Pharmaceutical Research, Vol. 25, No. 8, August 2008 (© 2008)
DOI: 10.1007/s11095-008-9584-5

Research Paper

New Predictive Models for Blood–Brain Barrier Permeability of Drug-like Molecules

Sandhya Kortagere,[1] Dmitriy Chekmarev,[1] William J. Welsh,[1] and Sean Ekins[1,2,3,4]

Received February 28, 2008; accepted March 27, 2008; published online April 16, 2008

Purpose. The goals of the present study were to apply a generalized regression model and support vector machine (SVM) models with Shape Signatures descriptors, to the domain of blood–brain barrier (BBB) modeling.
Materials and Methods. The Shape Signatures method is a novel computational tool that was used to generate molecular descriptors utilized with the SVM classification technique with various BBB datasets. For comparison purposes we have created a generalized linear regression model with eight MOE descriptors and these same descriptors were also used to create SVM models.
Results. The generalized regression model was tested on 100 molecules not in the model and resulted in a correlation r^2=0.65. SVM models with MOE descriptors were superior to regression models, while Shape Signatures SVM models were comparable or better than those with MOE descriptors. The best 2D shape signature models had 10-fold cross validation prediction accuracy between 80–83% and leave-20%-out testing prediction accuracy between 80–82% as well as correctly predicting 84% of BBB+ compounds (n= 95) in an external database of drugs.
Conclusions. Our data indicate that Shape Signatures descriptors can be used with SVM and these models may have utility for predicting blood–brain barrier permeation in drug discovery.

KEY WORDS: blood–brain barrier; principal component analysis; regression; shape signatures; support vector machine.

INTRODUCTION

Over the past decade we have witnessed a growing number of studies that have used computational methods to predict absorption, distribution, metabolism and excretion (ADME) properties (1–3). One of the key aspects of ADME profiling is to determine whether a molecule is likely to cross the blood–brain barrier (BBB) which may be desired or not depending on the therapeutic target (4) and traversing it is a major obstacle in drug discovery (5). The BBB is a complex physiological barrier that contains endothelial cells and helps in maintaining brain homeostasis. The BBB also expresses numerous efflux transporters such as P-glycoprotein, (P-gp) (6), multidrug resistance proteins (MRPs) as well as uptake transporters such as the glucose transporter and amino-acid transporters that can also influence whether a drug is absorbed in the brain and central nervous system (CNS). Experimentally testing libraries of compounds for BBB permeation very early on in drug development is essential but is very time consuming and expensive. Hence the development of *in silico* models of BBB penetration has gained considerable interest since the mid 1990s (7). Computational modeling of BBB data is an area of research which has been extensively studied with many techniques. These include the very simplest using a small number of interpretable physicochemical descriptors such as calculated log*P* and polar surface area, to those using large numbers of descriptors and statistical methods including linear regression techniques, neural networks and higher level classification models such as support vector machine (SVM) or other sophisticated machine learning approaches (Supplemental Table I). Several reviews have summarized the state of the art over the years for both *in vitro* (4) and *in silico* approaches to the BBB, including much of the earlier work (1,8,9). Most

Electronic supplementary material The online version of this article (doi:10.1007/s11095-008-9584-5) contains supplementary material, which is available to authorized users.

[1] Department of Pharmacology and Environmental Bioinformatics and Computational Toxicology Center (ebCTC), University of Medicine & Dentistry of New Jersey (UMDNJ)–Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, New Jersey 08854, USA.
[2] Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, Pennsylvania 19046, USA.
[3] Department of Pharmaceutical Sciences, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201, USA.
[4] To whom correspondence should be addressed. (e-mail: ekinssean@yahoo.com)

ABBREVIATIONS: ADME, absorption, distribution, metabolism and excretion; BBB, blood–brain barrier; CNS, central nervous system; MEP, molecular electrostatic potential; MOE, molecular operating environment; PCA, principal component analysis; P-gp, P-glycoprotein; QSAR, quantitative structure activity relationship; RFE, recursive feature elimination; SAS, solvent accessible surface; SVM, support vector machine; TPSA, topological polar surface area; UFS, unsupervised forward selection.

of the datasets used to date are primarily either those from rat or mouse *in vivo* studies with logBB data or using large datasets of drugs or drug-like molecules that are known to be active in the CNS (BBB+) or not active in the CNS (BBB−) of animals or humans. This binary data is also widely used to create classification models. Notwithstanding the fact that much of the BBB data have been accumulated over the years into slightly larger databases (Supplemental Table I) with subsequent mixing of data types, there have been some impressive attempts at model creation and testing (8,9). Our analysis of 32 of these studies, which is comprehensive to date, suggests that 19 of them utilize an external test set, while most perform some form of internal validation (such as leave 'n' out, or leave one out, Supplemental Table I).

The majority of BBB models include some descriptors relating to hydrogen bonding, lipophilicity, molecular size, molecular charge, shape and flexibility and in some cases these have been related as simple rules (8,10). The effect of molecular shape has been rarely assessed with different conclusions (11–14). A new approach called Shape Signatures has recently been proposed that utilizes molecular shape-dependent signatures as the basis for molecular recognition (15). The Shape Signatures method employs a customized ray-tracing algorithm to explore the volume enclosed by the surface of a molecule, then uses the output to construct compact histograms ('Shape Signatures') that encode for molecular shape, polarity, and other biorelevant properties (Fig. 1). The method has been successfully used for a number of drug discovery programs for database similarity searching (15–19) and has several advantages over other approaches including being alignment independent and enabling rapid 3D searching. The goals of the present study were to apply the Shape Signatures approach to the domain of BBB modeling using SVM and compare it to regression models using different test sets and, additionally, to validate the models with a database of FDA approved drugs.

MATERIALS AND METHODS

Data Compilation

The quality of computational models is directly influenced by the quality of the datasets. However, compiling diverse datasets with known experimental logBB values is complex due to different experimental conditions and measurements. Even more difficult is to derive a boundary condition to classify BBB+ and BBB− based on logBB values. Initially we have used the published datasets with our methods (20–25) to either create regression or classification models. We have also compiled meta-databases from this published literature (20–25). The first database was assembled using chemicals with measured values of logBB (20–22,24,25) carefully chosen from these multiple sources (datasets tabulated and summarized in Table I). For each of these datasets, the structures with experimental logBB ≥0 were labeled as BBB+ and those with logBB <0 as BBB−. In addition, since the original datasets contained several identical molecules, it was decided to retain a single copy of a compound in the process of building new databases for regression and classification analysis from different sources. The data for the same compound from different sources were generally comparable.

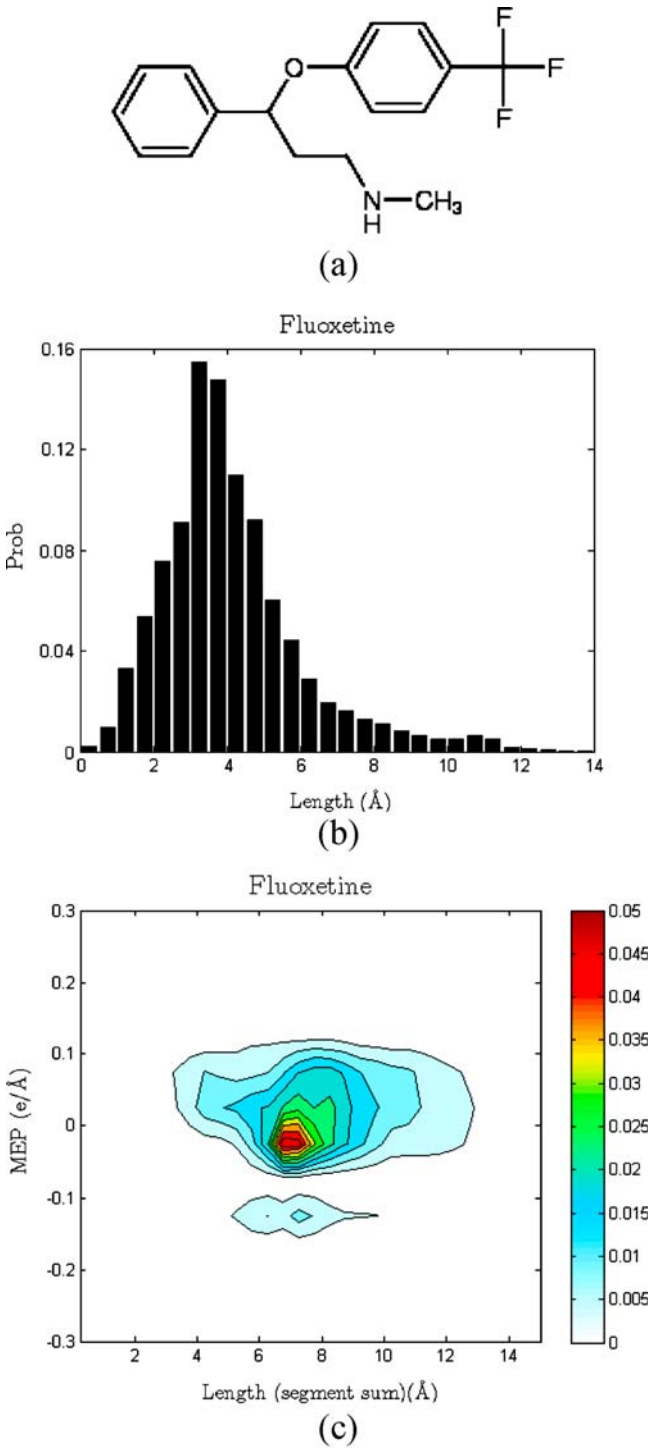

Fig. 1. 1D and 2D Shape Signatures of fluoxetine (BBB+). **a** Chemical structure. **b** 1D (shape only) signature histogram. **c** 2D (shape and polarity) signature plot.

The second database included a single dataset compiled by Li *et al.* (23). These authors assigned molecules with logBB ≥−1 to the class of BBB+ compounds and those with logBB <−1 to BBB−, and for each molecule its class attribute was reported in a binary format (either BBB+ or BBB−). The final database was used for making predictions with the

Table I. Datasets Used for this Study are Listed by the Author's Name along with the Total Number of Compounds

Dataset	Number of compounds	Number of BBB+	Number of BBB−	Reference
Xu-training	78	41	37	(21)
Kitchen-100	100	45	55	(25)
Kitchen-181	181	91	90	(25)
Garg	159	83	76	(20)
KC291	269	155	114	(22)
Liu	61	26	35	(24)
Li	376	250	126	(23)

models and was assembled from a database of the FDA approved drugs derived from the *Clinician's Pocket Drug Reference* (26) (SCUT database) that has been used for several pharmacophore database searching projects (27,28). All of the above databases have been provided as supplemental files.

Molecular Descriptors

The chemical composition of the lipid bilayer imposes certain characteristic features among molecules that have to penetrate through these membranes. Several published models include a variety of descriptors ranging from those that account for hydrophobicity to hydrophilicity, volume and mass (8). No single molecular descriptor has been solely shown to reliably influence the model for drug transport across the BBB. Therefore in this study, we have evaluated the performance of a number of molecular descriptors on their ability to be used to predict logBB values and further to classify the compounds into BBB+ and BBB− based on these values. We have chosen to use a set of molecular descriptors for a simple linear regression model that aims at predicting the logBB values. For the regression model, eight independent molecular descriptors namely logP, TPSA, log*S*, mass, volume, number of rotatable bonds, number of oxygen atoms, and number of nitrogen atoms were chosen. The values for the molecular descriptors were calculated for all the compounds in the three databases using the Molecular Operating Environment (MOE, Chemical Computing Group, Montreal, Canada) quantitative structure activity relationship (QSAR) and modeling program. In addition we have used a shape based descriptor method called "Shape Signatures" for an SVM based classification model for the BBB+ and BBB− set. The performance of both these sets has been validated using published datasets compiled from literature.

Regression Model

A simple linear regression model was developed using the Xu-training dataset (21) that consisted of 78 unique chemicals with continuous logBB data. The regression analysis was performed using routines from the Statistical Toolbox of MATLAB (Version 6). The model was validated using the Kitchen-100 dataset (25) (Table I) that contained 100 unique chemical compounds with continuous logBB data. Further, the regression model was used to predict the logBB values for other published datasets listed in Table I.

Shape Signatures Method

The Shape Signatures method relies on a customized ray-tracing algorithm (15), which explores the volume enclosed by the solvent accessible surface of a molecule. During the first step of the algorithm, the three-dimensional structure of a single lowest energy conformer of the molecule is generated by CORINA (Molecular Networks GmbH, Nägelsbachstr. 25, 91052 Erlangen, Germany. http://www.mol-net.de) and partial charges for each atom are assigned according to the Gasteiger–Marsili scheme (29). The second stage consists of constructing a solvent accessible surface (SAS) around the molecule and generating its triangulated representation by the SMART algorithm (30). In the third step, the ray-tracing process is executed inside the cavity bound by the SAS which encompasses the molecule. The ray of light, emitted initially from a randomly chosen point on the interior lining of the molecular compartment, travels inside the cavity and as it strikes the opposite face, is reflected and propagated in the direction determined by the law of optical reflection. For each reflection point, the value of the truncated Coulomb potential or the molecular electrostatic potential (MEP), and the lengths of the incident and reflected ray segments are recorded. The procedure terminates after 100,000 reflections. According to our previous work (31), this number was found sufficient for the trajectory of the ray to fully explore the entire volume of a typical drug-like molecule. The output is then used to construct two compact one- and two-dimensional histograms ('signatures') that encode molecular shape and polarity respectively (Fig. 1). In particular, all recorded ray segments are binned by their length into a one-dimensional histogram with the predefined bin width of 0.5 Å (Fig. 1b). In addition, a second histogram is also constructed, for the values of MEP (with a step of 0.05e/Å) and the associated total length of the two path segments joined by the reflection point, resulting in a two-dimensional histogram (Fig. 1c). Both the histograms are normalized. Once generated, these histogram based fingerprints ('signatures') can be used to compare any two small molecules. Shape similarity between a pair of molecules is assessed by comparing their 1D signature (Fig. 1b), whereas matching the 2D signatures of the two structures compares their overall molecular shapes and MEP (Fig. 1c). This process is fast and efficient, and it eliminates tedious and subjective atom-based alignment of the molecules.

The Shape Signatures method benefits from its ability to capture the true three-dimensional structure of the molecules. The method has already proven successful for a number of drug discovery programs when used for database similarity searching (15–19,31). Recently, the Shape Signatures method has been extended into the domain of predictive modeling. In particular, it was demonstrated that Shape Signatures can be employed to generate ensembles of three-dimensional molecular descriptors useful for classifying compounds with respect to their experimentally tested activity at the 5-HT_{2B} receptor and the hERG channel (31). It was found that the Shape Signatures based models performed as well as or even better than more traditional classification models with 2D molecular descriptors. The fact that one- and two-dimensional Shape Signatures collectively account for the molecular characteristics of shape and polarity that are key for successful transport across the blood–brain barrier, invites

examination of this methodology as a potential predictor of the blood–brain barrier permeation capability of virtually any drug-like chemical.

Shape Signatures Molecular Descriptors

Following our previous work (31), for each compound in this study, the heights of the bins of the associated 1D and 2D Shape Signatures histograms constituted two sets of distinct molecular descriptors related to this particular structure: the first based exclusively on molecular shape and the second reflecting both molecular shape and polarity. Despite being represented as 1D and 2D histograms, these Shape Signatures fingerprints are inherently three-dimensional molecular descriptors since they encode the 3D conformation and polarity of the molecule.

Support Vector Machine (SVM) Classification Procedure

In recent years, SVM has become a method of choice among different supervised classification methods for a broad variety of binary classification problems. This technique was built on the structural risk minimization principle (32,33), and is now widely recognized for its ability to solve highly non-trivial classification problems (23,31,34–37). The central idea of the method is to project the original descriptor vectors to a higher dimensional feature space where a clearer division between the two classes of data becomes feasible. In such a high-dimensional feature space, a linear SVM routine is applied next to optimally position the separating hyperplane between the instances from the two classes. Minimization of the expected generalization error for the test datasets is achieved by finding a separating hyperplane with the maximal margin. In this work, we used a well tested and freely available program LIBSVM (C-SVM) (38). We utilized the radial basis function kernel, whose parameter γ and the penalty term C was determined in each case via a simple grid search procedure by the 10-fold cross validation.

For every dataset, the associated library of Shape Signatures was generated and prepared for the classification analysis. Several SVM classification models trained on the data from the two BBB databases were applied to predict the blood–brain penetration capabilities of the molecules in the SCUT database. Prior to performing this, the SCUT derived dataset was screened for redundancy against the other two training sets (the full lists of compounds for each data class are available in the Supplementary Table II).

Data Analysis

The prediction power of both the regression model and each SVM model was evaluated by computing the following statistical indicators. The average number of correctly predicted BBB+ compounds in the test set $\langle BBB+\rangle = \langle BBB+_{true}/BBB+_{tot}\rangle$, the average number of correctly predicted BBB− molecules in the test set $\langle BBB-\rangle = \langle BBB-_{true}/BBB-_{tot}\rangle$, and the total prediction accuracy $\langle Q\rangle = \langle (BBB+_{true} + BBB-_{true})/(BBB+_{tot} + BBB-_{tot})\rangle$. These measures are equivalent to the standard statistical indicators: sensitivity (SE), specificity (SP) and overall accuracy (Q), respectively (34). In addition, following our previous study (31) and the work of Ung *et al.* (39), we report the values of Matthew's correlation coefficient (40)

$$C = \frac{\left[BBB+_{true} \times BBB-_{true} - BBB+_{fal} \times BBB-_{fal}\right]}{\left[(BBB+_{tot})(BBB+_{true} + BBB+_{fal})(BBB-_{tot})(BBB-_{true} + BBB-_{fal})\right]^{1/2}} \quad (1)$$

The Matthew's correlation coefficient is another measure of the overall prediction performance. For a perfect classification, when $BBB+_{fal}$ and $BBB-_{fal}$ are both zero, the value of $C=1$, while for a random performance, C would be close to zero since in this case, $BBB+_{true} \approx BBB+_{fal}$ and $BBB-_{true} \approx BBB-_{fal}$. A negative value of C would suggest worse than random performance.

RESULTS

Development of Generalized BBB Regression Models

The linear regression BBB model developed with eight interpretable molecular descriptors, calculated with MOE, is described below:

$$\begin{aligned}\log BB_{(pred)} &= 0.3408 * \log P - 0.0192 * TPSA + 0.2503 \\ &\quad * \text{a\_}nN + 0.1467 * \text{a\_}nO + 0.1069 * \log S \\ &\quad - 0.0011 * mass - 0.0001 * volume \\ &\quad - 0.0602 * \#rot.bonds \end{aligned} \quad (2)$$

Where: a_nN is number of nitrogen atoms, a_nO is number of oxygen atoms, TPSA is topological polar surface area, log$S$ is solubility and log$P$ is a water / octanol partition coefficient and measure of hydrophobicity, and # rot. bonds is the number of rotatable bonds.

The model was built using the data for 78 molecules from the training set of Hou and Xu (21), with an r^2=0.70 (Fig. 2). The model was further validated using the dataset from

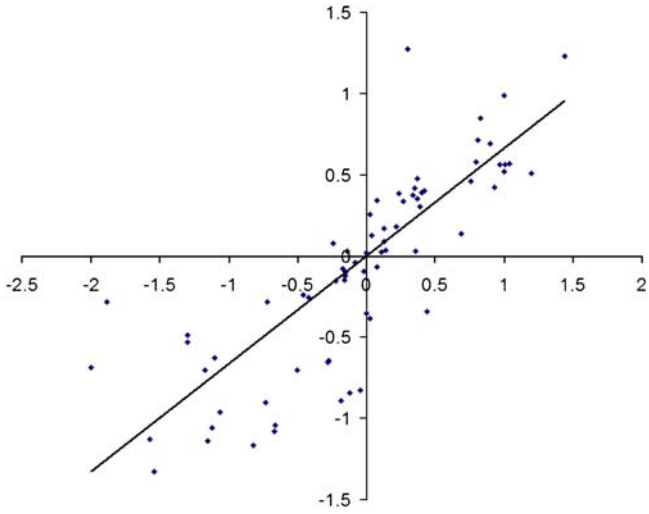

Fig. 2. Correlation of predicted logBB values (*x-axis*) *versus* the experimental logBB values (*y-axis*) for compounds from Xu-training set. The regression equation model resulted in an r^2=0.70.

Table II. Prediction of BBB+ and BBB− Compounds Based on the Eight Molecular Descriptors Implemented in the Generalized Regression Model

Test set	Total no. of compounds	Total prediction	No. of BBB+	No. of BBB+ predicted	No. of BBB−	No. of BBB− predicted	*C* (Matthews correlation coefficient)
Xu-training	78	69 (88%)	41	34 (83%)	37	35 (95%)	0.776
Kitchen-100	100	90 (90%)	45	37 (82%)	55	53 (96%)	0.802
Kitchen-181	181	120 (66%)	91	43 (47%)	90	77 (86%)	0.355
KC291	269	190 (71%)	155	104 (67%)	114	86 (75%)	0.420
Liu	61	57 (93%)	26	24 (92%)	35	33 (94%)	0.866
Li	376	225 (60%)	250	113 (45%)	126	112 (89%)	0.340
Combined	351	250 (71%)	186	122 (66%)	165	128 (78%)	0.433

The prediction accuracy of each group is reported in parenthesis

Kitchen *et al.* (25) on a set of 100 molecules with an r^2=0.65. These results are comparable to the respective test set correlations in earlier publications (r=0.79 [r^2=0.62] (21) and r~0.7 [r^2=0.49] (25)). The generalized regression model described here was also used to predict the BBB permeation of other published molecules (Table II). The overall prediction accuracy ranged from 59% to 93% irrespective of the number of compounds in the set. The values of the Matthews correlation coefficient were greater than zero (*C* value ranged between 0.340 and 0.866), showing that the model performed very well and better than random in all cases. Subsequent classification using the generalized model on the BBB+ and BBB− datasets was also performed. The results from this classification showed that the model performed well for the BBB− datasets with a classification rate between 75% and 96%. However, the results from the BBB+ sets were moderate, between 45% and 92%.

In order to further understand these results, we performed a Principal Component Analysis (PCA) of the datasets based on the eight molecular descriptors, with reference to the molecules from the Xu training set. PCA is a useful tool in exploratory data analysis. Principal components (PC) are linear combinations of the original variables constructed and organized in such a way so that the first principal component PC1 attempts to maximally explain the variance in the data. Geometrically, it defines the direction in which the data is maximally spread. The next PC2 is orthogonal to PC1, and tries to maximally explain the residual variance not explained by PC1. PC3, which is now orthogonal to both PC1 and PC2, in turn is set to maximally explain the variance not explained by the first two principal components, and so forth. Based on the PCA comparison, it is clear that there is a partial overlap of the chemical space covered by the Xu set used to initially derive the regression model (Fig. 3b). The PCA analysis shows a number of the BBB+ molecules outside the area covered by the Xu set, whereas the BBB− molecules seem to be generally closer to these initial training set molecules (Fig. 3b). Although this PCA analysis is rather qualitative, it provides confirmation that predictions outside the chemical space of a model could be unreliable. The poor performance of some of the BBB+ molecules could also be due to the molecules' larger size and higher flexibility that might well influence the logBB values in the regression model.

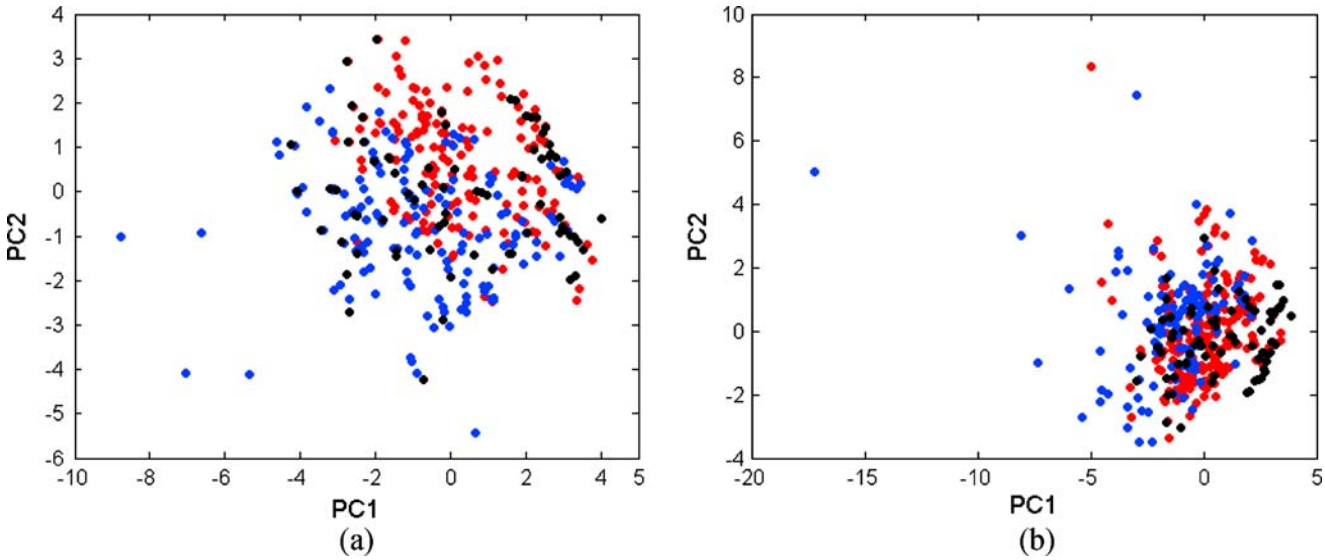

Fig. 3. **a** Results of the PCA analysis on the Xu–Combined dataset conducted in the space of eight molecular descriptors computed with MOE (PC1=54%, PC2=26%, PC3=10%). **b** Results of the PCA analysis performed on the Xu–Li dataset conducted in the space of eight molecular descriptors computed with MOE (PC1=52%, PC2=27%, PC3=11%). *Black circles*: molecules from Xu's dataset. *Red circles*: BBB+ compounds from Combined (a) and Li's (b) datasets. *Blue circles*: BBB− compounds from Combined (**a**) and Li's (**b**) datasets.

Generalized BBB Regression Models Used to Predict the SCUT Database

The performance of the generalized regression model was further assessed by predicting the BBB permeability of molecules from the SCUT database of FDA approved drugs. After removing those molecules in the model training set, the compounds were first classified based on their functionality (knowledge based method) as possible BBB+ and BBB− categories (for example an antidepressant would be categorized as BBB+ while an antihypertensive drug would be classified as BBB−). Similarly, for the rule based classification, using the five simple rules namely, (a) if $\sum$(N + O) atoms ≤ 5, (b) $C\log P - (N+O) > 0$, (c) PSA < 60–90 Å^2, (d) mass ≤ 450 and (e) $1 \leq \log D \leq 3$ (8,9) the SCUT database of compounds were classified into BBB+ and BBB− molecules. Only molecules that strictly obeyed all the five rules were categorized as BBB+ (74 molecules) and the remaining compounds in the SCUT database were classified as BBB− (315 molecules). Of the 389 total molecules, the knowledge based scheme found 95 compounds to be BBB+ and the rest of the compounds (293) to be BBB−. Further, the generalized regression model was applied to classify the SCUT database of compounds (Supplemental Table II). The logBB values were predicted using the generalized regression model and a cutoff of logBB=0 was used to classify the compounds into BBB+ and BBB− categories. The model performed with an overall accuracy of 77% and a correct prediction rate of 88% for BBB− and 45% for BBB+ molecules, when compared to the knowledge based classification of the molecules based on the known therapeutic indications. However, these results could be severely biased due to the nature of the compound classification.

SVM Classification Models for BBB

The datasets used to generate a number of the SVM models presented in this study are detailed in Table I. In the process of constructing the Shape Signatures histograms for each molecule from the aforementioned datasets, it was observed that 2D Shape Signatures normally included several hundred non-zero bins/descriptors and the resulting data matrix usually had a high degree of redundancy. Therefore, based on our previous experience (31), before building the SVM models we reduced the dimensionality of the original data matrices using the unsupervised forward selection (UFS) method of Livingstone and co-workers (41). The UFS routine was designed specifically to eliminate redundancy and decrease multicollinearity of the input data, and has been demonstrated to be useful for a number of QSAR (41) and SVM classification (41) studies. In each case, the output data matrix contained less than 100 data columns. For each dataset in Table III, two types of the SVM models were built and validated. The first set included models resulting from a straightforward 10-fold cross validation conducted on the entire datasets. The prediction accuracy of these models were assessed first by calculating the overall accuracy rates Q_{cross}, which show the average fractions of correctly predicted molecules (combined BBB+ and BBB−) from the test sets. Second, the classification models produced in a series of leave-20%-out SVM runs were assessed as follows. For each Shape Signatures database, approximately 20% of the compounds from the database were randomly selected and assigned to the hold-out test set while the remainder of the data (~80%) constituted the training set. The selection was carried out to approximately preserve the correct proportion of BBB+ and BBB− chemicals in both sets. Each SVM classification model was then generated with the training set and applied to predict class attributes of the compounds in the test set. Next, a set of statistical indicators of prediction accuracy were computed and stored. This procedure was repeated 100 times, each time with a different composition of the test and training sets. For each model, the reported final statistical measures were averaged over the number of repetitions. The predictive power of each SVM model in this group was evaluated by computing the statistical indicators such as the average Q value and the Matthews correlation coefficient C (Table III). It was found that both models performed similarly in terms of 10-fold cross validation prediction accuracy 80–83%, leave-20%-out testing prediction accuracy 80–82% and C values 0.53–0.63.

For comparison, we have also used the same eight MOE descriptors derived in the generalized regression model (described above), to generate SVM models with the Li and combined datasets. These generally performed comparably well although with lower Matthew's correlations than observed with Shape Signatures descriptors (Table III). Perhaps equally interesting is the comparison between the regression model (Table II) and the SVM model (Table III) using the same MOE descriptors. This analysis reveals that the SVM models produce a dramatic improvement in the predictions for the BBB+, BBB− and Matthews correlation, especially for the combined dataset.

Table III. SVM Classification of BBB+ and BBB− Compounds from Combined (351 Molecules) and Li (378 Molecules) Datasets

Dataset	Molecular descriptors	10-fold cross validation[b] (%)	Leave-20%-out testing[c] ⟨BBB+⟩ (%)	⟨BBB−⟩ (%)	⟨Q⟩ (%)	C
Combined	2D (shape + charges) Shape Signatures	83	84	79	82	0.635
Combined	MOE[a]	80	80	79	80	0.595
Li	2D (shape + charges) Shape Signatures	80	89	62	80	0.533
Li	MOE[a]	80	89	51	76	0.435

[a] Eight molecular descriptors computed with MOE also used in the regression equation: b_rotN, Weight, a_nN, a_nO, logs, TPSA, Vol and log$P$ (o/w)
[b] This column lists prediction accuracies Q estimated from 10-fold cross validations performed on the entire dataset
[c] The tabulated values of ⟨BBB+⟩, ⟨BBB−⟩, ⟨Q⟩ and C were averaged over the results of 100 different hold-out test set experiments

Table IV. Results of SVM Classifications Based on 1D (Shape Only) and 2D (Shape + Charges) Shape Signatures Molecular Descriptors

Dataset	Q_{cross}[a] (%)	
	1D Shape Signatures (shape only)	2D Shape Signatures (shape + charges)
Combined (continuous)	77	83
Li (discrete)	73	80

[a] For each dataset, Q_{cross} was estimated from 10-fold cross validations performed on the entire dataset

Comparison between SVM Classification Results Based on 1D and 2D Shape Signatures

Based on the Q_{cross} values we have demonstrated that the classification models based on 2D Shape Signatures descriptors (Q_{cross} 80–83%) which encode for molecular shape and polarity, performed slightly better than those constructed using 1D Shape Signatures descriptors (Q_{cross} 73–79%) which account exclusively for molecular size and shape (Table IV). This would also be expected based on past studies of molecular requirements for BBB penetration. Due to the unique physicochemical structure of the blood–brain membrane, for a molecule to penetrate the BBB the right balance between the molecular shape and distribution of atomic charges is required. Hence, the models that take into account both of these properties are expected to be generally more accurate.

Classification of the SCUT Database Using Shape Signature Based SVM Models

Finally, we attempted to classify molecules from the reduced SCUT database (27,28) (389 structures) using the Shape Signatures SVM models described in the previous sections (Supplemental Table II). As was noted before, we ensured this dataset did not contain structures present in either of the training sets and represents an application of the models to a group of molecules of medical importance. It should be considered that the experimental logBB values for many of these structures have not been documented so far in the literature, therefore the reported predictions for these compounds using the generalized regression model, the rule based model and SVM classifications model is the first effort to classify the SCUT database compounds as BBB+ or BBB− chemicals. Using our knowledge of the therapeutic targets and reported side effects of these molecules, we were able to ascertain the likely BBB+ or BBB− nature of the molecules. But as mentioned above, it is certainly possible that BBB− may be misclassified. Prior to using the classification models we assessed the chemical space covered by the structures from the training and test (SCUT) sets to evaluate whether they overlapped. As described above, we subjected the utilized molecular descriptors from both sets to PCA using the 2D Shape Signatures descriptors for the two mixed datasets, namely Combined-SCUT (Fig. 4a) and Li-SCUT (Fig. 4b). The PCA analysis shows that ~80% of the variance is explained in the space of the first three principal components and there is a significant overlap between the regions of chemical space occupied by the molecules from these three datasets using these descriptors. When analyzing the SVM results for the SCUT database we also need to consider that for the Combined and Li datasets the dividing boundaries between BBB+ and BBB− were set differently. For the Combined dataset (logBB=0) and for the Li dataset (logBB=−1). Based on the SVM models, the 10-fold cross validation models performed better in predicting the BBB+ category of compounds using either the combined or the Li datasets for training. However, for the prediction of the BBB− category, the leave-20%-out models performed marginally better than the 10-fold cross validation for both training sets.

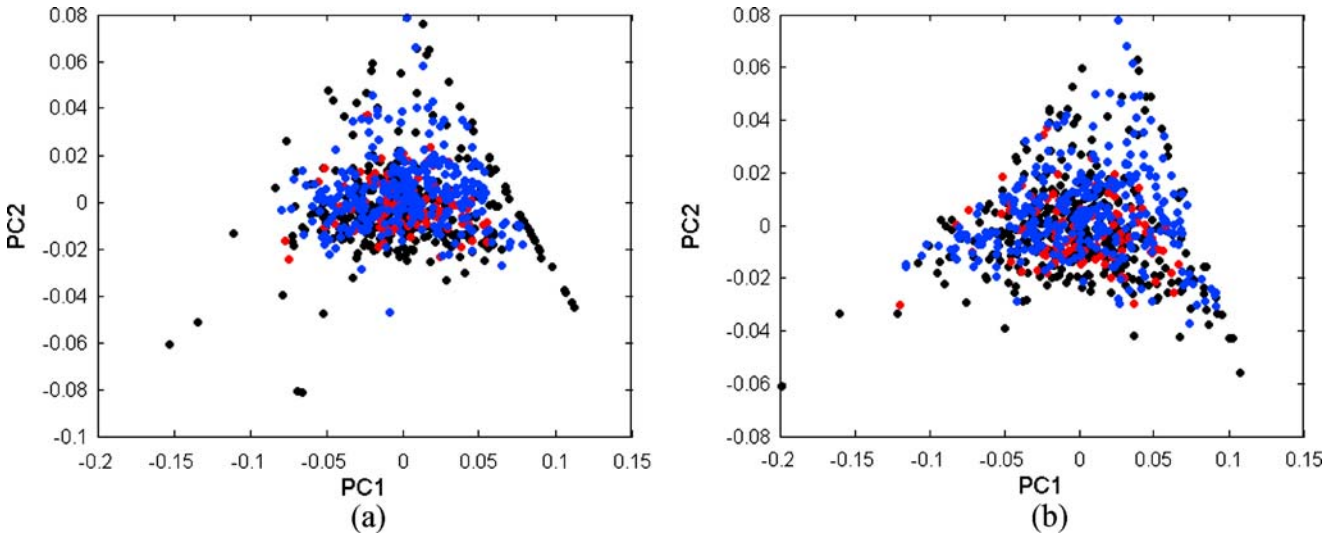

Fig. 4. Results of the PCA analysis conducted in the space of 2D Shape Signatures (shape + charges) molecular descriptors on the Combined–SCUT and Li–SCUT datasets. **a** PC1 *vs* PC2 for the Combined SCUT dataset (PC1=55%, PC2=12%, PC3=9%). *Black circles*: 351 compounds from Combined dataset. *Red circles*: 95 BBB+ compounds from SCUT. *Blue circles*: 294 BBB− compounds from SCUT. **b** PC1 *vs* PC2 for the Li-SCUT dataset (PC1=63%, PC2=11%, PC3=8%). *Black circles*: 378 compounds from Li dataset. *Red circles*: 95 BBB+ compounds from SCUT. *Blue circles*: 294 BBB− compounds from SCUT.

Table V. Predictions of BBB Permeation for Molecules from the SCUT Database of Known Drugs with BBB Permeation Classified Based on Known Therapeutic Use

			SVM models				
Knowledge based model	Regression model	Rule-based model	10-fold CV (Combined dataset)	Leave-20%-out (Combined dataset)	10-fold CV (Li dataset)	Leave-20%-out (Li dataset)	Consensus model
BBB+ (95 cmpds)	39/95 (41%)	32/95 (34%)	57/95 (60%)	51/95 (54%)	80/95 (84%)	76/95 (80%)	53/95 (56%)
BBB− (294 cmpds)	262/294 (89%)	253/294 (86%)	193/294 (67%)	204/294 (69%)	133/294 (45%)	149/294 (51%)	204/294 (69%)
Total (389 cmpds)	301/389 (77%)	285/389 (73%)	250/389 (64%)	255/389 (66%)	213/389 (55%)	225/389 (58%)	257/389 (66%)

Consensus Prediction for SCUT Database

Finally, a consensus 'model' was built based on the six different models (described above) for prediction of BBB permeation of the SCUT database. The results from the consensus model are described in Table V. In order to arrive at a consensus, all the models were assessed with equal weight and a decision was made based on a majority vote (4/6) (Supplemental Table II). Based on the consensus model, 53 of the 95 compounds (56%) were correctly categorized as BBB+ and 204 of 295 compounds (69%) were correctly categorized as BBB−. In all, 257 out of 389 (66%) compounds were correctly classified for BBB permeation in comparison with the knowledge based classification scheme.

DISCUSSION

The use of regression based BBB models was first proposed by Van de Waterbeemd and Kansy (42) followed by many others with varying molecular descriptors (Supplemental Table I). However, all these models perform well with their respective small training and test datasets and generally fail (or are less predictive) when tested against other more diverse datasets. This could be due to the fact that the compounds belong to a different region of chemical space and the models have an inherent descriptor independent value that is tuned to span only the chemical space of the original test sets (8). In order to overcome the inherent disadvantages of regression models, we first propose a simple generalized regression model (see Eq. (2)) that has been built based only on the values of eight standard molecular descriptors with no added constants.

The choice of the molecular descriptors was based on a few simple rules derived from the physiological features governing cellular permeability. These are:

(a) Inclusion of hydropathic descriptors that span both the hydrophilic and hydrophobic nature of bilayers (log*P*, TPSA, a_nN and a_nO in Eq. (2)) (43,44).
(b) Inclusion of log*S* based on the hypothesis that water soluble compounds have a high probability of passing across the BBB (45,46).
(c) Molecular weight and size of the compound should be a rate limiting factor for BBB permeation (47).
(d) A high level of flexibility (large number of rotatable bonds) of the compound should be a deterrent for BBB permeation (44).

The influence of each of the above molecular descriptors has been validated in previous regression models (22,24,25,43,48–50). However, we have assessed the total effect of all these descriptors in our generalized regression model. When used to predict an external set of 100 molecules the correlation was very comparable to those described for other more sophisticated models (10). Moreover, extensive testing of the equation with different datasets (Table II) suggested that the prediction accuracy would indicate the regression equation is generalizable. However this model performed less well with BBB+ molecules which may be due to the chemical space these represent (with the eight descriptors used) in the test sets compared to the training set. These results were in sharp contrast to the general prediction trend for BBB− molecules (including the SVM based models reported here), since they tend to be biased towards BBB+ molecules. This generalized model was further applied to classify compounds from an independent dataset of FDA approved drugs (SCUT database). The regression model performed better than the rule based model for both the BBB+ and BBB− categories (Table V), although again the prediction rate was better for the BBB− category. Overall, the regression model could correctly predict 77% of the compounds from the SCUT database for BBB permeation. The predictions with the regression model were also slightly better than using the simple rule base model (73% correct overall), which also only predicted 34% of BBB+ molecules. This would suggest the additional value of using regression or SVM methods which perform far better at predicting this class.

In order to understand the effect of molecular shape and size in more detail we further classified the datasets using a more sophisticated statistical method namely SVM, using the shape signature based descriptors. We found that 2D Shape signature descriptors slightly outperformed 1D Shape descriptors with the SVM algorithm. Additionally Shape Signatures also performed slightly better than SVM models developed with the MOE descriptors used in the regression model (as we have shown previously with other datasets (31)). SVM models with these eight descriptors were also superior to the regression models at classification of the two datasets. Using either 10-fold cross validation or leave-20%-out testing the Shape Signatures SVM models had greater than 80% prediction accuracies. Li *et al.* (23) reported a number of classification studies using a range of classifiers from logistic regression to SVM. Two types of SVM procedures were

presented which differ in the way the set of molecular 1D to 3D molecular descriptors were selected from the original pool of 199 (41). The first group of SVM models used all 199 descriptors while the second set utilized the advanced recursive feature elimination (RFE) program. This procedure selects the most informative subset of molecular descriptors. Upon comparing our results (Table III) with the predictions of Li *et al.*, we note the following. Our SVM models based on 2D Shape Signatures molecular descriptors (shape + charges) perform at the same level as their SVM classifications when used without the RFE feature selection (SE=89.9%, SP=64.3%, Q=79.1% and C=0.52). This observation certainly validates the applicability of the Shape Signatures derived molecular descriptors for predicting BBB permeation capability. However, according to Li *et al.* the best performing SVM model is the RFE–SVM approach which provided slightly better results on average SE=88.6%, SP=75.0%, Q=83.7% and C=0.64 than the less advanced UFS data reduction scheme which we have used.

Both the generalized regression model and the Shape Signatures SVM models were used to classify the FDA approved small molecule drugs from the SCUT database. The shape signature descriptor space for this set of molecules was compared to the SVM model training set and found to overlap closely (Fig. 4), providing some confidence in the applicability or domain of this model to this particular test set. The performance of the rule based and the regression models for the BBB+ category was low as opposed to the BBB− category, which had an average success rate of ~88% (Table V). The results from the SVM prediction was opposite to the regression and rule based methods, with the predictions in the BBB+ category faring better than the BBB− category. If we were only interested in the BBB+ compound prediction accuracy (which we have the most confidence in as these molecules are known to be centrally active based on their therapeutic use, enabling us to create the knowledge based model), the 10-fold CV (Li dataset) model performs very well with 84% correct predictions for the 95 molecules.

A consensus model was built for prediction of the SCUT database classifications. The results from the equally weighted consensus model show that 56% of the BBB+ and 69% of the BBB− category of compounds could be predicted correctly. These results essentially average the predictions across the models and do not improve upon the individual models as has been noted before (25), however we could envisage the use of more sophisticated scoring or weighting schemes (or the use of the leave-20%-out Li dataset SVM model and the regression model alone) to predict BBB+ and BBB−, respectively.

An objective of our research was to examine the quality of a novel set of molecular descriptors derived from molecular Shape Signatures (15–19). These descriptors are inherently three-dimensional and a relatively new addition to the other 2D/3D descriptor collections used in predictive QSAR modeling (51,52). We have now extended the Shape Signatures methodology to molecular classifiers for a physicochemical property, namely BBB penetration. Given the simplicity and physical transparency of the Shape Signatures representation, our results described herein are encouraging for the applicability of this method. The Shape Signatures method is capable of encoding of these main features in a compact and practical form, which underlies the versatility of its usage. Because the procedure does not require either a direct 3D molecular alignment or grid generation, the algorithm is also relatively fast and efficient. Models based on Shape Signatures histograms can accommodate various chemical compositions. Due to the universal character of the Shape Signatures histograms, once generated they can be used for a variety of tasks which require molecular recognition and at the molecular level no model refitting is necessary in going from one problem to another.

In summary, the present study suggests new approaches for assigning drugs to BBB classifications using (either in combination or alone) a generalized regression equation with MOE descriptors or SVM models using the novel Shape Signatures descriptors. These models may be valuable for providing predictions of BBB permeability that may overcome some of the limitations of previous models in terms of their generalizability and the chemical space covered.

ACKNOWLEDGMENTS

Support for this work has been provided by the USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under STAR Grant number GAD R 832721-010. WJW gratefully acknowledges support for this work provided by the Defense Threat Reduction Agency, under contract number HDTRA-BB07TAS020. This work was also funded in part by NIH R21-GM081394 from the National Institute of General Medical Sciences and by NIH Integrated Advanced Information Management Systems (IAIMS) Grant # 2G08LM06230-03A1 from the National Library of Medicine. This work has not been reviewed by and does not represent the opinions of the funding agencies. The authors are sincerely grateful to Randy Zauhar, Ph.D., of the University of the Sciences in Philadelphia, for useful discussions on technical aspects of Shape Signatures.

REFERENCES

1. S. Ekins, C. L. Waller, P. W. Swaan, G. Cruciani, S. A. Wrighton, and J. H. Wikel. Progress in predicting human ADME parameters *in silico*. *J. Pharmacol. Toxicol. Methods* **44**:251–272 (2000).
2. H. van de Waterbeemd, and E. Gifford. ADMET *in silico* modelling: towards prediction paradise? *Nat. Rev.* **2**:192–204 (2003).
3. S. Ekins, and P. W. Swaan. Computational models for enzymes, transporters, channels and receptors relevant to ADME/TOX. *Rev. Comp. Chem.* **20**:333–415 (2004).
4. R. Cecchelli, V. Berezowski, S. Lundquist, M. Culot, M. Renftel, M. P. Dehouck, and L. Fenart. Modelling of the blood–brain barrier in drug discovery and development. *Nat. Rev.* **6**:650–661 (2007).
5. A. George. The design and molecular modeling of CNS drugs. *Curr. Opin. Drug. Disc. Dev.* **2**:286–292 (1999).
6. K. M. Mahar Doan, J. E. Humphreys, L. O. Webster, S. A. Wring, L. J. Shampine, C. J. Serabjit-Singh, K. K. Adkison, and J. W. Polli. Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs. *J. Pharmacol. Exp. Ther.* **303**:1029–1037 (2002).
7. F. Lombardo, J. F. Blake, and W. J. Curatolo. Computation of brain–blood partitioning of organic solutes via free energy calculations. *J. Med. Chem.* **39**:4750–4755 (1996).
8. U. Norinder, and M. Haeberlein. Computational approaches to the prediction of the blood–brain distribution. *Adv. Drug Del. Rev.* **54**:291–313 (2002).

9. D. E. Clark. *In silico* prediction of blood–brain barrier permeation. *Drug Discov. Today.* **8**:927–933 (2003).
10. J. T. Goodwin, and D. E. Clark. *In silico* predictions of blood–brain barrier penetration: considerations to "keep in mind". *J. Pharmacol. Exp. Ther.* **315**:477–483 (2005).
11. M. Iyer, R. Mishru, Y. Han, and A. J. Hopfinger. Predicting blood–brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm. Res*. **19**:1611–1621 (2002).
12. M. Iyer, E. J. Reschly, and M. D. Krasowski. Functional evolution of the pregnane X receptor. *Expert Opin. Drug Metab. Toxicol*. **2**:381–397 (2006).
13. M. Lobell, L. Molnar, and G. M. Keseru. Recent advances in the prediction of blood–brain partitioning from molecular structure. *J. Pharm. Sci*. **92**:360–370 (2003).
14. F. Ooms, P. Weber, P. A. Carrupt, and B. Testa. A simple model to predict blood–brain barrier permeation from 3D molecular fields. *Biochim. Biophys. Acta*. **1587**:118–125 (2002).
15. R. J. Zauhar, G. Moyna, L. Tian, Z. Li, and W. J. Welsh. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem*. **46**:5674–5690 (2003).
16. K. Nagarajan, R. Zauhar, and W. J. Welsh. Enrichment of ligands for the serotonin receptor using the Shape Signatures approach. *J. Chem. Inf. Model.* **45**:49–57 (2005).
17. C. Y. Wang, N. Ai, S. Arora, E. Erenrich, K. Nagarajan, R. Zauhar, D. Young, and W. J. Welsh. Identification of previously unrecognized antiestrogenic chemicals using a novel virtual screening approach. *Chem. Res. Toxicol*. **19**:1595–1601 (2006).
18. S. Kortagere, and W. J. Welsh. Development and application of hybrid structure based method for efficient screening of ligands binding to G-protein coupled receptors. *J. Comput-Aided Mol. Des*. **20**:789–802 (2006).
19. P. J. Meek, Z. Liu, L. Tian, C. Y. Wang, W. J. Welsh, and R. J. Zauhar. Shape signatures: speeding up computer aided drug discovery. *Drug. Discov. Today* **11**:895–904 (2006).
20. P. Garg, and J. Verma. *In silico* prediction of blood brain barrier permeability: an artificial neural network model. *J. Chem. Inf. Model* **46**:289–297 (2006).
21. T. J. Hou, and X. J. Xu. ADME evaluation in drug discovery. 3. Modeling blood–brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci*. **43**:2137–2152 (2003).
22. D. A. Konovalov, D. Coomans, E. Deconinck, and Y. V. Heyden. Benchmarking of QSAR models for blood–brain barrier permeation. *J. Chem. Inf. Model* **47**:1648–1656 (2007).
23. H. Li, C. W. Yap, C. Y. Ung, Y. Xue, Z. W. Cao, and Y. Z. Chen. Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model* **45**:1376–1384 (2005).
24. R. Liu, H. Sun, and S. S. So. Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery. 2. Blood–brain barrier penetration. *J. Chem. Inf. Comput. Sci*. **41**:1623–1632 (2001).
25. G. Subramanian, and D. B. Kitchen. Computational models to predict blood–brain barrier permeation and CNS activity. *J. Comput-Aided Mol. Des*. **17**:643–664 (2003).
26. L. Gomella, and S. Haist. *Clinician's pocket drug reference*. McGraw-Hill, New York, 2004.
27. C. Chang, P. M. Bahadduri, J. E. Polli, P. W. Swaan, and S. Ekins. Rapid identification of P-glycoprotein substrates and inhibitors. *Drug Metab. Dispos.* **34**:1976–1984 (2006).
28. S. Ekins, J. S. Johnston, P. Bahadduri, V. M. D'Souzza, A. Ray, C. Chang, and P. W. Swaan. *In vitro* and pharmacophore based discovery of novel hPEPT1 inhibitors. *Pharm. Res*. **22**:512–517 (2005).
29. J. Gasteiger, and M. Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron.* **36**:3219–3228 (1980).
30. R. J. Zauhar. SMART: a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput-Aided Mol. Des*. **9**:149–159 (1995).
31. D. S. Chekmarev, V. Kholodovych, K. V. Balakin, Y. Ivanenkov, S. Ekins, and W. J. Welsh. Shape signatures: new descriptors for predicting cardiotoxicity *in silico*. *Chem. Res. Toxicol*., in press (2008).
32. C. Cortes, and V. Vapnik. Support vector networks. *Mach. Learn*. **20**:273–293 (1995).
33. V. Vapnik. *Statistical learning theory*. Wiley, New York, 1998.
34. A. H. Fielding. *Cluster and classification techniques for the biosciences*. Cambridge University Press, New York, 2007.
35. D. Plewczynski, S. A. Spieser, and U. Koch. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model* **46**:1098–1106 (2006).
36. M. Tobita, T. Nishikawa, and R. Nagashima. A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. *Bioorg. Med. Chem. Lett.* 15:2886–2890 (2005).
37. Y. Xue, C.W. Yap, L. Z. Sun, Z. W. Cao, J. F. Wang, and Y. Z. Chen. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **44**:1497–1505 (2004).
38. C. C. Chang, and C. J. Lin. LIBSVM: A library for support vector machines, 2001.
39. C. Y. Ung, H. Li, C. W. Yap, and Y. Z. Chen. *In silico* prediction of pregnane X receptor activators by machine learning approaches. *Mol. Pharmacol.* **71**:158–168 (2007).
40. B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **405**:442–451 (1975).
41. D. C. Whitley, M. G. Ford, and D. J. Livingstone. Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **40**:1160–1168 (2000).
42. H. Van de Waterbeemd, and M. Kansy. Hydrogen-bonding capacity and brain penetration. *Chimia.* **46**:5 (1992).
43. M. H. Abraham, H. S. Chadha, and R. C. Mitchell. Hydrogen-bonding. Part 36. Determination of blood brain distribution using octanol–water partition coefficients. *Drug Des. Discov.* **13**:123–131 (1995).
44. D. E. Clark. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration. *J. Pharm. Sci.* **88**:815–821 (1999).
45. W. L. Jorgensen, and E. M. Duffy. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **10**:1155–1158 (2000).
46. U. Norinder, P. Sjoberg, and T. Osterberg. Theoretical calculation and prediction of brain–blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. *J. Pharm. Sci.* **87**:952–959 (1998).
47. R. M. M. Kaliszan. Brain/blood distribution described by a combination of partition coefficient and molecular mass. *Int. J. Pharm.* **145**:8 (1996).
48. X. C. Fu, C. X. Chen, W. Q. Liang, and Q. S. Yu. Predicting blood–brain barrier penetration of drugs by polar molecular surface area and molecular volume. *Acta Pharmacol. Sin.* **22**:663–668 (2001).
49. H. Sun. A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci.* **44**:748–757 (2004).
50. S. Van Damme, W. Langenaeker, and P. Bultinck. Prediction of blood–brain partitioning: a model based on ab initio calculated quantum chemical descriptors. *J. Mol. Graph. Model.*, in press (2007).
51. S. Ekins, M. J. Embrechts, C. M. Breneman, K. Jim, and J.-P. Wery. Novel applications of Kernel-partial least squares to modeling a comprehensive array of properties for drug discovery. In S. Ekins (ed.), *Computational toxicology: risk assessment for pharmaceutical and environmental chemicals*, Wiley, Hoboken, 2007, pp. 403–432.
52. R. Todeschini, and V. Consonni. *Handbook of molecular descriptors*. Wiley, Weinheim, 2000.