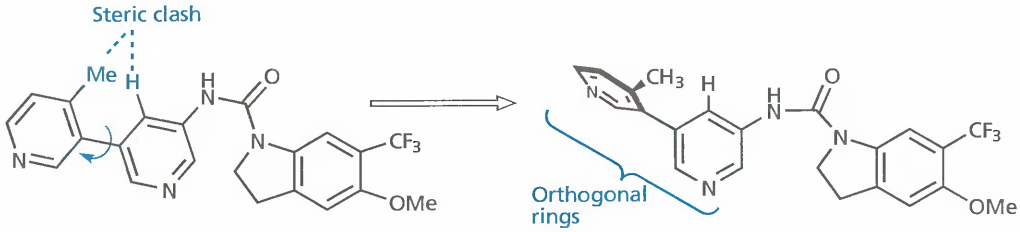


FIGURE CS6.21 Conformation constraint due to the *ortho* methyl group. The molecule has two possible conformations where the rings are orthogonal. One conformation is as drawn. The other has the terminal pyridine ring rotated 180°.

ring (Fig. CS6.21). The methyl group is therefore acting as a conformational blocker (section 13.3.10) and has forced the molecule into a conformation which is favoured by the 5-HT_{2C} receptor rather than the 5-HT_{2A} receptor.

Structure III had slightly increased affinity for the 5-HT_{2C} receptor and moderate oral activity. However, it still inhibited P450 enzymes although the level of inhibition had fallen.

Placing the heteroaromatic substituent at position 6 of the pyridine ring was also tried. as this should have a greater shielding effect and cause a greater reduction in cytochrome P450 inhibition. This was indeed observed. Selectivity between the 5-HT_{2C} and 5-HT_{2A} receptors was also increased 2000-fold, but the compound was orally inactive.

The pyridine nitrogen was identified as the cause of cytochrome P450 inhibition, so it was decided to remove it altogether and to replace the pyridine ring with an aromatic ring (structure IV) (Fig. CS6.22). This led to a fall in water solubility, but this could be restored by adding a pyridine ring as a substituent (structure V). This structure had improved selectivity, higher 5-HT_{2C} affinity, and potent oral activity, but was still observed to inhibit cytochrome P450 enzymes.

Moving the methyl group in the pyridine substituent to the *ortho* position (structure VI) resulted in restricted rotation between the two rings and forced them into a non-planar conformation with respect to each other (*conformational restraint*). This resulted in increased selectivity. Placing an *ortho* methyl in the aromatic ring as well (structure VII), increased 5-HT_{2C} affinity yet further and effectively abolished affinity for the 5-HT_{2A} receptor. Molecular modelling showed that the aromatic and pyridine rings were at right angles to each other, implying that this is a good conformation for binding to the 5-HT_{2C} receptor but not the 5-HT_{2A} receptor. Structure VII also had good oral activity, although it still retained some cytochrome P450 activity. Nevertheless, structure VI had the better overall *in vitro* profile with high *in vivo* potency and was chosen for further modification.

One problem with structure VI was its short duration of action—an indication that it was being metabolized, presumably at the electron-rich phenyl ring. A series of analogues was therefore prepared with electron-withdrawing substituents on the aromatic ring (structure VIII). Substituents at positions 2 and 6 were found to be bad for affinity and selectivity. Substitutions at this position are

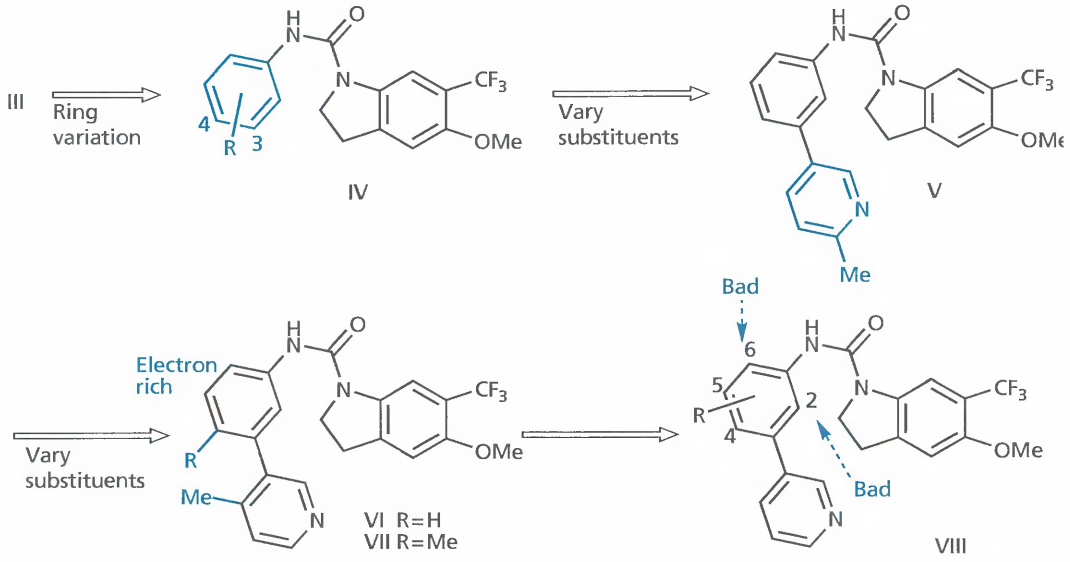


FIGURE CS6.22 Development of an agent with 5HT_{2C} affinity and no 5-HT_{2A} affinity (structure VII).

likely to twist the ring out of plane with the urea group. Substituents were better tolerated at positions 4 and 5 resulting in good affinity and selectivity. As mentioned previously, substitution at position 4 is particularly good for selectivity as it forces the neighbouring pyridine ring out of the same plane. However, the crucial feature at this stage was to increase the duration of activity, and SB 228357 with a fluorine at position 5 was found to have activity lasting 6 hours (Fig. CS6.23).

Further receptor-ligand modelling studies suggested that there was still space in the hydrophobic pocket to allow further extension. An oxygen linker was therefore placed between the pyridine and phenyl rings to allow a deeper occupation of the pocket (structure IX; Fig. CS6.23). However, this structure had poor oral activity, probably due to reduced solubility. As a result, the phenyl ring was replaced by the original pyridine ring (structure X). An *ortho* methyl group was added as R^2 (structure XI) to increase the torsional angle between the two rings, resulting in increased selectivity and affinity for the 5-HT_{2C} receptor. Moreover, selectivity over the 5-HT_{2B} receptor was found for the first time (80-fold). Structure XI had low cytochrome P450 activity and good *in vivo* activity, but there were worries that the 5-methoxy group on the indoline ring would be metabolically labile and so this was replaced by a methyl group. This compound (**SB 243213**) had a good profile with negligible cytochrome P450 activity. The receptor-ligand model also showed that the pyridine ring substituent was well inserted into the hydrophobic pocket Figs. CS6.24 and CS6.25).

SB 243213 and SB 228357 were both tested to make sure that they were inactive against a range of other receptors, ion channels and enzymes. Long-term studies were also carried out to see whether chronic dosing would affect receptor density or sensitivity. One of the key advantages of SB 243213 over SB 228357 is its much improved solubility, so the former was put forward for phase 1 clinical trials as a non-sedating antidepressant/anxiolytic.

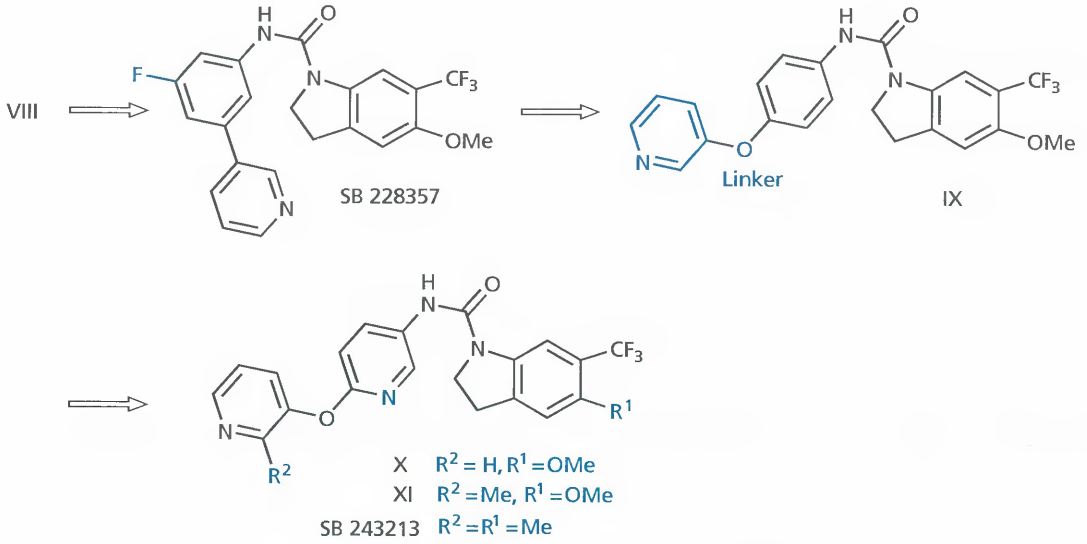


FIGURE CS6.23 Development of SB 228357 and SB 243213.

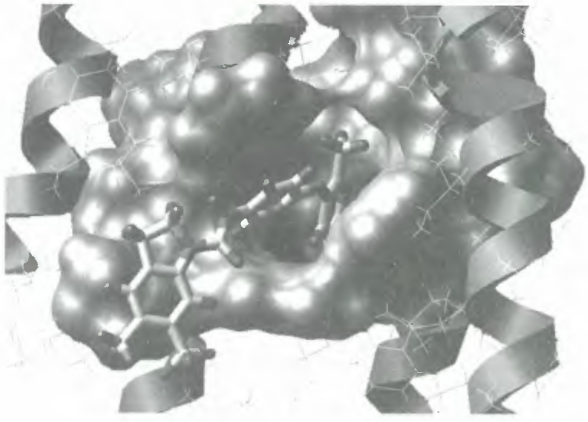


FIGURE CS6.24 Receptor–ligand model for SB 243213.

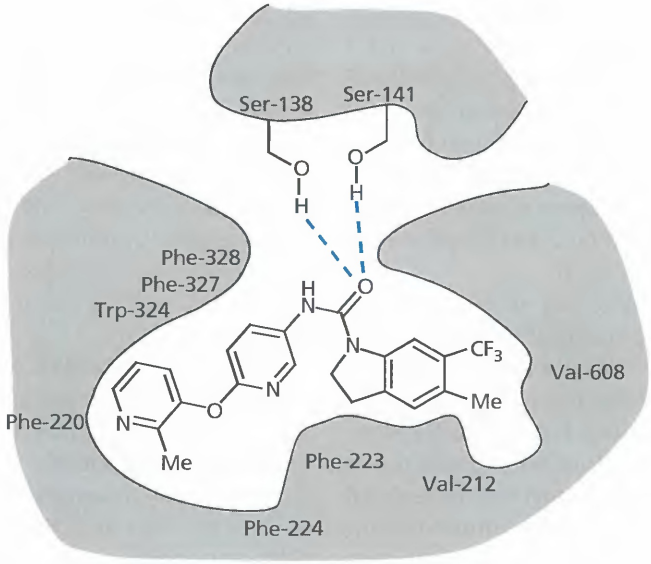


FIGURE CS6.25 Binding of SB 243213.

FURTHER READING

Bromidge, S. M., *et al.* (1997) 6-Chloro-5-methyl-1-[[2-[(2-methyl-3-pyridyl)oxy]-5-pyridyl]carbamoyl]-indoline (SB-242084): The first selective and brain penetrant 5-HT_{2C} receptor antagonist. *Journal of Medicinal Chemistry*, **40**, 3494–3496.

Bromidge, S. M., Dabbs, S., Davies, D. T., *et al.* (1998) Novel and selective 5-$HT_{2C/2B}$ receptor antagonists as potential anxiolytic agents: Synthesis, quantitative structure-activity relationships, and molecular modeling of substituted 1-(3-pyridylcarbamoyl)indolines. *Journal of Medicinal Chemistry*, **41**, 1598–1612.

Bromidge, S. M., Dabbs, S., Davies, D. T., *et al.* (1999) Model studies on a synthetically facile series of N-substituted phenyl-N-pyridin-3-yl ureas leading to 1-(3-pyridylcarbamoyl)indolines that are potent and selective 5-$HT_{2C/2B}$ receptor antagonists. *Bioorganic and Medicinal Chemistry*, **7**, 2767–2773.

Bromidge, S. M., Dabbs, S., Davies, D. T., *et al.* (2000) Biarylcarbamoylindolines are novel and selective 5-HT2C receptor inverse agonists: identification of 5-methyl-1-[[2-[(2-methyl-3-pyridyl)oxy]-5-pyridyl]carbamoyl]-6-trifluoromethylindoline (SB 243213) as a potential antidepressant/anxiolytic agent. *Journal of Medicinal Chemistry*, **43**, 1123–1134.

PART E

Selected topics in medicinal chemistry

In section E, we shall look at seven particular topics in medicinal chemistry. The topics which have been included are of the author's choosing, and demonstrate a personal preference or interest. There are many other fascinating topics which could have been included, but sadly there is a limit to what can be included in a textbook of this kind.

The different chapters illustrate different methods of classifying drugs, and demonstrate the advantages and disadvantages of such classifications. For example, there are four chapters where drugs are classified by their pharmacological effect, mainly those with antibacterial, antiviral, anticancer and antiulcer activities. The advantage of this classification is that it gives an overall view of the many different types of drug that can be used to treat these diseases. The disadvantage is the volume of information that has to be imparted. Since these diseases have many causes or mechanisms, there are many different possible targets for drugs. This means that the types of drugs which can be used in each of these fields are extremely varied in their structure and mechanism of action.

In contrast, the chapters on cholinergic and adrenergic agents concentrate on drugs which interact with specific biological systems, mainly the cholinergic and adrenergic nervous system. Since the target systems are more focused, there are fewer targets to consider. This has the advantage that it is easier to rationalize the drug structures involved and to study their mechanisms of action. It is also possible to understand why drugs having an effect on these systems can be used in particular fields of medicine, such as in the treatment of asthma or in cardiovascular medicine. The disadvantage in concentrating on a particular biological system is that it ignores drugs that could be anti-asthmatic or cardiovascular by acting on a different biological system.

A study of opioid analgesics is included where the drugs have been classed by their chemical structure. This is a useful classification for medicinal chemists since the drugs involved have the same pharmacological activity and targets, making their study well focused. The disadvantage with this classification is that analgesics with different structures and mechanisms of action are not included.

Topics have also been chosen to include several traditional fields of medicinal chemistry as well as those which are more recent. For example, the opioid analgesics were discovered more than 100 years ago, while most antibacterial agents were discovered in the mid twentieth century. In contrast, progress in antiviral therapy has been relatively recent. A comparison of these chapters illustrates the changing face of medicinal chemistry from one of trial and error to a more scientific approach where diseases are understood at the molecular level and drugs are then designed accordingly.

Case Study 7 looks at an example of a research project undertaken to find a novel antidepressant agent, but also includes an overview of clinically useful antidepressants and their mechanism of action. As such, it acts as a link to relevant material within other chapters in the book.

