Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 292683, 13 pages
http://dx.doi.org/10.1155/2015/292683

Research Article

A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction

Daqing Zhang,[1] Jianfeng Xiao,[2] Nannan Zhou,[3] Mingyue Zheng,[2] Xiaomin Luo,[2] Hualiang Jiang,[2,3] and Kaixian Chen[2]

[1] *Center for Systems Biology, Soochow University, Suzhou 215006, China*
[2] *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China*
[3] *School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*

Correspondence should be addressed to Mingyue Zheng; myzheng@simm.ac.cn, Xiaomin Luo; xmluo@simm.ac.cn, and Hualiang Jiang; hljiang@simm.ac.cn

Received 24 February 2015; Revised 7 May 2015; Accepted 19 May 2015

Academic Editor: Sílvia A. Sousa

Blood-brain barrier (BBB) is a highly complex physical barrier determining what substances are allowed to enter the brain. Support vector machine (SVM) is a kernel-based machine learning method that is widely used in QSAR study. For a successful SVM model, the kernel parameters for SVM and feature subset selection are the most important factors affecting prediction accuracy. In most studies, they are treated as two independent problems, but it has been proven that they could affect each other. We designed and implemented genetic algorithm (GA) to optimize kernel parameters and feature subset selection for SVM regression and applied it to the BBB penetration prediction. The results show that our GA/SVM model is more accurate than other currently available log *BB* models. Therefore, to optimize both SVM parameters and feature subset simultaneously with genetic algorithm is a better approach than other methods that treat the two problems separately. Analysis of our log *BB* model suggests that carboxylic acid group, polar surface area (PSA)/hydrogen-bonding ability, lipophilicity, and molecular charge play important role in BBB penetration. Among those properties relevant to BBB penetration, lipophilicity could enhance the BBB penetration while all the others are negatively correlated with BBB penetration.

1. Introduction

The blood-brain barrier (BBB) plays important roles in separating the central nervous system (CNS) from circulating blood and maintaining brain homeostasis. BBB penetration, which may be desired or not depending on the therapeutic target, is a critical character in chemical toxicological studies and in drug design. Compounds can cross the BBB by passive diffusion or by means of a variety of catalyzed transport systems that can carry compounds into the brain (carrier-mediated transport, receptor-mediated transcytosis) or out of the brain (active efflux). Various parameters are used for predicting BBB penetration such as CNS+/−, log *BB*, and log *PS*. CNS+/− is a qualitative property denoting the compound's activity (CNS+) or inactivity (CNS−) against a CNS target with its BBB penetration [1]. The problem with CNS+/− datasets is that CNS activity implies BBB permeation, while CNS inactivity might be due to factors other than nonpermeation, such as the fact that compounds might be rapidly metabolized or effluxed from the brain. Log *BB*, which is defined as logarithm of Brain/Blood partitioning ratio at steady state [2], is by far the most widely used parameter for BBB penetration. However, this parameter may also result in misleading conclusions because it ignores the main parts of process of permeability [3]. Log *PS*, which is defined as the logarithm of permeability-surface area product reflecting the rate of brain permeation, is superior to but more difficult to measure compared to log *BB* [4]. *In vivo* brain uptake methods may be the most reliable evaluation of BBB penetration. However, the low-throughput, expensive, and labor-intensive characteristics make these methods inapplicable in early drug discovery stages. For these reasons, *in vitro* and *in silico* methods have been introduced. As there is no one *in vitro* model which can mimic all properties of the *in vivo* BBB, developing more reliable models remains challenging [4]. So far, a great number of *in silico* BBB models have been developed and thoroughly reviewed [2, 3, 5–9]. Because of the high complex nature of the BBB, most computational models only account for passive diffusion.

Initial studies were focusing on making correlation between BBB permeability of small set of compounds and simple descriptors and then revealed "rules of thumb." These rough models reflect some important relationships between BBB penetration and properties of compounds but have a problem of oversimplification [10, 11]. As the accumulation of new data, various more sophisticated models were reported to predict BBB permeability. Classification models [12] which were used widely explored for distinguishing between the molecules capable of being across the BBB and those restricted to periphery. These models often developed by using the same dataset of about 1500 drugs compiled by Adenot and Lahana [13], which is the largest single homogeneous up-to-date source of qualitative data published. Some others [12, 14] distinguish molecules based on a certain log *BB* threshold. However, the main problem is the threshold which is subjectively determined and not unified. Most quantitative models were developed by building QSAR models [10, 15–18]. Since different datasets and validation methods were used, it is difficult to compare the performance of these models [19]. Recently, Carpenter et al. [20] developed a new model predicting the BBB penetration using molecular dynamic simulations and received good results, providing new thread of BBB permeability prediction. Here we focused on log *BB* models of BBB penetration by passive diffusion.

Various data mining methods have been employed in BBB penetration models, such as multiple linear regression [21, 22], partial least squares (PLS) regression [13, 23], recursive partitioning [23, 24], neural network [25–27], and support vector machine (SVM) [28–30]. SVM, which was originally developed by Vapnik and coworkers [31], has been extensively used and consistently achieves similar or superior performance compared to other machine learning methods [32]. Its main idea is to map data points to a high dimension space with a kernel function, and then these data points can be separated by a hyper plane.

For a successful SVM model, kernel parameters of SVM and feature subset selection are the two most important factors affecting the prediction accuracy. Various strategies have been adopted for the two problems. Grid-based algorithm is one of the most straightforward strategies for parameter optimization, which discretizes the parameters and then systematically searches every grid point to find a best combination of the parameters [33]. However, its use is limited due to the computational complexity and time-consumption. Gradient-based methods [34, 35] are also widely used, which require the kernel function and the scoring function differentiable to assess the performance of the parameters. Evolutionary method [36] has also been used and achieved promising results. As for the feature selection, genetic algorithms- (GA-) based [37–41], *F*-score based feature recursive elimination [42], and many other methods [43–47] have been employed. Most of these methods focus on feature selection or parameters optimization separately [45]. However, the choice of feature subset influences the appropriate kernel parameters and vice versa [48]. Hence the proper way seems to address the two problems simultaneously. GA [41], immune clonal algorithm (ICA) [49], and Bayesian approach [50] have been recently used for simultaneously feature selection and parameters optimization for SVM on general classification problems. In our study, GA was used to do parameter optimization and feature subset selection simultaneously, and an SVM regression model was developed for the blood-brain barrier penetration prediction.

2. Methods

The workflow used in this study for BBB penetration prediction is illustrated in Figure 1.

2.1. Dataset and Molecular Descriptors. The log *BB* dataset used in this study was compiled by Abraham et al. [51], which was a combination of both *in vivo* and *in vitro* data, including 302 substances (328 data points). Abraham et al. applied linear free energy relationship (LFER) to the dataset and obtained good correlation between log *BB* values and LFER descriptors plus two indicator variables [51]. CODESSA [52] could not calculate descriptors for the first 5 gases ([Ar], [Kr], [Ne], [Rn], and [Xe]) of the original dataset, and they were excluded from the dataset. The final dataset contained 297 compounds (323 data points). The indicator variables of I_v and AbsCarboxy used in Abraham's study [51] were retained in this study. I_v was defined as $I_v = 1$ for the *in vitro* data and $I_v = 0$ for the *in vivo* data. AbsCarboxy was an indicator for carboxylic acid (AbsCarboxy = 1 for carboxylic acid, otherwise AbsCarboxy = 0).

The initial structures in SMILES format were imported to Marvin [53] and exported in MDL MOL format. AM1 method in AMPAC [54] was used for optimization plus frequencies and thermodynamic properties calculation. The generated output files were used by CODESSA to calculate a large number of constitutional, topological, geometrical, electrostatic, quantum-chemical, and thermodynamic descriptors. Marvin was also used to calculate some physicochemical properties of the compound, including log *P*, log *D*, polar surface area (PSA), polarizability, and refractivity. All these descriptors and properties were used as candidate features in later modeling.

Features with missing values or having no change across the data set were removed. If the correlation coefficient of two features is higher than a specified cutoff value (0.999999 used here), then one of them is randomly chosen and removed. The cutoff value used here is very high because very high variable correlation does not mean absence of variable complementarity [55]. A total number of 326 descriptors were left for further analysis. However, many highly correlated features have very similar physicochemical meanings. In our final analysis, similar features were put together by their physicochemical meaning, which we hope could unveil some underlying molecular properties that determine the BBB penetration.

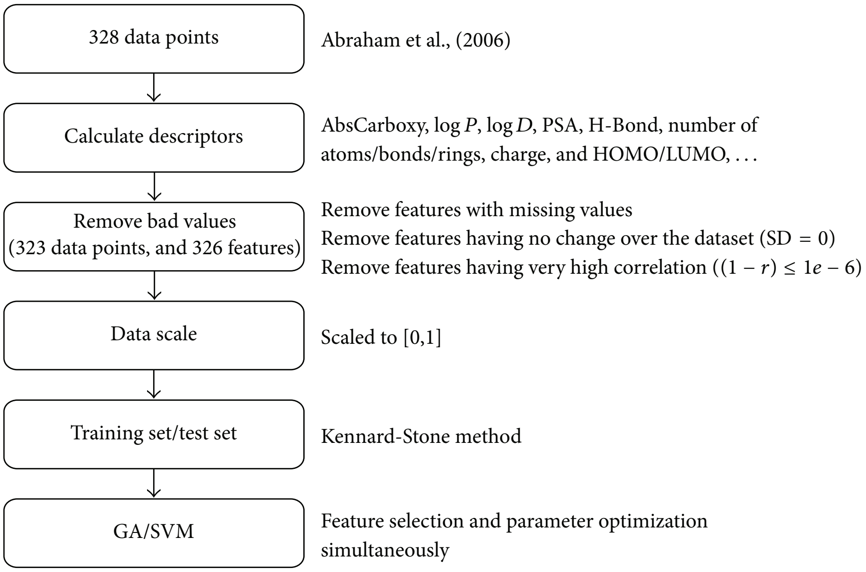

328 data points
Abraham et al., (2006)
Calculate descriptors
AbsCarboxy, log P, log D, PSA, H-Bond, number of atoms/bonds/rings, charge, and HOMO/LUMO, . . .
Remove bad values (323 data points, and 326 features)
Remove features with missing values
Remove features having no change over the dataset (SD = 0)
Remove features having very high correlation ((1 − r) ≤ 1e − 6)
Data scale
Scaled to [0,1]
Training set/test set
Kennard-Stone method
GA/SVM
Feature selection and parameter optimization simultaneously

Figure 1: Workflow of GA/SVM model for BBB penetration prediction.

The dataset was then split into training set and test set using the Kennard-Stone method [56], which selects a subset of representative data points uniformly distributed in the sample space [57]. At start, the Kennard-Stone method chooses the data point that is the closest to the center of the dataset measured by Euclidean distance. After that, from all remaining data points, the data point that is the furthest from those already selected is added to the training set. This process continues until the size of the training set reaches specified size. 260 data points were selected as training set and the other 63 were used as test set.

2.2. SVM Regression. Details about SVM regression can be found in literatures [58–60]. As in other multivariate statistical models, the performance of SVM regression depends on the combination of several parameters. In general, C is a regularization parameter that controls the tradeoff between training error and model complexity. If C is too large, the model will have a high penalty for nonseparable points and may store too many support vectors and get overfitting. If it is too small, the model may have underfitting. Parameter ε controls the width of the ε-insensitive zone, used to fit the training data. The value of ε can affect the number of the support vectors used to construct the regression function. The bigger ε is, the fewer support vectors are selected. On the other hand, bigger ε-values result in more flat estimates. Hence, both C and ε-values affect model complexity (but in a different way). The kernel type is another important parameter. In SVM regression, radial basis function (RBF) (1) was the most commonly used kernel function for its better generalization ability, less number of parameters, and less numerical difficulties [33] and was used in this study. Parameter γ in RBF controls the amplitude of the RBF kernel and therefore controls the generalization ability of SVM regression. The LIBSVM package (version 2.81) [61] was used in this study for SVM regression calculation, taking the form

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \quad \gamma > 0, \tag{1}$$

where x_i and x_j are training vectors ($i \neq j$, $x_i \neq x_j$) and γ is kernel parameter.

2.3. Genetic Algorithms. Genetic algorithms (GA) [41] are stochastic optimization and search method that mimics biological evolution as a problem-solving strategy. They are very flexible and attractive for optimization problems.

Given a specific problem to solve, the input to the GA is a set of potential solutions to that problem, encoded in some fashion, and a fitness function that allows each candidate to be quantitatively evaluated (Figure 2). Selection, mating, and mutation just mimic the natural process. For each generation, individuals are selected for reproduction according to their fitness values. Favorable individuals have a better chance to be selected for reproduction and the offspring have chance to mutate to keep diversity, while the unfavorable individuals are less likely to survive. After each generation, whether the evolution is converged or the termination criteria are met is checked; if yes, job is done; if not, the evolution goes into next generation. After many generations, good individuals will dominate the population, and we will get solutions that are good enough for our problem.

First, in order to solve a problem with GA, each individual in the population should be represented by a chromosome. In our study, since the parameter optimization and feature subset selection should be addressed simultaneously, the chromosome is a combination of parameter genes and feature gene (Figure 3), where f_n is an integer in the range of $[1, N]$ and N is the number of candidate features for model construction. A chromosome represents an individual in genetic algorithms and parameters contained in chromosome could be used for SVM modeling. Left part of the chromosome is the parameter genes, of which C, γ, and ε all are float genes.

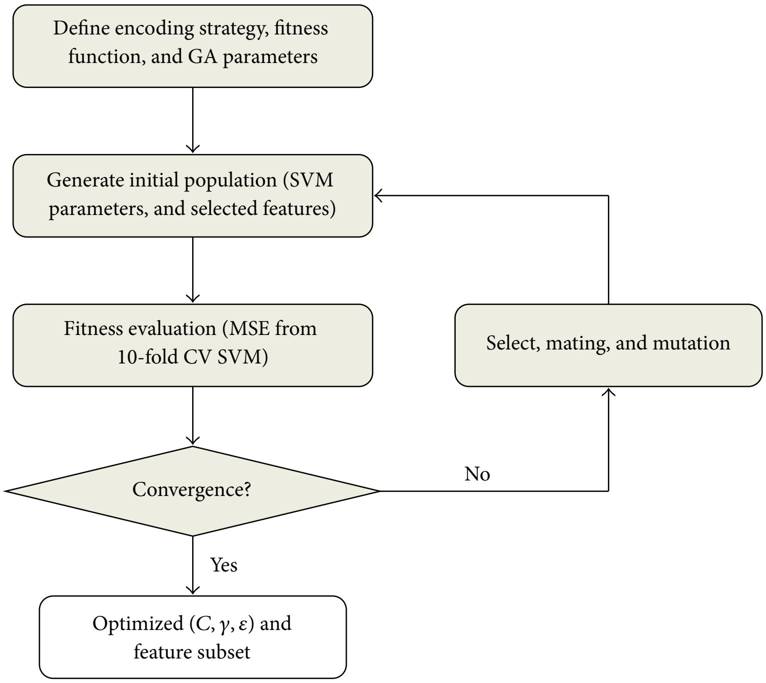

Define encoding strategy, fitness function, and GA parameters
Generate initial population (SVM parameters, and selected features)
Fitness evaluation (MSE from 10-fold CV SVM)
Select, mating, and mutation
Convergence?
No
Yes
Optimized (C, γ, ε) and feature subset

Figure 2: Workflow of genetic algorithms.

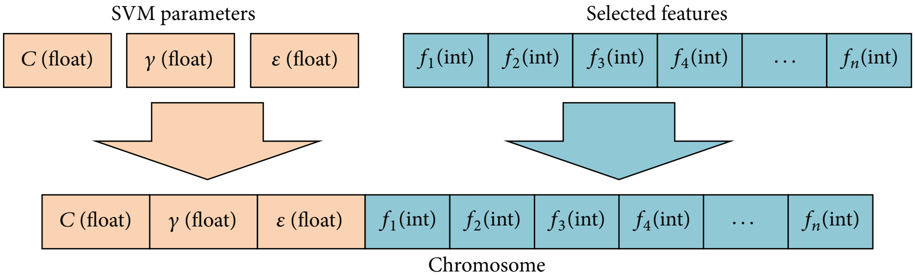


Figure 3: Encoding of the chromosome.

The feature gene is an array of integers, and each integer represents a feature.

Fitness function can be seen as a ruler, which was used to quantitatively evaluate and compare each candidate. In our study, the mean squared error (MSE) of 10-fold cross validation (CV) for SVM was used as fitness function, and smaller fitness value indicated better individual. Given a training set containing n compounds, $(x_1, y_1), \ldots, (x_n, y_n)$, x_i is descriptor vector of compound i and $x_i \in R^n$. y_i is the log BB value and $y_i \in \{-1, +1\}$. The objective function can be calculated by

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b. \quad (2)$$

α_i and α_i^* are Lagronia factors, $K(x_i, x)$ is kernel of radial basis function. The MSE of 10-fold CV for SVM was calculated by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{X_i} - X_i\right)^2, \quad (3)$$

where n is the number of all data points, $\widehat{X_i}$ is the predicted value, and X_i is the experiment value.

Tournament selection was used as the selection strategy in GA, which selected the best 1 from 3 randomly chosen candidates. The advantage of tournament selection over roulette wheel selection is that tournament selection does not need to sort the whole population by fitness value.

Since there are different types of genes in a chromosome, different mating strategies were used for different types of genes (Figure 4):

$$V_{\text{new}} = \beta p_1 + (1 - \beta) p_2, \quad (4)$$

where β uniformly distributed random number on the interval $[-0.25, 1.25]$, p_n is the value of parent gene, and V_{new} is the value of child gene.

For float genes, the new value is a linear combination of the parents (4). For feature gene, uniform crossover is used: each element of the child gene is selected randomly from the corresponding items of parents.

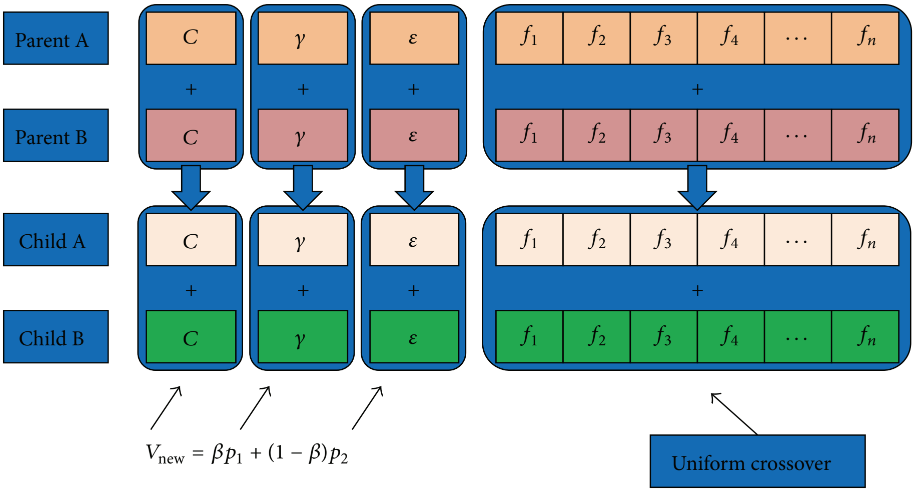


FIGURE 4: Mating strategy of GA.

TABLE 1: Performance comparison of models with different number of features.

Number of features	Training (CV = 10)		Prediction/r^2		Parameters of SVM		
	MSE	r^2	Test set	Training set	C	γ	ε
4	0.1197	0.674	0.722	0.740	38.8833	0.6081	0.1491
5	0.1042	0.715	0.770	0.805	16.3419	0.7973	0.2743
6	0.0945	0.744	0.840	0.829	13.3573	0.7158	0.1513
7	0.0959	0.74	0.821	0.843	34.3067	0.5218	0.1595
8	0.0883	0.761	0.834	0.883	60.9596	0.5871	0.2357
9	0.0815	0.777	0.847	0.864	3.7770	0.8764	0.1663
10	0.0823	0.776	0.858	0.903	15.2236	0.6247	0.1434
11	0.0714	0.804	0.861	0.891	5.6937	0.6531	0.1573
12	0.0780	0.787	0.864	0.905	7.2787	0.7428	0.1515
13	0.0817	0.778	0.862	0.922	4.1957	0.7791	0.1574
14	0.0812	0.778	0.882	0.917	14.8391	0.5002	0.2054
15	0.0734	0.799	0.870	0.919	4.9915	0.5231	0.1077

Again, different mutation strategies were used for different types of genes. For float genes, the values were randomly mutated upward or downward. The new value was given by

$$V_{\text{new}} = \begin{cases} V - \beta\left(V - V_{\min}\right) & \text{if random}\,() < 0.5 \\ V + \beta\left(V_{\max} - V\right) & \text{if random}\,() \geq 0.5, \end{cases} \quad (5)$$

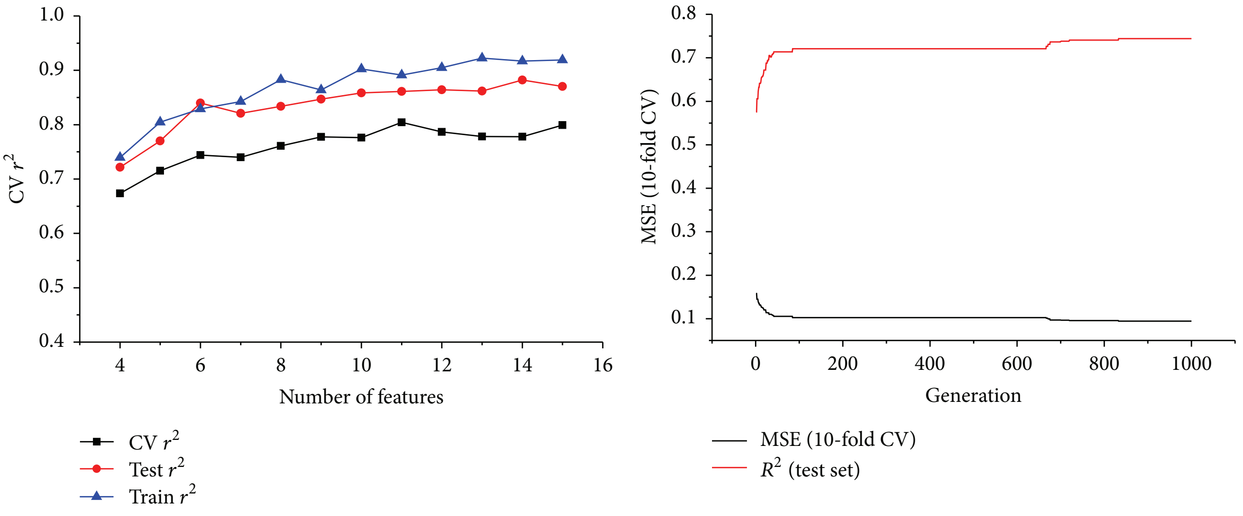
where β was a random number distributed in $[0,1]$, V and V_{new} are values before and after mutation, and $V_{\min}$ and $V_{\max}$ are the minimum and maximum values allowed for a gene.

For feature gene, several points were first randomly chosen for mutation, and then a random number in $[1, N]$ (N is the total number of features) was chosen as new feature while avoiding duplicate features. The GA was terminated when the evolution reached 1000 generations. In our pilot study (data not shown), 1000 generations were enough for the GA to converge. The other parameters for GA were as follows: population size 100, cross rate 0.8, mutation rate 0.1, elite size 2, and number of new individuals in each generation 8.

3. Results and Discussion

3.1. GA/SVM Performance. GA was run with different number of features from 4 to 15. For each number of features, GA was run 50 times, and the best model was chosen for further analysis. From Table 1 and Figure 5, the overall trend of the GA showed the following: (1) the accuracy of the model increased with the number of the features; (2) the accuracy of the model on training set was better than the accuracy on the test set, which was then better than the accuracy of cross validation.

As the feature number increases, the complexity also increases, which will often increase the probability of overfitting. A complex model is also difficult to interpret and apply in practical use, so generally speaking, we need to find a balance between the accuracy and complexity of the model. It is observed that (Table 1, Figure 5(a)) the prediction accuracy ($r^2 = 0.744$) of cross validation ($n = 10$) of the 6-feature model was similar to that of the Abraham's model ($r^2 = 0.75$) which used all 328 data points (Table 2). As the number of features increased from 6 to 15, the prediction accuracy



(a)

(b)

Figure 5: (a) Performance comparison of models with different number of features. (b) Evolution of the best 6-feature model.

Table 2: Comparison of most relevant QSAR studies on BBB permeability.

Descriptors	N_{train}	N_{test}	Methods	r_{train}^2	Predictive accuracy on test set	Reference
Δlop P, log P,and log P_{cyc}	20	—	Linear Regression	0.69	—	Young et al. [77]
Excess molar refraction, dipolarity/polarisability, H-bond acidity, and basicity Solute McGowan volume	148	30	LFER	0.75	$r_{\text{test}}^2 = 0.73$	Platts et al. [66]
ΔG°_{W}	55	—	Linear Regression	0.82	—	Lombardo et al. [78]
PSA, the octanol/water partition coefficient, and the conformational flexibility	56	7	MLR	0.85	$r_{\text{test}}^2 = 0.80$	Iyer et al. [79]
CODESSA/DRAGON (482)	200	110	PLS SVM	0.83 0.97	$r_{\text{test}}^2 = 0.81$ $r_{\text{test}}^2 = 0.96$	Golmohammadi et al. [62]
Molecular (CODESSA-PRO) descriptors (5)	113	19	MLR	0.78	$r_{\text{test}}^2 = 0.77$	Katritzky et al. [15]
Molecular fragment (ISIDA) descriptors	112	19	MLR	0.90	$r_{\text{test}}^2 = 0.83$	Katritzky et al. [15]
PSA, log P, the number of H-bond acceptors, E-state, and VSA	144	10	Combinatorial QSAR (KNN SVM)	0.91	$r_{\text{test}}^2 = 0.8$	Zhang et al. [17]
Abraham solute descriptors and indicators	328	—	LFER	0.75	—	Abraham et al. [51]
Abraham solute descriptors and indicators	164	164	LFER	0.71	$s = 0.25$, MAE $= 0.20$	Abraham et al. [51]
CODESSA/Marvin/indicator (6)	260	63	GA based SVM	0.83	$r_{\text{test}}^2 = 0.84$, RMSE $= 0.23$	This research, GA/SVM, final model $C = 13.3573$, $\gamma = 0.715761$, $\varepsilon = 0.151289$
CODESSA/Marvin/indicator (236)	260	63	GA based SVM	0.97	$r_{\text{test}}^2 = 0.55$, RMSE $= 0.31$	This research, Grid/SVM $C = 8.0$, $\gamma = 0.015625$, $\varepsilon = 0.0625$
CODESSA/Marvin/indicator (6)	260	63	GA based SVM	0.86	$r_{\text{test}}^2 = 0.58$, RMSE $= 0.29$	This research, Grid/SVM $C = 8.0$, $\gamma = 1.0$, $\varepsilon = 0.125$

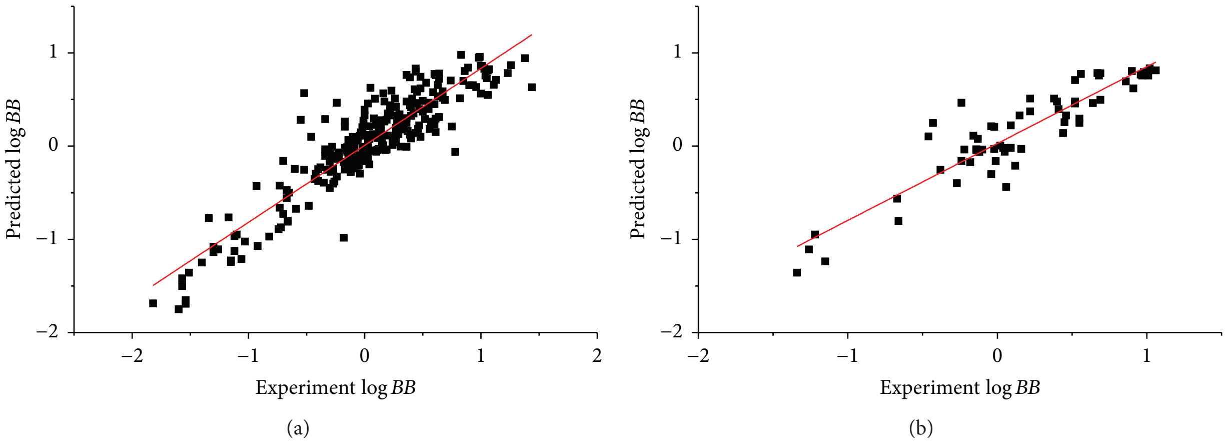


Figure 6: Prediction accuracy of the final model on training set (a) and test set (b).

on training set increased from 0.829 to 0.919, while the prediction accuracy on test set only slightly increased from 0.840 to 0.870 accordingly. Take all these into consideration, the 6-feature model (Table 1, Figure 5(a)) was chosen as our final model (Figure 6), of which the prediction accuracy on both test set and training set was similar ($r_{\text{train}}{}^2 - r_{\text{test}}{}^2 = 0.11$) and high enough (>0.82).

Figure 5(b) showed the evolution of the prediction performance of the model (MSE and $r_{\text{test}}{}^2$). In the first 100 generations, the MSE decreased very fast, followed with a platform stage from about 100 to 650 generations. Another decrease occurred at about 650 generations, and then the evolution became stable at about 850 generations. The models did not improve much in the last 150 generations, which may imply a convergence.

A tabular presentation of relevant studies regarding the prediction of the blood-brain distribution is shown in Table 2. These models were constructed by using different statistical learning methods, yielding different prediction capability with $R_{\text{test}}{}^2$ ranging from 0.5 to over 0.9. Generally, regression by SVM appears to be more robust than traditional linear approaches such as PLS and MLR, with respect to the nonlinear effects induced by multiple potentially cooperative factors governing the BBB permeability. For example, the SVM model by Golmohammadi et al. [62] yielded the highest $R_{\text{test}}{}^2$ on a test set containing 110 molecules. However, it should be noted that direct comparison with results from previous studies is usually inappropriate because of differences in their datasets. In this study, a combination of both *in vivo* and *in vitro* data compiled by Abraham et al. [51] was used for developing BBB prediction model, which is of high data quality and covers large chemical diversity space. In addition to the data source, kernel parameter optimization and feature selection are two crucial factors influencing the prediction accuracy of SVM models. To reduce the computational cost, most of the existing models addressed the feature selection and parameter optimization procedures separately. In this study, we used a GA scheme to perform the kernel parameter optimization and feature selection simultaneously, which is more efficient at searching the optimal feature subset space.

Abraham's model [51] is the best model that is currently publicly available. A comparison of our models with Abraham's models was shown in Table 2. In Table 2, the last 3 rows are models in our study. The same dataset was used in Abraham's research and this study, but the data set was split into different training set and test set (our model: train/test = 260/63; Abraham's model: training/test = 164/164). 7 variables were used in Abraham's model, compared to 6 in our final model. The r^2 values for training set in Abraham's 164/164 model and 328/0 model were 0.71 and 0.75, respectively, compared with 0.83 for our model. It has to be noted that the size of our training set (260) is bigger than Abraham's (164).

Our model was also compared with grid method implemented with Python toolkit (grid.py) shipped with libsvm [61] for parameter optimization. First, since grid method cannot be used to select feature subset, all 326 features were used to construct a BBB prediction model. The prediction accuracy of the training set was very high ($r_{\text{train}}{}^2 = 0.97$) but that of the test set was disappointing ($r_{\text{test}}{}^2 = 0.55$). Then we used the same feature set as our final model (6 features). The result was slightly better, but still too bad for test set prediction ($r_{\text{train}}{}^2 = 0.86$, $r_{\text{test}}{}^2 = 0.58$).

So compared with the grid-based method, our GA-based method could get better accuracy with fewer features, which suggested that GA could get much better combination of parameters and feature subset. This was also observed in other's study [48].

3.2. Feature Analysis. An examination of the descriptors used in the model could provide an insight into the molecular properties that are most relevant to BBB penetration. Table 3 showed the features used in the final 6-feature model and their meanings. In order to explore the relative importance and the underlying molecular properties of the descriptors, the most frequently used features in all 50 6-feature models were analyzed. Table 4 showed the top 10 most frequently used features. Interestingly, AbsCarboxy, an indicator of the existence of carboxylic acid was the most significant property. Some features with similar meaning were also found to occur in the models, such as PSA related features (M_PSA_7.4,

Table 3: Features used in the final model.

Name	Meaning
M_log *P*	log *P* (Marvin)
HA_dependent_HDSA-2_[Zefirov's_PC]	H-bond donor surface area related (CODESSA)
M_PSA_7.4	PSA at pH 7.4 (Marvin)
AbsCarboxy	Carboxylic acid indicator (Abraham)
HA_dependent_HDCA-2/SQRT(TMSA)_[Zefirov's_PC]	H-bond donor charged area related (CODESSA)
Average_Complementary_Information_content_(order_0)	Topology descriptor (CODESSA)

Table 4: The most frequently used features for all 6-feature models[a].

Number	Feature name	Occurrence (50 models)	Meaning
11	AbsCarboxy	36	Indicator for carboxylic acid[†]
268	ESP-FHASA_Fractional_HASA_(HASA/TMSA)_Quantum-Chemical_PC	14	H-acceptor surface area/total molecular surface area[#]
101	Topographic_electronic_index_(all_bonds)_Zefirov's_PC	12	Topological electronic index for all bonded pairs of atoms[b‡]
8	M_PSA_7.4	11	PSA at pH 7.4[c$]
267	ESP-HASA_H-acceptors_surface_area_Quantum-Chemical_PC	10	H-acceptor surface area[#]
5	delta_log *D*	9	log *D* (pH 6.5) – log *D* (pH 7.4)[d,e∧]
7	M_PSA_7.0	9	PSA at pH 7.0[$]
138	HA_dependent_HDCA-2_[Zefirov's_PC]	9	H-donors charged surface area[#]
6	M_PSA_6.5	8	PSA at pH 6.5[$]
1	M_log *P*	7	log *P*[∧]

[a] Rows with the same symbol could be categorized into the same group.

[b] Topological electronic index is a feature to characterize the distribution of molecular charge: $T = \sum_{(i<j)}^{N_B}(|q_i - q_j|/r^2{}_{ij})$, where q_i is net charge on ith atom and r_{ij} is the distance between two bonded atoms.

[c] 7.4 is the pH in blood.

[d] 6.5 is the pH in intestine.

[e] *D* is the ratio of the sum of the concentrations of all species of a compound in octanol to the sum of the concentrations of all species of the compound in water. For neutral compounds, log *D* is equal to log *P*.

M_PSA_7.0, and M_PSA_6.5), and H-bond related descriptors (number 267, 268, and 138). So we decided to put similar descriptors into the same group (Figure 7) to find the underlying properties affecting BBB penetration.

According to the associated molecular properties of the features, those frequently used features are categorized into 5 groups: AbsCarboxy (indicator of carboxylic acid), H-bonding (H-bonding ability, including H-bond donor/acceptor related features), PSA (molecular polar surface area related features), lipophilicity (including M_log *P* and delta_log *D*), and molecular charge (including charge and topological electronic index related features).

The following is observed:

(1) Interestingly, AbsCarboxy was also the most significant feature, which occurred 36 times in total 50 models. This may indicate that the carboxylic acid group plays an important role in the BBB penetration, which is consistent with the study of Abraham et al. [51].

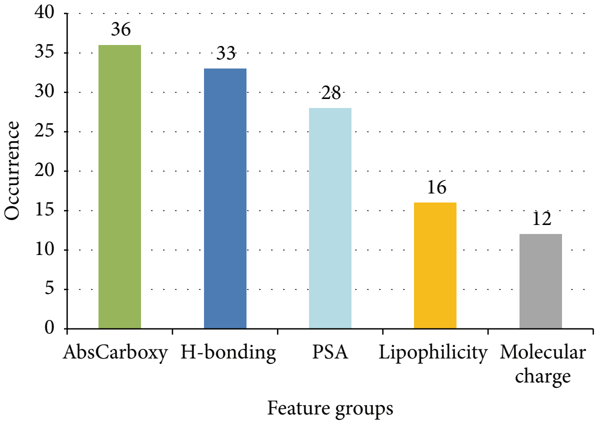


Figure 7: Top features for all 6-feature models (50 in all).

(2) H-bonding (H-bond donor/acceptor) related surface area, polar surface area, log *P* related features, and

Table 5: Most frequently used features for all top models (number of features range from 4 to 15)*.

Number	Descriptor name	Occurrence (120 models)	Meaning
11	AbsCarboxy	84	Indicator for carboxylic acid†
5	delta_log *D*	35	log *D* (pH 6.5) – log *D* (pH 7.4)$^{\wedge}$
7	M_PSA_7.0	32	PSA at pH 7.0$^{\$}$
1	M_log *P*	28	log *P*$^{\wedge}$
138	HA_dependent_HDCA-2_[Zefirov's_PC]	27	H-donors charged surface area$^{\#}$
268	ESP-FHASA_Fractional_HASA_(HASA/TMSA)_Quantum-Chemical_PC	27	H-acceptor surface area/total molecular surface area$^{\#}$
6	M_PSA_6.5	25	PSA at pH 6.5$^{\$}$
167	PPSA-1_Partial_positive_surface_area_[Quantum-Chemical_PC]	25	Partial positive surface area$^{\$}$
101	Topographic_electronic_index_(all_bonds)_Zefirov's_PC	23	Topological electronic index for all bonded pairs of atoms‡
8	M_PSA_7.4	21	PSA at pH 7.4$^{\$}$

*Rows with the same symbol could be categorized into the same group.

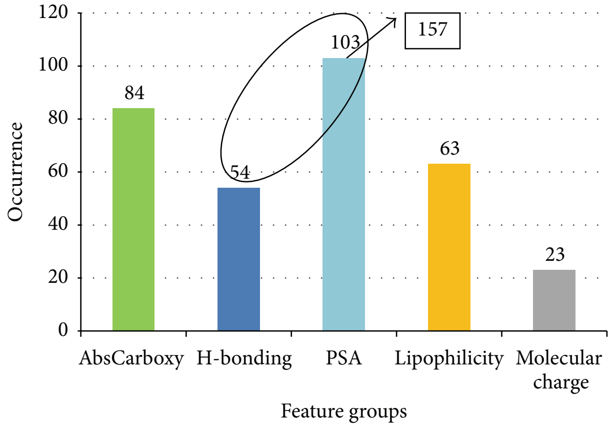


Figure 8: The most frequently used features for all top models.

topological electronic index related features are also significant.

In order to further confirm the previous finding, all top 10 models from features = 4 to 15 were further analyzed, and the result was almost the same (Table 5, Figure 8). The top 10 most frequently used feature sets were very similar. Compared to the 6-feature models, the only difference was that H-bond related feature number 267 is replaced with PSA related feature number 167.

Again, the descriptors were analyzed by group. The composites of the groups were almost the same. Given that there are 4 descriptors in PSA group, AbsCarboxy was also the most significant property, followed by PSA, lipophilicity, and H-bonding having similar occurences, then followed by molecular charge with relatively low frequencies. The high consistency suggested that these groups of features had signficant impact on the BBB penetration ability.

PSA and H-bonding descriptors are highly relevant properties: PSA is the molecular areas contributed by polar atoms (nitrogen, sulphur, oxygen, and phosphorus), and most of the time these polar atoms can be H-bond acceptor or donor. If PSA and H-bonding were merged into one group (n = 157), they will become the most significant property group of features.

3.3. Properties Relevant to BBB Penetration

3.3.1. Carboxylic Acid Group. It was proposed by Abraham et al. [51] that carboxylic acid group played an important role in BBB penetration. While it was commonly believed that the most important molecular properties related to BBB penetration were H-bonding ability, lipophilicity, and molecular charge [63, 64]. However, our study confirmed Abraham's conclusion and showed that the importance of carboxylic acid group in BBB penetration could be underestimated.

In the models by Abraham et al. [51], the indicator variable of carboxylic acid group has the largest negative coefficient, indicating its importance in BBB penetration, and is consistent with observations in our model that the indicator of carboxylic acid group is the most frequently used descriptor. Zhao et al. [23] tried to classify compounds into BBB positive or BBB negative groups using H-bonding related descriptors, and the indicator of carboxylic acid group was also found to be important in their model. Furthermore, our results are consistent with the fact that basic molecules have a better BBB penetration than the acid molecules [10].

The carboxylic acid group may affect the BBB penetration through molecular charge interactions since in most cases the carboxylic acid group will exist in the ionized form carrying a negative charge. The carboxylic acid group could also affect the BBB penetration by forming H-bond with BBB and hence weaken the BBB penetration abilities of molecules.

Abraham et al. suggested that the presence of carboxylic acid group which acted to hinder BBB penetration was not only simply due to the intrinsic hydrogen bonding and polarity properties of neutral acids [51]. There were some other ways in which the carboxylic acid groups could affect the BBB penetration, such as acidic drugs which could bind to albumin [65], the ionization of the carboxylic acid groups which could increase the excess molar refraction and hydrogen bonding basicity, and finally the carboxylic acid groups which may be removed from brain by some efflux mechanism [66].

3.3.2. Polar Surface Area and H-Bonding Ability. As pointed out in our previous analysis, PSA and H-bonding ability are actually two highly correlated properties. If these two groups are merged, they will be the most significant group of properties, even more significant than the carboxylic acid indicator (Figure 8). Norinder and Haeberlein [6] concluded that hydrogen bonding term is a cornerstone in BBB penetration prediction. In Zhao et al.'s study [23], PSA, AbsCarboxy, number of H-bonding donors, and positively charged form fraction at pH 7.4 were all treated as H-bonding descriptors and were found to be important in the final model.

Furthermore, almost all published models make use of molecular polar and/or H-bonding ability related descriptors, such as PSA [23, 67, 68], high-charged PSA [22], number of hydrogen donors and acceptors [23, 69], and hydrogen bond acidity/basicity [23]. And, in these models, PSA and/or H-bond ability are all negatively correlated with $\log BB$, which is in agreement with Abraham et al.'s study [51] in which the coefficients of the H-bond acidity and H-bond basicity are both negative.

After review of many previous works, Norinder and Haeberlein [6] proposed that if the sum of number of nitrogen and oxygen atoms (N + O) in a molecule was five or less, it had a high chance of entering the brain. As we all know, nitrogen and oxygen atoms have great impact on PSA and H-bonding. Norinder and Haeberlein [6] also concluded that BBB penetration could be increased by lowering the overall hydrogen bonding ability of a compound, such as by encouraging intramolecular hydrogen bonding. After an analysis of the CNS activity of 125 marketed drugs, van de Waterbeemd et al. [70] suggested that the upper limit for PSA in a molecule that is desired to penetrate the brain should be around 90 Å^2, while Kelder et al. [71] analyzed the PSA distribution of 776 orally administered CNS drugs that have reached at least phase II studies and suggested that the upper limit should be 60–70 Å^2.

Having in mind that molecules mainly cross the BBB by passive diffusion, we think it may be because molecules with strong H-bonding ability have a greater tendency to form H-bonds with the polar environment (the blood), hence weakening their ability to cross the BBB by passive diffusion.

3.3.3. Lipophilicity. Lipophilicity is another property widely recognized as being important in BBB penetration, and most of the current models utilize features related to lipophilicity [22, 64, 67, 72]. Lipophilicity was thought to be positively correlated with $\log BB$; that is, increase the lipophilicity of a molecule will increase the BBB penetration of the molecule. Norinder and Haeberlein [6] also proposed that if $\log P-(\mathrm{N}+\mathrm{O}) > 0$, then $\log BB$ was positive. Van de Waterbeemd et al. [70] suggested that the $\log D$ of the molecule should be in [1, 3] for good BBB penetration. These observations are consistent with the fact that the lipid bilayer is lipophilic in nature, and lipophilic molecules could cross the BBB and get into the brain more easily than hydrophobic molecules.

3.4. Molecular Charge. From the viewpoint of computational chemistry, the distribution of molecular charge is a very important property that affects the molecule properties greatly. It is the uncharged form that can pass the BBB by passive diffusion. Fischer et al. [73] have shown that acid molecules with $\mathrm{p}K_\mathrm{a} < 4$ and basic molecules with $\mathrm{p}K_\mathrm{a} > 10$ could not cross the BBB by passive diffusion. Under physiological conditions (pH = 7.4), acid molecules with $\mathrm{p}K_\mathrm{a} < 4$ and basic molecules with $\mathrm{p}K_\mathrm{a} > 10$ will be ionized completely and carry net charges. As mentioned in our previous analysis, the carboxylic acid group may affect the BBB penetration through molecular charge interactions. Mahar Doan et al. [74] compared physicochemical properties of 93 CNS ($n = 48$) and non-CNS ($n = 45$) drugs and showed that 0 of 48 CNS drugs have a negative charge and CNS drugs tend to have less positive charge. These are reasonably consistent with the study of Abraham et al. [51] in which the coefficient of the carboxylic acid indicator is negative.

It has to be noted that, all these molecular properties are not independent, and they are related to each other. For example, the carboxylic acid group is related to both PSA and H-bonding ability, for the O atom in the carboxylic acid group is a polar atom and has a strong ability to form H-bonds; the carboxylic acid group, which carries charge under most conditions, is also related to molecular charge. The PSA/H-bonding ability is also correlated to molecular charge, because in many cases atoms could contribute to PSA or form H-bonds which could probably carry charges. Lipophilicity is also related to molecular charge.

We can get a conclusion that the most important properties for a molecule to penetrate BBB are carboxylic acid group, PSA/H-bonding ability, lipophilicity, and charge. BBB penetration is positively correlated with the lipophilicity and negatively correlated with the other three properties. A comparison of the physicochemical properties of 48 CNS drugs and 45 non-CNS suggested that compared to non-CNS drugs, CNS drugs tend to be more lipophilic and more rigid and have fewer hydrogen-bond donors, fewer charges, and lower PSA (<80 Å^2) [74], which is in reasonable consistency with our finding except that the molecular flexibility is not important in our model.

There are some other properties utilized in some existing models, such as molecular weight, molecular shape, and molecular flexibility. It is suggested by van de Waterbeemd et al. [70] that molecular weight should be less than 450 for good BBB penetration, while Hou and Xu [22] suggested that the influence of molecular bulkiness would be obvious when the size of the molecule was larger than a threshold and found that molecular weight made a negative contribution to the BBB penetration when the molecular weight is greater 360. This is not widely observed in other studies. In Zhao et al.'s study [23], molecular weight was found to be not important compared to hydrogen bond properties.

Lobell et al. [68] proposed that spherical shapes have a small advantage compared with rod-like shapes with regard to BBB penetration, they attributed this to the membranes that are largely made from rod-shaped molecules and rod-like shape may become more easily trapped within membrane without exiting into the brain compartment. However, in Rose et al.'s model [75] based on electrotopological state descriptors showed that BBB penetration increased with less sketch branching. Crivori et al. [76] tried to correlate descriptors derived from 3D molecular fields and BBB penetration and concluded that the size and shape descriptors had no marked impact on BBB penetration.

Iyer et al. [67] found that increasing the solute conformational flexibility would increase log *BB*, while in the study of Mahar Doan et al. [74], CNS drugs tend to be more rigid. However, the roles of molecular weight, molecular shape, and molecular flexibility in BBB penetration seem to be still unclear and not well received. Further studies are still needed.

4. Conclusion

In this study, we have developed a GA/SVM model for the BBB penetration prediction, which utilized GA to do kernel parameters optimization and feature selection simultaneously for SVM regression. The results showed that our method could get better performance than addressing the two problems separately. The same GA/SVM method can be extended to be used on other QSAR modeling applications.

In addition, the most important properties (carboxylic acid group, PSA/H-bond ability, lipophilicity, and molecular charge) governing the BBB penetration were illustrated through analyzing the SVM model. The carboxylic acid group and PSA/H-bond ability have the strongest effect. The existence of carboxylic acid group (AbsCarboxy), PSA/H-bonding and molecular charge is all negatively correlated with BBB penetration ability, while the lipophilicity enhances the BBB penetration ability.

The BBB penetration is a highly complex process and is a result of many cooperative effects. In order to clarify the factors that affect the BBB penetration, further efforts are needed to investigate the mechanistic nature of the BBB, and, as pointed out by Goodwin and Clark [7], the most fundamental need is for more high quality data, both *in vivo* and *in vitro*, upon which the next generation of predictive model can be built.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Daqing Zhang, Jianfeng Xiao, and Nannan Zhou contributed equally to this work.

Acknowledgments

The authors gratefully acknowledge financial support from the Hi-Tech Research and Development Program of China (Grant 2014AA01A302 to Mingyue Zheng and Grant 2012AA020308 to Xiaomin Luo), National Natural Science Foundation of China (Grants 21210003 and 81230076 to Hualiang Jiang and Grant 81430084 to Kaixian Chen), and National S&T Major Project (Grant 2012ZX09301-001-002 to Xiaomin Luo).

References

[1] K. Lanevskij, J. Dapkunas, L. Juska, P. Japertas, and R. Didziapetris, "QSAR analysis of blood-brain distribution: the influence of plasma and brain tissue binding," *Journal of Pharmaceutical Sciences*, vol. 100, no. 6, pp. 2147–2160, 2011.

[2] K. Lanevskij, P. Japertas, and R. Didziapetris, "Improving the prediction of drug disposition in the brain," *Expert Opinion on Drug Metabolism & Toxicology*, vol. 9, no. 4, pp. 473–486, 2013.

[3] A. R. Mehdipour and M. Hamidi, "Brain drug targeting: a computational approach for overcoming blood-brain barrier," *Drug Discovery Today*, vol. 14, no. 21-22, pp. 1030–1036, 2009.

[4] J. Bicker, G. Alves, A. Fortuna, and A. Falcão, "Blood-brain barrier models and their relevance for a successful development of CNS drug delivery systems: a review," *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 87, no. 3, pp. 409–432, 2014.

[5] K. Nagpal, S. K. Singh, and D. N. Mishra, "Drug targeting to brain: a systematic approach to study the factors, parameters and approaches for prediction of permeability of drugs across BBB," *Expert Opinion on Drug Delivery*, vol. 10, no. 7, pp. 927–955, 2013.

[6] U. Norinder and M. Haeberlein, "Computational approaches to the prediction of the blood-brain distribution," *Advanced Drug Delivery Reviews*, vol. 54, no. 3, pp. 291–313, 2002.

[7] J. T. Goodwin and D. E. Clark, "In silico predictions of blood-brain barrier penetration: considerations to 'keep in mind'," *Journal of Pharmacology and Experimental Therapeutics*, vol. 315, no. 2, pp. 477–483, 2005.

[8] J. A. Nicolazzo, S. A. Charman, and W. N. Charman, "Methods to assess drug permeability across the blood-brain barrier," *Journal of Pharmacy and Pharmacology*, vol. 58, no. 3, pp. 281–293, 2006.

[9] L. Di, E. H. Kerns, and G. T. Carter, "Strategies to assess blood-brain barrier penetration," *Expert Opinion on Drug Discovery*, vol. 3, no. 6, pp. 677–687, 2008.

[10] Y. Fan, R. Unwalla, R. A. Denny et al., "Insights for predicting blood-brain barrier penetration of CNS targeted molecules using QSPR approaches," *Journal of Chemical Information and Modeling*, vol. 50, no. 6, pp. 1123–1133, 2010.

[11] M. H. Abraham, "The factors that influence permeation across the blood-brain barrier," *European Journal of Medicinal Chemistry*, vol. 39, no. 3, pp. 235–240, 2004.

[12] I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, "A bayesian approach to in silico blood-brain barrier penetration modeling," *Journal of Chemical Information and Modeling*, vol. 52, no. 6, pp. 1686–1697, 2012.

[13] M. Adenot and R. Lahana, "Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates," *Journal of Chemical*

Information and Computer Sciences, vol. 44, no. 1, pp. 239–248, 2004.

[14] M. Muehlbacher, G. M. Spitzer, K. R. Liedl, and J. Kornhuber, "Qualitative prediction of blood-brain barrier permeability on a large and refined dataset," *Journal of Computer-Aided Molecular Design*, vol. 25, no. 12, pp. 1095–1106, 2011.

[15] A. R. Katritzky, M. Kuanar, S. Slavov et al., "Correlation of blood-brain penetration using structural descriptors," *Bioorganic & Medicinal Chemistry*, vol. 14, no. 14, pp. 4888–4917, 2006.

[16] O. Obrezanova, J. M. R. Gola, E. J. Champness, and M. D. Segall, "Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility," *Journal of Computer-Aided Molecular Design*, vol. 22, no. 6-7, pp. 431–440, 2008.

[17] L. Zhang, H. Zhu, T. I. Oprea, A. Golbraikh, and A. Tropsha, "QSAR modeling of the blood–brain barrier permeability for diverse organic compounds," *Pharmaceutical Research*, vol. 25, no. 8, pp. 1902–1914, 2008.

[18] S. van Damme, W. Langenaeker, and P. Bultinck, "Prediction of blood-brain partitioning: a model based on ab initio calculated quantum chemical descriptors," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 8, pp. 1223–1236, 2008.

[19] D. A. Konovalov, D. Coomans, E. Deconinck, and Y. V. Heyden, "Benchmarking of QSAR models for blood-brain barrier permeation," *Journal of Chemical Information and Modeling*, vol. 47, no. 4, pp. 1648–1656, 2007.

[20] T. S. Carpenter, D. A. Kirshner, E. Y. Lau, S. E. Wong, J. P. Nilmeier, and F. C. Lightstone, "A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations," *Biophysical Journal*, vol. 107, no. 3, pp. 630–641, 2014.

[21] J. Zah, G. Terre'Blanche, E. Erasmus, and S. F. Malan, "Physicochemical prediction of a brain-blood distribution profile in polycyclic amines," *Bioorganic and Medicinal Chemistry*, vol. 11, no. 17, pp. 3569–3578, 2003.

[22] T. J. Hou and X. J. Xu, "ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 2137–2152, 2003.

[23] Y. H. Zhao, M. H. Abraham, A. Ibrahim et al., "Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes," *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 170–175, 2007.

[24] S. R. Mente and F. Lombardo, "A recursive-partitioning model for blood-brain barrier permeation," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 7, pp. 465–481, 2005.

[25] P. Garg and J. Verma, "In silico prediction of blood brain barrier permeability: an artificial neural network model," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 289–297, 2006.

[26] A. Guerra, J. A. Páez, and N. E. Campillo, "Artificial neural networks in ADMET modeling: prediction of blood-brain barrier permeation," *QSAR & Combinatorial Science*, vol. 27, no. 5, pp. 586–594, 2008.

[27] Z. Wang, A. Yan, and Q. Yuan, "Classification of blood-brain barrier permeation by Kohonen's self-organizing neural network (KohNN) and support vector machine (SVM)," *QSAR & Combinatorial Science*, vol. 28, no. 9, pp. 989–994, 2009.

[28] A. Yan, H. Liang, Y. Chong, X. Nie, and C. Yu, "In-silico prediction of blood-brain barrier permeability," *SAR and QSAR in Environmental Research*, vol. 24, no. 1, pp. 61–74, 2013.

[29] J. Shen, F. Cheng, Y. Xu, W. Li, and Y. Tang, "Estimation of ADME properties with substructure pattern recognition," *Journal of Chemical Information and Modeling*, vol. 50, no. 6, pp. 1034–1041, 2010.

[30] H. Golmohammadi, Z. Dashtbozorgi, and W. E. Acree Jr., "Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine," *European Journal of Pharmaceutical Sciences*, vol. 47, no. 2, pp. 421–429, 2012.

[31] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[32] K. Heikamp and J. Bajorath, "Support vector machines for drug discovery," *Expert Opinion on Drug Discovery*, vol. 9, no. 1, pp. 93–104, 2014.

[33] C. W. Hsu, C. C. Chang, and C. J. Lin, *A Practical Guide to Support Vector Classication*, National Taiwan University, Taipei, Taiwan, 2006.

[34] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 131–159, 2002.

[35] K.-M. Chung, W.-C. Kao, C.-L. Sun, L.-L. Wang, and C.-J. Lin, "Radius margin bounds for support vector machines with the RBF kernel," *Neural Computation*, vol. 15, no. 11, pp. 2643–2681, 2003.

[36] F. Friedrichs and C. Igel, "Evolutionary tuning of multiple SVM parameters," *Neurocomputing*, vol. 64, no. 1–4, pp. 107–117, 2005.

[37] J. H. Yang and V. Honavar, "Feature subset selection using genetic algorithm," *IEEE Intelligent Systems & Their Applications*, vol. 13, no. 2, pp. 44–48, 1998.

[38] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 2, pp. 164–171, 2000.

[39] S. Salcedo-Sanz, M. Prado-Cumplido, F. Perez-Cruz, and C. Bousono-Calzon, "Feature selection via genetic optimization," in *Artificial Neural Networks—Icann 2002*, vol. 2415, pp. 547–552, Springer, 2002.

[40] Z. Q. Wang and D. X. Zhang, "Feature selection in text classification via SVM and LSI," in *Advances in Neural Networks—ISNN 2006*, vol. 3971 of *Lecture Notes in Computer Science*, part 1, pp. 1381–1386, Springer, Berlin, Germany, 2006.

[41] M. Fernandez, J. Caballero, L. Fernandez, and A. Sarai, "Genetic algorithm optimization in drug design QSAR: bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM)," *Molecular Diversity*, vol. 15, no. 1, pp. 269–289, 2011.

[42] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[43] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[44] K. Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 60–67, 2004.

[45] K.-Q. Shen, C.-J. Ong, X.-P. Li, and E. P. V. Wilder-Smith, "Feature selection via sensitivity analysis of SVM probabilistic outputs," *Machine Learning*, vol. 70, no. 1, pp. 1–20, 2008.

[46] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, vol. 207 of

Studies in Fuzziness and Soft Computing, pp. 315–324, Springer, Berlin, Germany, 2006.

[47] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, and V. Vapnik, “Feature selection for SVMs,” in *Advances in Neural Information Processing Systems 13*, pp. 668–674, MIT Press, 2000.

[48] C.-L. Huang and C.-J. Wang, “A GA-based feature selection and parameters optimizationfor support vector machines,” *Expert Systems with Applications*, vol. 31, no. 2, pp. 231–240, 2006.

[49] X. R. Zhang and L. C. Jiao, “Simultaneous feature selection and parameters optimization for SVM by immune clonal algorithm,” in *Advances in Natural Computation*, vol. 3611 of *Lecture Notes in Computer Science*, pp. 905–912, Springer, Berlin, Germany, 2005.

[50] C. Gold, A. Holub, and P. Sollich, “Bayesian approach to feature selection and parameter tuning for support vector machine classifiers,” *Neural Networks*, vol. 18, no. 5-6, pp. 693–701, 2005.

[51] M. H. Abraham, A. Ibrahim, Y. Zhao, and W. E. Agree Jr., “A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data,” *Journal of Pharmaceutical Sciences*, vol. 95, no. 10, pp. 2091–2100, 2006.

[52] A. R. Katritzky, V. S. Lobanov, and M. Karelson, *CODESSA: Reference Manual*, University of Florida, 1996.

[53] Marvin 4.0.5, ChemAxon, 2006, http://www.chemaxon.com/.

[54] AMPAC 8, 1992–2004 Semichem, Shawnee, Kan, USA.

[55] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[56] R. W. Kennard and L. A. Stone, “Computer aided design of experiments,” *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[57] M. Daszykowski, B. Walczak, and D. L. Massart, “Representative subset selection,” *Analytica Chimica Acta*, vol. 468, no. 1, pp. 91–103, 2002.

[58] R. Collobert and S. Bengio, “SVMTorch: support vector machines for large-scale regression problems,” *Journal of Machine Learning Research*, vol. 1, no. 2, pp. 143–160, 2001.

[59] A. J. Smola and B. Scholkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[60] R. G. Brereton and G. R. Lloyd, “Support Vector Machines for classification and regression,” *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.

[61] C. C. Chang and C. J. Lin, “LIBSVM: a library for support vector machines,” 2001.

[62] H. Golmohammadi, Z. Dashtbozorgi, and W. E. Acree Jr., “Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine,” *European Journal of Pharmaceutical Sciences*, vol. 47, no. 2, pp. 421–429, 2012.

[63] D. E. Clark, “In silico prediction of blood-brain barrier permeation,” *Drug Discovery Today*, vol. 8, no. 20, pp. 927–933, 2003.

[64] S. Vilar, M. Chakrabarti, and S. Costanzi, “Prediction of passive blood-brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors,” *Journal of Molecular Graphics and Modelling*, vol. 28, no. 8, pp. 899–903, 2010.

[65] F. Herve, S. Urien, E. Albengres, J.-C. Duche, and J.-P. Tillement, “Drug binding in plasma. A summary of recent trends in the study of drug and hormone binding,” *Clinical Pharmacokinetics*, vol. 26, no. 1, pp. 44–58, 1994.

[66] J. A. Platts, M. H. Abraham, Y. H. Zhao, A. Hersey, L. Ijaz, and D. Butina, “Correlation and prediction of a large blood-brain distribution data set—an LFER study,” *European Journal of Medicinal Chemistry*, vol. 36, no. 9, pp. 719–730, 2001.

[67] M. Iyer, R. Mishra, Y. Han, and A. J. Hopfinger, “Predicting blood-brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis,” *Pharmaceutical Research*, vol. 19, no. 11, pp. 1611–1621, 2002.

[68] M. Lobell, L. Molnár, and G. M. Keserü, “Recent advances in the prediction of blood-brain partitioning from molecular structure,” *Journal of Pharmaceutical Sciences*, vol. 92, no. 2, pp. 360–370, 2003.

[69] Y. N. Kaznessis, M. E. Snow, and C. J. Blankley, “Prediction of blood-brain partitioning using Monte Carlo simulations of molecules in water,” *Journal of Computer-Aided Molecular Design*, vol. 15, no. 8, pp. 697–708, 2001.

[70] H. van de Waterbeemd, G. Camenisch, G. Folkers, J. R. Chretien, and O. A. Raevsky, “Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors,” *Journal of Drug Targeting*, vol. 6, no. 2, pp. 151–165, 1998.

[71] J. Kelder, P. D. J. Grootenhuis, D. M. Bayada, L. P. C. Delbressine, and J.-P. Ploemen, “Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs,” *Pharmaceutical Research*, vol. 16, no. 10, pp. 1514–1519, 1999.

[72] F. Ooms, P. Weber, P.-A. Carrupt, and B. Testa, “A simple model to predict blood-brain barrier permeation from 3D molecular fields,” *Biochimica et Biophysica Acta—Molecular Basis of Disease*, vol. 1587, no. 2-3, pp. 118–125, 2002.

[73] H. Fischer, R. Gottschlich, and A. Seelig, “Blood-brain barrier permeation: molecular parameters governing passive diffusion,” *Journal of Membrane Biology*, vol. 165, no. 3, pp. 201–211, 1998.

[74] K. M. Mahar Doan, J. E. Humphreys, L. O. Webster et al., “Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs,” *Journal of Pharmacology and Experimental Therapeutics*, vol. 303, no. 3, pp. 1029–1037, 2002.

[75] K. Rose, L. H. Hall, and L. B. Kier, “Modeling blood-brain barrier partitioning using the electrotopological state,” *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 3, pp. 651–666, 2002.

[76] P. Crivori, G. Cruciani, P.-A. Carrupt, and B. Testa, “Predicting blood-brain barrier permeation from three-dimensional molecular structure,” *Journal of Medicinal Chemistry*, vol. 43, no. 11, pp. 2204–2216, 2000.

[77] R. C. Young, R. C. Mitchell, T. H. Brown et al., “Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists,” *Journal of Medicinal Chemistry*, vol. 31, no. 3, pp. 656–671, 1988.

[78] F. Lombardo, J. F. Blake, and W. J. Curatolo, “Computation of brain-blood partitioning of organic solutes via free energy calculations,” *Journal of Medicinal Chemistry*, vol. 39, no. 24, pp. 4750–4755, 1996.

[79] M. Iyer, R. Mishra, Y. Han, and A. J. Hopfinger, “Predicting blood–brain barrier partitioning of organic molecules using membrane–interaction QSAR analysis,” *Pharmaceutical Research*, vol. 19, no. 11, pp. 1611–1621, 2002.