Effect of Selection of Molecular Descriptors on the Prediction of Blood−Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods

Hu Li,[†] Chun Wei Yap,[†] Choong Yong Ung,[†] Ying Xue,[†,‡] Zhi Wei Cao,[§] and Yu Zong Chen*[,†,§]

Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, College of Chemistry, Sichuan University, Chengdu 610064, P. R. China, and Shanghai Center for Bioinformation Technology, Shanghai 201203, P. R. China

Received April 14, 2005

The ability or inability of a drug to penetrate into the brain is a key consideration in drug design. Drugs for treating central nervous system (CNS) disorders need to be able to penetrate the blood−brain barrier (BBB). BBB nonpenetration is desirable for non-CNS-targeting drugs to minimize potential CNS-related side effects. Computational methods have been employed for the prediction of BBB-penetrating (BBB+) and -nonpenetrating (BBB−) agents at impressive accuracies of 75∼92% and 60∼80%, respectively. However, the majority of these studies give a substantially lower BBB− accuracy, and thus overall accuracy, than the BBB+ accuracy. This work examined whether proper selection of molecular descriptors can improve both the BBB− and the overall accuracies of statistical learning methods. The methods tested include logistic regression, linear discriminate analysis, *k* nearest neighbor, C4.5 decision tree, probabilistic neural network, and support vector machine. Molecular descriptors were selected by using a feature selection method, recursive feature elimination (RFE). Results by using 415 BBB+ and BBB− agents show that RFE substantially improves both the BBB− and the overall accuracy for all of the methods studied. This suggests that statistical learning methods combined with proper feature selection is potentially useful for facilitating a more balanced and improved prediction of BBB+ and BBB− agents.

INTRODUCTION

Optimization of the pharmacokinetic as well as the pharmacodynamic properties of a drug candidate is an important consideration in the drug design process.[1] Good pharmacokinetic properties are required to achieve sufficient drug concentration at a target site while possibly limiting its distribution elsewhere to reduce potential side effects.[2] One important pharmacokinetic property of a drug is its ability or inability to penetrate the blood−brain barrier (BBB).[3] BBB penetration is important for drugs that target receptors in the brain. Examples of these drugs are antipsychotics, antiepileptics, and antidepressants.[4] For drugs not directed at targets in the brain, BBB penetration is undesirable as it would lead to unwanted CNS-related side effects.

A variety of experimental techniques have been employed for BBB-penetration study and screening.[5] Moreover, computational methods have been introduced as potential prescreening tools with the aim to reduce the cost and enhance the speed of BBB-penetration analysis.[3] The most widely explored computational methods are regression methods[6−18] and statistical learning methods.[19−24] As shown in Table 1, statistical learning methods achieve a high prediction accuracy of 75∼92% for BBB-penetrating (BBB+) agents and 60∼80% for BBB nonpenetrating (BBB−) agents. The corresponding overall accuracy is in the range of 71∼81%, which is substantially lower than that of the BBB+ accuracy as a result of the lower BBB− accuracy. A more balance prediction with an improved BBB− accuracy is needed for facilitating the prediction of BBB− agents as well as for reducing the false-positive rate of the prediction of BBB+ agents.

Table 1. Prediction Accuracies of BBB-Penetrating (BBB+) and -Nonpenetrating Agents (BBB−) from Different Studies Reported in the Literature[a]

study (reference)	methods[b]	no. of agents	BBB^+ accuracy (%)	BBB^- accuracy (%)	overall accuracy *Q* (%)
Ajay et al.[21]	BNN	275	92.0	71.0	81.8
Crivori et al.[19]	PCA	120	90.0	65.0	71.7
Cruciani et al.[20]	PCA	35	>75.0	>75.0	>75%
Trotter et al.[23]	SVM	304	78.9	60.4	76.0
Doniger et al.[22]	NN	324	81.5	69.9	75.7
Doniger et al.[22]	SVM	324	82.7	80.2	81.5
this work	LR	415	83.9	46.4	71.0
	LDA	415	78.2	58.3	71.2
	C4.5 DT	415	80.3	62.8	74.3
	k-NN	415	85.5	61.4	77.1
	PNN	415	84.3	62.1	76.5
	SVM	415	88.6	75.0	83.7

[a] It is cautioned that a direct comparison of these results may not be appropriate because of the use of different sets of agents, molecular descriptors, classification methods and parameters, and methods for generating testing sets. [b] BNN (Bayesian neural network), PCA (principal component analysis), NN (neural network), LR (linear regression), LDA (linear discriminate analysis), C4.5 DT (C4.5 decision tree), PNN (probabilistic neural network), *k*-NN (k-nearest neighbors), SVM (support vector machine).

* Corresponding author tel.: 65−6874−6877, fax: 65−6774−6756, e-mail: csccyz@nus.edu.sg.
[†] National University of Singapore.
[‡] Sichuan University.
[§] Shanghai Center for Bioinformation Technology.

10.1021/ci050135u CCC: $30.25
Published on Web 08/10/2005

In earlier studies of the prediction of agents of different pharmacokinetic and toxicological properties by using a statistical learning method, it had been found that the selection of appropriate molecular descriptors can give a substantially more balanced prediction accuracy and enhance the overall accuracy.[25,26] For instance, the prediction accuracies for human intestine absorption and nonabsorption agents are improved from 83.4 and 63.2% to 90.0 and 80.7%, and those for torsades-de-pointes-causing and -noncausing agents are improved from 54.5 and 90.6% to 66.8 and 89.3%, respectively. It is, therefore, of interest to examine whether the proper selection of molecular descriptors can also improve the BBB− accuracy and the overall BBB+/BBB− accuracy of statistical learning methods.

Several statistical learning methods were tested in this work, which include logistic regression (LR), linear discriminate analysis (LDA), *k* nearest neighbor (*k*-NN), C4.5 decision tree (DT), probabilistic neural network (PNN), and support vector machine (SVM). A total of 415 BBB+ and BBB− agents reported from the literature were collected and used to test these methods. A widely used feature selection method, recursive feature elimination (RFE),[27,28] was used for selecting molecular descriptors relevant to the prediction of BBB+ and BBB− agents. RFE has recently gained popularity because of its effectiveness for discovering informative features or attributes in a variety of drug activity analyses.[25−29] A 5-fold cross validation was conducted to properly assess the prediction accuracy of these methods.

METHODS

Selection of BBB+ and BBB− Agents. In this work, a total of 415 agents with known BB ratios (the ratio of the steady-state concentrations of a drug in the brain and blood) were selected from Micromedex,[30] American Hospital Formulary Service,[31] and a number of publications.[13,15,16,19,32−34] The 2D structure of each of the compounds studied was generated using ChemDraw[35] and DS ViewerPro 5.0,[36] which were subsequently converted into 3D structures using CONCORD[37] and optimized using the semiempirical AM1 method.[38] All the geometries had been fully optimized without symmetry restrictions. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent was properly represented.

Agents were divided into BBB+ and BBB− groups according to whether the BB ratio was ≥ 0.1 or <0.1, respectively.[15,39] Under this definition, there are a total of 276 BBB+ and 139 BBB− agents, which are given in Table S1 of the Supporting Information. These 415 agents were 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 is repeated five times so that every subset is used once as a testing set. To evaluate the models, representative training and validation sets were constructed from the data sets according to their distribution in the chemical space. Here, chemical space is defined by the structural and chemical descriptors used to represent a compound.[40] Each compound occupies a particular location in this chemical space. All possible pairs of these compounds were generated, and a similarity score was computed for each pair. These pairs were then ranked in terms of their similarity scores, on the basis of which compounds of similar structural and chemical features were evenly assigned into separate data sets. Those compounds without enough structurally and chemically similar counterparts were assigned to the training set.

Molecular Descriptors. Molecular descriptors have been routinely used in the quantitative description of structural and physicochemical properties of molecules in the statistical study of drugs and small molecules.[20,40−43] In this study, a set of 199 molecular descriptors were selected from the more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or unrelated to the problem studied here. These descriptors, given in Table 2, include 18 descriptors in the class of simple molecular properties, 28 descriptors in the class of molecular connectivity and shape, 97 descriptors in the class of electrotopological state, 31 descriptors in the class of quantum chemical properties, and 25 descriptors in the class of geometrical properties. They were computed from the 3D structure of each compound using our own designed molecular descriptor computing program. The remaining redundant and unrelated descriptors were further reduced by using a feature selection method.[25,26,28,29]

Examples of topological descriptors include numbers of rings and rotatable bonds, numbers of hydrogen bond acceptors and donors, molecular connectivity χ indices, molecular shape κ indices, electrotopological state indices, and atom type electrotopological state indices. Molecular connectivity χ indices and shape κ indices encode information about molecular size, shape, branching, unsaturation, heteroatom content, and cyclicity.[44,45] The electrotopological state indices are numerical values computed for each atom in a molecule, which encode information about both the topological environment of that atom and the electronic interactions due to all other atoms in the molecule.[46,47]

Quantum chemical descriptors are used to describe electrostatic and electronic properties of a molecule. These descriptors are calculated using molecular orbital energies and wave functions of electronic motion in a molecule, which can be obtained by solving the Schrödinger equation of electronic motion.[38] The computed quantum chemical descriptors include partial atomic charges, the highest occupied and lowest unoccupied molecular orbital energies, dipole moment, polarizability, and other descriptors derived from them.[48]

Geometric descriptors encode the 3D-structural features of molecules. These include the van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, and the related properties from combining them with partial atomic charges.[49,50]

Feature Selection Method. The feature selection method used in this work is the RFE method, which has gained popularity because of its effectiveness for discovering informative features or attributes in the analysis of drug activity[27,29] and pharmacokinetic and toxicological properties.[25,26] Here, an agent is represented by a vector $\mathbf{x}_i$, with its molecular descriptors (or features) as the components. The task of selecting appropriate molecular descriptors can be conducted by ranking and selecting those with higher contributions to a particular drug classification problem.

Table 2. Molecular Descriptors Used in This Work

descriptor class	number of descriptor in class	descriptors
simple molecular properties	18	molecular weight, numbers of rings, rotatable bonds, H-bond donors, and H-bond acceptors, element counts,
molecular connectivity and shape	28	molecular connectivity indices, valence molecular connectivity indices, molecular shape κ indices, κ α indices, flexibility index,
electrotopological state	97	electrotopological state indices, and atom type electrotopological state indices, Weiner index, centric index, Altenburg index, Balaban index, Harary number, Schultz index, PetitJohn R2 index, PetitJohn D2 index, mean distance index, PetitJohn I2 index, information Weiner, Balaban RMSD index, graph distance index
quantum chemical properties	31	polarizability index, hydrogen bond acceptor basicity (covalent HBAB), hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index, most positive charge on H, C, N, O atoms, most negative charge on H, C, N, O atoms, most positive and negative charge in a molecule, sum of squares of charges on H,C,N,O and all atoms, mean of positive charges, mean of negative charges, mean absolute charge, relative positive charge, relative negative charge
geometrical properties	25	length vectors (longest distance, longest third atom, 4th atom), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively charged atoms, molecular rugosity, molecular globularity, hydrophilic region, hydrophobic region, capacity factor, hydrophilic−hydrophobic balance, hydrophilic intery moment, hydrophobic intery moment, amphiphilic moment

It has been suggested that the ranking criterion for feature selection can be based on the change in the objective function upon removing each feature.[51] To improve the efficiency of training, this objective function is represented by a cost function J computed by using the training set only. When a given feature is removed or its weight is brought to zero, the change DJ(i) in the cost function J is computed by

$$\mathrm{DJ}(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2 \quad (1)$$

where w_i is the weight of the feature I and the change in weight $Dw_i = \mathrm{w}_i$ corresponds to the removed feature i. In this work, the cost function to be minimized is

$$J = (1/2)\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} \quad (2)$$

where **H** is the matrix with elements such as that of $y_i y_j \exp[-||x_i - x_j||^2/(2\sigma^2)]$ in SVM and **1** is an m dimensional identity vector (m is the number of compounds in training set).

To compute the change in cost function caused by removing input component i, the $\boldsymbol{\alpha}$'s are kept unchanged and the matrix **H** is recomputed. The resulting ranking coefficient is

$$\mathrm{DJ}(i) = {}^1/_2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} - {}^1/_2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}(-i)\boldsymbol{\alpha} \quad (3)$$

where $\mathbf{H}(-i)$ is the matrix computed by the same method as that for matrix **H** but with its i-h component removed. One or more of features with the smallest DJ(i) are, thus, eliminated.

The RFE feature selection computation procedure[25] is outlined as follows: The prediction accuracy of a statistical learning prediction system during the training process was evaluated by means of a 5-fold cross validation. In the first step, for a fixed parameter, the statistical learning prediction system is trained by using the complete set of features (molecular descriptors) described in the previous section. The second step is to compute the ranking criterion score DJ(i) for each feature in the current set by using eq 3. All of the computed DJ(i) values are subsequently ranked in descending order. The third step is to remove the m features with the smallest criterion scores. In this work, m was chosen to be 5 as that value was used in earlier studies.[29] In the fourth step, the prediction system is retrained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of a 5-fold cross validation. The first to fourth steps are then repeated for other parameter values. After the completion of these procedures, the set of features and parameters that give the best prediction accuracy are selected.

Statistical Learning Methods. *I. Logistic Regression Method.* LR[52] is based on the assumption that a logistic relationship exists between the probability of class membership and one or more descriptors:

$$Y = \frac{1}{1 + \exp[-(\beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta kXk)]} \quad (4)$$

where Y is the probability that vector **x** belongs to the positive class, $\beta 0$ is the regression model constant, and $\beta 1 - \beta k$ are the coefficients corresponding to the descriptors $X1 - Xk$. A Y value of 0.5 or greater indicates that the vector **x** belongs

to the positive class, while a value of below 0.5 classifies vector **x** as negative.

II. Linear Discriminate Analysis Method. LDA[53] separates two different classes of vectors by constructing a hyperplane that is defined by the following linear discriminant function:

$$L = \sum_{i}^{k} w_i x_i \tag{5}$$

where L is the resultant classification score and w_i is the weight associated with the corresponding descriptor x_i. A positive or negative L value indicates that the vector **x** belongs to the positive or negative class, respectively.

III. k Nearest Neighbor Method. In *k*-NN, the Euclidean distance between an unclassified vector **x** and each individual vector $\mathbf{x}_i$ in the training set is measured.[54,55] A total of k number of vectors nearest to the unclassified vector **x** are used to determine the class of that unclassified vector. The class of the majority of the k nearest neighbors is chosen as the predicted class of the unclassified vector **x**.

IV. C4.5 Decision Tree Method. C4.5 DT is a branch-test-based classifier.[56] A branch in a decision tree corresponds to a group of classes, and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes of the test. C4.5 DT uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector **x** is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each nonleaf decision node, a test is conducted and the classification process proceeds to the branch selected by the test. Upon reaching the destination leaf, the class of the vector **x** is predicted to be that associated with the leaf.

V. Probabilistic Neural Network Method. PNN is a form of neural network that uses Bayes' optimal decision rule for classification.[57] Traditional neural networks such as the feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparameteric estimator. Thus, the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

VI. Support Vector Machine Method. In linearly separable cases, SVM constructs a hyperplane that separates two different classes of vectors with a maximum margin.[58] This is done by finding another vector **w** and a parameter b that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions:

$$\mathbf{w}\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \qquad \text{Class 1 (positive)} \tag{6}$$

$$\mathbf{w}\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \qquad \text{Class 2 (negative)} \tag{7}$$

where y_i is the class index, **w** is a vector normal to the hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin, and $||\mathbf{w}||^2$ is the Euclidean norm of **w**. After the determination of **w** and b, a given vector $\mathbf{x}_i$ can be classified by

$$sign[(\mathbf{w}\mathbf{x}) + b] \tag{8}$$

In nonlinearly separable cases, SVM maps the input variable into a high dimensional feature space using a kernel function such as $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-||\mathbf{x}_i-\mathbf{x}_i||^2/2\sigma^2)$. Linear SVM is then applied to this feature space, and then, the decision function is given by

$$f(\mathbf{x}) = sign[\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x},\mathbf{x}_i) + b] \tag{9}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{10}$$

under the following conditions: $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$. A positive or negative value from eq 8 or 9 indicates that the vector **x** belongs to the positive or negative class, respectively.

Prediction Accuracy. As in the case of all discriminative methods,[24,59] the performance of a statistical learning method can be measured by the quantity of true positives TP, true negatives TN, false positives FP, false negatives FN, sensitivity SE = TP/(TP + FN) (BBB+ accuracy in this work), and specificity SP = TN/(TN + FP) (BBB− accuracy in this work). The overall prediction accuracy (Q) and Matthews correlation coefficient (C)[60] are also frequently used to measure the prediction accuracies:

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

$$C = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FN}\cdot\mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{12}$$

RESULTS AND DISCUSSION

Molecular Descriptors Selected for BBB Penetration Prediction. Table 2 lists the 199 descriptors used in this study, and Table 3 gives the descriptors selected for the prediction of BBB+ and BBB− agents from the feature selection method RFE. Six of the selected descriptors are VolSurf descriptors.[20] In VolSurf, information from 3D molecular field maps are extracted into 1D descriptors specially designed for pharmacokinetic applications.[19,20] Thus, it is not surprising that VolSurf descriptors are selected for the prediction of BBB penetration and BBB nonpenetration properties of agents. The six selected VolSurf descriptors are molecular rugosity, the capacity factor, the hydrophilic and hydrophobic regions, and the hydrophobic and amphiphilic moments. Those from the class of electrotopological state constitute the largest percentage of the descriptors selected. A large variety of descriptors in this class, such as those of different functional groups and hydrophilic and hydrophobic properties, are important in the characterization of pharmacokinetic properties. There are also a substantial number of descriptors from the quantum chemical, connectivity, and geometric classes. The selected quantum chemical descriptors determine molecular dipole moment, chemical potential, electronegativity, hydrogen bond donor acidity, and the atomic charge on H and O atoms in a

Table 3. Thirty-Seven Molecular Descriptors Selected from the RFE (Recursive Feature Elimination) Feature Selection Method for Classification of BBB-Penetrating and -Nonpenetrating Agents

descriptors	description	class
Rugty	molecular rugosity	geometrical properties
Shpl	hydrophilic region	geometrical properties
Shpb	hydrophobic region	geometrical properties
Capty	capacity factor	geometrical properties
Hiwpb	hydrophobic intery moment	geometrical properties
Hiwpa	amphiphilic moment	geometrical properties
dis1, dis2, dis3	length vectors (longest distance, longest third atom, 4th atom)	geometrical properties
Sanc	sum of solvent accessible surface areas of negatively charged atoms	geometrical properties
$^5\chi^v_{CH}$	valence molecular connectivity χ index for cycle of five atoms	connectivity and shape
$^6\chi^v_{CH}$	valence molecular connectivity χ index for cycle of six atoms	connectivity and shape
ϵ_a	hydrogen bond donor acidity (covalent hbda)	quantum chemical properties
m	molecular dipole moment	quantum chemical properties
μ_{cp}	chemical potential	quantum chemical properties
χ_{en}	electronegativity index	quantum chemical properties
$Q_{H,Max}$	most positive charge on H atoms	quantum chemical properties
$Q_{O,Max}$	most positive charge on O atoms	quantum chemical properties
$Q_{H,Min}$	most negative charge on H atoms	quantum chemical properties
S_{hal}	sum of estate indices of halogen atoms	electrotopological state
S(1)	atom-type H estate sum for −OH	electrotopological state
S(4)	atom-type H estate sum for $-NH_2$	electrotopological state
S(8)	atom-type H estate sum for $=CH_2$ (sp^2)	electrotopological state
S(10)	atom-type H estate sum for :CH: (sp^2, aromatic)	electrotopological state
S(13)	atom-type H estate sum for CH_n (unsaturated)	electrotopological state
S(16)	atom-type estate sum for $-CH_3$	electrotopological state
S(17)	atom-type estate sum for $=CH_2$	electrotopological state
S(25)	atom-type estate sum for =C<	electrotopological state
S(26)	atom-type estate sum for :C:−	electrotopological state
S(27)	atom-type estate sum for :C::	electrotopological state
S(30)	atom-type estate sum for =NH	electrotopological state
S(34)	atom-type estate sum for =N−	electrotopological state
S(35)	atom-type estate sum for :N:	electrotopological state
S(36)	atom-type estate sum for >N−	electrotopological state
S(37)	atom-type estate sum for $-N\ll(NO_2)$	electrotopological state
S(39)	atom-type estate sum for −OH	electrotopological state
S(41)	atom-type estate sum for −O−	electrotopological state

molecule. The selected molecular connectivity descriptors are the valence molecular connectivity χ indices for cycles of five and six atoms. The rest of geometrical descriptors, except VolSurf descriptors, describe the molecular size and the sum of solvent accessible surface areas of negatively charged atoms. These descriptors are also important in describing electrostatic, topological, and geometric properties of chemical compounds.

These selected descriptors are also overlapped with those used in QSAR studies of BBB penetrating and nonpenetrating agents. Iyer and co-workers[34] showed that both hydrogen bonding and ClogP (which measures overall hydrophobicity) are the primary descriptors for the BBB classification in their data set. Pan and co-workers[61] found that ClogP, polar surface areas, the number of hydrogen bond donors, and molecular flexibility, such as the number of rotatable bonds, are the most important descriptors in their model. Ooms et al.[32] showed that polarity inversely correlated with the BBB permeability while hydrophobicity directly correlated to BBB permeability. Similarly, Rose et al.[62] showed that increasing lesser skeletal branching of the carbon chain and hydrogen bond donors enhances BBB permeability.

Hydrophobicity is collectively described by the two geometrical (Shpb and Hiwpb) and eight electrotopological state [S(8), S(10), S(13), S(16), S(17), S(25), S(26), and S(27)] descriptors selected by RFE. The selected descriptors related to hydrogen bonds are four quantum chemical ($Q_{H,Max}$, $Q_{O,Max}$, $Q_{H,Min}$, and ϵ_a) and nine electrotopological state [S(1), S(4), S(30), S(34), S(35), S(36), S(37), S(39), and S(41)] descriptors, which collectively describe hydrogen bond donor acidity and hydrogen bond acceptor basicity properties. Polarity is collectively described by the three geometrical (Shp1, Hiwpa, and Sanc) and two quantum chemical (μ and χ_{en}) descriptors selected by RFE. Molecular flexibility and skeletal branching of the carbon chain are collectively described by the two RFE-selected molecular connectivity and shape descriptors ($^5\chi^v_{CH}$ and $^6\chi^v_{CH}$).

Because composite descriptors encode multiple physicochemical and structural aspects of the molecule, it is difficult to extract from these descriptors information about which specific molecular characteristics are important for the BBB+ and BBB− prediction. Nonetheless, it is possible to infer some information from noncomposite descriptors. As many descriptors are overlapping and some of them are redundant, it is more appropriate to group them into classes of descriptors of similar properties and discuss their contribution to the BBB penetration predictions at the class level. Table 4 gives the classes of noncomposite descriptors selected by our computations. It is found that the hydrophobic is the dominant factor involved in BBB penetration. This is not surprising because the hydrophobic is the primary descriptor in the BBB penetration. In addition to the shape and size descriptors, the electrostatic descriptor and hydrogen bonding are found to be other dominant forces involved in BBB penetration, which are consistent with the findings that the electrostatic interaction between the solute molecule and the membrane and the ability of a molecule to form hydrogen bonds are important descriptors involved in BBB penetration.[61]

Table 4. Important Descriptor Classes Selected for the Prediction of BBB-Penetrating and -Nonpenetrating Agents

dataset	electrostatic (%)	hydrogen bond acceptors and donors (%)	hydrophobic (%)	shape (%)	size (%)
BBB+/BBB−	18.9	27.1	35.1	8.1	10.8

Table 5. Differences in the Values of Descriptors Important for Distinguishing between BBB-Penetrating (BBB+) and -Nonpenetrating (BBB−) Agents

descriptors	average value[a] BBB+	average value[a] BBB−
μ	1.89 (1.90)	2.92 (3.48)
ϵ_a	0.01 (0.87)	−0.34 (1.36)
Shpb	2.44 (2.85)	2.17 (3.96)
Rugty	1.18 (2.31)	1.54 (2.84)
dis1	−4.52 (0.68)	−4.37 (1.01)

[a] Values in parentheses are the standard deviations.

It is also possible to roughly distinguish between BBB+ and BBB− from the values of five selected descriptors—μ, ϵ_a, Shpb, Rugty, and dis1. These descriptors are representative of the four dominant interaction forces—electrostatic, hydrogen bond acceptor, hydrogen bond donor and hydrophobicity, and size and shape of the compounds, respectively. μ is the molecular dipole moment, ϵ_a is the hydrogen bond donor acidity, Shpb is the hydrophobic region, Rugty is the molecular wrinkled surface, and dis1 is the length vector of the longest distance. The average values of these five descriptors for BBB+ and BBB− compounds of all of the various datasets are given in Table 5. In general, BBB penetration agents are generally smaller in size, lesser in molecular wrinkled surface shape, more hydrophobic, lesser in the molecular dipole moment, and have more hydrogen bonding sites than BBB nonpenetration agents. This conclusion is consistent with the findings that less polar, more lipophilic compounds partition more readily into the brain and the greater binding of the solute to the membrane, the higher the BBB partitioning.[34]

Prediction Accuracy for BBB+ and BBB− Agents. The BBB+ and BBB− prediction accuracies of different statistical learning methods by using both the full molecular descriptor set and the RFE-selected descriptors are given in Table 6. The listed accuracies are the average accuracies from the 5-fold validation study by using each method. The detailed 5-fold cross-validation results of one of the methods, SVM with and without the use of RFE, are given in Table 7. It is found that the use of RFE substantially improves the BBB− accuracy and the overall accuracy of all of the methods studied in this work. The BBB− accuracy is improved from 42.8∼64.3% to 46.4∼75.0%, the overall accuracy Q is improved from 46.8∼79.1% to 71.0∼83.7%, and the Matthews correlation coefficient C is improved from −0.067∼0.524 to 0.321∼0.645 by using the full molecular descriptor set and the RFE-selected descriptor set.

Of the statistical learning methods studied, SVM gives the highest BBB+, BBB−, Q, and C values 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. The overall accuracy of SVM with RFE-selected descriptors is found to be slightly higher than those from previous studies listed in Table 1. 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, which is consistent with the results from an earlier study of BBB+ and BBB− agents[22] and other studies of different chemical and protein systems.[63−65] This suggests that SVM is capable of the prediction of BBB+ and BBB− agents at a comparable or perhaps better accuracy with respect to that from other classification methods without requiring either the knowledge of a mechanism or the intrinsic structure−activity relationships.

To compare with previous binary classification models, prediction accuracies of BBB+ and BBB− agents from different studies reported in the literature are provided in Table 2. However, it should be cautioned that direct comparison of these results may not be appropriate because of the use of different sets of agents, molecular descriptors, classification methods and parameters, and methods for generating testing sets. A meaningful comparison can be achieved if all the classification models are tested by using the same set of BBB+ and BBB− agents not used in training

Table 6. Comparison of the Prediction Accuracy of the BBB-Penetrating (BBB+) and -Nonpenetrating (BBB−) Agents by Different Statistical Learning Methods[a]

	BBB+ accuracy (%)		BBB− accuracy (%)		prediction performance			
					overall accuracy Q (%)		Matthews correlation coefficient C	
method[b]	all descriptors	RFE-selected descriptors	all descriptors	RFE-selected descriptors	all descriptors	RFE-selected descriptors	all descriptors	RFE-selected descriptors
LR	63.6	83.9	42.8	46.4	57.1	71.0	0.063	0.321
LDA	40.0	78.2	58.4	58.3	46.8	71.2	−0.067	0.360
C4.5 DT	83.7	80.3	54.9	62.8	73.8	74.3	0.398	0.433
k-NN	77.0	85.5	58.0	61.4	70.8	77.1	0.348	0.477
PNN	76.2	84.3	57.8	62.1	70.3	76.5	0.357	0.481
SVM	89.9	88.6	64.3	75.0	79.1	83.7	0.524	0.645

[a] The accuracy of each method is taken from the average accuracy of a 5-fold cross validation by using 276 BBB+ and 139 BBB− agents. [b] LR (logistic regression), LDA (linear discriminate analysis), C4.5 DT (C4.5 decision tree), *k*-NN (*k* nearest neighbor), PNN (probabilistic neural network), SVM (support vector machine), RFE (recursive feature elimination).

Table 7. SVM and SVM with RFE (SVM + RFE) Prediction Accuracy of the BBB-Penetrating (BBB+) and -Nonpenetrating Agents (BBB−) by Using a 5-Fold Cross Validation[a]

		BBB+			BBB−				
method	cross validation	TP	FN	SE (%)	TN	FP	SP (%)	*Q* (%)	*C*
SVM	1	45	7	86.5	18	6	75	82.9	0.609
	2	45	10	81.8	15	14	51.7	71.4	0.349
	3	54	8	87.1	22	4	84.6	83.4	0.690
	4	50	6	89.3	14	11	56	79.0	0.485
	5	46	5	90.2	19	16	54.3	75.6	0.487
	average			**89.9**			**64.3**	**79.1**	**0.524**
	SD			3.16			13.07	4.53	0.117
	SE			1.29			5.34	1.85	0.048
SVM + RFE	1	39	13	75.0	22	2	91.7	80.3	0.622
	2	49	6	89.1	24	5	82.8	86.9	0.713
	3	58	4	93.5	21	5	80.8	89.8	0.752
	4	51	5	91.1	15	10	60.0	81.5	0.547
	5	48	3	94.1	21	14	60.0	80.2	0.593
	average			**88.6**			**75.0**	**83.7**	**0.645**
	SD			7.01			12.83	3.90	0.08
	SE			2.86			5.24	1.59	0.03

[a] Predicted results are given in TP (true positive); FN (false negative); TN (true negative); FP (false positive); SE (sensitivity), which is the prediction accuracy for BBB+; SP (specificity), which is the prediction accuracy for BBB−; *Q* (overall prediction accuracy); and *C* (Matthews correlation coefficient). Statistical significance is indicated by SD (standard deviation) and SE (standard error).

Table 8. Comparison of Accuracy of BBB-Penetrating (BBB+) and -Nonpenetrating (BBB−) Agents by Using Cross Validation with Independent Validation Set[a]

5-fold cross validation				independent validation set							
SE (%)	SP (%)	*Q* (%)	*C* (%)	TP	FN	SE (%)	TN	FP	SP (%)	*Q* (%)	*C*
88.6	75.0	83.7	0.645	57	7	89.1	25	7	78.1	85.4	0.672

[a] Predicted results are given in TP (true positive); FN (false negative); TN (true negative); FP (false positive); SE (sensitivity), which is the prediction accuracy for BBB+; SP (specificity), which is the prediction accuracy for BBB−; *Q* (overall prediction accuracy); and *C* (Matthews correlation coefficient).

any of these models. Such a testing set should ideally include some of the newly discovered BBB+ agents, particularly the novel ones. Examples of such agents are AD4 [thiol antioxidant and the *N*-acetyl cysteine (NAC) related compound],[66] 2,5-bis(4-amidinophenyl)furan (DB75) and 2,5-bis-(4-amidinophenyl)furan-bis-O-methylamidoxime (DB289),[67] glucosamine-kynurenic acid,[68] (*R*)-3′-(5-chlorothiophen-2-yl)spiro-1-azabicyclo[2.2.2]octane-3,5′-[1′,3′]oxazolidin-2′-one,[69] GPI 15427,[70] and (1*R*,2*R*,3*R*,5*R*,6*R*)-2-amino-3-(3,4-dichlorobenzyloxy)-6-fluorobicyclo[3.1.0]hexane-2,6-dicarboxylic acid (−)-11be (MGS0039).[71]

On the other hand, the binary classification models developed in this and other works cannot be directly compared with the continuous BBB-value QSAR models without incorporating algorithms that provide quantitative activity values. Binary classification models have primarily been explored with the intention of analyzing agents of diverse structures, in many cases, more diverse than those of QSAR models that are required to provide accurate activity values. The SVM regression model has been introduced to provide activity values for some drug classification problems including the prediction of COX-2 selective inhibitors;[72] the prediction of binding affinities to human serum albumin;[73] and the case studies of drug likeness, agrochemical likeness, and enzyme inhibition predictions.[74] Such a regression model can be developed for the prediction of BBB+ and BBB− agents, which can be directly compared to the continuous BBB-value QSAR models.

It has been shown that chance correlations may occur during descriptor selection, especially if the number of descriptors available for selection is large.[75,76] *y* randomization has been frequently used to determine the probability of chance correlation during descriptor selection processes.[77,78] In *y* randomization, a portion of BBB+ compounds in the data set was randomly selected and converted to BBB− compounds. Another portion of BBB− compounds was also randomly selected and converted to BBB+ compounds. The ratios of BBB+ to BBB− compounds were kept unchanged during *y* randomization. The "scrambled" data set was then used for the descriptor selection process. The process of the scrambling of the data set and descriptor selection process was repeated 30 times. This *y* randomization analysis was conducted on the SVM model that consistently gave the better classification accuracies in this and other studies.[79−81] Unless overfitting is found in the SVM model, no further analysis on other models is to be conducted. The average Matthews correlation coefficient of these scrambled SVM classification systems derived by using the 5-fold cross-validation sets was found to be 0.1176, which is significantly lower than that of the original SVM classification system, which is 0.645. This suggests that the original SVM classification system is relevant and unlikely to arise as a result of chance correlation.

A frequently used method for checking whether a prediction system is overfitted is to compare the prediction accuracies determined by using cross-validation methods with

those determined by using independent validation sets.[82] Since descriptor selection was performed by using the cross-validation method as the modeling testing sets, an overfitted classification system is expected to have a much higher prediction accuracy for the cross-validation sets than that for the independent validation sets. As shown in Table 8, the prediction accuracies of the SVM systems based on the cross-validation method and those based on independent validation sets are similar. This suggests that the SVM classification systems in this work are unlikely to overfit.

The prediction of BBB+ or BBB− agents may need to be correlated to pharmacodynamic properties in order to determine their clinical significance. This is because a drug with high BB ratio may not have effects in the brain either because of the absence of target receptors or because of insufficient potencies toward the target receptors in the brain. Conversely, a drug with a relatively low BB ratio may still have effects in the brain because of its high potency toward specific receptors. An example of such a drug is diazepam, which is an antidepressant that exerts its effects in the brain even though its BB ratio is low.[83]

CONCLUSION

The feature selection method appears to be helpful in improving the performance of statistical learning methods for the prediction of the blood−brain barrier penetration potential of chemical agents. Of the six statistical learning methods tested, SVM appears to give a slightly higher prediction accuracy than other methods for both BBB+ and BBB− agents. Recent efforts are directed at the improvement of the efficiency and speed of feature selection methods,[84] which can further help to optimally select molecular descriptors and enable the development of more accurate and efficient prediction tools. The prediction accuracy of statistical learning methods may be further improved by consideration of factors such as hydrogen bonding, active transport, and the relationship with pharmacodynamic properties.

ACKNOWLEDGMENT

This work was supported, in part, by grants from Shanghai Commission for Science and Technology (04DZ19850, 04QMX1450, 04DZ14005) and the "973" National Key Basic Research Program of China (2004CB720103, 2004CB715901).

Supporting Information Available: A table giving the 415 agents, 276 BBB+ and 139 BBB−, used in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Eddershaw, P. J.; Beresford, A. P.; Bayliss, M. K. ADME/PK as part of a rational approach to drug discovery. *Drug Discovery Today* **2000**, *5* (9), 409−414.
(2) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties in silico: Methods and models. *Drug Discovery Today* **2002**, *7* (Suppl. 11), S83−S88.
(3) Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood−brain distribution. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 291−313.
(4) Hardman, J. G.; Limbird, L. E.; Gilman, A. G. *Goodman and Gilman's the pharmacological basis of therapeutics*, 10th ed.; McGraw-Hill: New York, 2002.
(5) Li, A. P. Early ADME/Tox studies and in silico screen. *Drug Discovery Today* **2002**, *7* (1), 25−27.
(6) Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffith, R.; Jones, M.; Rana, K. K.; Saundesr, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *J. Med. Chem.* **1988**, *31*, 656−671.
(7) Abraham, M. H.; Chadha, H. S.; Mitchell, R. Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci.* **1994**, *83* (9), 1257−1268.
(8) Lombardo, F.; Blake, J. F.; Curatolo, W. J. Computation of brain−blood partitioning of organic solutes via free energy calculations. *J. Med. Chem.* **1996**, *39* (24), 4750−4755.
(9) Clark, D. E. 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.* **1999**, *88* (8), 815−821.
(10) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J. P. Polar molecular surface area as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16* (10), 1514−1519.
(11) Liu, R.; Sun, H.; So, S. S. Development of quantitative structure−property relationship models for early ADME evaluation in drug discovery. 2. Blood−brain barrier penetration. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1623−1632.
(12) Keserü, G. M.; Molnár, L. High-throughput prediction of blood-brain partitioning: a thermodynamic approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (1), 120−128.
(13) Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. Correlation and prediction of a large blood−brain distribution data set − an LFER study. *Eur. J. Med. Chem.* **2001**, *36* (9), 719−730.
(14) Norinder, U.; Sjöberg, P.; Österberg T. Theoretical calculation and prediction of brain−blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. *J. Pharm. Sci.* **1998**, *87* (8), 952−959.
(15) Luco, J. M. Prediction of the brain−blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (2), 396−404.
(16) Feher, M.; Sourial, E.; Schmidt, J. M. A simple model for the prediction of blood−brain partitioning. *Int. J. Pharm.* **2000**, *201* (2), 239−247.
(17) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood−brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2137−2152.
(18) Liu, X.; Tu, M.; Kelly, R. S.; Chen, C.; Smith, B. J. Development of a computational approach to predict blood−brain barrier permeability. *Drug Metab. Dispos.* **2004**, *32* (1), 132−139.
(19) Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. Predicting blood−brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43* (11), 2204−2216.
(20) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29−S39.
(21) Ajay; Bemis, G. W.; Murcko, M. A. Designing libraries with CNS activity. *J. Med. Chem.* **1999**, *42* (24), 4942−4951.
(22) Doniger, S.; Hofman, T.; Yeh, J. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. *J. Comput. Biol.* **2002**, *9* (6), 849−864.
(23) Trotter, M. W. B.; Buxton, B. F.; Holden, S. B. Support vector machines in combinatorial chemistry. *Meas. Control* **2001**, *34* (8), 235−239.
(24) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16* (5), 412−424.
(25) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of p-glycoprotein substrates by support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497−1505.
(26) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630−1638.
(27) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learning* **2002**, *46* (1−3), 389−422.
(28) Degroeve, S.; De Baets, B.; Van de Peer, Y.; Rouzé, P. Feature subset selection for splice site prediction. *Bioinformatics* **2002**, *18*, S75−S83.
(29) Yu, H.; Yang, J.; Wang, W.; Han, J. In *Discovering Compact and Highly Discriminative Features or Feature Combinations of Drug Activities Using Support Vector Machines*, IEEE Computer Society Bioinformatics Conference (CSB '03), Stanford, California, August 11−14, 2003; IEEE: Stanford, California, 2003; pp 220−228.
(30) *MICROMEDEX*. Micromedex: Greenwood Village, Colorado. Edition expires 12/2003.

(31) McEvoy, G. K.; Litvak, K.; Welsh, O. H. *AHFS Drug Information*; American Society of Health-System Pharmacists, Inc: Bethesda, MD, 2001.
(32) Ooms, F.; Weber, P.; Carrupt, P. A.; Testa, B. A simple model to predict blood−brain barrier permeation from 3D molecular fields. *Biochim. Biophys. Acta* **2002**, *1587* (2−3), 118−125.
(33) Lobell, M.; Molnár, L.; Keserü, G. M. Recent advances in the prediction of blood−brain partitioning from molecular structure. *J. Pharm. Sci.* **2003**, *92* (2), 360−370.
(34) Iyer, M.; Mishru, R.; Han, Y.; Hopfinger, A. J. Predicting blood−brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm. Res.* **2002**, *19* (11), 1611−1621.
(35) *ChemDraw*, 7.0.1; CambridgeSoft Corporation: Cambridge, MA, 2002.
(36) *DS ViewerPro*, 5.0; Accelrys: San Diego, CA.
(37) Pearlman, R. S. *CONCORD User's Manual*; Tripos: St. Louis, MO.
(38) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Steward, J. P. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(39) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 1. Applications of genetic algorithms to the prediction of blood−brain partitioning of a large set of drugs. *J. Mol. Model.* **2002**, *8* (12), 337−349.
(40) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(41) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs electronic, geometrical and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33* (6), 835−857.
(42) Kier, L. B.; Hall, L. H. *Molecular structure description: The electrotopological state*. Academic Press: San Diego, CA, 1999.
(43) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96* (3), 1027−1043.
(44) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*. Research Studies Press: Wiley: Letchworth, Hertfordshire, England, 1986; Vol. 9.
(45) Hall, L. H.; Kier, L. B. *The molecular connectivity chi indices and kappa shape indices in structure−property modeling*. VCH publishers: New York, 1991; Vol. 2, pp 367−412.
(46) Hall, L. H.; Mohney, B. K.; Kier, L. B. The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(47) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(48) Thanikaivelan, P.; Subramanian, V.; Rao, J. R.; Nair, B. A. Application of quantum chemical descriptors in quantitative structure activity and structure property relationship. *Chem. Phys. Lett.* **2000**, *323* (1−2), 59−70.
(49) Hopfinger, A. J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.
(50) Tsodikov, O. V.; Record, M. T. J.; Sergeev, Y. V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.* **2002**, *23*, 600−609.
(51) Kohavi, R.; John, G. H. Wrappers for Feature Subset Selection. *Artif. Intell. Med.* **1997**, *97*, 273−324.
(52) Hosmer, D. W.; Lemeshow, S. *Applied logistic regression*; Wiley: New York, 1989.
(53) Huberty, C. J. *Applied discriminant analysis*; John Wiley & Sons: New York, 1994.
(54) Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*; Prentice Hall: Englewood Cliffs, NJ, 1982.
(55) Fix, E.; Hodges, J. L. *Discriminatory analysis: Nonparametric discrimination: Consistency properties.* USAF School of Aviation Medicine: Randolph Field, TX, 1951; pp 261−279.
(56) Quinlan, J. R. *C4.5: programs for machine learning*. Morgan Kaufmann: San Mateo, CA, 1993.
(57) Specht, D. F. Probabilistic neural networks. *Neural Networks* **1990**, *3* (1), 109−118.
(58) Vapnik, V. N. *The nature of statistical learning theory*. Springer: New York, 1995.
(59) Roulston, J. E. Screening with tumor markers. *Mol. Pharmacol.* **2002**, *20*, 153−162.
(60) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405* (2), 442−451.
(61) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. J. Constructing optimum blood brain barrier QSAR models using a combination of 4D-molecular similarity measures and cluster analysis. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 2083−2098.
(62) Rose, K.; Hall, L. H.; Kier, L. B. Modeling blood−brain barrier partitioning using the electrotopological state. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (3), 651−666.
(63) Trotter, M. W. B.; Holden, S. B. Support vector machines for ADME property classification. *QSAR Comb. Sci.* **2003**, *22* (5), 533−548.
(64) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16* (10), 906−914.
(65) Bock, J. R.; Gough, D. A. Predicting protein−protein interactions from primary structure. *Bioinformatics* **2001**, *17* (5), 455−460.
(66) Bahat-Stroomza, M.; Gilgun-Sherki, Y.; Offen, D.; Panet, H.; Saada, A.; Krool-Galron, N.; Barzilai, A.; Atlas, D.; Melamed, E. A novel thiol antioxidant that crosses the blood brain barrier protects dopaminergic neurons in experimental models of Parkinson's disease. *Eur. J. Neurosci.* **2005**, *21* (3), 637−646.
(67) Sturk, L. M.; Brock, J. L.; Bagnell, C. R.; Hall, J. E.; Tidwell, R. R. Distribution and quantitation of the anti-trypanosomal diamidine 2,5-bis(4-amidinophenyl)furan (DB75) and its N-methoxy prodrug DB289 in murine brain tissue. *Acta Tropica* **2004**, *91* (2), 131−143.
(68) Robotka, H.; Nemeth, H.; Somlai, C.; Vecsei, L. T. J. Systemically administered glucosamine-kynurenic acid, but not pure kynurenic acid, is effective in decreasing the evoked activity in area CA1 of the rat hippocampus. *Eur. J. Pharm.* **2005**, *513* (1−2), 75−80.
(69) Tatsumi, R.; Fujio, M.; Satoh, H.; Katayama, J.; Takanashi, S.; Hashimoto, K.; Tanaka, H. Discovery of the alpha7 nicotinic acetylcholine receptor agonists. (R)-3′-(5-Chlorothiophen-2-yl)spiro-1-azabicyclo[2.2.2]octane-3, 5′-[1′, 3′]oxazolidin-2′-one as a novel, potent, selective, and orally bioavailable ligand. *J. Med. Chem.* **2005**, *48* (7), 2678−2686.
(70) Tentori, L.; Leonetti, C.; Scarsella, M.; Vergati, M.; Xu, W.; Calvin, D.; Morgan, L.; Tang, Z.; Woznizk, K.; Alemu, C.; Hoover, R.; Lapidus, R.; Zhang, J.; Graziani, G. Brain distribution and efficacy as chemosensitizer of an oral formulation of PARP-1 inhibitor GPI 15427 in experimental models of CNS tumors. *Int. J. Oncol.* **2005**, *26* (2), 415−422.
(71) Nakazato, A.; Sakagami, K.; Yasuhara, A.; Ohta, H.; Yoshikawa, R.; Itoh, M.; Nakamura, M.; Chaki, S. Synthesis, in vitro pharmacology, structure−activity relationships, and pharmacokinetics of 3-alkoxy-2-amino-6-fluorobicyclo[3.1.0]hexane-2,6-dicarboxylic acid derivatives as potent and selective group II metabotropic glutamate receptor antagonists. *J. Med. Chem.* **2004**, *47* (18), 4570−4587.
(72) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR and classification models of a novel series of COX-2 selective inhibitors: 1,5-diarylimidazoles based on support vector machines. *J. Comput.-Aided Mol. Des.* **2004**, *18* (6), 389−399.
(73) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1693−1700.
(74) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2048−2056.
(75) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure−activity relationships. *J. Med. Chem.* **1979**, *22*, (10), 1238−1244.
(76) Jouan-Rimbaud, D.; Massart, D. L.; de Noord, O. E. Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemom. Intell. Lab. Syst.* **1996**, *35* (2), 213−220.
(77) Manly, B. F. J. *Randomization bootstrap and Monte Carlo methods in biology*, 2nd ed.; Chapman and Hall: London, 1997.
(78) Leardia, R.; González, A. L. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41* (2), 195−207.
(79) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26* (1), 5−14.
(80) Czerminski, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.−Act. Relat.* **2001**, *20* (3), 227−240.
(81) Meyer, D.; Leischa, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55* (1−2), 169−186.
(82) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 1−12.
(83) Hallstrom, C.; Lader, M. H. Diazepam and N-desmethyldiazepam concentrations in saliva, plasma and CSF. *Br. J. Clin. Pharmacol.* **1980**, *9* (4), 333−339.
(84) Furlanello, C.; Serafini, M.; Merler, S.; Jurman, G. An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks* **2003**, *16* (5−6), 641−648.

CI050135U