Themed Issue: Enabling Drug Developability - Challenges in Optimizing Biopharmaceutical Properties of New Drug Candidates
Guest Editor - Saeho Chong

Recent Progress in the Computational Prediction of Aqueous Solubility and Absorption

Submitted: October 25, 2005; Accepted: December 5, 2005; Published: February 3, 2006

Stephen R. Johnson[1] and Weifan Zheng[2]

[1]Computer-Assisted Drug Design, Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, NJ; stephen.johnson@bms.com
[2]Division of Medicinal Chemistry, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC; weifan_zheng@unc.edu

ABSTRACT

The computational prediction of aqueous solubility and/or human absorption has been the goal of many researchers in recent years. Such an *in silico* counterpart to the biopharmaceutical classification system (BCS) would have great utility. This review focuses on recent developments in the computational prediction of aqueous solubility, P-glycoprotein transport, and passive absorption. We find that, while great progress has been achieved, models that can reliably affect chemistry and development are still lacking. We briefly discuss aspects of emerging scientific understanding that may lead to breakthroughs in the computational modeling of these properties.

KEYWORDS: Solubility, permeability, P-glycoprotein, BCS, in silico, prediction

INTRODUCTION

The biopharmaceutical classification system (BCS) presents a basis for categorizing a drug based on its aqueous solubility and permeability with an eye toward drug absorption.[1] The BCS has had great impact in early development of potential therapeutics by providing significant insight into the formulation of new drugs, in addition to its primary role in the waiver of in vivo bioequivalence studies for BCS class I drugs. More recently, there has been significant interest in considering the BCS properties in the discovery stages of research.

The BCS guidance considers 2 primary facets of a drug's properties. The solubility designation is based upon the lowest solubility determined for a compound in the pH range of 1.0 to 7.5. A compound is considered highly soluble if the highest immediate-release dose is soluble in 250 mL of aqueous media in this entire pH range. Otherwise, the compound falls into the poor solubility category. It is worth noting that, while the U.S. Food and Drug Administration's guidance[2] is clear, there is considerable debate in the literature on whether this definition of high and low solubility is too conservative.[3] The alternate volume of 500 mL has been proposed as a better choice because it reflects the average volume of the small intestines in the fasted state. There is also debate about the pH range of interest, with the range of 1.0 to 6.8 being suggested, given that a compound should be fully dissolved prior to reaching the ileum, assuming rapid dissolution. The high permeability criterion is defined as 90% or higher intestinal absorption. Below 90% absorption, the compound is considered to have low permeability. In practice, this is somewhat difficult to determine in discovery, as human data are only sparingly determined.

This report reviews the current status of computational tools in predicting the base properties of the BCS. Although we have not been truly comprehensive, we have sought to provide as complete an understanding of the status of the field as possible. In addition, we provide our perspective on the progress of research into an in silico equivalent to the BCS.

SOLUBILITY PREDICTION

As solubility prediction has been the focus of several reviews over recent years,[4-11] we will focus on work from only the last few years, with a particular focus on related methodologies that may positively affect solubility prediction in the future. We turn the reader's attention to an excellent review of solubility prediction by Delaney,[7] who discusses some of the major obstacles faced in empirical solubility modeling.[12] Most of the previous reviews point out that solubility modeling efforts have suffered from some basic concerns, such as training sets that are not druglike, unknown or high experimental error, lack of structural diversity, incorrect tautomers or structures, neglect of ionization, no consideration of salt and/or common ion issues, avoidance of crystal packing effects, and range in solubility data that is not pharmaceutically relevant. A small number of publications are starting to address many of these issues, especially the issues relating to data quality and scope, though no publication has tried to address all of them. Practically speaking, a model that addressed all of these issues

Corresponding Author: Stephen R. Johnson, Bristol-Myers Squibb Co, PO Box 4000, Princeton, NJ 08543. Tel: 609-252-3003; E-mail: stephen.johnson@bms.com

might be so complicated as to be of limited value to most discovery organizations.

Delaney[7] mentions the need for a "fit-for-purpose" (FFP) metric to compensate for the broad range of solubilities over which most reports model performance. By presenting statistics over ranges of up to 13 orders of magnitude, the statistics often portray a rosier picture of the model's utility than is often experienced in a pharmaceutical discovery project. We are most interested in the performance of models in the key 0.1 μg/mL to 250 μg/mL range for discovery, which corresponds to roughly –7 to –3.5 logS in molarity for a compound with a molecular weight (MW) of ~500. When higher than this range, solubility is rarely at the forefront of issues in a discovery project. Solubility measurements below 0.1 μg/mL are very difficult analytically, making it practically impossible to verify such predictions. In this review, we will report statistics for solubility models for compounds in this range, when possible, referring to it as the "FFP range." When only a qualitative statement can be made based on a figure from the original paper, we will identify it as such.

Several manuscripts have appeared with models of varying complexity for the prediction of solubility.[4,8-10,13-36] We will discuss only a few in depth; several others are listed in Table 1.

Bergstrom et al used carefully measured solubility data for druglike compounds to develop a series of predictive models.[28] Both general models and a series of more localized models were generated and compared. The consensus model used a combination of topological, physicochemical, and surface area descriptors and had an $R^2 = 0.80$. In the FFP range, the consensus was $R^2 = 0.62$, which is encouraging. The authors also reported predictions for the oft-used data from Huuskonen et al[38] and Jorgensen and Duffy,[39] providing a nice example of a common phenomenon in absorption, distribution, metabolism, and excretion (ADME) modeling: the lack of transferability of empirically parameterized models. The correlation for compounds in this external test set in the FFP range was a very disappointing $R^2 = 0.22$. To the authors' credit, the paper contains a significant discussion of the suitability of this external set compared with the training data. Much of this discussion echoes concerns raised previously.[12,22]

Yan and Gasteiger have published several recent reports of solubility models.[22,25,29] In their initial reports,[22,25] the authors used the Huuskonen et al[38] data set for training but found that it was of limited applicability when the resulting models were tested by a large data set provided by Merck KGaA. In their most recent report,[29] the authors used the

Table 1. Summary of Recent Aqueous Solubility Models*

Authors	N	Modeling Methods	Some Relevant Statistics	Descriptors
Engkvist and Wrede[35]	3042	CNN	Train $R^2 = 0.91$ Valid $R^2 = 0.86$	Topological and constitutional
Cheng and Merz[14]	809	GA/MLR	Train $R^2 = 0.84$	Topological
Wegner and Zell[34]	1016	CNN	Train $R^2 = 0.94$ Valid $R^2 = 0.82$	Topological, electronic
Manallack et al[30]	788	CNN	87.18% correct	BCUT
Schaper et al[37]	787	MLR	Train $R^2 = 0.94$ Valid $R^2 = 0.92$	HYBOT
Yan and Gasteiger[22]	793	CNN	Train $R^2 = 0.92$ Valid $R^2 = 0.94$	Topological
Yan and Gasteiger[25]	797	CNN	Train $R^2 = 0.93$ Valid $R^2 = 0.92$	3D
Yan et al[29]	1217	CNN	Train $R^2 = 0.86$ Valid $R^2 = 0.81$	3D
Hou et al[23]	1290	MLR	Train $R^2 = 0.92$ Valid $R^2 = 0.88$	Atom types
Votano et al[19]	3343	CNN	Train $R^2 = 0.88$ Valid $R^2 = 0.77$	Topological, constitutional
Bergstrom et al[28]	85	PLS	Train $R^2 = 0.80$ Valid $R^2 = 0.59$	2D and 3D
Raevsky et al[36]	1063	MLR	Train $R^2 = 0.90$	HYBOT, nearest neighbor similarity

*CNN indicates computational neural network; Train, training set statistics; Valid, external validation set statistics; GA/MLR, genetic algorithm/multiple linear regression; BCUT, Pearlman's BCUT descriptors; HYBOT, hydrogen bond thermodynamics descriptors; PLS, partial least squares.

Merck data set for training and the Huuskonen data for prediction. The outcome was quite reasonable for the Huuskonen test data, although both the correlation coefficient ($R^2 = 0.83$ compared with $R^2 = 0.92$ previously, 3D neural network models) and the mean absolute error (0.66, compared with 0.49 in the earlier work) degraded substantially compared with those of the earlier report. The authors note that the Merck data set contains a more diverse collection of compounds than the Huuskonen data, which may partially explain why the model trained on the Merck data performed somewhat better in predicting the Huuskonen data than vice versa. Not enough data were given to determine the results in the FFP range, although based on the figures in the earlier papers,[22,25] the statistics for the test data in the FFP range were not as good as for the entire data set.

Two very interesting recent papers applied the HYBOT descriptors to solubility prediction of liquids[37] and solids.[36] They are both discussed in greater detail below, in the section about crystal effects on solubility prediction.

A few reports of models for the prediction of aqueous solubility of pharmaceutical salts have emerged in recent years.[40-43] Parshad et al developed a model for the solubility of benzylamine salts with a combination of experimental and theoretical descriptors.[42] The best purely computational model employed the Charton steric parameter, Hansch parameters, and the MW of the salt, resulting in $R^2 = 0.73$ for the training data and $R^2 = 0.70$ for the validation data. Tantishaiyakul[41] reused the benzylamine data set to develop a computational neural network model using only calculated descriptors. The model used the number of H-bond acceptor oxygens, the total H-bond number, the clogP, the surface area, MW, and the calculated binding energy between the salt and benzylamine. The 6-2-1 network had an overall $R^2 = 0.87$ and a root-mean-square error (RMSe) = 0.17. An earlier report by Tantishaiyakul[40] reported a simple partial least-square (PLS) model for the solubility prediction for a series of diclofenac salts. Though limited in scope, these reports represent an interesting application of modeling.

To forecast the exposure of a potential drug, the entire pH range of the gastrointestinal tract must be accounted for in the prediction of aqueous solubility.[44-46] Indeed, fundamental to the BCS is the dose number, or the ratio of the dose to the amount of compound that will dissolve in 250 mL at the minimum solubility in a pH range of 1 to 8.[47] Several models have appeared that account for ionization, typically through the use of the Henderson-Hasselbach (HH) equation with predicted pK_a values. Bergstrom et al[48] performed an elegant study of the appropriateness of the HH equation to 25 druglike monoprotolytic cationic compounds in divalent buffer systems. They show a range of slopes from 0.5 to 8.6 for the linear portion of the experimental pH-solubility curve, compared with the assumed value of 1 used in the HH equation. They conclude that this variability in slopes is due to a combination of low-MW aggregation and salting out effects of the phosphate counter-ion. Furthermore, they show that the range of solubility over the pH-solubility curve varies from 1.1 log units to 6.3 log units, most likely as a result of the common ion effect. In total, the HH estimated solubility at pH 6.5 deviated from the measured solubility by 10-fold on average, and up to 776-fold in the extreme.

Interestingly, Schaper et al attempted to correct their training data for the effect of ionization.[37] A relatively small portion of their training data were ionizable compounds with no reported pH information. Their assumption was that the solubility measurements were done in unbuffered solution. To determine the pH of the solution (and, therefore, the fraction of compound un-ionized at that pH), they numerically solved a third- or fourth-order equation using an observed solubility measurement and known pK_a. They then used the resulting fraction un-ionized term as a feature in their model.

The aqueous solubility of potential drugs is typically measured in a buffered salt solution, which further complicates solubility prediction. The Setschenow equation[49] describes the ratio of the solubilities of an organic solute in an aqueous salt solution versus pure water:

$$\log\left(\frac{S}{S_o}\right) = -K_{salt}C_{salt} \qquad (1)$$

where S is the solubility of the solute in the aqueous salt solution, S_o is the solubility in water, C_{salt} is the molar concentration of the electrolyte, and K_{salt} is the solute-specific Setschenow constant. There have been 2 recent examples of models reported to predict Setschenow constants.[50,51] Ni and Yalkowsky[50] demonstrate a fairly simple relationship between clogP and the Setschenow constant ($K_{salt} = 0.040$ clogP + 0.114, $r^2 = 0.60$, n = 101, SE = 0.041). They also show that this relationship is superior to previously hypothesized links between K_{salt} and aqueous solubility[52] or the molar volume calculated by the Le Bas method.[53] Li et al[51] generate results of similar quality using topological connectivity indices. A simplified mechanistic look at Setschenow constants was among the properties investigated by the "Mercedes Benz" model of water.[54] The direct application of any of these approaches in the in silico prediction of aqueous solubility is still lacking.

Yet another factor in solubility prediction that is often ignored is the role of the crystal form of a druglike solute (an early notable exception is that of Abraham and Le[55]). An excellent recent review discusses the importance of solid-state properties on a range of developability considerations in drug discovery and development.[56] Most interesting out of this review, with respect to solubility, is the relative rarity

of large solubility differences among polymorphs. Indeed, the authors state that the "solubility difference between different polymorphs is typically less than 10 times" (p.323). Pudipeddi and Serajuddin[57] go further, providing data showing that the ratio is typically less than 2. Contrasting this, however, is the extreme difference in solubility often observed between amorphous material and crystalline material.[56] In aggregate, these data could lead to the alluring possibility that, while the ability to identify the most stable crystal lattice may not be necessary, the incorporation of some realistic measure of crystal packing is crucial to practical solubility prediction.

Conversely, Nielsen et al[58] attempted to quantitatively link aspects of crystal packing to solubility for a series of N-alkyl bupivacaine salts. They did see a trend of decreasing solubility with increasing crystal lattice density. However, their analysis was frustrated by unpredictable changes in packing modes due to anionic counter-ions and the alkyl substituent. They conclude that reliable lattice energy calculations are required for deriving relationships between solubility and solid-state characteristics, even for a series of closely related analogs. This opinion is supported by Romero and Rhodes, who find that even quantifying the impact of crystals of different enantiomers on solubility is nontrivial.[59]

Probably the best-known solubility model that includes crystal forces is the general solubility equation (GSE) popularized by Yalkowsky[60]:

$$Log(S_o) = -0.01(MP - 25°C) - \log P + 0.50 \qquad (2)$$

where S_o is the intrinsic aqueous solubility in the units of moles/L, logP is the octanol-water partition coefficient, and MP is the melting point in °C. Derived from a theoretical basis with no fit parameters, the GSE is based on a few simple assumptions, including the applicability of Walden's rule (entropy of melting) and the fact that organic neutral liquids are completely miscible with octanol. Still, the GSE performs admirably in prediction for simple organics. The requirement of a measured melting point has greatly limited the use of the GSE within the pharmaceutical industry, as melting point measurements are no longer routine in modern medicinal chemistry. Not surprisingly, several reports relating to the prediction of melting point from structure have emerged in recent years.[10,61-65] We will not review melting point prediction in depth here; instead, we refer the reader to the fairly recent perspective by Katritzky et al.[66] More recently, Yalkowsky has derived new parameters for predicting the entropy of melting, which presumably could be used in the GSE to remove the assumption of the applicability of Walden's rule.[63] Bergstrom and coworkers[61] generated a consensus model using 277 diverse drugs for melting point prediction that focuses on descriptors of polarity and molecular flexibility. The RMSe of the validation data was 44.6°C. Karthikeyan et al used 4173 compounds to develop a computational neural network model of melting point.[64] The best resulting model employed 26 2-dimensional descriptors in a 26-12-1 neural network, resulting in $R^2 = 0.66$, with RMSe ranging from 41.4 to 49.3°C depending on which external validation set was used.

In 2 papers Raevsky and colleagues build on the GSE by deconvoluting the logP component into separate descriptors.[36,37] Using a collection of liquid chemicals and drugs,[37] they demonstrate that descriptors previously identified[67-70] as being relevant for logP prediction perform admirably in the GSE in place of the logP term. This is an interesting result, as hydrogen bond donor strength was not an important descriptor for logP prediction but appears critical for modeling solubility. By adding H-bond donor strength to their model, they improve the solubility prediction significantly.

In the follow-up paper that considers solid chemicals and drugs,[36] the authors apply the GSE to 1063 neutral solid-state compounds with known melting point. The statistical performance ($R^2 = 0.85$, RMSe = 0.81) was substantially less impressive than for previous data sets that included a large number of liquid-state compounds.[71] Predictions were more than a log unit off the observed values for ~19% of the compounds. Nonetheless, in our opinion this result (based on experimentally determined melting point and logP) is more than competitive with most external tests of purely in silico models reported in the literature. Based on a qualitative view of the figure, the predictions in the FFP range were worse than those for the data set as a whole. Of the 26 outliers noted for the entire 13 orders of magnitude, 12 fell in the 3.5 log units designated here as the FFP range.

The authors then take a novel approach to a computational model of the solubility of solids. They implicitly capture the crystal lattice energy using the observed solubility of several highly similar compounds with an adjustment based on the differences in the physicochemical descriptors identified as important for predicting the solubility of liquids.[36] The resulting model performance is very competitive with those obtained for the GSE, but without the requirement for a measured melting point. Roughly 7 of the 14 outliers in this model fall in the FFP range, based on a plot in the paper. Based on the figures in the paper, the fit seems better in this range for the computational model than for the experimentally based GSE.

Several methods[72-80] for predicting crystal lattices or quantifying lattice energy, with possible use in solubility prediction, are on the horizon. We particularly highlight the excellent recent review by Datta and Grant,[77] who discuss in comfortable detail recent efforts to predict crystal structures of druglike compounds. The molecular dynamics method proposed by Gavezzotti[78] to assess the relative stability of crystal lattices is particularly appealing, although difficult to implement as a fully computational method in

practice. In general, these approaches hold great promise but, due to their complexity and predictive limitations, are not yet practical solutions for a general treatment of crystal packing limited solubility prediction.

In summary, substantial research has been underway in predicting the aqueous solubility of druglike compounds. Of particular note is the increased awareness of the many confounding aspects of solubility that must be considered for a model to yield highly accurate predictions. We expect that in the coming years more sophisticated models will emerge that begin to more tightly integrate measures of crystal packing, salt effects, and ionization alongside solvation considerations.

Absorption Modeling

Modeling of P-Glycoprotein Substrates and Inhibitors

P-glycoprotein (Pgp), which belongs to the adenosine triphosphate binding cassette transporter family, is found in all cells in every species. The effect of Pgp-mediated drug efflux limits intestinal absorption and oral bioavailability of drugs. Because of this critical role in oral absorption and bioavailability, extensive research has been conducted to uncover the molecular features required for the substrates and/or inhibitors of Pgp transporter. Such models are expected to play an important role in early drug discovery and help reduce the attrition rate in later-stage drug development.

In recent years there have been many research articles that examined the molecular determinants of Pgp substrates and/or inhibitors. Both qualitative and quantitative molecular models have been developed to offer insights into the molecular mechanisms as well as to provide predictive tools. Some selected modeling work is listed in Table 2.

In 1998 Pajeva and Wiese[81] published their work on comparative molecular field analysis (CoMFA) studies of phenothiazines. Later, they conducted pharmacophore modeling research of drugs involved in Pgp multidrug resistance using the genetic algorithm similarity program (GASP) pharmacophore modeling tool.[82] In a widely cited study, Seelig identified 2 recognition elements for Pgp that were composed of hydrogen bond acceptors with distinct spatial arrangements.[90] Seelig referred to 2 hydrogen bond acceptors separated by ~2.5 Å as a Type I pattern; Type II patterns are formed by 2 hydrogen bond acceptors separated by ~4.6 Å, or 3 hydrogen bond acceptors separated by ~2.5 Å with a 4.6 Å separation of the outer 2 acceptor groups.

In 2002, Ekins et al reported their studies on both Pgp substrates and inhibitors using the catalyst pharmacophore modeling method.[83,91] Although their work has not provided detailed predictive models tested by extensive external compounds, they seem to have provided some molecular insights consistent with experimental observations. Penzotti et al described their work on classification models using a special pharmacophore ensemble approach to classify Pgp substrates with an overall accuracy of 63%.[84] In addition to classification models, this method revealed the molecular pharmacophores that underlie the interaction between Pgp substrates and the protein.

In the past 2 to 3 years, more work has been published concentrating on applying machine learning methods and descriptor-based quantitative structure-property relationship (QSPR) methods. Xue et al used support vector machine (SVM) to study 201 compounds, including 116 Pgp substrates and 85 nonsubstrates.[85] This method gave a prediction accuracy of at least 81.2% for Pgp substrates. The prediction accuracy for nonsubstrates was 79.2% using a cross-validation. A data set of 57 flavonoid Pgp inhibitors was studied using a Bayesian-regularized neural network in which Molconn-Z and clogP descriptors were used.[86] Wang et al also used Kohonen self-organizing maps to develop a classification model to discriminate substrates and inhibitors with an average accuracy of 82.3%.[87] The data set has 206 chemicals: 96 substrates, 78 inhibitors, and 32 overlapping compounds. Although these models may not have practical utility because of the nature of the data set collected from the literature, they do represent some new development in applying machine learning methods to the study of Pgp substrates/inhibitors classification.

Two recent publications are especially interesting from a practical standpoint. Gombar et al published their work using what they called "information-rich descriptors" and linear discriminate analysis.[88] They analyzed a set of 95 compounds, including GlaxoSmithKline proprietary and known drug molecules. They paid special attention to the data consistency with unified experimental protocol. In addition to a regular QSPR classification model, they proposed a simple rule for Pgp substrates. Another recent work described a novel approach to performing alignment-independent 3D quantitative structure-activity relationship (QSAR).[89] This method not only afforded statistically tested predictive models but also revealed potential pharmacophore requirements. In the following text, we provide some more detailed discussion of these 2 publications as well as that published by Penzotti et al.[84]

One of the early efforts in gathering and modeling a large set of diverse Pgp compounds was published by Penzotti et al.[84] They assembled a data set of 195 compounds from the literature. In the study, they used a holdout set of ~25% of the compounds selected randomly from the 195 compounds. The test set contained 32 Pgp substrates and 19 nonsubstrates. The remaining 144 compounds, including 76 substrates and 68 nonsubstrates, were used as the training data set to derive the computational model.

Table 2. Summary of Selected Computational Modeling Work on P-Glycoprotein Substrates/Inhibitors*

Authors	Data Set Size	Modeling Methods	Some Relevant Statistics	Descriptors
Pajeva and Wiese[81]	40	CoMFA	$R^2 = 0.79$; $Q^2 = 0.35$ $R^2 = 0.90$; $Q^2 = 0.84$	CoMFA
Pajeva and Wiese[82]	25	Pharmacophore GASP		GASP scoring
Ekins et al[83]	<20	Catalyst Pharmacophore		Catalyst
Penzotti et al[84]	195	Pharmacophore ensemble	CR: 63%	Pharmacophore descriptors
Xue et al[85]	201	Classification	CR: 79%-81%	Properties, Molconn-Z, quantum, ES, and geometric
Wang et al[86]	57	QSAR	$R^2 = 0.73$–0.75	ClogP, Molconn-Z
Wang et al[87]	206	Kohonen SOM classification	CR: 82.3%	
Gombar et al[88]	95	LDA classification	Training CR: 100%; 90% Test CR: 94%; 78%	Estate; topological etc
Cianchetta et al[89]	129	PLS	$R^2 = \sim 0.7$–0.8 $Q^2 = \sim 0.5$–0.7	VolSurf; GRIND

*CoMFA indicates comparative molecular field analysis; GASP, genetic algorithm similarity program; CR, classification rate; ES, electrotopological state; QSAR, quantitative structure-activity relationship; SOM, self-organizing map; LDA, linear discriminant analysis; PLS, partial least-square; GRIND, grid independent descriptors.

One interesting fact about their method is that they characterized the chemical similarity using the Daylight Tanimoto similarity measure. The average pairwise Daylight Tanimoto similarity of all 195 compounds is 0.18. The training set and the test set compounds also have comparable pairwise similarities. This is an important metric because it indicates that their method may be used to predict compounds from different chemical classes. The classification rates of the final pharmacophore ensemble model are 80% and 63% for the training and test sets, respectively.

In addition to the classification model, which can be used as a virtual library filter, the authors derived molecular pharmacophores. Contained in the significant pharmacophores are examples of the Type I and Type II patterns composed of hydrogen bond acceptors described by Seelig.[90] This indicates that their method can successfully provide molecular models in addition to classification models.

Gombar et al obtained one of the most remarkable results on the modeling of Pgp substrates.[88] They used a training set of 95 compounds (63 substrates and 32 nonsubstrates) based on the results from in vitro monolayer efflux assays. (All 95 compounds were uniformly assayed for their Pgp activity.) The use of uniform assay conditions is an important requirement for reliable model development. They derived a 2-group linear discriminant analysis (LDA) model. Remarkably, the model computed the probability that a structure is a Pgp substrate with a sensitivity of 100% (ability to correctly identify substrates) and a specificity of 90.6% (ability to correctly identify nonsubstrates) in the cross-validation test. A prediction accuracy of 86.2% was obtained on an additional test set of 58 compounds (35 substrates + 23 nonsubstrates) with sensitivity 94.3% and specificity 78.3%.

Another interesting aspect of this work is that they derived a simple rule: those molecules with MolES > 110 were predominantly Pgp substrates (18/19: 95%), and those with MolES < 49 (11/13: 84.6%) were predominantly nonsubstrates. Here, MolES is the molecular bulk calculated as the sum of atomic electrotopological states (ES) values.

In 2005 Cianchetta et al described another exciting effort.[89] They derived a pharmacophore hypothesis based on a set of 129 compounds (100 Sanofi-Aventis compounds and 29 publicly available compounds). The data set was divided into 4 classes based on compound activity values. They proposed a novel set of descriptors based on GRID calculated interaction fields followed by modified autocorrelation calculation. This generated 940 descriptors similar to pharmacophore keys. The results included both predictive models and pharmacophore descriptions. The models using VolSurf descriptors achieved training set $R^2 = 0.72$ and $Q^2 = 0.52$; for combined pharmacophore descriptors and 94 VolSurf descriptors, the training $R^2 = 0.80$ and $Q^2 = 0.72$. The models derived using only pharmacophore descriptors after some preselection gave a training $R^2 = 0.82$ and $Q^2 > 0.72$. The pharmacophore hypothesis includes 2 hydrophobic groups 16.5 Å apart and 2 hydrogen bond acceptor groups 11.5 Å apart.

To develop reliable models for Pgp substrates/inhibitors identification, future work should attend more to data

consistency. Model validation using a true holdout set should always be attempted. In addition to traditional QSPR models, which can be useful in virtual screening and database search exercises, more interpretable models or rules can be very useful for medicinal chemists to consider in the lead optimization work. These are represented by Gombar's work and Cianchetta's work, reviewed above. Only when we achieve this can such tools have a broad impact on lead optimization projects in assisting chemists in morphing the chemical series at hand to increase potency and decrease Pgp liability.

Modeling of Intestinal Permeability

Intestinal drug permeability, together with aqueous solubility, is one of the most important factors influencing drug absorption. In recent years, various in vitro permeability models have been developed to help predict oral drug absorption and bioavailability. These include Caco-2, Madin-Darby canine kidney (MDCK), and 2/4/A1 cell culture models, as well as the parallel artificial membrane permeability assay (PAMPA) and immobilized artificial membrane (IAM) and physicochemical models. While the former can reflect both transcellular and paracellular routes of permeability, as well as active transport to some degree, the latter models mostly mimic the transcellular route of drug permeability.[92]

As the in vitro models continue to develop and mature in terms of their throughput and quality, computational methods that can correlate chemical structures and their experiment permeability measurements are highly desirable. These in silico models, when carefully developed and rigorously validated, have the potential to be used in early screening set or library design. They can also play a role in lead optimization and preclinical candidate selection. Several computational models for permeability are shown in Table 3. Recently, Bergstrom reviewed the application of polar surface area (PSA) as a descriptor to model permeability and solubility data.[93] Malkia et al looked at various physicochemical factors underlying the permeation process, in vitro experiment models as well as in silico modeling of permeability.[94] These reviews covered mostly articles published prior to 2002. We will look at some more recent work on this topic.

Kulkarni et al analyzed 38 compounds and tested the model using 8 additional compounds.[95] They created a model monolayer to represent the membrane structure. They calculated both intramolecular properties and intermolecular properties, which characterized the interaction with the membrane as well as solute dissolution and solvation. A series of models were developed using 1 to 6 terms using multiple linear regression (MLR). Most models gave $R^2 > 0.8$ and $Q^2 > 0.70$. This method also provided some mechanistic interpretation of the model. That is, the Caco-2 permeability depended on several factors: solubility, drug-membrane binding, and conformational flexibility.

Yamashita et al developed a model of Caco-2 permeability using genetic algorithm with partial least squares (GA-PLS).[96] Data for 73 compounds were collected from the literature. Molconn-Z descriptors were employed. The model achieved $R^2 = 0.886$ for the entire data set and a predictive $R^2 = 0.825$. In a related work, they also developed a new concept—the latent membrane permeability concept—to model permeability data from different sources.[97] Eighty-one compounds were analyzed based on this concept using an iterative calculation method. It analyzed the Caco-2 permeability from different sources simultaneously, assuming that all the data sets share a hidden, common relationship between their permeability and their physicochemical properties. The model achieved R^2 ranging from 0.75 to 0.88.

Using a combination of molecular orbital (MO)-calculation and neural network analysis, Fujiwara et al developed a model of Caco-2 permeability.[98] The data set had 87 compounds. They used dipole moment, polarizability, sum of charges on nitrogen and oxygen atoms, and so on. A feed-forward back-propagation neural network with a configuration of 5-4-1 was developed. The predictive root mean square error was 0.507 in cross-validation.

Marrero et al developed a new set of topological descriptors called quadratic indices.[99] They applied these descriptors to analyze Caco-2 permeability data for 33 compounds. The model could distinguish the high-absorption compounds from those with moderate-to-poor absorption. A global classification rate was 87.87% using LDA. In a test experiment, they used the model to assess 18 compounds; the classification rates were 80.00% and 94.44% for the moderate-to-poor and high-absorption groups, respectively.

Hou et al[100] analyzed a data set of 77 compounds with Caco-2 permeability data collected from literature sources. The descriptors included logD at pH = 7.4, highly charged PSA (HCPSA), and radius of gyration (rgyr), representing the shape and volume of the compounds. They found that logD had the largest impact on diffusion through Caco-2. The comparison among HCPSA, PSA, and topological PSA (TPSA) demonstrates the importance of the highly charged atoms to the interactions between Caco-2 cells and drugs. The results also indicate that lipophilicity, H-bonding, bulk properties, and molecular flexibility can improve the correlation. The model statistics are n = 77, $R^2 = 0.82$, and $Q^2 = 0.79$. The authors also validated the model using an external set of 23 compounds and achieved similar predictions.

Fujikawa et al developed QSPR models for data based on PAMPA assay.[101] Descriptors included logP, $|pK_a - pH|$ and the surface areas occupied by the hydrogen-bond acceptor (SAHA) and donor atoms (SAHB). The data set had 35

Table 3. Selected Work on Computational Modeling of Permeability*

Authors	Size of Data Sets	Methodology	Descriptors	Statistics
Refsgaard et al[102]	712	Near neighbor method	ClogP, rotbond, HBD/HBA, PSA, MW, etc	CR: 85%
Fujikawa et al[101]	35	PLS	$\|pK_a\text{-}pH\|$, logP, SAHA, SAHB, PSA	Model 1: $R^2 = 0.84$; $Q^2 = 0.79$ Model 2: $R^2 = 0.78$; $Q^2 = 0.74$
Hou et al[100]	77	MLR	LogD, HCPSA, rgyr, MW, vol, etc	$R^2 = 0.82$; $Q^2 = 0.79$
Marrero et al[99]	33	LDA	Quadratic indices	Training CR: 87.87% Test CR: 80% and 94.44%
Fujiwara et al[98]	87	CNN	Dipole, polarizability, charges on N, O, etc	RMSe: 0.507
Yamashita et al[96]	73	GA-PLS	Molconn-Z	$R^2 = 0.89$; $Q^2 = 0.83$
Yamashita et al[97]	81	MLR and latent permeability concept	Dipole, polarizability, charges on N, O	$R^2 = 0.75$-0.88
Kulkarni et al[95]	38	MLR	Membrane interaction descriptors	$R^2 > 0.8$; $Q^2 > 0.70$

*HBD/HBA indicates number of hydrogen bond donors and acceptors; PSA, polar surface area; MW, molecular weight; CR, classification rate; PLS, partial least-square; SAHA, surface area of hydrogen bond acceptors; SAHB, surface area of hydrogen bond donors; MLR, multiple linear regression; HCPSA, highly charged polar surface area; LDA, linear discriminant analysis; CNN, computational neural network; RMSe, root-mean-square error; GA-PLS, genetic algorithm–partial least squares.

compounds. The authors used the PLS correlation method. The model yielded $R^2 = 0.84$ and $Q^2 = 0.79$. In another data set that had 57 compounds, the authors obtained a model with $R^2 = 0.78$ and $Q^2 = 0.74$. They also demonstrated that PAMPA and Caco-2 data were well correlated for this data set (n = 27; $R^2 = 0.81$; $Q^2 = 0.78$). Based on these results, they suggested a procedure by which one could in theory predict the oral absorption of compounds.

In a recent publication, Refsgaard et al developed in silico models of membrane permeability based on the largest self-consistent data set published so far.[102] They analyzed 712 compounds (380 nonpermeable, 332 permeable) to develop a 2-class classification model. The permeability data were binned into 2 classes based on apparent permeability: those below 4×10^{-6} cm/s were classified as low permeability, and those with 4×10^{-6} cm/s or higher were classified as high permeability. Nine molecular descriptors were calculated for each compound. A 5-descriptor Near Neighbor model (number of flexible bonds, number of hydrogen bond acceptors and donors, and molecular surface area and PSA) was built. The model was tested using an external set of 112 compounds. The misclassification rate was 15%, and no compounds were falsely predicted in the nonpermeable class.

Most of the published work is based on small sets of permeability data (<100 compounds) that are collected from literature sources. Since the validity of any model depends on the data set used, a large and self-consistent data set is required for the development of global models with general applicability. It is highly desirable to develop the models using data generated from the same lab using the same protocols. For example, Artursson et al compared 4 calibration curves relating the fraction absorbed in human and Caco-2 permeability. They demonstrated that the curves were shifted relative to one another by ~0.25 to 1.75 log permeability units.[103] Local models may solve this problem by building compound class-specific models using, again, self-consistent data. It is exciting to see such work as Refsgaard et al's publication,[102] which was based on a large set of >700 compounds with data from the same laboratory.

Model validation is another critically important step in building robust QSPR models. Ideally, one should report results based on the training set, the cross-validation set, and an external data set to increase the users' confidence level. Consistent reporting of model statistics is highly desirable so that readers can objectively evaluate the model quality and applicability in a real-life drug discovery setting. For example, for quantitative QSPR models, one may want to report the (training set) R^2, leave-one-out (leave-some-out) R^2 or Q^2, external holdout R^2, and RMSe between the experimental values and the predicted values. For classification models, a "confusion matrix" may need to be published in addition to an overall classification rate, which may be biased by extreme values. The confusion matrix also provides information on false-positive and false-negative in addition to correct classifications.

We hope that more permeability modeling work will be based on larger self-consistent data sets and will adopt rigorous validation procedures, as discussed above. Only by doing so can we move the field forward from the proof-of-concept stage to having real-life impact in drug discovery setting.

Modeling of Bioavailability and Fraction Dose Absorbed

Bioavailability is used to describe the fraction of an administered drug that reaches the systemic circulation and its site of action. Oral bioavailability is the result of a complex series of events: chemical and enzymatic stability, solubility and dissolution, and intestinal permeability. Because of incomplete absorption and first pass metabolism and other stability factors, bioavailability is usually <100%. To lower the attrition rate of drug development, we need to develop robust and accurate in silico models that can predict and prioritize compounds before they are synthesized or moved forward to preclinical and clinical development.

There are several approaches to the development of computational models for bioavailability and fraction dose absorbed. Rules derived from statistical analysis of known oral drugs have been popularized by Lipinski et al's seminal work.[104] More predictive mechanistic models are appearing in the literature.[105-107] QSPR methods that employ molecular descriptors and machine learning techniques continue to develop. Here we review some of the work in this area, most of which was published after 2002.

Table 4. Selected Rules or Alerts Derived Statistically for Absorption/Bioavailability*

Authors	Rules or Alerts
Palm et al[119]	FA > 90% if PSA ≤ 60 Å^2; FA < 10% if PSA ≥ 140 Å^2
Lipinski et al[104]	Compounds are more likely to be bioavailable, if logP ≤ 5; HBD ≤ 5; HBA ≤ 10; MW ≤ 500
Veber et al[108]	Compounds are likely to be bioavailable, if (1) rotatable bonds ≤ 10 AND (2) PSA ≤ 140 Å^2 or total HB count ≤ 12
Martin[111]	Anions ABS = 0.11 if PSA >150 ABS = 0.56 if 75 Å^2 ≤ PSA ≤ 150 Å^2 ABS = 0.85 if PSA is < 75 Å^2 Cationic and neutral compounds ABS = 0.55 if Lipinski's rule passes ABS = 0.17 if Lipinski's rule fails

*FA indicates fraction absorbed; PSA, polar surface area; HBD, number of hydrogen bond donors; HBA, number of hydrogen bond acceptors; MW, molecular weight; HB, number of hydrogen bond donors and acceptors; ABS, a bioavailability score.

Statistically Derived Rules

Since Lipinski et al's influential work[104] on the analysis of orally active drugs to derive the widely publicized rule (ie, the "rule of 5"), several groups have studied this issue, trying to uncover other factors that may also be important for oral bioavailability. Veber et al described their analysis of oral bioavailability data in rats for over 1100 drug candidates.[108] A number of these rules and alerts are shown in Table 4. They found that reduced molecular flexibility and low PSA (or total hydrogen bond counts) are important predictors of good oral bioavailability. They suggested that compounds that meet only the 2 criteria of (1) 10 or fewer rotatable bonds and (2) PSA equal to or less than 140 A^2 (or 12 or fewer total H-bonds) will have a high probability of good oral bioavailability in the rat.

Lu et al studied the relationship of rotatable bond count and PSA with oral bioavailability in rats and compared their results with Veber et al's.[109] They examined 434 Pharmacia compounds. Although the general trend of Veber's finding was still seen, the resulting correlations depended on the calculation method and the therapeutic class of the compounds. Thus, Lu et al suggested that any generalization must be used with caution. Later, Vieth et al analyzed 1729 marketed drugs and found that oral drugs tended to be lighter and had fewer H-bond donors, acceptors, and rotatable bonds than did drugs with other routes of administration, especially when oral and injectable drugs were compared.[110] Again, this observation was in general consistent with Veber et al's finding. However, they also cautioned that these were general statistical observations and might not be applicable for a particular drug or a class of drugs.

Recently, Martin developed "a bioavailability score" (ABS), a method to estimate the bioavailability of potential drugs.[111] This work was based on a diverse set of 553 compounds (99 different rings and 180 different side chains) with rat bioavailability data. They further used the ABS to examine the human data for 449 compounds. The most interesting discovery was that the rule of 5 could identify poorly bioavailable compounds that are neutral and positively charged but could not predict anionic compounds. On the other hand, the PSA rule worked for anionic compounds, not for neutral and cationic ones. When they separated the anionic compounds from the neutral and cationic compounds, a set of interesting rules emerged, from which the ABS can be estimated.

The ABS indicates the probability that a compound will have >10% bioavailability in rat or measurable Caco-2 permeability. For example, for anions, the ABS is 0.11 if PSA >150 Å^2, the ABS is 0.56 if PSA is ≥75 and ≤150 Å^2, and the ABS is 0.85 if PSA is <75 Å^2. For cationic and neutral compounds, the ABS is 0.55 if it passes the rule of 5 and 0.17 if it fails the rule of 5.

Mechanism-Based Pharmacokinetics Models

One advantage of mechanism-based pharmacokinetics (PK) models is that they can provide mechanistic details and hypotheses that may guide the chemist's work in optimizing

the PK properties of compounds. In the past couple of years, a few published articles have indicated the potential predictive power of these models as well.

Usansky and Sinko developed an absorption-disposition model that predicted bioavailability values (%F) well for both highly and poorly absorbed drugs.[106] For 49 of the 51 compounds in the study, the residuals between predicted and experimental %F values ranged from –17% to 22%. This model offers a quantitative approach for predicting human oral absorption from in vitro permeability and perhaps from computationally predicted permeability as well.

Obata et al developed a theoretical passive absorption model (TPAM).[107] It used logD at pH 6.0, intrinsic logP, pK_a, and MW that were calculated from the chemical structures. The data set had 258 compounds with observed %F values. Only 4 coefficients in the model needed to be optimized based on experimental data. The TPAM predicted the %F values with RMSe of 15% to 21% and a correlation coefficient of 0.78 to 0.88. The possibility of overlearning was low, because only 4 coefficients in the model were optimized by fitting with hundreds of %F values.

Willmann et al developed a physiologically based model.[105] The model can be used to study the dependency of the fraction dose absorbed on the 2 main physicochemical parameters (the intestinal permeability and the solubility) as well as physiological parameters such as the gastric emptying time and the intestinal transit time. The model parameters were optimized using 126 compounds with known %F values. The model was used to predict the human %F values with permeability-limited absorption; the cross-validation RMSe was 7% for passively absorbed compounds. This model required experimentally measured permeability data. However, it is conceivable that well-validated computational data may also be applicable.

QSPR Approaches to Modeling Bioavailability

In the past 3 years, we have seen various efforts to model oral bioavailability using the descriptor-based QSPR approaches. Both linear and nonlinear learning methods have been applied in either quantitative modeling or classification modeling of bioavailability. Table 5 provides a summary of recent QSAR models of bioavailability.

Liu et al[112] analyzed 169 compounds (113 in training and 56 in test set) using SVM and the regression method implemented in comprehensive descriptors for structural and statistical analysis (CODESSA). Both the linear and nonlinear models can give satisfactory prediction results. Bai et al used classification regression trees to analyze a large set of 1261 structures and their human oral bioavailability and absorption data.[113] They used 899 compounds as the training set and 362 as the test set. Compounds were divided into 6 classes. On 2 test sets, their model achieved correct classification rates ranging from 79% to 86%.

Table 5. Selected QSPR Approaches to Bioavailability Modeling*

Authors	Size of Data Sets	Methodology	Descriptors	Statistics
Liu et al[112]	169	CODESSA regression; SVM	Constitutional, topological descriptors	Training R^2: 0.78, 0.86 Test R^2: 0.70, 0.73
Bai et al[113]	1260	Classification Regression Tree	LogP, PSA, HBD, HBA, intramolecular HB count	79%–86% correct classification
Turner et al[114]	137	CNN	Solubility, topological, constitutional, geometric	RMSe training: 19%; test set: 16%; validation: 20%
Zmuidinavicius et al[115]	1000	RP	Physicochemical, structural descriptors	Classification rate: 85%
Klopman et al[116]	417	CASE	Substructure	RMSe training: 12%; test set: 12%
Yoshida and Topliss[117]	232	Adaptive least-square	LogD; differential LogD; constitutional	Classification rate training: 71%; test set: 67%; cross-val: 60%
Andrews et al[118]	591	RP; stepwise regression	Substructure counts	RMSe: ~18%

*CODESSA indicates comprehensive descriptors for structural and statistical analysis; SVM, support vector machine; PSA, polar surface area; HBD, number of hydrogen bond donors; HBA, number of hydrogen bond acceptors; HB, number of hydrogen bond acceptors and donors; CNN, computational neural network; RMSe, root-mean-square error; RP, recursive partitioning; CASE, computer automated structure evaluation.

Turner et al[114] used artificial neural networks to analyze the human bioavailability data for 167 compounds. The model was trained with 137 compounds and tested with a further 15. An additional 15 compounds were used as a validation set. This model could distinguish compounds with low and high bioavailability. Zmuidinavicius et al[115] analyzed 1000 druglike compounds with experimental human intestinal absorption values using recursive partitioning. They were able to achieve 15% false-positives and 3% false-negatives classification rates. Klopman et al[116] developed a model based on a modified group contribution method using the computer automated structure evaluation (CASE) program. The data set contained 417 compounds. The model was able to predict the percentage of drug absorbed with an R^2 of 0.79 and a standard deviation of 12.3% for the compounds from the training set. The standard deviation for an external test set for 50 drugs was 12.3%.

In an earlier work, Yoshida and Topliss[117] analyzed 232 compounds with human bioavailability data. They discovered that acids generally had better bioavailability than bases, with neutral compounds in between. One interesting idea that came out of this observation was the formulation of a new parameter, the differential logD (logD6.5-logD7.4), which contributed significantly to the bioavailability. The model had a correct classification rate of 71%, a cross-validation rate of 67%, and a validation rate of 60% for 40 compounds. Similarly, Andrews et al analyzed 591 compounds with human oral bioavailability.[118] They used substructure count descriptors, recursive partitioning, and stepwise regression. The model achieved predictions with an RMSe of ~18%.

All 3 types of approaches to the modeling of bioavailability data have their pros and cons. The rules/alerts often provide intuitive and easy ways to think about the issues and help guide the chemist's efforts in designing new molecules. However, these rules/alerts do not often perform as well as the well-validated QSPR models. Thus, QSPR models can play a critically important role in virtual screening, compound collection, or library design work. The mechanism-based PK models can provide mechanistic insights and help decipher the components of the oral absorption process. Thus, such models often play an important role in the lead optimization and candidate selection process. It is highly likely that these in silico models will become more robust and accurate in the future as more self-consistent data of bioavailability become available. Until then, researchers have to use these rules/alerts/models with caution and should always validate these models with known experimental data before making any critical decisions.

CONCLUSION

We have attempted to provide a summary of the progress in computational modeling of aqueous solubility, Pgp efflux, and absorption. While significant effort continues in modeling these critical components of the BCS, much work remains in making predictions reliable and thus of impact to discovery. During our survey of the literature, very few applications of either solubility models or permeability models in the prospective design of new compounds were uncovered, with the exception of the use of clogP or PSA. While it is impossible to identify what has caused this dearth of application, we speculate that the primary reasons are low prediction reliability and the lack of model interpretability that would provide guidance for chemical synthesis. Our personal experience would imply that striving for high prediction accuracy, while scientifically appealing, is often at odds with assembling a model that will appeal to chemists. A better practice is to present interpretable models that can provide several testable hypotheses for advancing a chemical series. Models of this sort are still absent from the literature.

To this point, a reliable in silico analog to the BCS is still lacking, although great progress has been made at understanding the underlying properties. With typical prediction accuracies of 1 to 2 orders of magnitude in solubility, and comparably high errors of prediction for absorption-related parameters, trustworthy computational estimates of the maximum absorbable dose still appear out of reach generally. Nonetheless, as a tool alongside in vitro assays for early parameterization of physiologically based pharmacokinetic (PBPK) models, or as a starting point for refined models of a constrained series of chemical analogs, solubility and absorption models play an increasingly important role in drug discovery.

REFERENCES

1. Amidon GL, Lennernaes H, Shah VP, Crison JR. A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharm Res*. 1995;12:413-420.

2. Center for Drug Evaluation and Research. *Guidance for Industry.* Rockville, MD: CDER/FDA; 2000.

3. Yu LX, Amidon GL, Polli JE, et al. Biopharmaceutics classification system: the scientific basis for biowaiver extensions. *Pharm Res*. 2002;19:921-925.

4. Valko K. Measurements and predictions of physicochemical properties. *High-Throughput ADMETox Estimation*. 2002;1-24:A21-A25.

5. Lipinski C. Aqueous solubility in discovery, chemistry, and assay changes. *Methods Principles Med Chem*. 2003;18:215-231.

6. Lombardo F, Gifford E, Shalaeva MY. In silico ADME prediction: data, models, facts and myths. *Mini Rev Med Chem*. 2003;3:861-875.

7. Delaney JS. Predicting aqueous solubility from structure. *Drug Discov Today*. 2005;10:289-295.

8. Jorgensen WL, Duffy EM. Prediction of drug solubility from structure. *Adv Drug Deliv Rev*. 2002;54:355-366.

9. Eros D, Keri G, Kovesdi I, Szantai-Kis C, Meszaros G, Orfi L. Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS, and ANN methods. *Mini Rev Med Chem*. 2004;4:167-177.

10. Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modelling. *Adv Drug Deliv Rev*. 2003;55:1163-1183.

11. McFarland JW, Du CM, Avdeef A. Factors influencing the water solubilities of crystalline drugs. *Methods Principles Med Chem*. 2003;18:232-242.

12. Stouch TR, Kenyon JR, Johnson SR, Chen X-Q, Doweyko A, Li Y. In silico ADME/Tox: why models fail. *J Comput Aided Mol Des*. 2003;17:83-92.

13. Bergstrom CA, 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;44:1477-1488.

14. Cheng A Jr, Merz KM Jr. Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. *J Med Chem*. 2003;46:3572-3580.

15. Bergstrom CA. In silico predictions of drug solubility and permeability: two rate-limiting barriers to oral drug absorption. *Basic Clin Pharmacol Toxicol*. 2005;96:156-161.

16. Bergstrom CAS, Norinder U, Luthman K, Artursson P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm Res*. 2002;19:182-188.

17. Chen X-Q, Cho SJ, Li Y, Venkatesh S. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J Pharm Sci*. 2002;91:1838-1852.

18. Lobell M, Sivarajah V. In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pKa and AlogP98 values. *Mol Divers*. 2003;7:69-87.

19. 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;1:1829-1841.

20. Hilal SH, Karickhoff SW, Carreira LA. Prediction of the solubility, activity coefficient and liquid/liquid partition coefficient of organic compounds. *QSAR Combinator Sci*. 2004;23:709-720.

21. Nohair M, Zakarya D. Prediction of solubility of aliphatic alcohols using the restricted components of autocorrelation method (RCAM). *J Mol Model (Online)*. 2003;9:365-371.

22. Yan A, Gasteiger J. Prediction of aqueous solubility of organic compounds by topological descriptors. *QSAR Combinator Sci*. 2003;22:821-829.

23. Hou TJ, Xia K, Zhang W, Xu XJ. ADME evaluation in drug discovery, IV: prediction of aqueous solubility based on atom contribution approach. *J Chem Inf Comput Sci*. 2004;44:266-275.

24. Xia X, Maliski E, Cheetham J, Poppe L. Solubility prediction by recursive partitioning. *Pharm Res*. 2003;20:1634-1640.

25. Yan A, Gasteiger J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J Chem Inf Comput Sci*. 2003;43:429-434.

26. Yin C, Liu X, Guo W, Lin T, Wang X, Wang L. Prediction and application in QSPR of aqueous solubility of sulfur-containing aromatic esters using GA-based MLR with quantum descriptors. *Water Res*. 2002;36:2975-2982.

27. Klamt A, Eckert F, Hornig M, Beck ME, Burger T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J Comput Chem*. 2002;23:275-281.

28. Bergstrom CA, 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;44:1477-1488.

29. 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 Des*. 2004;18:75-87.

30. 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;43:674-679.

31. Ma W, Zhang X, Luan F, et al. Support vector machine and the heuristic method to predict the solubility of hydrocarbons in electrolyte. *J Phys Chem A*. 2005;109:3485-3492.

32. Tetko IV, Bruneau P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J Pharm Sci*. 2004;93:3103-3110.

33. Thompson JD, Cramer CJ, Truhlar DG. Predicting aqueous solubilities from aqueous free energies of solvation and experimental or calculated vapor pressures of pure substances. *J Chem Phys*. 2003;119:1661-1670.

34. 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;43:1077-1084.

35. 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;42:1247-1249.

36. 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 Combinator Sci*. 2004;23:327-343.

37. 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 Combinator Sci*. 2003; 22:943-958.

38. 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;38:450-456.

39. Jorgensen WL, Duffy EM. Prediction of drug solubility from Monte Carlo simulations. *Bioorg Med Chem Lett*. 2000;10:1155-1158.

40. Tantishaiyakul V. Prediction of aqueous solubility of organic salts of diclofenac using PLS and molecular modeling. *Int J Pharm*. 2004;275:133-139.

41. Tantishaiyakul V. Prediction of the aqueous solubility of benzylamine salts using QSPR model. *J Pharm Biomed Anal*. 2005;37:411-415.

42. Parshad H, Frydenvang K, Liljefors T, Larsen CS. Correlation of aqueous solubility of salts of benzylamine with experimentally and theoretically derived parameters: a multivariate data analysis approach. *Int J Pharm*. 2002;237:193-207.

43. Ikeda H, Chiba K, Kanou A, Hirayama N. Prediction of solubility of drugs by conductor-like screening model for real solvents. *Chem Pharm Bull (Tokyo)*. 2005;53:253-255.

44. Hendriksen BA, Felix MV, Bolger MB. The composite solubility versus pH profile and its role in intestinal absorption prediction. *AAPS PharmSci*. 2003;5:E4.

45. Horter D, Dressman JB. Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. *Adv Drug Deliv Rev*. 2001;46:75-87.

46. Agoram B, Woltosz WS, Bolger MB. Predicting the impact of physiological and biochemical processes on oral drug bioavailability. *Adv Drug Deliv Rev*. 2001;50:S41-S67.

47. Rinaki E, Valsami G, Macheras P. Quantitative biopharmaceutics classification system: the central role of dose/solubility ratio. *Pharm Res*. 2003;20:1917-1925.

48. Bergstrom CAS, Luthman K, Artursson P. Accuracy of calculated pH-dependent aqueous drug solubility. *Eur J Pharm Sci*. 2004;22:387-398.

49. Setschenow JZ. Uber Die Konstitution Der Salzlosungenauf Grund Ihres Verhaltens Zu Kohlensaure. *Z Physik Chem*. 1889;4:117-125.

50. Ni N, Yalkowsky SH. Prediction of Setschenow constants. *Int J Pharm*. 2003;254:167-172.

51. Li Y, Hu Q, Zhong C. Topological modeling of the Setschenow constant. *Ind Eng Chem Res*. 2004;43:4465-4468.

52. Gould PL. Salt selection for basic drugs. *Int J Pharm*. 1986;33:201-217.

53. Reid RC, Pransnitz JM, Poling BE. *The Properties of Gases and Liquids.* New York, NY: McGraw Hill; 1984.

54. Hribar B, Southall NT, Vlachy V, Dill KA. How ions affect the structure of water. *J Am Chem Soc*. 2002;124:12302-12311.

55. 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;88:868-880.

56. Huang L-F, Tong W-Q. Impact of solid state properties on developability assessment of drug candidates. *Adv Drug Deliv Rev*. 2004;56:321-334.

57. Pudipeddi M, Serajuddin ATM. Trends in solubility of polymorphs. *J Pharm Sci*. 2005;94:929-939.

58. Nielsen AB, Frydenvang K, Liljefors T, Buur A, Larsen C. Assessment of the combined approach of N-alkylation and salt formation to enhance aqueous solubility of tertiary amines using bupivacaine as a model drug. *Eur J Pharm Sci*. 2005;24:85-93.

59. Romero AJ, Rhodes CT. Stereochemical aspects of the molecular pharmaceutics of ibuprofen. *J Pharm Pharmacol*. 1993;45:258-262.

60. Sanghvi T, Jain N, Yang G, Yalkowsky SH. Estimation of aqueous solubility by the general solubility equation (GSE) the easy way. *QSAR Combinator Sci*. 2003;22:258-262.

61. Bergstrom CA, Norinder U, Luthman K, Artursson P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J Chem Inf Comput Sci*. 2003;43:1177-1185.

62. Jain A, Yang G, Yalkowsky SH. Estimation of melting points of organic compounds. *Ind Eng Chem Res*. 2004;43:7618-7621.

63. Johnson JLH, Yalkowsky SH. Two new parameters for predicting the entropy of melting: eccentricity (e) and spirality (m). *Ind Eng Chem Res*. 2005;44:7559-7566.

64. Karthikeyan M, Glen RC, Bender A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J Chem Inf Model*. 2005;45:581-590.

65. Thomas E, Rubino J. Solubility, melting point and salting-out relationships in a group of secondary amine hydrochloride salts. *Int J Pharm*. 1996;130:179-185.

66. Katritzky AR, Jain R, Lomaka A, Petrukhin R, Maran U, Karelson M. Perspective on the relationship between melting points and chemical structure. *Cryst Growth Des*. 2001;1:261-265.

67. Raevsky OA, Schaper K-J, Seydel JK. H-bond contribution to octanol-water partition coefficients of polar compounds. *Quant Struct-Activ Relat*. 1995;14:433-436.

68. Raevsky OA. Hydrogen bond strength estimation by means of the HYBOT program package. In: van de Waterbeemd H, Testa B, Folkers G, eds. *Computer-Assisted Lead Finding and Optimization.* Basel, Switzerland: Verlag Helvetica Chimica Acta; 1997:369-378.

69. Raevsky OA, Grigor'ev VG. Quantitative description of lipophilicity of organic chemicals on the basis of polarizability and H-bond acceptor factors. *Chem-Pharm Z (Rus)*. 1999;33:46-49.

70. Raevsky OA, Schaper KJ, van de Waterbeemd H, McFarland JW. Hydrogen bond contributions to properties and activities of chemicals and drugs. In: Gundertofte K, Jorgensen FS, eds. *Molecular Modeling and Prediction of Bioactivity.* New York, NY: Springer; 2000:221-227.

71. 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;48:487-509.

72. Ouvrard C, Mitchell JBO. Can we predict lattice energy from molecular structure? *Acta Crystallogr B*. 2003;59:676-685.

73. Copley RCB, Deprez LS, Lewis TC, Price SL. Computational prediction and X-ray determination of the crystal structures of 3-oxauracil and 5-hydroxyuracil—an informal blind test. *CrystEngComm*. 2005;7:421-428.

74. Dey A, Kirchner MT, Vangala VR, Desiraju GR, Mondal R, Howard JAK. Crystal structure prediction of aminols: advantages of a supramolecular synthon approach with experimental structures. *J Am Chem Soc*. 2005;127:10545-10559.

75. Price SL. Quantifying intermolecular interactions and their use in computational crystal structure prediction. *CrystEngComm*. 2004;6:344-353.

76. Price SL. The computational prediction of pharmaceutical crystal structures and polymorphism. *Adv Drug Deliv Rev*. 2004;56:301-319.

77. Datta S, Grant DJ. Crystal structures of drugs: advances in determination, prediction and engineering. *Nat Rev Drug Discov*. 2004;3:42-57.

78. Gavezzotti A. A molecular dynamics test of the different stability of crystal polymorphs under thermal strain. *J Am Chem Soc*. 2000;122:10724-10725.

79. Gavezzotti A. The chemistry of intermolecular bonding: organic crystals, their structures and transformations. *Synlett*. 2002;2002:0201-0214.

80. Gavezzotti A. Quantitative ranking of crystal packing modes by systematic calculations on potential energies and vibrational amplitudes of molecular dimers. *J Chem Theory Computat*. 2005;1:834-840.

81. Pajeva I, Wiese M. Molecular modeling of phenothiazines and related drugs as multidrug resistance modifiers: a comparative molecular field analysis study. *J Med Chem*. 1998;41:1815-1826.

82. Pajeva IK, Wiese M. Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (hypothesis). *J Med Chem*. 2002;45:5671-5686.

83. Ekins S, Kim RB, Leake BF, et al. Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol Pharmacol*. 2002;61:974-981.

84. Penzotti JE, Lamb ML, Evensen E, Grootenhuis PD. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J Med Chem*. 2002;45:1737-1740.

85. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ. Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci*. 2004;44:1497-1505.

86. Wang YH, Li Y, Yang SL, Yang L. An in silico approach for screening flavonoids as P-glycoprotein inhibitors based on a Bayesian-regularized neural network. *J Comput Aided Mol Des*. 2005;19:137-147.

87. Wang YH, Li Y, Yang SL, Yang L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J Chem Inf Model*. 2005;45:750-757.

88. Gombar VK, Polli JW, Humphreys JE, Wring SA, Serabjit-Singh CS. Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. *J Pharm Sci*. 2004;93:957-968.

89. Cianchetta G, Singleton RW, Zhang M, et al. A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J Med Chem*. 2005;48:2927-2935.

90. Seelig A. A general pattern for substrate recognition by P-glycoprotein. *Eur J Biochem*. 1998;251:252-261.

91. Ekins S, Kim RB, Leake BF, et al. Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein. *Mol Pharmacol*. 2002;61:964-973.

92. Balimane PV, Chong S. Cell culture-based models for intestinal permeability: a critique. *Drug Discov Today*. 2005;10:335-343.

93. Bergstrom CA. In silico predictions of drug solubility and permeability: two rate-limiting barriers to oral drug absorption. *Basic Clin Pharmacol Toxicol*. 2005;96:156-161.

94. Malkia A, Murtomaki L, Urtti A, Kontturi K. Drug permeation in biomembranes: in vitro and in silico prediction and influence of physicochemical properties. *Eur J Pharm Sci*. 2004;23:13-47.

95. Kulkarni A, Han Y, Hopfinger AJ. Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J Chem Inf Comput Sci*. 2002;42:331-342.

96. Yamashita F, Wanchana S, Hashida M. Quantitative structure/property relationship analysis of Caco-2 permeability using a genetic algorithm-based partial least squares method. *J Pharm Sci*. 2002;91:2230-2239.

97. Yamashita F, Fujiwara S, Hashida M. The "latent membrane permeability" concept: QSPR analysis of inter/intralaboratory variable Caco-2 permeability. *J Chem Inf Comput Sci*. 2002;42:408-413.

98. Fujiwara S, Yamashita F, Hashida M. Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *Int J Pharm*. 2002;237:95-105.

99. Marrero Ponce Y, Cabrera Perez MA, Romero Zaldivar V, Gonzalez Diaz H, Torrens F. A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell culture. *J Pharm Pharm Sci*. 2004;7:186-199.

100. Hou TJ, Zhang W, Xia K, Qiao XB, Xu XJ. ADME evaluation in drug discovery, V: correlation of Caco-2 permeation with simple molecular properties. *J Chem Inf Comput Sci*. 2004;44:1585-1600.

101. Fujikawa M, Ano R, Nakao K, Shimizu R, Akamatsu M. Relationships between structure and high-throughput screening permeability of diverse drugs with artificial membranes: application to prediction of Caco-2 cell permeability. *Bioorg Med Chem*. 2005;13:4721-4732.

102. Refsgaard HH, Jensen BF, Brockhoff PB, Padkjaer SB, Guldbrandt M, Christensen MS. In silico prediction of membrane permeability from calculated molecular parameters. *J Med Chem*. 2005;48:805-811.

103. Artursson P, Palm K, Luthman K. Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Adv Drug Deliv Rev*. 1996;22:67-84.

104. 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;23:3-25.

105. Willmann S, Schmitt W, Keldenich J, Lippert J, Dressman JB. A physiological model for the estimation of the fraction dose absorbed in humans. *J Med Chem*. 2004;47:4022-4031.

106. Usansky HH, Sinko PJ. Estimating human drug oral absorption kinetics from Caco-2 permeability using an absorption-disposition model: model development and evaluation and derivation of analytical solutions for k(a) and F(a). *J Pharmacol Exp Ther*. 2005;314:391-399.

107. Obata K, Sugano K, Saitoh R, et al. Prediction of oral drug absorption in humans by theoretical passive absorption model. *Int J Pharm*. 2005;293:183-192.

108. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem*. 2002;45:2615-2623.

109. Lu JJ, Crimin K, Goodwin JT, et al. Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *J Med Chem*. 2004;47:6104-6107.

110. Vieth M, Siegel MG, Higgs RE, et al. Characteristic physical properties and structural fragments of marketed oral drugs. *J Med Chem*. 2004;47:224-232.

111. Martin YC. A bioavailability score. *J Med Chem*. 2005;48:3164-3170.

112. Liu HX, Hu RJ, Zhang RS, et al. The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. *J Comput Aided Mol Des*. 2005;19:33-46.

113. Bai JP, Utis A, Crippen G, et al. Use of classification regression tree in predicting oral absorption in humans. *J Chem Inf Comput Sci*. 2004;44:2061-2069.

114. Turner JV, Maddalena DJ, Agatonovic-Kustrin S. Bioavailability prediction based on molecular structure for a diverse series of drugs. *Pharm Res*. 2004;21:68-82.

115. Zmuidinavicius D, Didziapetris R, Japertas P, Avdeef A, Petrauskas A. Classification structure-activity relations (C-SAR) in prediction of human intestinal absorption. *J Pharm Sci*. 2003;92:621-633.

116. Klopman G, Stefan LR, Saiakhov RD. ADME evaluation, II: a computer model for the prediction of intestinal absorption in humans. *Eur J Pharm Sci*. 2002;17:253-263.

117. Yoshida F, Topliss JG. QSAR model for drug human oral bioavailability. *J Med Chem*. 2000;43:2575-2585.

118. Andrews CW, Bennett L, Yu LX. Predicting human oral bioavailability of a compound: development of a novel quantitative structure-bioavailability relationship. *Pharm Res*. 2000;17:639-644.

119. Palm K, Stenberg P, Luthman K, Artursson P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm Res*. 1997;14:568-571.