

Recent Advances in Computational Prediction of Drug Absorption and Permeability in Drug Discovery

Tingjun Hou^a, Junmei Wang^b, Wei Zhang^c, Wei Wang^{*,a} and Xiaojie Xu^{*,b}

^aDepartment of Chemistry and Biochemistry, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093, USA

^bCollege of Chemistry and Molecular Engineering, Peking University, Beijing 100871, P.R. China

^cDepartment of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Abstract: Approximately 40%-60% of developing drugs failed during the clinical trials because of ADME/Tox deficiencies. Virtual screening should not be restricted to optimize binding affinity and improve selectivity; and the pharmacokinetic properties should also be included as important filters in virtual screening. Here, the current development in theoretical models to predict drug absorption-related properties, such as intestinal absorption, Caco-2 permeability, and blood-brain partitioning are reviewed. The important physicochemical properties used in the prediction of drug absorption, and the relevance of predictive models in the evaluation of passive drug absorption are discussed. Recent developments in the prediction of drug absorption, especially with the application of new machine learning methods and newly developed software are also discussed. Future directions for research are outlined.

Keywords: ADME, drug adsorption, permeability, Caco-2 monolayer, blood-brain partitioning (BBB), logBB, QSAR.

In the traditional drug design paradigm, the central stage focuses on the activity and the specificity of a drug candidate, while some other properties, especially those related to absorption, distribution, metabolism, excretion (ADME) and even toxicity (Tox), are only considered at a later stage. The *in vitro* screening using the traditional strategy may usually lead to potent ligands but not necessarily good drug candidates, since lead compounds that have high molecular weight and increased lipophilicity usually tend to have high potency but poor absorption. It has been estimated that about 40%-60% of such failures are caused by ADME/Tox deficiencies [1-3]. The significant failure rate of drug candidates in late stage development is driving the need for development of new *in vitro*, *in vivo*, and *in silico* tools that can eliminate inappropriate compounds before substantial resources are wasted. Accordingly, a paradigm shift has occurred in the initial phases of drug discovery. In addition to potency and selectivity towards the biological target of interest, ADME/Tox properties of a drug are now taken into account at an early stage. Beginning in the early- to mid-1990s, many pharmaceutical companies took steps to integrate the functions of discovery and development scientists. Development scientists are involved in the early stage of the drug discovery program and provide input to *in vitro* and *in vivo* optimization process [3]. Furthermore, advances in automation technology and experimental ADME/Tox techniques, such as the Caco-2 permeability screening based on the three-day Caco-2 culture system, the metabolic stability screening using microsomes or hepatocytes, and the P450 inhibition assay, have enabled the assaying of much larger

numbers of compounds than those using traditional strategies [4]. ADME/Tox properties of molecules can be possibly optimized in parallel with assays for potency and selectivity, making lead optimization a truly multi-parametric procedure.

In addition to the development of experimental assays with greater throughput, there is an urgent need for effective computational methods for predicting ADME/Tox-related properties. Compared to experimental approaches, these *in silico* methods have advantages: they do not initially require the compounds to be synthesized and experimentally tested; compound databases can be virtually screened rapidly in a high-throughput fashion when the calculations are computationally efficient. Until now, many computational approaches have been developed for the ADME/Tox properties, such as bioavailability, aqueous solubility, intestinal absorption, blood-brain barrier penetration, drug-drug interactions, transporter, plasma-protein binding and toxicity [5].

In this review, we will survey the computational methods that have been developed for the prediction of drug absorption, with special focus on three specific properties: intestinal absorption, Caco-2 permeability, and blood-brain barrier penetration. Considering that the mechanisms of the passive diffusion through different biological barriers are quite similar and the major difference may be the extent, the discussion is based on the important physicochemical determinants in drug absorption and not the specific properties. Furthermore, since the prediction of drug absorption is a very active and rapidly developed research area, and substantial progress has been made in recent years, resulting in a broad spectrum of models for estimation of drug absorption, we only emphasize the new developments in this field.

1. ADSORPTION AND *IN VIVO* BARRIERS

Oral administration is the most convenient way for patients to receive medication. When the drug is administrated

*Address correspondence to these authors at the Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093, USA; E-mail: wei-wang@ucsd.edu and College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, P.R. China; E-mail: xiaojxu@pku.edu.cn

orally, it has to be absorbed across the epithelium of the small intestine. The intestinal adsorption of a drug molecule includes two stages: dissolution and membrane transfer. In the first stage, the drug molecule is dissolved in the aqueous contents of the gastrointestinal tract. The dissolved molecule is then transferred across the actual barrier of the gastrointestinal tract to reach the blood circulation. In addition to the gastrointestinal tract, there are other several *in vivo* barriers of high interests, such as the blood–brain barrier (BBB) and the stratum corneum of the skin. The compositions of all *in vivo* barriers are similar and involve the crossing of biological membranes. Biological membranes are composed of a lipid-bilayer that results from the orientation of the lipids (phospholipids, glycolipids and cholesterol) in the aqueous medium. A wide variety of proteins such as selective ion channels (Na^+ , K^+ , Ca^{2+} and Cl^-) are embedded within the membrane. Tight junctions between cells which occur as a result of the interaction of membrane proteins at the contact surface between single cells generate a regulated barrier with small aqueous-filled pores. The dimensions of these pores are estimated to be in the range of 8–20 Å, depending on the cell type [6].

The major route for the drug permeability through the barrier is passive diffusion that is driven by a concentration gradient. Two types of passive diffusion mechanism can be distinguished: paracellular transport and transcellular transport. For hydrophobic molecules with weight molecule smaller than 200, the paracellular transport between the cell junctions is favorable. Lipophilic molecules preferably use the transcellular transport. For both intestinal epithelium and the blood brain barrier, the transcellular transport is more important and the prediction of drug absorption and permeability concentrates on this pathway. In addition to passive diffusion, some molecules, such as amino acid and glucose, can be actively transported by specific transporters, such as peptide transporter PepT1, P-glycoprotein (P-gp), and multidrug resistance-associated protein 2 (MRP2) [7]. PepT1 has been classified as a low-affinity high-capacity transporter belonging to the proton oligopeptide transporter (POT) superfamily, which can enhance the transcellular permeability of many peptide-like molecules from the apical to basolateral direction, such as β -lactam antibiotics, ACE inhibitors, the antineoplastic agent bestatin, and so on. Adversely, some efflux proteins localized in the apical or basolateral cell membranes have the potential to pump drugs out from the cell into the apical or basolateral extracellular fluids. A well-studied example of the efflux proteins is P-gp. A portion of the P-gp substrate, such as vinblastine, entering the intestinal mucosal cells *via* passive diffusion, is transported out of the cell and into the intestinal lumen by P-gp. In experiment, the intestinal absorption is usually measured by fraction absorption, %FA, which is defined by the total mass absorbed ($m(\infty)$) divided by the given dose of the drug (dose):

$$\%FA = \frac{m(\infty)}{\text{dose}} \quad (1)$$

For accurate and effective prediction of intestinal absorption, several *in vitro* methods have been developed. Among them, the most popular cell-based model for intestinal permeability is the Caco-2 cell system [8,9]. Caco-2 cells, derived from colorectal carcinoma cells, display many of the

morphological and functional properties of the *in vivo* intestinal epithelial cell barrier. Extensive studies have shown that human oral drug absorption and Caco-2 permeability coefficient have a good sigmoidal relationship [10], suggesting that the human absorption can be well predicted by this *in vitro* model. Caco-2 culture models have many advantages. First of all, it measures the transport of the drug across a cell membrane, rather than an interaction of the drug with the lipid bilayer. Secondly, it can measure the parallel transport routes, both passive and active. However, it has several limitations, such as long preparation time, very slow absorption times compared to human intestine and large interlaboratory differences in quantitative results. Furthermore, it cannot quantitatively predict the level of active drug transport *in vivo* [11]. The transport across the Caco-2 cell is measured by the apparent permeability coefficient (P_{app}), which is calculated as the steady-state appearance rate of the compound on the receiver side, divided by the initial concentration on the donor side and the surface area of the monolayer:

$$P_{app} = \frac{\text{amount transported}}{\text{area} \times \text{initial concentration} \times \text{time}} \quad (2)$$

Another interesting epithelial barrier is that separating the brain and central nervous system (CNS) from the blood stream, called blood brain barrier (BBB). In the case of effective CNS acting drugs, the knowledge of the penetration of drugs through BBB is critical to screen potential therapeutic agent and improve the side effect profile of drugs with peripheral activity. The extent to which drug molecules across from the blood into the brain is governed by two physiologically and anatomically related systems, BBB and the blood-cerebral spinal fluid (CSF) barrier, which forms two pathways through which drug compounds partition between plasma and brain tissue. At the molecular level, the principal component of the barrier is the lipid bilayer of the capillary endothelial cell membrane, through which compounds diffuse to reach the brain. The membranes involved are tight junction membranes of brain parenchymal cells. Tight junction membranes limit the size of hydrophilic molecules that can across the membrane by paracellular diffusion. Similar to the permeability across the intestinal epithelium, the vast majority of substances that penetrate a tight junction barrier are lipophilic molecules that cross by a transcellular route. Experimental data has shown that lipophilic compounds, along with water and small polar molecules, can cross both the blood-brain and blood-CSF barriers. Hydrophilic organic molecules, including plasma proteins and larger polar molecules, cannot penetrate well. The relative affinity for the blood or brain tissue can be expressed in terms of the blood-brain partition coefficient, $\log BB$, the partition between the equilibrium concentrations of the drug in the brain (C_{brain}) and the blood (C_{blood}):

$$\log BB = \frac{C_{brain}}{C_{blood}} \quad (3)$$

2. PREDICTION MODEL OF DRUG ABSORPTION AND PERMEABILITY

(a). **The rules to define “drug-like” molecules.** Rather than trying to predict specific absorption-related quantities, researchers have tried to find general principles to distinguish drug-like from non-drug-like molecules by analyzing

databases of drugs and non-drugs. Generally, these rules obtained from database analysis can be used to distinguish well-absorbed molecules from poorly-absorbed molecules. The most popular ADME-concerned filters may be “rule of 5” proposed by Lipinski and coworkers in 1997 from analysis of 2245 drugs from the World Drug Index (WDI) [12]. They found that poor absorption and permeation are more likely to occur when

- (1) molecular weight > 500
- (2) calculated $\log P$ > 5 (CLOGP) or > 4.15 (MLOGP)
- (3) number of hydrogen-bond donors (OH and NH groups) > 5
- (4) number of hydrogen-bond acceptors (N and O atoms) > 10

The fast estimations of $\log P$ allow the “rule of 5” screening of library prior to enumeration. Moreover, based on screening results from Merck and Pfizer, Lipinski argues that it is much easier to optimize pharmacokinetic properties early in the process of drug discovery, and attempt to optimize the receptor binding affinity at a later stage.

Following the work of Lipinski, other researchers proposed some similar drug-like rules. By analyzing the Comprehensive Medicinal Chemistry (CMC) database with 7183 compounds, Ghost and coworkers found the distributions of several important molecular properties (covering more than 80% of the compounds) [13]:

- (1) $-0.4 \leq \text{calculated } \log P \text{ (ALOGP)} \leq 5.6$ (average value: 2.52)
- (2) $160 \leq \text{molecular weight} \leq 480$ (average value: 357)
- (3) $40 \leq \text{molar refractivity} \leq 130$ (average value: 97)
- (4) $20 \leq \text{total number of atoms} \leq 70$ (average value: 48).

Opera examined the property distribution in several databases containing drug-like and non-drug-like compounds [14]. The examined properties include molecular weight (MW), the calculated octanol/water partition coefficient (CLOGP), the number of rotatable (RTB) and rigid bonds (RGB), the number of rings (RNG), and the number of hydrogen bond donors (HDO) and acceptors (HAC). Skewed distributions can be exploited to focus on the ‘drug-like space’: 62.68% of “non-drug-like” compounds have $0 \leq \text{RNG} \leq 2$, and $\text{RGB} \leq 17$, while 28.88% of ‘non-drug-like’ compounds have $3 \leq \text{RNG} \leq 13$, and $18 \leq \text{RGB} \leq 56$. By contrast, 61.22% of “drug-like” compounds have $\text{RNG} \geq 3$, and $\text{RGB} \geq 18$, and only 24.73% “drug-like” compounds have $0 \leq \text{RNG} \leq 2$ rings, and $\text{RGB} \leq 17$. The author concludes that the probability of identifying “drug-like” structures increases with molecular complexity.

Wenlock and coworkers analyzed 594 compounds from Physicians Desk Reference 1999 (PDR) and obtained limited distributions of molecular weight, lipophilicity, and hydrogen bonding for oral drugs (covering more than 90% of the compounds) [15]. The distributions are generally consistent with ‘rule-of-five’, but more stringent:

- (1) molecular weight < 473
- (2) calculated $\log P$ (ACD $\log P$) < 5 or calculated $\log D_{7.4}$ (ACD $\log D$) < 4.3

- (3) number of hydrogen-bond donors < 4
- (4) number of hydrogen-bond acceptors < 7

After analysis of the difference between current development oral drugs and marketed oral drugs, Wenlock *et al.* found that the mean molecular weight of orally administered drugs in development decreases on passing through each of the different clinical phases and gradually converges toward the mean molecular weight of marketed oral drugs. Meanwhile, the most lipophilic compounds are being discontinued from development

Recently, Vieth and coworkers analyzed 1193 “oral” drugs and found that the oral drug molecules show property distributions essentially identical to those of the set of 594 oral drugs examined by Wenlock, with minor differences in the distribution for the 90th percentiles for hydrogen-bond donors and acceptors [16]. Furthermore, the authors found that with respect to other routes of administration, oral drugs tend to be lighter and have fewer H-bond donors, acceptors, and rotatable bonds than drugs with other routes of administration.

However, “rule of five” and other general rules are only the minimum criterion of a molecule to be drug-like. It is very easy for a compound to fall within the “rule of five” but has no potential to lead to a drug. As a matter of fact, 68.7% compounds in ACD (Available Chemical Directory) Screening Database (2.4 million compounds) and 55 % compounds in ACD (240 thousand compounds) have no violation of “rule of five” at all. Therefore, more stringent criteria should be built up to discriminate drug-like compounds from the others.

(b). The prediction models of drug absorption. Certainly, the rules to predict “drug-likeness” are too general, and it is necessary to develop prediction models for specific absorption properties. The predictions of the AMDE properties are involved in two aspects of modeling methods: data modeling and molecular modeling. For molecular modeling, molecular mechanics, pharmacophore modeling, molecular docking, and quantum mechanics are used to explore the potential interactions between the small molecules under consideration and proteins known to be involved in ADME processes, such as cytochrome P450s [17]. For data modeling, quantitative structure-property relationship (QSPR) approaches are typically applied. Based on appropriate descriptors, QSPR exploiting from simple multiple linear regression (MLR) to modern multivariate analysis techniques or machine-learning methods are now being applied to the analysis of ADME data. Most prediction methods applied in the estimation of drug absorption belong to the category of data modeling. Data modeling can be applied with great efficiency to large number of molecules, but require a significant quantity of high quality data to deduce a relationship between the structures and the modeled property.

The reported prediction models for intestinal absorption, Caco-2 permeability and blood-brain partitioning are listed in Table 1 [18-40], Table 2 [24,27,41-51] and Table 3 [24,27,45,52-76]. The prediction models can be divided into two categories: correlation models and classification models. The important methods involved are briefly introduced as follows:

Table 1. The prediction Models for Intestinal Absorption

| Reference | Method | Model | Descriptors | Dataset | | Performance | |
|-------------------------|---|------------------------------|--|--|--------------------------|---|---|
| | | | | Training set | Prediction set | Correlation or classification | Prediction |
| Palm [18] | Nonlinear regression | Correlation | PSA and other several calculated descriptors | 20 | | $r^2=0.94$ | |
| Wessel [19] | GA and ANN | Correlation | 162 calculated molecular descriptors | 76 | 10 | RMSE=9.5, MAE=6.7 | RMSE=16.0, MAE=11.0 |
| Ghuloum [20] | ANN | Correlation | Molecular hashkeys | 20 | | $r=0.83$ | |
| Clark [21] | MLR | Correlation | PSA | 20 | 74 | $r^2=0.94$ | 68/74 correctly classified |
| Norinder [22] | PLS | Correlation | MolSurf parameters | 13 | 7 | $r^2=0.90$, $q^2=0.69$ | RMSE=0.49 |
| Raevsky [23] | MLR and nonlinear fit | Correlation | 4 H-bond descriptors and other several molecular descriptors | 32 | | $r=0.94$ | |
| Österberg [24] | PLS | Correlation | 3 H-Bond descriptors and logP | 20 | | $r^2=0.81$, $q^2=0.73$ | |
| Egan [25] | Classification | Classification | PSA and AlogP98 | 199 (well-absorbed)+ 35 (poorly-absorbed) | | Good classification | |
| Zhao [26] | MLR | Correlation | Abraham descriptors | 38 | 131 | $r^2=0.83$, RMSE=14 | RMSE =14 |
| Norinder [27] | PLS | Correlation | Electrotopological state indices and some other calculated molecular descriptors | 13 | 7 | $r^2=0.93$, $q^2=0.86$, RMSE=0.44 | RMSE=0.55 |
| Agatonovic-Kustrin [28] | GA and ANN | Correlation | 57 calculated molecular descriptors | 66 (training set) + 10 (test set) | 10 | RMSE=0.59 (training set), RMSE=0.90 (test set) | $R^2=0.80$, RMSE=0.42 |
| Klopman [29] | MLR | Correlation | 37 molecular groups | 417 | 50 | $r^2=0.79$ | $R^2=0.79$ |
| Abraham [30] | MLR | Correlation | Abraham descriptors | 127 | | $r^2=0.80$ | |
| Derety [31] | Nonlinear fit | Correlation | Multiple calculated molecular descriptors | 93 | 31 | $r^2=0.80$, RMSE=14 | RMSE=12 |
| Zmuidinavicius [32] | Recursive partitioning analyses | Classification | Multiple calculated molecular descriptors | > 1000 | | 5% false-positives and 3% false-negatives | |
| Niwa [33] | General Regression neural network (GRNN) and Probabilistic Neural Network (PNN) | Correlation + Classification | PSA, ClogP, CMR and some topological descriptors | 67 (training set) + 9 (test set) | 10 | GRNN: RMSE=6.5 (training set), RMSE=27.7 (test set) PNN: 100% correctly classified (training set), 88.9% correctly classified (test set) | GRNN: RMSE = 22.8 PNN: 80% correctly classified |
| Perez [34] | LDA | Classification | TOPS-MODE descriptors | 82 | 127(set1) + 109(set2) | 89.0 % correctly classified | 88.9 % and 93.6% correctly classified for set1 and set2 |

(Table 1) contd.....

| Reference | Method | Model | Descriptors | Dataset | | Performance | |
|----------------|--|----------------------------|---|--|----------------|--|----------------------------|
| | | | | Training set | Prediction set | Correlation or classification | Prediction |
| Wegner [35] | GA based on Shannon Entropy Cliques (GA-SEC) | Correlation classification | 3387 calculated molecular descriptors | 172 | 24 | | |
| Sun [36] | PLS-discriminant analysis (PLS-DA) | Classification | Atom types | 169 | | 167/169 correctly classified | |
| Xue [37] | Recursive feature elimination (RFE) and support vector machine (SVM) | Classification | 159 molecular descriptors | 131 (well-absorbed) + 65 (poorly-absorbed) | | SVM: 83.4% correctly classified for well-absorbed molecules and 63.2% for poorly-absorbed molecules, SMV+RFE: 90.0 % for well-absorbed molecules and 80.7% for poorly-absorbed molecules | |
| Deconinck [38] | Classification and regression trees (CART) | Classification | Multiple calculated molecular descriptors | 141 | 27 | 138/141 are classified correctly | 23/27 correctly classified |
| Liu [39] | Heurist method (HM) and SVM | Correlation | Multiple calculated molecular descriptors | 113 | 56 | $r^2=0.86$ | $r^2=0.73$ |
| Jones [40] | MLR | Correlation | 6-Moment Descriptors | 38 | 131 | RMSE=12.5 | RMSE=15 |

- Multiple linear regression (MLR):** MLR is the most widely-used linear correlation method, which can model the relationship between two or more explanatory variables (X) and a response variable (Y) by fitting a linear equation to the observed data. As a general rule, the samples (N) should be larger than 2^m (m is the number of descriptors used in correlation). As the number of descriptors increase, however, MLR becomes problematic, for example, redundancy of information when descriptors are correlated.
- Partial least square (PLS):** PLS combines features from principal component analysis (PCA) and MLR, which is based on linear transformation from a large number of original descriptors to a new variable space based on small number of orthogonal factors (latent variables). It is especially useful in quite common cases where the number of descriptors (X) is comparable to or greater than the number of compounds (samples) and/or there exist other factors leading to correlations between variables.
- Linear discriminant analysis (LDA):** LDA calculates discriminant functions or hyperplanes that partition the space of chemical descriptors to give the best separation between different classes. When the discriminating function is parameterized, it has to be tested either by using an independent set of test data, or by performing cross-validation.
- Artificial neural networks (ANNs):** ANNs are a class of machine learning methods inspired by the way of biological nervous systems, such as the brain, to process information. ANNs are typically used when there are a large number of observations (X) and when the problem is not understood well enough to write a procedural program or expert system. Using ANNs, the solution to the problem can be sought as follows: an answer is calculated by multiplying each input by the connection weight; products are summed at each hidden unit; and the output of each hidden unit is then multiplied by the connection weight, summed, and then interpreted. ANNs are very powerful in dealing with non-linear correlation or classification. The network although can overfit or memorize the data if too many hidden units are used. A sufficiently large test set is necessary to supervise the training of the ANNs models.
- Genetic algorithms (GAs):** GAs are a class of heuristic optimization algorithms inspired by the mechanism of biological evolution. In GAs, a group of individuals with the predicted property (Y) and a set of descriptors (X) represent 'chromosome' for the

Table 2. The prediction models for Caco-2 permeability

| Reference | Method | Model | Descriptors | Dataset | | Performance | |
|-------------------------|--------|-------------|--|--------------|------------------------|--|---|
| | | | | Training set | Test set | Correlation or classification | Prediction |
| Palm [41] | MLR | Correlation | PSA | 6 | | $r^2=0.99$ | |
| van de Water-Beemd [42] | MLR | Correlation | PSA, MW | 17 | | $r=0.83$ | |
| Norinder [43] | PLS | Correlation | MolSurf parameters | 9 | 8 | $r^2=0.93, q^2=0.74$ (eq1) $r^2=0.94, q^2=0.85$ (eq2) | RMSE=0.45 (eq1) RMSE=0.41 (eq2) |
| Kraru [44] | MLR | Correlation | Surface-related parameters | 11 | | $r^2=0.98, q^2=0.93$ | |
| Segarra [45] | PLS | Correlation | Grid-based descriptors | 6 | | $r=0.93$ | |
| Cruciani [46] | PLS | Correlation | VolSurf parameters | 11 | | Good | |
| Österberg [24] | PLS | Correlation | $\log P$ and H-bonding parameters | 11 | | $r^2=0.92, q^2=0.74$ | |
| Norinder [27] | PLS | Correlation | Electrotopological state indices and some other calculated molecular descriptors | 9 | 8 | $r^2=0.93, q^2=0.79$, RMSE=0.32 | RMSE=0.51 |
| Fujiwara [47] | ANN | Correlation | Multiple calculated molecular descriptors | 87 | | RMSE=0.51 | |
| Kulkarni [48] | MLR | Correlation | Multiple calculated molecular descriptors and several intermolecular interaction descriptors based on molecular dynamics simulations | 30 | 8 | $r^2=0.86, q^2=0.77$ | $r=0.89$ |
| Ponce [49] | MLR | Correlation | Topological descriptors. | 17 | 20 | $r=0.96, q=0.93$ | RMSE=0.52 |
| Hou [50] | MLR | Correlation | Multiple calculated molecular descriptor | 77 | 23 | $r=0.82, q=0.79$ | $r=0.78$ |
| Nordqvist [51] | PLS | Correlation | Multiple calculated molecular descriptor | 46 | 5(set1) + 125(set2) | $r^2=0.79, q^2=0.65$ | Set1: RMSE=0.45 Set2: 82% correctly classified |

population. The individuals are scored according to the fitness score. Then the population is evolved using three basic evolution operations: selection, crossover and mutation. In principle, GAs can be combined with any correlation or classification approaches, such as MLR, PLS or ANNs. The fitness score is estimated using the MLR, PLS or ANNS models. QSAR analysis based on GAs can find a group of prediction models from a large numbers of samples efficiently, rather than one.

- **Support vector machines (SVMs):** SVMs are based on the structural risk minimization principle (SRM) from computational learning theory. SVMs construct a hyperplane that separates two classes (this can be extended to multiclass problems). Separating the classes with a large margin minimizes a bound on the expected generalization error. Moreover, SVMs are relatively insensitive to variation in the parameters and are not prone to overfitting when, for example, using high degree polynomial kernels. In many cases SVM has been found to be consistently superior to other supervised learning methods and less prone to overfitting.

The prediction models in Table 1, 2 and 3 are organized in chronological order with the descriptors and methods used and the predictive power of the models. In the following section, we only discuss some representative studies of the models. The discussions here are only based on the important physicochemical descriptors, because most descriptors are universally applicable to most permeability process through biological barriers. The review may be treated as the supplementary material to the reported reviews on the *in silico* predictions of drug absorptions or ADME properties [5,7,77-81].

3. IMPORTANT PHYSICOCHEMICAL DESCRIPTORS

It is well known that many factors are related to membrane permeability, including lipophilicity, H-bonding capability, solute size, and the ionization state of solute. In order to consider these factors, many physicochemical descriptors are introduced into the prediction of drug absorption. However, drug absorption may not be determined by a single defined descriptor, but rather by the combination of different physicochemical characteristics. Moreover, the descriptors introduced here are not completely independent, and some of

Table 3. The Prediction Models of Blood-Brain Partitioning

| Reference | Method | Model | Descriptors | Dataset | | Performance | |
|------------------------|--------------------------|-------------------------------|--|-----------------------|-------------------------------------|--|---|
| | | | | Training set | Test set | Correlation or classification | Prediction |
| Young [52] | MLR | Correlation | $\Delta\log P$ | 6 | | $r^2=0.96$ | |
| van de Waterbeemd [53] | MLR | Correlation | PSA | 20 | | $r=0.84$ | |
| Abraham [54] | MLR | Correlation | Abraham descriptors | 57 | | $r^2=0.92$ | |
| Lombardo [55] | MLR | Correlation | Free energy of solvation | 55 | 6 | $r=0.82$ | |
| Norinder [56] | PLS | Correlation | MolSurf parameters | 28 | 28 (test set1) + 6 (test set2) | $r^2=0.862$ RMSE=0.29 | RMSE=0.35 (test set1) RMSE=0.47 (test set2) |
| Clark [57] | MLR | Correlation | PSA and MlogP | 57 | 5 (test set1) + 5 (test set2) | $r=0.82$ (n=57) $r=0.89$ (n=55) | MAE=0.13 (test set1) MAE=0.23 (test set2) |
| Luco [58] | PLS | Correlation | Multiple calculated molecular descriptors | 56 | 14 (test set1) + 25 (test set2) | $r=0.92$ | RMSE=0.24 (test set1) RMSE=0.54 (test set2) |
| Segarra [45] | PLS | Correlation | Grid-based descriptors | 20 | | $r=0.85$ | |
| Ajay [59] | ANN | Classification | Multiple calculated molecular descriptors | 275 | | 92.0% correctly classified for the high BBB compounds, and 71.0% for low BBB compounds | |
| Crivori [60] | PCA and discriminant PLS | Classification | VolSurf descriptors | 110 | 120 | | 90% correctly classified for the high BBB compounds (40 out of 44), and about 65% for low BBB compounds |
| Österberg [24] | PLS | Correlation | 3 H-bonding descriptors and logP | 69 (Tr1)+ 45 (Tr2) | | Tr1: $r^2=0.76$, $q^2=0.75$ Tr2: $r^2=0.76$, $q^2=0.75$ | |
| Platts [61] | MLR | Correlation | Abraham descriptors | 148 | | $r^2=0.74$, $q^2=0.72$ | |
| Liu [62] | MLR and ANN | Correlation | electrotopological state indices and several calculated molecular descriptors | 55 | 11 | $r^2=0.79$ (MLR) $r^2=0.81$ (ANN) | $r^2=0.84$ (MLR) |
| Kaznessis [63] | MLR | Correlation | Multiple calculated molecular descriptors | 85 | | | |
| Norinder [27] | PLS | Correlation | Electrotopological state indices and some other calculated molecular descriptors | 28 | 31 | $r^2=0.78$, $q^2=0.73$, RMSE=0.35 | RMSE=0.42 |
| Hou [64] | GA | Correlation | Multiple calculated molecular descriptors | 59 | 14 (test set1) + 23 (test set2) | $r=0.87$ | RMSE=0.26 (test set1) RMSE=0.55 (test set2) |
| Rose [65] | MLR | Correlation | electrotopological state indices | 106 | 20 (test set1) + 28 (test set2) | $r^2=0.66$ $q^2=0.62$ | RMSE=0.36 (test set1) 27 compounds in test set2 are correctly classified |
| Doniger [66] | ANN and SVM | Classification | Multiple calculated molecular descriptors | 274 | 50 | | 81.5% correctly classified for SMV and 75.7 % correctly classified for ANN |
| Subramanian [67] | G/PLS | Correlation Classification | Multiple calculated molecular descriptors | 58 | 39 (test set1) + 181 (test set2) | $r=0.92$ | $r^2=0.62$ (test set1) > 70 % and 60 % correctly classified for CNS permeable and impermeable drugs in the test set 2. |

(Table 3) contd.....

| Reference | Method | Model | Descriptors | Dataset | | Performance | |
|-----------------|---|----------------|--|--------------|--|--|--|
| | | | | Training set | Test set | Correlation or classification | Prediction |
| Hutter [68] | MLR | Correlation | Quantum chemically derived descriptors and some other calculated molecular descriptors | 90 | 23 | $r^2=0.87$, $q^2=0.84$ | |
| Hou [69] | MLR | Correlation | Multiple calculated molecular descriptors | 72 | 14 (test set1) + 23 (test set2) | $r=0.89$ | RMSE=0.26 (test set1) RMSE=0.46 (test set2) |
| Dorronsoro [70] | ANN | Correlation | Topological parameters | 35 | | $r=0.94$ | |
| Stanton [71] | PLS | Correlation | Hydrophobic surface area parameters and some other calculated molecular descriptors | 97 | | $r^2=0.78$ | |
| Winkler [72] | Bayesian neural nets | Correlation | Multiple calculated molecular descriptors | 85 | 21 | $r^2=0.74$ | $r^2=0.65$ |
| Cabrera [73] | MLR | Correlation | TOPS-MODE topological descriptors | 114 | 28 | $r=0.86$ | MAE=0.33 |
| Li [74] | Logistic regression (LR), linear discriminate analysis (LDA), k nearest neighbor (KNN), C4.5 decision tree (C4.5 DT), probabilistic neural network (PNN), and SVM | Classification | 199 calculated molecular descriptors | 415 | | 71.0 % (LR) 71.2% (LDA) 71.2% (C4.5 DT) 77.1% (KNN) 76.5% (PNN) 83.7% (SVM) | |
| Narayanan [75] | MLR | Correlation | Electrotopological state indices and some other molecular descriptors | 88 | 13 (test set1) + 15 (test set2) + 92 (test set3) | $r=0.86$, $q=0.85$ | |
| Yap [76] | General regression neural network (GRNN) | Correlation | Multiple calculated molecular descriptors | 129 | 30 | | $r^2=0.70$, RMSE=0.13 |

them may have high correlation. For example polar surface area is partially correlated to some hydrogen-bonding descriptors.

a. Polar surface area (PSA). Polar surface area is defined as the surface area associated with the hydrogen-bonding acceptor atoms nitrogen and oxygen and the hydrogen atoms bound to these heteroatoms. Sometimes, sulfur atoms and hydrogen atoms attached to sulfur may also (be included. In 1992, van de Waterbeemd and Kansy correlated the PSA of a series of CNS drugs to $\log BB$ firstly [53]. Thenceforward, PSA has become the most popular parameter for the prediction of molecular transport properties. van de Waterbeemd found that a good correlation could be obtained using PSA together with the calculated molar volume (V_M):

$$\log BB = -0.021 \times PSA - 0.003 \times V_M + 1.643 \quad (4)$$

$$(n=20, r=0.835, s=0.448, F=19.5)$$

Waterbeemd and Kansy also note that the V_M term can be replaced by non-polar surface area (NPSA) while retaining good statistics ($r = 0.845$).

Another pioneering research relating to PSA was that of van de Waterbeemd *et al.* in which a quantitative structure-absorption relationship was derived for the passage of 17 compounds across Caco-2 monolayers [42]:

$$\log P_{app} = -0.043 \times PSA + 0.008 \times MW - 5.165 \quad (5)$$

$$(n=17, r=0.833)$$

The method used by van de Waterbeemd *et al.* to calculate PSA is only based on a single conformation of the molecule of interest. By contrast, in 1996, Palm and co-workers found that excellent correlation could be obtained between the dynamic polar van der Waals surface areas

(PSA_d) and Caco-2 permeabilities ($r^2=0.99$) [41]. Furthermore, using PSA_d, Palm and coworkers found that an excellent sigmoidal relationship could be established between %FA and PSA_d ($r^2=0.94$) for a set of 20 drugs covering a wide range of fractional absorption values (%FA) in humans [18]. Drugs that are completely absorbed (FA > 90%) had a PSA_d ≤ 60 Å² while drugs that are less than 10% absorbed had a PSA_d ≥ 140 Å². The dynamic polar surface area is a statistical average in which the surface area of each conformation is weighted by its probability to exist. Dynamic surface properties of each compound were calculated considering all low-energy conformations within 2.5 kcal/mol of the global minimum based on Monte Carlo conformational search and energy minimizations.

The major drawback of PSA_d is that it is computationally expensive, which makes PSA_d inappropriate for database screening. Clark compared the performance of PSA and PSA_d and found that PSA is not very sensitive to the different conformation of small organic molecule and can be simply computed based on a single well-generated 3-D structure [21,57]. As suggested by Clark, the criterion for poor absorption of PSA > 140 Å² appears to be an efficient and robust method of computationally screening large numbers of compounds prior to synthesis, which is consistent with the rules proposed by Palm *et al.* [18].

Both of PSA_d and PSA require the 3-D structures of a molecule. Is it possible to develop a procedure to roughly estimate the PSA only based on the 2-D topology connection information of a molecule? Ertl and coworkers have developed such a method to generate a topological PSA (TPSA) based on 3D PSA values for 43 fragments resulting from analysis of 34,810 compounds taken from the WDI database [82]. The correlation between PSA and TPSA is very high ($r^2=0.98$), while the computation speed is 2-3 orders of magnitude faster. But it should be noted that the basic assumption of TPSA is that all defined atom types expose to solvent. This assumption is true for small molecules. But for relatively large and flexible molecules, the conformational dependencies may bury parts of the polar atoms, thus possibly resulting in an overestimation of the computed TPSA.

Recently, Hou and coworker proposed the concept of "high-charged polar atom". According to the definition, only polar atoms with high charge densities belong to high-charged polar atoms. The Gasteiger method was used to calculate the partial charges, and the PSA surrounding those polar atoms with absolute partial charges larger than 0.1 |e| was treated as the high-charged polar surface area (HCPSA). Compared with PSA, HCPSA obtained a better correlation with logBB [69] and logP_{app} [50].

PSA can of course be combined with other molecular descriptors to develop further improved models compared with using PSA alone. For example, in an effort to account for hydrophobic contributions, Clark introduced logP as an additional descriptor [57]:

$$\log BB = 0.139 - 0.148PSA + 0.152C \log P \quad (6)$$

($n=55, r^2=0.79, s=0.35, F=95.8$)

$$\log BB = 0.131 - 0.145PSA + 0.172M \log P \quad (7)$$

($n=55, r^2=0.77, s=0.37, F=86.0$)

Egan *et al.* published a model for intestinal absorption based on PSA and logP descriptors alone [25]. Extensive validation of the model on known orally delivered drugs, drug-like molecules, and compounds assayed by Pharmacocepta, Inc. for Caco-2 cell permeability demonstrated a reasonably good rate of successful predictions (74-92%, depending on dataset and criterion).

(b). Lipophilicity parameters: According to Fick's first law of diffusion, passive drug transport across a biological membrane is proportional to the membrane-water partition coefficient, assuming that the membrane interior is homogenous and that the drug concentration on the receiver side is much lower than the concentration on the donor side. Since membrane-water partition coefficients are not readily available, partition coefficients between water and an organic solvent such as *n*-octanol are normally used. True partition coefficient, *P*, is the easiest lipophilicity parameter that can be used in drug absorption because logP can be precisely computed by using atomic or fragment-addition approaches. In fact, logP values can only be a first estimate of the lipophilicity of a compound in a biological environment. Since many organic molecules that have different ionizable state in different pH have different logP values. The presence of more than one species results in an average partition coefficient: the apparent partition coefficient or distribution coefficient *D*, which is pH-dependent in case of the existence of ionizable compounds. Several investigators have reported a correlation between lipophilicity parameters (logP, logD, ΔlogP) and drug absorption.

Krämer compared the experimentally determined logD of 14 structurally diverse drug and potential drug compounds with their apparent Caco-2 permeability coefficient (*P*_{app}), and with the fraction absorbed in humans after oral administration [25]. The bell-shaped relationship between logP_{app} and the experimental logD can be observed. logD gave a better correlation than both logP of the neutral species or logP of the neutral species with a correction for the molar fraction of the neutral species. Generally, compounds with low logD are poorly absorbed, whereas compounds with logD < -1 offer satisfactory absorption.

Hou and coworkers compared the correlation between Caco-2 permeability and logP and that between Caco-2 permeability and logD (Eqs 8 and 9) [50]. A direct fitting of logP values with logP_{eff} values of the compounds in the training set produced an *r* value of approximately 0.47. If logD was used instead of logP, the correlation coefficient was improved to $r=0.71$. Obviously, for partition processes in the body, the distribution coefficient *D*, for which an aqueous buffer at pH=7.4 (blood pH) is used in the experimental determination, often provides a more meaningful description of lipophilicity, especially for ionizable compounds.

$$\log P_{eff} = -5.469 + 0.236 \log P \quad (8)$$

$$(n=77, r=0.471, s=0.657, F=24.0)$$

$$\log P_{eff} = -5.265 + 0.311 \log D \quad (9)$$

$$(n=77, r=0.708, s=0.536, F=75.2)$$

In equation 9, the experimental logD values were used. In the development of a theoretical predictive model, it is cer-

tantly expected that all variables in a model are theoretically-derived, so the prediction is experimentally irrespective. It is interesting to compare the performance of the experimental and calculated $\log D$ values, and thus the authors [50] performed a correlation between $\log P_{eff}$ and the predicted $\log D$ values on 44 compounds in the training set. If the experimental $\log D$ was replaced by the predicted $\log D$ (ACD $\log D$), the correlation was decreased from 0.69 to 0.51. It is obvious that the performance of the calculated $\log D$ values was not satisfactory. The deviations caused by calculations may be mainly caused by the pKa prediction. Several approaches have been developed for pKa predictions, including ACD/pKa (ACD), Pallas/pKa (Compudrug) and SPARC [83]. But until now, these methods cannot provide very reliable prediction for some complicated organic molecules.

Besides n-octanol/water partition system, other solvent/water partition system is used to gain addition information on barrier permeability. Young *et al.* proposed a correlation between $\log BB$ and $\Delta \log P$ (see equation 10) [52]. $\Delta \log P$ is defined as the difference between the n-octanol/water partition coefficient ($\log P_{ow}$) and the cyclohexane/water partition coefficient ($\log P_{cychw}$). The $\Delta \log P$ can be treated as a measurement of the hydrogen-bonding capability, because, unlike cyclohexane, n-octanol permits hydrogen-bonding.

$$\log BB = 1.889 - 0.485\Delta \log P \quad (10)$$

($n=20$, $r=0.831$, $s=0.439$, $F=40.23$)

(c). The Abraham descriptors. The simplest way of calculating the hydrogen bonding capacity is to count the number of hydrogen bond donor and acceptor atoms or to count the number of lone pairs of electrons on certain kinds of atoms. Certainly, these simplified models are not highly accurate descriptions of the hydrogen bonding properties of the molecules, but they have in some cases provided reasonable predictions of membrane permeability. PSA belongs to H-bonding descriptor. Österberg and Norinder analyzed the relationship between PSA and three H-bonding descriptors (number of H-bond nitrogen atoms, number of H-bond oxygen atoms and number of H-bond donor atoms on nitrogen and oxygen), and found high linear correlation between the hydrogen-bonding descriptors and PSA of five chemically diverse sets of drugs ($r^2 > 0.93$, $q^2 > 0.69$) [24].

Abraham and coworkers developed a set of parameters to model solvation and H-bonding properties of organic molecules [84]. These solute descriptors are based on the physically meaningful theoretical cavity model of solute-solvent interactions, and widely applied in the prediction of a variety of physicochemical and pharmacokinetic properties, such as solubility, blood-brain partitioning, skin permeability, and human intestinal absorption according to Equation 11:

$$\log SP = c + rR_2 + s\pi_2^H + a \sum \alpha_2^H + b \sum \beta_2^H + vV_x \quad (11)$$

where SP is a solute property in a given system; R_2 represents excess molar refraction which models dispersion force interactions arising from the greater polarizability of π - and n- electrons; π_2^H represents solute dipolarity/polarizability due to solute-solvent interactions between bond dipoles and induced dipoles; $\sum \alpha_2^H$, hydrogen-bond acidity, relates to

the strength and number of H-bonds formed by donor groups in the solute when they interact with lone pairs of acceptor groups in solvent molecules; $\sum \beta_2^H$, hydrogen-bond basicity, relates to the strength and number of H-bonds formed by the lone pairs of acceptor groups in the solute when they interact with donor solvents; V_x represents McGowan characteristic molar volume. π_2^H , $\sum \alpha_2^H$ and $\sum \beta_2^H$ can be obtained from partition studies in different biphasic systems with known c , r , s , a , b and v factors or, where applicable, from gas-liquid chromatography analysis on a polar, non-acidic stationary phase for π_2^H and highly basic or acidic stationary phases for $\sum \alpha_2^H$ and $\sum \beta_2^H$, respectively. These descriptors can also be calculated from existing databases for the respective molecule fragments. V_x can be calculated from the molecular structure of the solute.

The first application of Abraham descriptors in drug absorption is reported in the prediction of blood-brain partitioning [54]. Using a dataset of 65 drug or drug-like molecules, Abraham and coworkers obtained the following equation:

$$\log BB = -0.038 - 0.715 \sum \alpha_2^H - 0.698 \sum \beta_2^H + 0.198R_2 - 0.687\pi_2^H + 0.995V_x \quad (12)$$

($n=57$, $r^2=0.916$, $s=0.197$, $F=99.2$)

It should be noted that Equation 12 was obtained after eliminating eight outliers from the training set. Equation 12 shows exactly the solute factors that govern BB values. From this equation, it can be deduced that factors relating to polarity and hydrogen bonding disfavor brain penetration while solute size (V_x) appears to promote partitioning into the brain.

Platts and coworkers reparameterized Equation 12 using a large dataset (148 molecules) (Equation 13) [61]:

$$\log BB = 0.044 + 0.511R_2 - 0.886s\pi_2^H - 0.724 \sum \alpha_2^H - 0.666b \sum \beta_2^H + 0.861V_x \quad (13)$$

($n=148$, $r=0.843$, $s=0.367$, $F=71.0$)

According to prediction using Equation 13, several large discrepancies were observed for carboxylic acid containing molecules such as salicylic acid and indomethacin, and an obvious improvement could be achieved by adding an indicator variable. Equation 13 indicates that size strongly enhances brain uptake, whereas polarity/polarizability, H-bond acidity, basicity, and the presence of carboxylic acid groups strongly reduce brain penetration.

Zhao and coworker reported a prediction model based on the Abraham descriptors to model the human intestinal absorption data of 169 drugs [26]. The obtained model possesses good correlation and external prediction ability. The step-wise analysis show that the two dominated descriptors are $\sum \alpha_2^H$ and $\sum \beta_2^H$. This is in agreement with previous work that suggests hydrogen-bond donors and hydrogen-bond acceptors or polar molecular surface are good descriptors with which to model human intestinal absorption.

$$\%FA = 90 + 2.11R_2 + 1.70\Pi_2^H - 20.7\sum\alpha_2^H - 22.3\sum\alpha_2^H + 15.0V_x \quad (14)$$

($n=38$, $r^2=0.83$, $q^2=0.75$, $s=16\%$, $F=31$)

(d). Volsurf parameters. Volsurf descriptors, developed by Cruciani and coworkers [46], were used to quantitatively characterize size, shape, polarity, hydrophobicity and the balance between them of organic molecules. The GRID forcefield was chosen to calculate energetically favorable interactions sites around a molecule, and produce 3D molecular interaction fields (MIFs). A MIF maps the chemical forces between an interacting partner and a target molecule onto a 3D grid. The water probe (OH2) was used to simulate solvation-desolvation processes, while the hydrophobic probe (called DRY in the GRID program) and the carbonyl probe (O) were used to simulate drug-membrane interactions. The DRY probe is a specific probe to compute the hydrophobic energy; the overall energy of the hydrophobic probe is computed at each grid point as $E_{\text{entropy}} + E_{\text{LJ}} - E_{\text{HB}}$, where E_{entropy} is the ideal entropic component of the hydrophobic effect in an aqueous environment, E_{LJ} the induction and dispersion interactions occurring between any pair of molecules; and E_{HB} the H-bonding interactions between water molecules and polar groups on the target surface. VolSurf has the nice advantage of producing 2D descriptors using the 3D information embedded in any map. Moreover, the VolSurf transformation is easy to be understood, and fast to be computed. The descriptors have a clear chemical meaning and are lattice independent, and some of them can be projected back into the original 3D grid map from which they were obtained. The VolSurf descriptors show good performance in the prediction of Caco-2 permeability [46] and blood-brain partitioning [60].

(e) MolSurf descriptors. Norinder and co-workers also analyzed the same set of compounds used by van de Waterbeemd *et al.* using a quantum chemistry-based approach and PLS multivariate data analysis [43]. The authors developed a protocol involving both semi-empirical as well as *ab initio* calculations followed by the final computation of molecular calculated descriptors by the MolSurf technology. The chemical behavior and the calculated descriptors depend on the distribution of electrons and energy in the valence region. The electrostatic potential, $V(r)$, and the local ionization energy, $I(r)$, are calculated at points evenly distributed on this surface. The computed 13 descriptors describe properties such as based strength, hydrophobicity, hydrogen bonding, polarity as well as polarizability [22,43,56].

Norinder *et al.* divided the 'Waterbeemd data set' into a training set of 9 compounds and a test set of 8 compounds. The derived PLS model has good statistical significance and good predictivity [43]:

$$\log P_{\text{eff}} \text{ PLS model:} \quad (15)$$

($n=9$, $r^2=0.935$, $q^2=0.849$, $s=0.33$, $F=40.88$)

($\text{rmse}^{\text{tr}}=0.270$, $\text{rms}^{\text{te}}=0.409$)

From the analysis of the PLS model, the authors found that the most important factors influencing the model are associated with hydrogen bonding. Thus variables such as the number of possible hydrogen donor atoms as well as the

number of hydrogen bond acceptor nitrogens have the greatest impact along with the actual strength of the hydrogen bond in the latter case. High lipophilicity and the presence of surface electrons, i.e. valence electrons, which are not tightly bonded to the molecule, were also found to have a favorable influence to achieve high Caco-2 monolayer permeability.

Norinder *et al.* used the MolSurf parameters to model the brain-blood partitioning of 57 organic molecules [56]. Norinder *et al.* divided the 57 molecules into three parts—a training set and two test sets (one compound was excluded). The second test set (test set 2), identical to the test set used by Lombardo *et al.* [55], was also used for evaluation of the derived PLS statistical models. The derived PLS models predicted the brain-blood partitioning of the second test set with greater precision (rmse 0.473 and 0.508, respectively) than the model reported by Lombardo *et al.* (rmse 1.244). The results from the PLS analyses are summarized as follows:

$$\log BB \text{ PLS model:} \quad (16)$$

($n=28$, $r^2=0.862$, $q^2=0.782$, $s=0.311$, $F=49.93$)

($\text{rmse}^{\text{tr}}=0.288$, $\text{rms}^{\text{te}}=0.473$)

The most important properties influencing the model were associated with polarity and Lewis base strength and should be kept to a minimum to promote high partitioning. The absence of atoms capable of hydrogen bonding interactions as well as high lipophilicity ($\log P$) and the presence of polarizable surface electrons, i.e., valence electrons, were also found to promote high $\log BB$.

Using the same molecular descriptors, Norinder *et al.* developed a prediction model of intestinal absorption based on dataset of 20 diverse drug-like molecules [22]. The PLS model for the training set of 13 molecules has excellent correlation:

$$\log(\%FA) \text{ PLS model:} \quad (17)$$

($n=13$, $r^2=0.903$, $q^2=0.685$, $s=0.628$, $F=28.05$)

($\text{rmse}^{\text{tr}}=0.523$, $\text{rms}^{\text{te}}=0.488$)

Properties associated with hydrogen bonding had the largest impact on drug absorption. The analyses show that the MolSurf parameters have high correlation with the polar surface area (PSA), which can be indicated by the high correlation coefficient ($r^2=0.98$, $q^2=0.93$). Although PSA and the MolSurf parameters are highly correlated, it is believed that the MolSurf descriptors have some advantages than PSA. Firstly, the MolSurf descriptors give a more comprehensive characterization of the molecule. Moreover, the MolSurf descriptors are also easier to be interpreted than dynamic polar surface area with respect to structural requirements that are of importance for intestinal absorption.

(f). Eletrotopological state index (E-state). The eletrotopological state index (E-state) is developed from chemical graph theory and uses the chemical graph (hydrogen-suppressed skeleton) for generation of atom-level structure indices. The index is based on the electronic effect of each atom on the other atoms in the molecule as modified by molecular topology. Each atom has an assigned intrinsic state value I_i calculated as follows:

$$I_i = ((2/N_i)2\delta_i^v + 1)/\delta_i \quad (18)$$

where N is the principal quantum number of the atom i , δ^v the number or valence electrons in the skeleton (Z^v-h), and δ the number of s electrons in the skeleton ($s-h$). For a skeleton atom, Z^v is the number of valence electrons, s the number of electrons in s orbitals, and h the number of bounded hydrogen atoms. The E-state $S(A_i)$ for the atom is the modified intrinsic value:

$$S(A_i) = I_i + \Delta I_i \quad (19)$$

where ΔI_i quantifies the perturbing effect on the intrinsic atom value. This perturbation is assumed to be a function of the difference in the intrinsic values I_i and I_j :

$$\Delta I_i = \sum_{j=1}^N (I_i - I_j) / r_{ij}^2 \quad (20)$$

where r_{ij} is the number of atoms in the shortest path between atoms i and j including both i and j . The difference in intrinsic values, ΔI_i , for a pair of skeletal atoms encodes both electronic and topological attributes that arise from electronegativity differences and skeletal connectivity. Derived from this electronegativity difference, the E-state value for an atom is related but not limited to the concept of atomic partial charge. In addition to an atom-level E-State value computed for each atom, an atom-type formalism has been developed. The atom type E-State index is defined as the sum of the individual atom level E-State values for a particular atom type.

Rose and Hall have developed a QSPR model to predict $\log BB$ based on a training set of 86 compounds and a test set of 20 molecules [65]. The model based on three variables yielded statistical information as follow:

$$\log BB = -0.202HS^T(\text{HBd}) + 0.00627[HS^T(\text{arom})]^2 - 0.105[d^2\chi^v]^2 - 0.425 \quad (21)$$

$$(n=102, r^2=0.66, q^2=0.62, s=0.45, F=62.4)$$

where $HS^T(\text{HBd})$ represents the sum of the hydrogen E-State values for groups that act as hydrogen bond donors. The negative coefficient on $HS^T(\text{HBd})$ indicates that hydrogen bond donor groups lead to negative or low value of $\log BB$. The second value in the model is the square of the atom-type hydrogen E-State descriptor for aromatic CH groups, $HS^T(\text{arom})$. Because of the positive coefficient on $HS^T(\text{arom})$, larger values are related to larger $\log BB$ value. The third variable in the model is the square of the second-order valence molecular connectivity difference chi index, $d^2\chi^v$. This variable increases with increased branching in the structure. Because of the negative coefficient on $d^2\chi^v$, larger values are related to more negative $\log BB$ values.

Norinder and Österberg developed several PLS models for the predictions of Caco-2 cell permeability, human intestinal absorption and blood-brain partitioning using CLOGP, CMR (calculated molar refraction) and the electrotopological state indices. Good statistical models were derived ($r^2=932$ and $q^2=0.790$ for $\log P_{\text{app}}$, $r^2=781$ and $q^2=0.729$ for $\log BB$, $r^2=933$ and $q^2=0.855$ for %HIA) that permit fast computational screening and prioritization of virtual libraries [27].

(g) Molecular group descriptors. In 2002, Klopman and coworkers developed a novel approach to predict human intestinal absorption [29]. Quite different from the other models, the molecular descriptors used in this model is molecular

groups, not usually used molecular descriptors. The calculated human intestinal absorption is summed by counting the frequencies of the defined groups as the following:

$$\%FA = C_0 + \sum_i c_i G_i \quad (22)$$

where C_0 is constant, c_i are the correlation coefficients of the presence (1) or (0) of a certain group.

The method developed by Klopman *et al.* is based on a modified contribution group method in which the basic parameters are structural descriptors identified by the CASE program, together with the number of hydrogen bond donors. The search for the basic molecular groups was performed using the MCASE program. The prediction model includes 36 structural groups derived from the chemical structures of a data set containing 417 drugs and one H-bond parameter, H_{donor} , the sum of all OH and NH groups. The model was able to predict the percentage of drug absorbed from the gastrointestinal tract with an r^2 of 0.79 and a standard deviation of 12.32% of the compounds from the training set. The standard deviation for an external test set (50 drugs) was 12.34%.

In fact, the atom or group addition methods have been widely applied in the prediction of $\log P$ and $\log S$ (solubility) [85,86]. The superiority of this class of approaches is that they do not need any descriptors from other theoretical models. Moreover, what this class of methods needs is to count the occurrence of functional groups in a molecule, so they are extremely computationally efficient. The shortcoming of this approach is also obvious. First, it requires a large data set to obtain contribution of each functional group. Second, it may suffer from the 'missing fragment' problem, which means that if a compound contains 'missing fragment' which can be considered by the group contribution model, its properties cannot be precisely predicted. Now, the experimental data for $\log P_{\text{app}}$, $\log BB$ or even %FA are very limited, so it is difficult to define many atom types or molecular groups and build a reliable addition model. But we believe that along with the increase of the experimental data, this class of approaches will become more important.

4. RECENT ADVANCES IN PREDICTION OF DRUG ABSORPTION

(a). The application of support vector machine (SVM). In recent years, one of the most exciting advances in this area is the introduction of some new statistical and machine-learning methods, especially support vector machine [37,39,66,74]. SVM is based on the "structural risk minimization, SRM" principle from statistical learning theory. In linearly separable cases, SVM constructs a hyperplane that separates two different classes of vectors with a maximum margin. In this case, a vector corresponds to a chemical agent, and this vector is represented by x_i , with structural and physicochemical descriptors of the chemical agent as its components. This is done by finding another vector w and a parameter b that minimizes $\|w\|^2$ and satisfies the following conditions:

$$\begin{cases} w \cdot x_i + b \geq -1, & \text{for } y_i = +1 \text{ class 1 (positive samples)} \\ w \cdot x_i + b \leq +1, & \text{for } y_i = -1 \text{ class 2 (negative samples)} \end{cases} \quad (23)$$

where y_i is the class index, w is a vector normal to the hyperplane. After the determination of w and b , a given vector x_i can be classified by:

$$\text{sign}(w \cdot x + b) \quad (24)$$

In nonlinearly separable cases, SVM maps the input variable into a high-dimensional feature space using a kernel function $K(x_i, x_j)$. An example of a kernel function is the Gaussian kernel, which has been extensively used in different studies with good results.

$$K(x_i, x_j) = e^{-\|x_j - x_i\|^2 / 2\sigma^2} \quad (25)$$

Linear support vector machine is applied to this feature space and then the decision function is given by:

$$f(x) = \text{sign}\left(\sum_{i=1}^l \alpha_i^0 y_i K(x, x_i) + b\right) \quad (26)$$

where the coefficient α_i^0 and b are determined by maximizing the following Lagrangian expression

$$\sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (27)$$

under the following conditions:

$$\alpha_i \geq 0 \text{ and } \sum_{i=1}^l \alpha_i y_i = 0 \quad (28)$$

Compared with the other correlation or classification methods, SVM possesses prominent advantages: (1) strong theoretical background provides SVM with high generalization capability and can avoid local minima; (2) SVM always has a solution, which can be quickly obtained by a standard algorithm (quadratic programming); (3) SVM need not determine network topology in advance, which can be automatically obtained when the training process ends; (4) SVM builds a result based on a sparse subset of training samples, which reduces the workload.

Doniger and coworkers applied two different machine-learning algorithms: artificial neural network and support vector machine, to predict the blood-brain barrier permeability of 324 drug and drug-like molecules [66]. For both the ANN and the SVM, the performance of the algorithm was measured by counting the number of molecules in the validation set that were correctly classified. Based on over 30 different validation sets, SVM can predict up to 96% of the molecules correctly, averaging 81.5% over 30 test sets, which comprised of equal numbers of CNS positive and negative molecules. This is quite favorable when compared with the neural network's average performance of 75.7% with the same 30 test sets.

Recently, Li and coworkers have applied some statistical learning methods to construct the classification models to distinguish 276 BBB-penetrating (BBB+) and 139 nonpenetrating (BBB-) agents [74] based on 199 molecular descriptors. The whole dataset was randomly divided into five subsets of approximately equal size for conducting a 5-fold cross-validation test of the prediction accuracy of each of the statistical learning methods. After training a statistical learning system with a collection of four subsets, the performance of the system was tested against the fifth subset. This process

was repeated five times so that every subset was used once as a testing set. The methods tested include logistic regression, linear discriminate analysis, k nearest neighbor, C4.5 decision tree, probabilistic neural network, and support vector machine. The recursive feature elimination (RFE) method was used for the descriptor selections. For all methods studied here, RFE substantially improves both the BBB- and the overall accuracy. Of the statistical learning methods studied, SVM gives the highest BBB+, BBB-, Q (the overall prediction accuracy), and C values (Matthews correlation coefficient) of 88.6%, 75.0%, 83.7%, and 0.645, respectively, by using RFE selected descriptors and of 89.9%, 64.3%, 79.1%, and 0.524, respectively, by using the full set of descriptors. For the other five methods tested in this work, their prediction accuracies for BBB+ agents are in the range of 78.2~85.5% by using RFE-selected descriptors and 40.0~83.7% by using the full set of descriptors, and those for BBB- agents are in the range of 46.4~62.8% by using RFE-selected descriptors and 42.8~58.4% by using the full set of descriptors. Thus, SVM appears to give a somewhat better prediction accuracy than the other statistical learning methods.

(b). The development of software system to predict drug absorption. Of course, the *in silico* prediction models discussed here are not simply of academic interest – they should play a vital role in focusing and accelerate drug discovery. We expect that the algorithms for the accurate prediction of passive permeation characteristics will be developed and available for routine use by medicinal chemists. So besides the development of prediction of higher confidence, another challenge is to develop *in silico* ADME-Tox prediction software system and integrate the existing tools into a single, consistent workflow environment. Very encouraging, a number of companies traditionally active in the field of molecular modeling have also recently began to develop software or modules to assist estimate ADME/Tox properties. At present, many commercial programs are available, including VolSurf (tripos, <http://www.tripos.com>), C2.ADME (accelrys, <http://www.accelrys.com>), and QikProp (schrodinger, <http://www.schrodinger.com>). It should be noted that at present the predicted computational models are limited to several properties, such as drug solubility, Caco-2 cell absorption, blood-brain barrier permeation and intestinal absorption. For example, in C2.ADME, the predicted properties include intestinal absorption, blood-brain barrier (BBB) penetration, and aqueous solubility at 25°C. In QikProp, the predicted properties include solubility, blood-brain barrier permeability, Caco-2 cell Permeability, MDCK cell permeability and skin permeability. Based on these software systems and the integrated molecular modeling environment, we can perform effect analysis on combinatorial and virtual libraries.

5. THE FUTURE DIRECTIONS

Drug absorption is a very difficult process because it arises from multiple physiological processes. At present, almost all developed models are related to passive diffusion. To deal with the biological complexity from which ADME properties arise, more experimental data with high quality, both *in vitro* or *in vivo*, are necessary. As shown in Table 1, 2 and 3, the largest logBB dataset used in correlation consist of about 150 compounds (Zhao's dataset) [26], and the largest

$\log P_{\text{eff}}$ dataset about 110 compounds (Hou's dataset) [50]. Clearly, the small size of these training sets will limit the general applicability of any models that are derived from them. A representative example is the correlation between PSA and drug absorption. Palm found that PSA could be well linearly correlated with $\log P_{\text{eff}}$ of six β -adrenoreceptor antagonists [41]. But if we used the Hou's dataset of 77 compounds, the correlation between PSA and $\log P_{\text{eff}}$ is not very good ($r=0.664$). It is obvious that the model based on PSA cannot be treated as a universal principle to predict caco-2 permeabilities [50]. So large dataset gives us opportunities to develop more reliable prediction models. Also, for increased effectiveness, approaches combining models considering both passive diffusion and active transport should be considered, as more data on these become available.

As presented in Table 1, 2 and 3, for the same property, there are many prediction models. Any single model used for *in silico* prediction of drug absorption may not be completely accurate. In fact, a consensus score, a combination of two or more models for the same property, based on different principles, may enable us to make a sound judgment on the quality and reliability of the predictions and to explore the source of uncertainty of the predictions. Actually, the concept 'consensus score' is not new in molecular modeling. In the estimation of the protein-ligand interaction, it has been validated that compared with the performances of a single scoring function, the hit rates can be effectively improved by using the consensus score [87]. But it seems that the concept 'consensus score' is only introduced to the field of ADME prediction recently. Hou *et al.* used eight best models from genetic algorithm, instead of a single one, to predict $\log BB$. The top model predicted the training set with $r=0.87$; by averaging its output with the eight best models, the correlation coefficient climbs to 0.88. Moreover, the prediction on the external test set using the multiple models was improved. The authors believed that selection of a single model and the discarding of the remaining models might not be the most advantageous course, and the outputs of the multiple models can be averaged to gain the most reliable results [64]. Recently, researchers in Bio-Rad Laboratories, Inc. has introduced the consensus score to the KnowItAll ADME/Tox software system (Bio Rad, <http://www.bio-rad.com>), and found that the benefits of employing multiple complementary models for the same ADME-Tox endpoint in a consensus modeling approach to provide significantly greater accuracy over that of any single model. In this software system, two types of consensus model are applied: Real variables and Boolean variables [88]. For real variable consensus models, a weighted average of the individual models is used. A real variable consensus model is trained against a set of experimental results. By comparing the actual values to the results predicted for each individual model, the software system can mathematically compare the models and create a weighted average that most closely matches the experimental values. The weighted average consensus model can then be used to screen large libraries of compounds in batch mode. Boolean variable consensus models work with predictors that classify compounds into one of two classes, for example, BBB+ or BBB-. In the near future, how to integrate multiple models and develop reliable consensus model is also a very interesting research direction.

REFERENCES

- Hou, T. J.; Xu, X. J. *Curr. Pharm. Des.*, **2004**, *10*, 1011.
- Kennedy, T. *Drug Discov. Today*, **1997**, *2*, 436.
- Venkatesh, S.; Lipper, R. A. *J. Pharm. Sci.*, **2000**, *89*, 145.
- Li, A. P. *Drug Discov. Today*, **2001**, *6*, 357.
- van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.*, **2003**, *2*, 192.
- Mitic, L. L.; Anderson, J. M. *Annu. Rev. Physiol.*, **1998**, *60*, 121.
- Stenberg, P.; Bergstrom, C. A. S.; Luthman, K.; Artursson, P. *Clin. Pharmacokinet.*, **2002**, *41*, 877.
- Artursson, P.; Ungell, A. L.; Lofroth, J. E. *Pharmaceut. Res.*, **1993**, *10*, 1123.
- Artursson, P. *Crit. Rev. Ther. Drug*, **1991**, *8*, 305.
- Artursson, P.; Karlsson, J. *Biochem. Biophys. Res. Commun.*, **1991**, *175*, 880.
- Lennernas, H.; Palm, K.; Fagerholm, U.; Artursson, P. *Int. J. Pharm.*, **1996**, *127*, 103.
- Lipinski, C. A. *J. Pharmacol. Toxicol.*, **2000**, *44*, 235.
- Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Comb. Chem.*, **1999**, *1*, 55.
- Oprea, T. I. *J. Comput. Aid. Mol. Des.*, **2000**, *14*, 251.
- Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. *J. Med. Chem.*, **2003**, *46*, 1250.
- Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipkind, P. A. *J. Med. Chem.*, **2004**, *47*, 224.
- Ekins, S.; De Groot, M. J.; Jones, J. P. *Drug. Metab. Dispos.*, **2001**, *29*, 936.
- Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. *Pharmaceut. Res.*, **1997**, *14*, 568.
- Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. *J. Chem. Inf. Comp. Sci.*, **1998**, *38*, 726.
- Ghuloum, A. M.; Sage, C. R.; Jain, A. N. *J. Med. Chem.*, **1999**, *42*, 1739.
- Clark, D. E. *J. Pharm. Sci.*, **1999**, *88*, 807.
- Norinder, U.; Osterberg, T.; Artursson, P. *Eur. J. Pharm. Sci.*, **1999**, *8*, 49.
- Raevsky, O. A.; Fetisov, V. I.; Trepalina, E. P.; McFarland, J. W.; Schaper, K. J. *Quant. Struct.-Act. Rel.*, **2000**, *19*, 366.
- Osterberg, T.; Norinder, U. *J. Chem. Inf. Comp. Sci.*, **2000**, *40*, 1408.
- Egan, W. J.; Merz, K. M.; Baldwin, J. J. *J. Med. Chem.*, **2000**, *43*, 3867.
- Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. *J. Pharm. Sci.*, **2001**, *90*, 749.
- Norinder, U.; Osterberg, T. *J. Pharm. Sci.*, **2001**, *90*, 1076.
- Agatonovic-Kustrin, S.; Beresford, R.; Yusof, A. P. M. *J. Pharmaceut. Biomed.*, **2001**, *25*, 227.
- Klopman, G.; Stefan, L. R.; Saiakhov, R. D. *Eur. J. Pharm. Sci.*, **2002**, *17*, 253.
- Abraham, M. H.; Zhao, Y. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Reynolds, D. P.; Beck, G.; Sherborne, B.; Cooper, I. *Eur. J. Med. Chem.*, **2002**, *37*, 595.
- Derety, E.; Feher, M.; Schmidt, J. M. *Quant. Struct.-Act. Rel.*, **2002**, *21*, 493.
- Zmuidinavicius, D.; Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. *J. Pharm. Sci.*, **2003**, *92*, 621.
- Niwa, T. *J. Chem. Inf. Comp. Sci.*, **2003**, *43*, 113.
- Perez, P. A. C.; Sanz, M. B.; Torres, L. R.; Avalos, R. C.; Gonzalez, M. P.; Diaz, H. G. *Eur. J. Med. Chem.*, **2004**, *39*, 905.
- Wegner, J. K.; Frohlich, H.; Zell, A. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 931.
- Sun, H. M. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 748.
- Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 1630.
- Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Vander Heyden, Y. *J. Pharmaceut. Biomed.*, **2005**, *39*, 91.
- Liu, H. X.; Hu, R. J.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Comput. Aid. Mol. Des.*, **2005**, *19*, 33.
- Jones, R.; Connolly, P. C.; Klamt, A.; Diedenhofen, M. *J. Chem. Inf. Model.*, **2005**, *45*, 1337.
- Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. *J. Pharm. Sci.*, **1996**, *85*, 32.
- vandeWaterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. *Quant. Struct.-Act. Rel.*, **1996**, *15*, 480.

- [43] Norinder, U.; Osterberg, T.; Artursson, P. *Pharmaceut. Res.*, **1997**, *14*, 1786.
- [44] Krarup, L. H.; Christensen, I. T.; Hovgaard, L.; Frokjaer, S. *Pharmaceut. Res.*, **1998**, *15*, 972.
- [45] Segarra, V.; Lopez, M.; Ryder, H.; Palacios, J. M. *Quant. Struct.-Act. Rel.*, **1999**, *18*, 474.
- [46] Cruciani, C.; Crivori, P.; Carrupt, P. A.; Testa, B. *J. Mol. Struct. Theochem.*, **2000**, *503*, 17.
- [47] Fujiwara, S.; Yamashita, F.; Hashida, M. *Int. J. Pharm.*, **2002**, *237*, 95.
- [48] Kulkarni, A.; Han, Y.; Hopfinger, A. J. *J. Chem. Inf. Comp. Sci.*, **2002**, *42*, 331.
- [49] Ponce, Y. M.; Perez, M. A. C.; Zaldivar, V. R.; Ofori, E.; Montero, L. A. *Int. J. Mol. Sci.*, **2003**, *4*, 512.
- [50] Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 1585.
- [51] Nordqvist, A.; Nilsson, J.; Lindmark, T.; Eriksson, A.; Garberg, P.; Kihlen, M. *Qsar Comb. Sci.*, **2004**, *23*, 303.
- [52] Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. *J. Med. Chem.*, **1988**, *31*, 656.
- [53] Vandewaterbeemd, H.; Kansy, M. *Chimia*, **1992**, *46*, 299.
- [54] Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. *J. Pharm. Sci.*, **1994**, *83*, 1257.
- [55] Lombardo, F.; Blake, J. F.; Curatolo, W. J. *J. Med. Chem.*, **1996**, *39*, 4750.
- [56] Norinder, U.; Sjoberg, P.; Osterberg, T. *J. Pharm. Sci.*, **1998**, *87*, 952.
- [57] Clark, D. E. *J. Pharm. Sci.*, **1999**, *88*, 815.
- [58] Luco, J. M. *J. Chem. Inf. Comp. Sci.*, **1999**, *39*, 396.
- [59] Ajay, Bemis, G. W.; Murcko, M. A. *J. Med. Chem.*, **1999**, *42*, 4942.
- [60] Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. *J. Med. Chem.*, **2000**, *43*, 2204.
- [61] Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. *Eur. J. Med. Chem.*, **2001**, *36*, 719.
- [62] Liu, R. F.; Sun, H. M.; So, S. S. *J. Chem. Inf. Comp. Sci.*, **2001**, *41*, 1623.
- [63] Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. *J. Comput. Aid. Mol. Des.*, **2001**, *15*, 697.
- [64] Hou, T. J.; Xu, X. J. *J. Mol. Model.*, **2002**, *8*, 337.
- [65] Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comp. Sci.*, **2002**, *42*, 651.
- [66] Doniger, S.; Hofmann, T.; Yeh, J. *J. Comput. Biol.*, **2002**, *9*, 849.
- [67] Subramanian, G.; Kitchen, D. B. *J. Comput. Aid. Mol. Des.*, **2003**, *17*, 643.
- [68] Hutter, M. C. *J. Comput. Aid. Mol. Des.*, **2003**, *17*, 415.
- [69] Hou, T. J.; Xu, X. J. *J. Chem. Inf. Comp. Sci.*, **2003**, *43*, 2137.
- [70] Dorronsoro, I.; Chana, A.; Abasolo, I.; Castro, A.; Gil, C.; Stud, M.; Martinez, A. *Qsar Comb. Sci.*, **2004**, *23*, 89.
- [71] Stanton, D. T.; Mattioni, B. E.; Knittel, J. J.; Jurs, P. C. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 1010.
- [72] Winkler, D. A.; Burden, F. R. *J. Mol. Graph. Model.*, **2004**, *22*, 499.
- [73] Cabrera, M. A.; Bermejo, M.; Perez, M.; Ramos, R. *J. Pharm. Sci.*, **2004**, *93*, 1701.
- [74] Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Cao, Z. W.; Chen, Y. Z. *J. Chem. Inf. Model.*, **2005**, *45*, 1376.
- [75] Narayanan, R.; Gunturi, S. B. *Bioorg. Med. Chem.*, **2005**, *13*, 3017.
- [76] Yap, C. W.; Chen, Y. Z. *J. Pharm. Sci.*, **2005**, *94*, 153.
- [77] Norinder, U.; Haeblerlein, M. *Adv. Drug Deliv. Rev.*, **2002**, *54*, 291.
- [78] Butina, D.; Segall, M. D.; Frankcombe, K. *Drug Discov. Today*, **2002**, *7*, S83.
- [79] Clark, D. E. *Drug Discov. Today*, **2003**, *8*, 927.
- [80] Lobell, M.; Molnar, L.; Keseru, G. M. *J. Pharm. Sci.*, **2003**, *92*, 360.
- [81] Ecker, G. F.; Noe, C. R. *Curr. Med. Chem.*, **2004**, *11*, 1617.
- [82] Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.*, **2000**, *43*, 3714.
- [83] Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. *Quant. Struct.-Act. Rel.*, **1995**, *14*, 348.
- [84] Abraham, M. H.; Chadha, H. S.; Martins, F.; Mitchell, R. C.; Bradbury, M. W.; Gratton, J. A. *Pestic. Sci.*, **1999**, *55*, 78.
- [85] Hou, T. J.; Xu, X. J. *J. Chem. Inf. Comp. Sci.*, **2003**, *43*, 1058.
- [86] Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. *J. Chem. Inf. Comp. Sci.*, **2004**, *44*, 266.
- [87] Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. *J. Med. Chem.*, **1999**, *42*, 5100.
- [88] Banik, G. M. *Curr. Drug Discov.*, **2004**, *31*.