ChemPubSoc Europe

DOI: 10.1002/cmdc.201800533

CHEMMEDCHEM Full Papers

In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods

Zhuang Wang, Hongbin Yang, Zengrui Wu, Tianduanyi Wang, Weihua Li, Yun Tang,* and Guixia Liu*[a]

The blood–brain barrier (BBB) as a part of absorption protects the central nervous system by separating the brain tissue from the bloodstream. In recent years, BBB permeability has become a critical issue in chemical ADMET prediction, but almost all models were built using imbalanced data sets, which caused a high false-positive rate. Therefore, we tried to solve the problem of biased data sets and built a reliable classification model with 2358 compounds. Machine learning and resampling methods were used simultaneously for the refinement of models with both 2D molecular descriptors and molecular fingerprints to represent the chemicals. Through a series of evaluation, we realized that resampling methods such as Synthetic Minority Oversampling Technique (SMOTE) and SMOTE+edited nearest neighbor could effectively solve the problem of imbalanced data sets and that MACCS fingerprint combined with support vector machine performed the best. After the final construction of a consensus model, the overall accuracy rate was increased to 0.966 for the final external data set. Also, the accuracy rate of the model for the test set was 0.919, with an excellent balanced capacity of 0.925 (sensitivity) to predict BBB-positive compounds and of 0.899 (specificity) to predict BBB-negative compounds. Compared with other BBB classification models, our models reduced the rate of false positives and were more robust in prediction of BBB-positive as well as BBB-negative compounds, which would be quite helpful in early drug discovery.

Introduction

During the process of drug design and discovery, a large number of drug candidates could not finally reach the market because of the poor ADMET properties.[1] Thus, finding and optimizing suitable structures of chemicals should be paid more attention to ensure that the research and development of new drugs proceed smoothly and quickly.[2] At present, numerous in vitro and in vivo assay methods have been developed to investigate drug ADMET properties.[3,4] However, it is time-consuming and costly to perform an experimental evaluation of these complex profiles.[5] Nowadays, with the high-speed development of computational technology, various in silico prediction models have been developed, which can help us filter and predict the required ADMET properties for individual compounds.[6]

The blood–brain barrier (BBB) as a part of absorption protects the central nervous system (CNS) by separating the brain tissue from the bloodstream. It is mainly formed by brain endothelium, which can prevent larger molecules (≈100%) and small molecules (≈98%) from entering into the CNS and allow transport of only water- and lipid-soluble molecules and selective transport molecules across itself.[7] Also, the channel expresses numerous active transporters such as P-glycoprotein and glucose transporters to prevent the entry of lipophilic potential neurotoxins.[8]

The most common in vitro methods used for BBB permeability prediction are parallel artificial membrane permeability assay[9] and immobilized artificial membrane technique.[10]

In recent years, with the development of artificial intelligence, there also appear some statistical methods or machine learning algorithms to find this prediction. In these models, values of CNS+/CNS−, LogBB, and surface permeability product (LogPS) were often used as reference values to describe the ability of BBB penetration in the study of a quantitative structure–activity relationship (QSAR) model.[11–14] Therein most public models use classical machine learning methods such as support vector machine (SVM),[15] random forest (RF),[16] *k*-nearest neighbor (KNN),[16] and artificial neural network[17] to build supervised classification models or regression models, in which BBB+ (compounds that can cross the BBB) or BBB− (compounds that cannot cross the BBB) is used as the property label of classification models and the value of LogBB is often used in regression prediction.

Here are some previous prediction instances from the public literature. Suenderhauf et al. used in vivo LogPS values as a quantitative parameter to predict the probability of BBB permeability of 153 compounds with the corrected classification rate of 90%.[18] Raevsky et al. used a method called "read-

[a] *Z. Wang, Dr. H. Yang, Dr. Z. Wu, T. Wang, Prof. W. Li, Prof. Y. Tang, Prof. G. Liu*
Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237 (China)
E-mail: ytang234@ecust.edu.cn
gxliu@ecust.edu.cn

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/cmdc.201800533.

across" to study the structure–activity relationship of the BBB, and the method showed that the hydrogen bond donor–acceptor would be the main factor of penetration.[19] Also, Martins et al. built a series of SVM and RF classification models using the Bayesian approach with improved results [accuracy rate (ACC) = 0.947, sensitivity (SE) = 0.826, and specificity (SP) = 0.712].[13] Furthermore, Shen et al. from our laboratory optimized SVM models with 1593 compounds (1283 BBB+ and 310 BBB−) by using different pattern selection methods and obtained the overall accuracy of 98.2%.[20] All these models indicated that classical machine learning methods performed well in the prediction of BBB penetration and the widely applied SVM had a powerful capacity to identify BBB+ structures. However, these high accuracy models have the same defect, which is the low SP value in prediction because of the imbalanced number between BBB+ and BBB− (a lack of BBB− data samples). This vital issue may have caused a high false-positive rate in prediction and cannot help chemists find whether the compound is BBB−. Meanwhile, it cannot provide helpful tactics to optimize potential compounds with BBB+ property.

Considering this vital defect in building a model, we used resampling methods such as Synthetic Minority Oversampling Technique (SMOTE),[21] SMOTE+ENN (edited nearest neighbor),[22] RandomUnderSampler, and adaptive synthetic sampling (ADASYN)[23] provided by the *imbalanced-learn* package (version 0.3.3) and cost-sensitive methods provided by the *scikit-learn* package (version 0.19.1) to solve the problem of imbalanced data sets.[24] To further improve the accuracy of our models, a consensus model combining well-performed single models was also created. According to the results of the prediction evaluations, our models performed well and can markedly decrease the false-positive rate in prediction compared with other previous models. Afterward, we chose three simple but important descriptors—molecular weight (MW), lipo-hydro partition coefficient (LogP), topological polar surface area (TPSA)—to define the applicability domain (AD) of our model and analyzed the significance of these descriptors in the mechanism of BBB penetration.[25,28] Also, substructure frequency analysis and information gain (IG) method were used to identify several representative substructures of BBB+ and BBB− by using SARpy tools. At last, we made the BBB prediction of 7179 compounds including approved, experimental, investigated, and withdrawn drugs derived from DrugBank. Our research results might be helpful for users to find, predict, and optimize the BBB permeability of individual compounds.

Computational Methods

Data collection and preparation

In this study, the data set was obtained and integrated from four recent studies.[13,20,26,27] First, all these compounds containing noncovalent, inorganic, mixtures, or only salt were deleted and those with MW greater than 1000 Da were also removed. Then, we used LogBB as the criterion to divide compounds into BBB+ and BBB− if LogBB ≥ -1 and LogBB < -1, respectively.

After the above processes, we used MacroModel 11.1 (Schrödinger 2016) to neutralize all compounds and generate the most populated neutral tautomer for each compound at pH 7.0 by using Epik 3.5.[29] Then, we used Open Babel to standardize dative bonds and generate canonical SMILES for each chemical structure.[30] Finally, all compounds were merged and represented by the canonical SMILES as the unique identifier for each chemical structure. Furthermore, some molecules with ambiguous values or contradicting data were eliminated.

With random selection of an external data set for evaluation, the remaining compounds were divided into a training set and a test set in a ratio of 0.85:0.15. To expand the scope of external validation, we added 92 CNS+ chemicals retrieved from a related article to the external data set for the final evaluation.[31] The full data set is presented in Table S1, which includes a self-generated set ID, the generic name as referred in the literature, the canonical SMILES derived from the literature, the canonical SMILES for each compound, the binary classification of the BBB ("p" stands for BBB+ and "n" for BBB), and the reference ID of the related articles. The added 92 CNS+ samples are listed in Table S9, which contains their PubChem CID, name, and canonical SMILES. In addition, we retrieved all kinds of small molecules from DrugBank (http://www.drugbank.ca/), including approved, experimental, investigated, and withdrawn drugs, for sample comparison and prediction.[32] These molecules were prepared using the same procedure as mentioned above.

Molecular descriptors

We tried to use 2D molecular descriptors (MDs) and six types of molecular fingerprints, namely, ECFP2, FCFP4, MACCS, PubChem, Substructure, and Klekota Roth, as molecular features to describe the molecules. The 2D descriptors were calculated by using PaDEL-Descriptor,[33] which contains 1437 types of 2D descriptors such as ALogP, CarbonTypes, ChiChain, TPSA, and other 2D molecular properties.[34] ECFP2 and FCFP4 were calculated by using a package of Discovery Studio software (version 3.5).[35,36] The other four types of fingerprints were calculated by using Chemistry Development Kit, where MACCS was the shortest substructure key-based fingerprint with 166 bits and PubChem, Substructure, and Klekota Roth contained 881, 307, and 4860 bits, respectively.[37]

Selection of molecular features

Feature selection has been widely used for feature preprocessing in machine learning. Removing irrelevant and redundant information can improve the performance of models. Herein five types of feature selection methods provided by the *scikit-learn* package were tested.[38]

1) Variance threshold (VT): For the variance of each descriptor, we deleted all features that were either one or zero in more than 98% of samples.
2) Univariate feature selection (UFE): We used the method using the *F*-test algorithm provided as an API class *f_classif*

module in *scikit-learn* to estimate the degree of dependency of each descriptor with label values. Then, we removed all but a high scoring independent percentage of descriptors, where the threshold of the score was set at 0.8.

3) Recursive feature elimination (RFE): This method selects descriptors by recursively using smaller and smaller sets of descriptors.[39] First, we chose a base estimator SVM. Second, we performed this base estimator with each prepared set of descriptors in the cross-validation loop to find the optimal number of features for models.
4) Pearson correlation coefficient (PCC):[40] If the correlation coefficient between two descriptors is higher than 0.65, we choose only one to keep. This method was used only for MDs.
5) Principal component analysis (PCA):[41] PCA is a dimensionality reduction method that is used, through a rotation of the original data set, to see the data from a new perspective.

In our modeling process, we designed a series of combinations of these selection methods for only MDs with continuous variables: 1) VT only, 2) VT + UFE, 3) VT + RFE, 4) VT + UFE + RFE, and 5) VT + PCC + PCA. These combination methods were used to build a classification model repeatedly, and both chosen features and performance were recorded.

Solution to the problem of imbalanced data sets

We evaluated five types of methods provided by the *imbalanced-learn* package to solve the problem of imbalanced data sets.[24]

1) Random undersampling (RUS): It is a nonheuristic method that aims to make class distribution balanced by randomly selecting samples from majority samples.[42] Although random selection will build a balanced training set, the training set is smaller than the original set and so loss of information may be caused.
2) SMOTE:[21] This method analyzes the similarity of the minority class in the feature space of near-neighbor samples and randomly synthesizes new fake minority data into the original set. But this method also has a drawback that it does not consider the distribution of the majority samples surrounding the minority samples so that it may be divided into majority samples.
3) ADASYN:[23] Considering the blindness of SMOTE, the ADASYN method can not only adaptively synthesize minority samples in terms of the distribution of minority samples but also adaptively shift the decision boundary to machine learning samples.
4) SMOTE + ENN:[22] It is an ensemble method to synthesize minority samples. First, it adopts the SMOTE algorithm to generate the balanced data set *T*. Second, to decrease the potential of overfitting, it uses the ENN method to predict each sample in the *T* set. If the prediction is not the same as that for the true class, then the synthetic sample is deleted.
5) Weight loss function (WLF): WLF is a method that sets the weight of loss function for each sample and tries to adjust the loss of the prediction error for the minority class larger than the loss of the prediction error for the majority class. We used a parameter *class weight* that accepts a key-value pair dictionary in *scikit-learn* to set weight.

Model building

Six types of machine learning algorithms from *scikit-learn* (version 0.19.1) were used to conceive the best classification model.

1) Logistic regression (LR):[43] We used default parameters set by *scikit-learn*, where *liblinear* was set as the solver method and parameters *penalty* and *C* were set as "L2 regularization" and "1.0," respectively.
2) AdaBoost classifier: AdaBoost is an iterative algorithm called *adaptive boosting* with the base estimator *DecisionTreeClassifier* as default in our study and the parameter *n estimators* set at 50 and *learning rate* set at 1.0.
3) RF: We used the grid search method to adjust initial parameters such as *n estimators* and *max features*. The parameter *n estimators* is the number of trees in the forest and *max features* is the number of features. The values of *n estimators* were tried as 5, 10, 25, 55, 85, and 110, whereas attempts of *max features* included 10, 30, 50, and 80.
4) KNN: In our classification model, we used the grid search method to adjust *n neighbors* in the range of 1 to 25. Weight is a variable to add weights for nearest neighbors, including *uniform* and *distance*, that had been searched. *Uniform* assigns the same weights to the nearest neighbors, whereas *distance* assigns weights according to the distance. For each descriptor, the appropriate values were searched. For the *k-nearest neighbor* parameter, we used the grid search method to find the most suitable values ranging from 1 to 10.
5) SVM: The parameters of SVM, including *kernel*, *C*, and *gamma*, were modified by using the grid search method on the basis of the classifier performance on the test model. The values of *C* and *gamma* were set from 2^{-5} to 2^{10}.
6) Multilayer perceptron neural network (MLP): MLP is a class of feedforward artificial neural network, which consists of at least three layers of nodes and is trained with stochastic gradient descent. In our model, MLP contained two hidden layers, each consisting of 128 neurons, with *relu* as the activation function and *learning rate* being 0.001. In addition, we searched the *batch size* as 50, 100, 200, and 300 as the potential parameter by using grid search written in Python scripts.
7) Consensus classifier model: As mentioned in previous studies, the consensus classifier methods combining multiple well-performed single classifier models had a better prediction accuracy than each single classifier model.[44] Thus, the construction of this classifier used in our study consisted of

a three-layer perceptron neural network of which the classified results are from single models. The parameter of the neural network was searched by using the grid search method in 10-fold cross-validation. The range of *learning rate* was selected from 0.0001, 0.001, and 0.01, and *batch sizes* were adopted as 50 and 100.

Model evaluation

First, we used our test set to evaluate the performance of different feature selection and imbalanced learning methods. In addition, 10-fold cross-validation was used to search hyperparameters for several well-performing algorithms that can improve the ability to classify the test data. In the selection of single models, because of the imbalanced problem, it is not enough to merely consider the accuracy of the classifier. Therefore, we chose a best feature combination according to evaluation values of both G-means and AUC (area under the receiver operating characteristic curve) of a model for our test set. The external data set was used to evaluate the capacity of a consensus model that takes account of difference contribution of each single model. The performance of the BBB classification models was evaluated using the following statistical parameters: TN (true negative), FN (false negative), TP (true positive), FP (false positive), SE (sensitivity), SP (specificity), ACC (accuracy), F1-score value (F), G-means, and AUC,[45] which are defined in Equations (1)–5, respectively:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

$$\mathrm{ACC} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{3}$$

$$\mathrm{F} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{G} - \mathrm{means} = \sqrt{\mathrm{SP}} \times \mathrm{SE} \tag{5}$$

Definition of the AD

The AD of a QSAR model defines the model's limitation in its structural domain and responsible space.[46] This restricts the applicability of a model to reliably predict those test chemicals that are structurally similar to the training chemicals used to build that model. Several approaches were proposed to analyze the AD of a QSAR model using descriptor spaces. Herein, 2D MDs were used to define AD, where MW, LogP, and TPSA were calculated as definition factors and the Euclidean distance was calculated for a query sample with the mean values of MW, LogP, and TPAS of the training samples. After the calculation of the average distance for all training samples, we used a predefined AD threshold, D_T, calculated in Equation (6):

$$D_\mathrm{T} = \gamma + Z\sigma \tag{6}$$

where γ equal to 132.68 is the average Euclidean distance between each compound in the training set with the mean values of MW (336), LogP (2.90), TPAS (75.52); σ equal to 69.54 is the standard deviation of this Euclidean distance; and Z is an arbitrary value varying from −1.49 to 5.45. If AD exceeds this threshold, the prediction of a compound would be regarded as potentially unreliable. If the Z value decreases, the compound would be much similar to the BBB molecules in the training set in space features.

Results

Data set analysis

After data preparation, we obtained a data set containing 2358 compounds with 546 BBB− and 1812 BBB+ (SMILES for each compound was publicly available, and all data are provided in Table S1). From the data set, we randomly selected a set of 145 molecules (36 BBB− and 109 BBB+) as the external validation set to evaluate the generalization ability and reliability of the consensus model. The remaining 2213 molecules were used to build the prediction models, which were randomly separated into a training set containing 1881 molecules and a test set containing 332 molecules in a ratio of 0.85:0.15. In addition, to explore the chemical space distribution of our training, test, and external validation sets, two simple descriptors—MW and partition coefficient of solutes in octanol/water—were used to define the chemical space, which is illustrated in Figure 1. From the scatter diagram, it can be concluded that the chemical space was relatively wide and the three separated sets shared similar chemical space.

The Tanimoto similarity index was calculated for each pair of molecules by using MACCS fingerprint, and the average similarity was 0.33. Meanwhile, 7179 diverse small molecules retrieved from DrugBank were also calculated for the Tanimoto similarity index. From the distribution plot in Figure 2A, we can conclude that the average similarity index of our BBB data is approximately equal to the average value (0.31) of the DrugBank compounds. This evidence indicated that our BBB chemi-

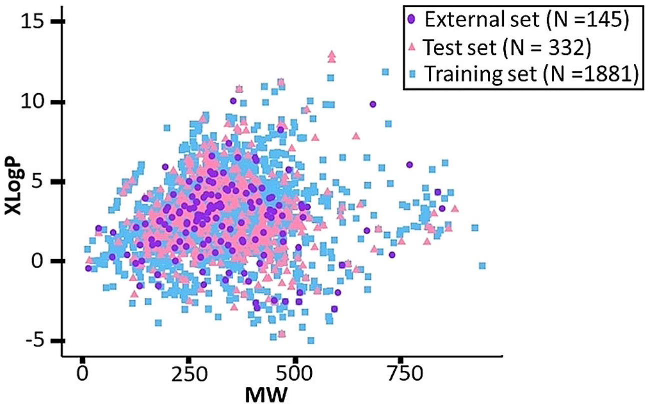


Figure 1. Chemical diversity analysis of the training, test, and external validation sets. MW = molecular weight; N = number of chemicals in each data set; XLogP = partition coefficient of solutes in octanol/water.

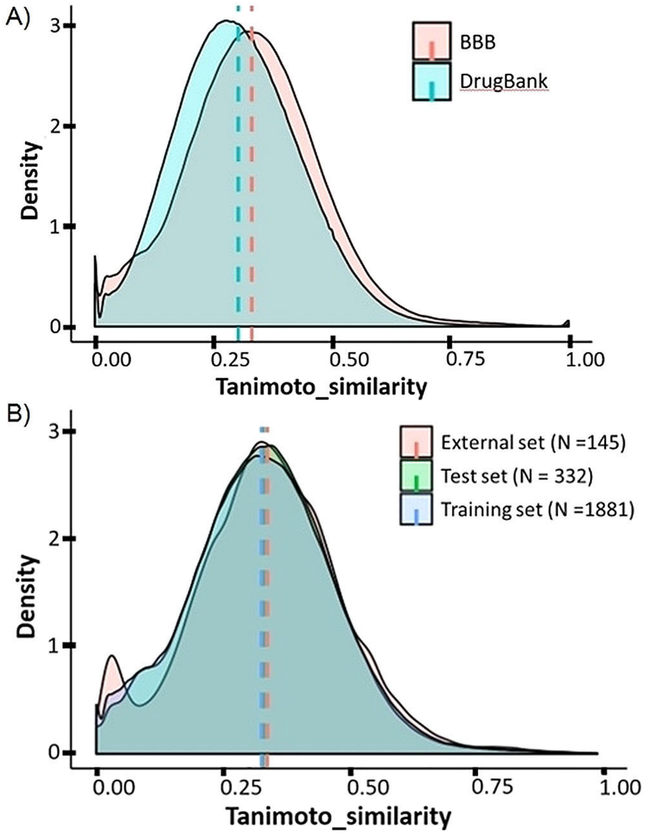


Figure 2. A) Tanimoto similarity index of blood–brain barrier (BBB) chemicals and DrugBank chemicals. B) Tanimoto similarity index of chemicals in training, test, and external validation sets. The *x* axis represents the Tanimoto similarity index, and the *y* axis represents the 10 times of the probability of the density of each Tanimoto similarity index.

cals were wide and structurally diverse. Furthermore, it showed that the chemical similarity indices of chemicals in the training, test, and external validation sets were low (Figure 2B), which is helpful to construct a reliable model.

Selection of molecular fingerprints and resampling methods

First, for the six types of molecular fingerprints (ECFP2, FCFP4, MACCS, Klekota Roth, PubChem, and Substructure), we used SVM as the base classifier to compare the performance of these fingerprints by using 10-fold cross-validation. The results are presented in Table 1. G-means values of most models were higher than 0.6, except for ECFP2 and FCFP4, for which G-means values were 0.000. We considered that ECFP2 and FCFP4 fingerprints may not be suitable for imbalanced data when SVM was chosen as the classifier. From the other four fingerprints, MACCS was chosen as the main fingerprint because it had relatively higher ACC, G-means, and AUC values than other fingerprints. However, although the SE value of MACCS approached nearly 1000, the SP value was only 0.533, which suggested that this model would have an excellent ability to predict the positive samples but would be incapable to classify any negative samples. Thus, four types of resampling methods were added to our modeling trials to balance the training data. With 30 various types of combinations of fingerprints and resampling methods, we tried to obtain the most stable model for both positive and negative samples. The detailed evaluation results obtained through 10-fold cross-validation are presented in Table S2.

Table 1. Tenfold cross-validation results of six types of fingerprints.[a]

Fingerprint	ACC	SE	SP	G-means	AUC
ECFP2	0.772	1.000	0.000	0.000	0.877
FCFP4	0.772	1.000	0.000	0.000	0.887
Klekota Roth	0.828	0.996	0.255	0.504	0.862
MACCS	0.885	0.983	0.553	0.737	0.895
PubChem	0.881	0.990	0.514	0.713	0.879
Substructure	0.860	0.992	0.412	0.638	0.882

[a] ACC = accuracy rate; AUC = area under the receiver operating characteristic curve; SE = selectivity; SP = specificity.

By comparison, the performance of MACCS was better than that of other models with the combination of resampling methods. For the top two resampling methods—SMOTE and SMOTE + ENN—the G-means and AUC values were 0.873–0.881 and 0.944–0.951, respectively. Moreover, after the addition of resampling methods, the cross-validation performance of all models, especially the SP values, had improved significantly. Although the SE values of most models were slightly reduced, considering their ability to classify different samples, we thought that the effect of the reduction in SE values was trivial for model construction. To further improve the performance, we selected SMOTE and SMOTE + ENN as resampling methods and MACCS as the base fingerprint for the next process. The total increased number of positive and negative chemicals in the training set after using resampling methods with MACCS is shown in Figure 3.

Selection of molecular features

In addition to molecular fingerprints, MDs were used as the features to predict. Herein we used a workflow of feature selection to retrieve the key features and choose the parameter *k-nearest neighbor* ($k = 1, 2, 3,\ldots, 10$) for SMOTE and SMOTE + ENN. The workflow is described in Figure 4. In our experiment, for the original MACCS training set samples (not using resampling methods), VT + UFE + RFE was adopted as the feature selection method, which finally retained 72 descriptors. For the resampling fingerprint samples, we adopted VT + RFE as our feature selection method and *k-nearest neighbor* ($k = 2$). After feature selection, the retrieved number bits of the resampling fingerprint were 112. For the original MDs (not using resampling methods), VT was adopted as the feature selection method, which finally retained 336 descriptors. Furthermore, for the resampling descriptor samples, VT + PCC + UFE was used as our feature selection method and *k-nearest neighbor* ($k = 2$) was used for both SMOTE and SMOTE + ENN. The retrieved 45 types of these descriptors are listed in Table S3, for which *P* values by analysis of variance were much lower than the statistical significance level of 0.01 and PCC was lower

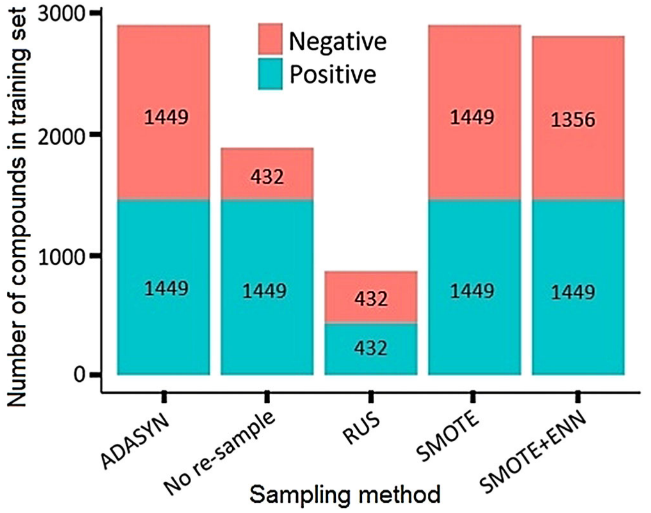


Figure 3. Distribution of number of positive and negative chemicals in the training set after using resampling methods with MACCS fingerprint. ADASYN = adaptive synthetic sampling; ENN = edited nearest neighbor; RUS = random undersampling; SMOTE = Synthetic Minority Oversampling Technique.

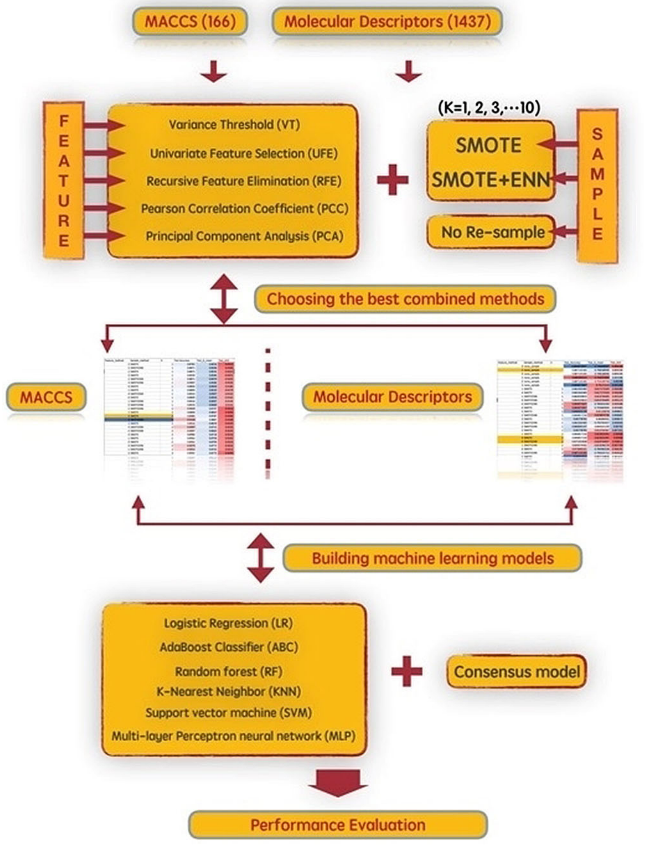

MACCS (166)
Molecular Descriptors (1437)
FEATURE
Variance Threshold (VT)
Univariate Feature Selection (UFE)
Recursive Feature Elimination (RFE)
Pearson Correlation Coefficient (PCC)
Principal Component Analysis (PCA)
(K=1, 2, 3,···10)
SMOTE
SMOTE+ENN
No Re-sample
SAMPLE
Choosing the best combined methods
MACCS
Molecular Descriptors
Building machine learning models
Logistic Regression (LR)
AdaBoost Classifier (ABC)
Random forest (RF)
K-Nearest Neighbor (KNN)
Support vector machine (SVM)
Multi-layer Perceptron neural network (MLP)
Consensus model
Performance Evaluation

Figure 4. Workflow of feature selection and *k-nearest neighbor* selection for MACCS fingerprint and molecular descriptors. ENN = edited nearest neighbor; SMOTE = Synthetic Minority Oversampling Technique.

than 0.7. The relationships between descriptors are presented as a heat map in Figure 5, and the actual PCC values are presented in Table S4.

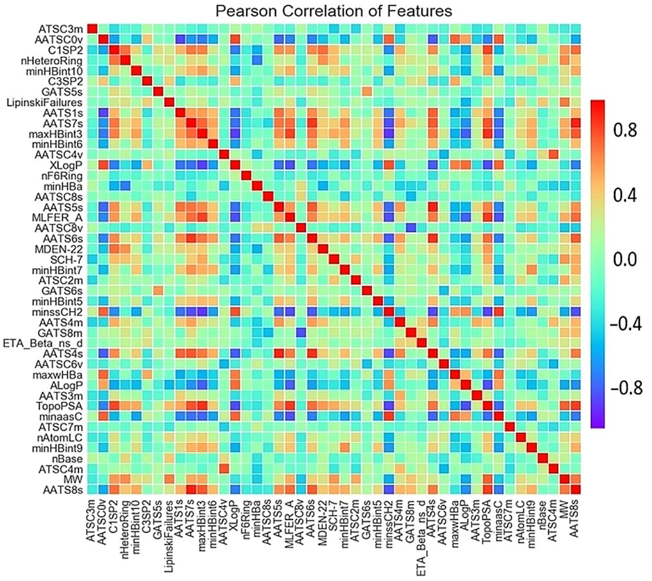


Figure 5. Pearson correlation coefficients for relationships between descriptors after feature selection.

Performance of single classification models

In total, there are six types of machine learning methods together with two types of selected features (MACCS and MDs) and two resampling methods (SMOTE and SMOTE + ENN). We also adopted a method called WLF to balance the weight of different samples. In this method, 44 models were constructed in total for MACCS and MDs after the essential features were selected. The grid search method was used in 10-fold cross-validation to search the hyperparameters for all models. The performance results for the training set and the test set are presented in Tables S5 and S6, respectively. As shown in these tables, the performance of all the models on the training set was excellent with AUC values approximately equal to 1, which indicated that these models had obtained useful information from the training set to distinguish samples. However, three of them had ACC lower than 0.90 for the training set, namely, MD + LR + SMOTE + ENN, MD + LR + no resampling, and MACCS + LR + SMOTE. From these models, it can be concluded that the resampling method is not suitable for LR classification. SVM combined with SMOTE and SMOTE + ENN performed slightly better than other classifiers such as MLP, RF, KNN, and AdaBoost. The top eight models were selected as shown in Tables S5 and S6 in orange. Their ACC values were 0.954–1.000 and AUC values were 0.988–1.000. The values of G-means increased to 0.954–1.000, which suggested that the resampling methods we used can improve the ability of a model to fit much more features of the positive and negative samples in training sets.

To ensure the reliability of our models, we further compared these top eight models on the basis of G-means and AUC values for the test set, as shown in Table 2 and Figure 6. The *class weight* parameter for each single model are shown in Table S8. The performance of five models MACCS + SVM + SMOTE + ENN + class weight, MACCS + SVM + no resampling +

Table 2. Performance evaluation of the top 8 models among 44 single models generated by combining different features, classifiers, resampling methods, and *class weight* parameters for the training set and test set.[a]

Model	*Class weight*	Training set					Test set				
		ACC	SE	SP	G-means	AUC	ACC	SE	SP	G-means	AUC
MD + KNN + SMOTE + ENN	–	1	1	1	1	1	0.877	0.874	0.886	0.88	0.951
MD + KNN + SMOTE	–	1	0.999	1	1	1	0.873	0.87	0.886	0.878	0.95
MACCS + SVM + SMOTE	yes	0.975	0.95	0.999	0.975	0.998	0.907	0.921	0.861	0.89	0.944
MACCS + SVM + SMOTE + ENN	yes	0.992	0.984	1	0.992	1	0.907	0.917	0.873	0.895	0.943
MACCS + SVM + no resampling	yes	0.966	0.959	0.988	0.974	0.993	0.91	0.933	0.835	0.883	0.94
MD + SVM + no resampling	yes	0.954	0.943	0.988	0.966	0.996	0.907	0.937	0.81	0.871	0.938
MD + MLP + SMOTE	yes	0.954	0.921	0.988	0.954	0.99	0.886	0.897	0.848	0.872	0.933
MACCS + MLP + SMOTE	–	0.954	0.947	0.96	0.954	0.988	0.907	0.937	0.815	0.874	0.928

[a] Entries are in the descending order of AUC values; ACC = accuracy rate; AUC = area under the receiver operating characteristic curve; ENN = edited nearest neighbor; KNN = *k*-nearest neighbor; MD = molecular descriptor; MLP = multilayer perceptron neural network; SE = selectivity; SMOTE = Synthetic Minority Oversampling Technique; SP = specificity; SVM = support vector machine. Dashes indicate the model without using class weight method.

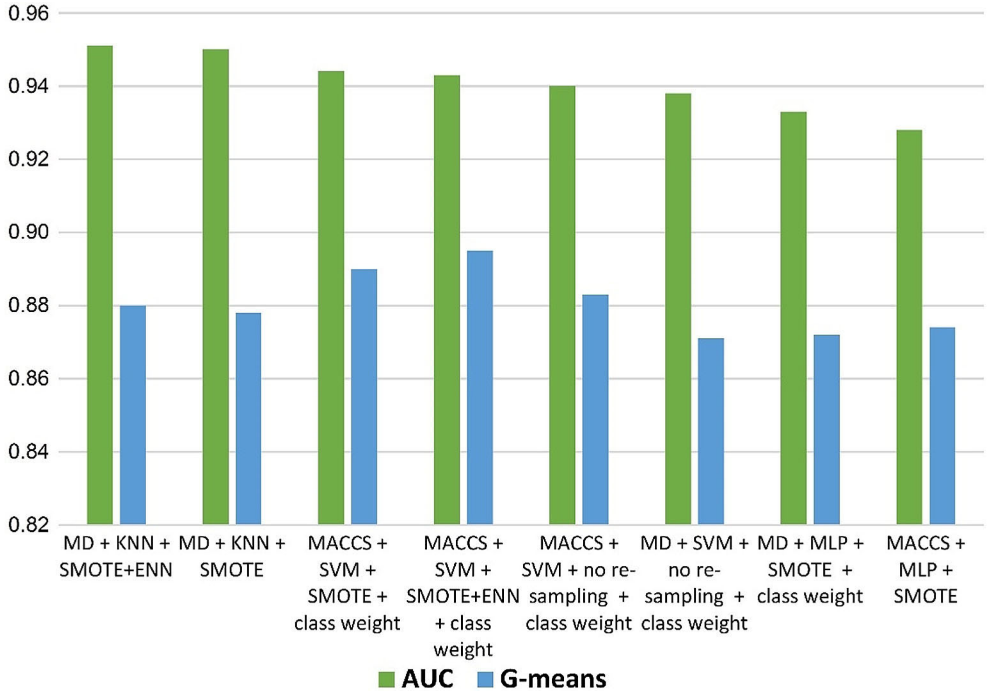


Figure 6. Performance evaluation of G-means and AUC (area under the receiver operating characteristic curve) values of each single model on the test set. ENN = edited nearest neighbor; KNN = *k*-nearest neighbor; MD = molecular descriptor; SMOTE = Synthetic Minority Oversampling Technique; SVM = support vector machine.

class weight, MACCS + SVM + SMOTE + class weight, MD + KNN + SMOTE + ENN, and MD + KNN + SMOTE was slightly better than that of other models. The G-means values of these five models for the test set were 0.895, 0.883, 0.890, 0.880, and 0.878, respectively. The comparison of their SE and SP values with those of other models is shown in Figure S1. Compared with models using no-resampling methods, SMOTE + class weight and SMOTE + ENN + class weight methods were truly useful to solve the imbalanced problem of BBB classification when combined with MD + KNN or MACC + SVM.

Moreover, as shown in Figure S1, the combination of the resampling method and WLF method performed better than only one of them. For instance, when using only MACCS + SVM without any resampling method or WLF method, the G-means and AUC values were 0.857 and 0.947, respectively. When adding only the resampling method, the G-means and AUC values of MACCS + SVM + SMOTE changed to 0.735 and 0.953, respectively. Similarly, when adding only the WLF method, the G-means and AUC values of MACCS + SVM + no resampling + class weight were 0.883 and 0.940, respectively. After the combination of resampling and WLF methods, the G-means value of MACCS + SVM + SMOTE + class weight approached to 0.890 and the AUC value was 0.944, which suggested that the combination of these strategies can improve the ability of a model to classify imbalanced positive and negative samples.

Performance of the consensus model

Although our single models could solve the imbalanced problem and increased G-means, AUC, and SP values, the ACC value was slightly reduced. To overcome the limitations of single models, we proposed a consensus model with the prediction of each single model as the input variable and the consensus prediction as the output variable.

The consensus model was consisted of the top five single models as mentioned above, namely, MACCS+SVM+SMOTE+ENN+class weight, MACCS+SVM+no re-sampling+class weight, MACCS+SVM+SMOTE+class weight, MD+KNN+SMOTE+ ENN, and MD+KNN+SMOTE. A classifier called a three-layer perceptron neural network (16, 32, 32) was used to fit the prediction results of singe models. As shown in Table 3, the consensus model performed better than most of the single models and showed improved results of some statistical values. The best fitted model exhibited an overall ACC value of 0.945 for the external data set, with an excellent balanced capacity of 0.982 (SE) to predict BBB+ compounds and of 0.833 (SP) to predict BBB− compounds. For test set data, we also used our consensus model to make classification and obtained fairly good evaluation results (ACC=0.912, SE=0.925, and SP=0.899). Afterward, because the external data set contains only 145 samples, we added 92 CNS+ chemicals retrieved from a related article to the external data set to validate the final consensus model; for the final consensus model, the ACC value increased to 0.966 and the SE value equaled 0.99 but the SP value did not change (0.833) because the added data sets contained only positive samples and the model could not evaluate more negative samples.

With this consensus model, we made a prediction of each of the 7179 compounds retrieved from DrugBank. The results are presented in Table S7, which indicated that 56.8% drugs were classified into the BBB+ class and 43.2% drugs were classified into the BBB− class; 94% of CNS drugs with Anatomical Therapeutic Chemical (ATC) codes starting with "N" were predicted correctly. Moreover, we analyzed the permeability of drugs with ATC codes starting with "A" (for diseases of the alimentary tract). The results indicated that 60% of those drugs were classified into the BBB− class.

AD of the consensus model

Previous studies have mentioned that BBB penetration was related to hydrogen bonding properties.[8,16,20] Multiple regression analyses revealed that TPSA was highly related to hydrogen bond acidity. So, hydrogen bond basicity, MW, and LogP are usually used in LogBB prediction. We selected these three MDs to define an AD for classification. The threshold of the AD, D_T [Eq. (6)], was determined by using the parameter Z (an arbitrary value varying from −1.49 to 5.45). In Equation (6), γ is the average Euclidean distance between each compound in the training set with the mean values of MW (336), LogP (2.90), and TPAS (75.52). To investigate their respective effects on the prediction accuracy, performance was compared at different thresholds with the reduction in Z value varying from −1.49 to 5.45. Herein, we assigned an approximated reliability boundary for the consensus model mentioned above and performed statistical evaluation on the external test set. The results are presented in Table 4, which indicated that the ACC value approached 0.956 and the AUC value equaled 0.931 when $Z=0.01$; this demonstrated that the model obtained the local optimum at this point and there were 92 compounds remaining in the external data set. Although the threshold became smaller and smaller, samples in the external test set were similar to the training set samples. The AUC value for external validation samples approached nearly 0.95, and the SP value was 1. Thus, we can depend on our desired level of reliability to adaptively adjust Z values and then identify prediction with errors of different degrees. Although the AUC value was 0.967 and other statistical evaluation values were also high when $Z=-1.18$, the AD was narrow and only 21 molecules remained in the external validation set. To consider the predicting ability for both positive and negative samples, we also used G-means values as the most important values when selecting the Z value. So considering both the diversity of data and statistical values of the model, we chose $Z=0.01$ as the parameter in the definition of AD with ACC=0.956, SE=0.972, SP=0.889, and G-means=0.93, which is the best point when Z changed. The confusion matrix between true labels and predict labels is depicted in Figure S2 with four outlier molecules, of which the number of both false-positive samples and false-negative samples is 2. Furthermore, we studied the outlier samples in detail, which are listed in Table S10.

Table 3. Comparison of single models with a combined consensus model for the classification of external set samples.[a]

Model	ACC	SE	SP	G-means	AUC
MACCS+SVM+SMOTE+class weight	0.890	0.908	0.833	0.756	0.871
MACCS+SVM+SMOTE+ENN+class weight	0.883	0.900	0.833	0.75	0.867
MACCS+SVM+no resampling+class weight	0.938	0.982	0.806	0.791	0.897
MD+KNN+SMOTE+ENN	0.931	0.991	0.750	0.743	0.870
MD+KNN+SMOTE	0.931	0.991	0.750	0.743	0.870
Consensus model	0.945	0.982	0.833	0.905	0.908
Consensus model (92 CNS+ chemicals added)	0.966	0.99	0.833	0.908	0.919

[a] ACC=accuracy rate; AUC=area under the receiver operating characteristic curve; CNS=central nervous system; ENN=edited nearest neighbor; KNN=*k*-nearest neighbor; MD=molecular descriptor; SE=selectivity; SMOTE=Synthetic Minority Oversampling Technique; SP=specificity; SVM=support vector machine.

Table 4. Statistical evaluation of different applicability domains adjusted by *Z* values for external validation sets.[a]

Z value	ACC	SE	SP	G-means	AUC	Count	Outlier (all)	Outlier (P)	Outlier (N)
5.45	0.945	0.982	0.833	0.904	0.907	145	8	2	6
3.13	0.943	0.981	0.818	0.896	0.9	141	8	2	6
2.1	0.941	0.981	0.806	0.889	0.894	136	8	2	6
1.88	0.938	0.98	0.793	0.882	0.887	131	8	2	6
1.17	0.944	0.98	0.815	0.893	0.897	125	7	2	5
1	0.942	0.978	0.815	0.893	0.897	119	7	2	5
0.91	0.948	0.978	0.846	0.909	0.912	115	6	2	4
0.76	0.945	0.977	0.833	0.902	0.905	112	6	2	4
0.61	0.943	0.976	0.818	0.894	0.897	105	6	2	4
0.36	0.94	0.975	0.81	0.888	0.892	100	6	2	4
0.23	0.937	0.973	0.81	0.887	0.891	98	6	2	4
0.01	0.956	0.972	0.889	0.93	0.931	90	4	2	2
−0.1	0.953	0.97	0.889	0.929	0.93	85	4	2	2
−0.25	0.95	0.968	0.882	0.924	0.925	80	4	2	2
−0.31	0.947	0.966	0.882	0.923	0.924	75	4	2	2
−0.4	0.943	0.962	0.882	0.921	0.922	70	4	2	2
−0.47	0.938	0.958	0.882	0.92	0.92	65	4	2	2
−0.53	0.933	0.957	0.857	0.905	0.907	62	4	2	2
−0.61	0.927	0.951	0.857	0.903	0.904	55	4	2	2
−0.71	0.92	0.944	0.857	0.9	0.901	50	4	2	2
−0.8	0.911	0.939	0.833	0.885	0.886	45	4	2	2
−0.88	0.9	0.931	0.818	0.873	0.875	40	4	2	2
−0.98	0.886	0.923	0.778	0.847	0.85	35	4	2	2
−1.02	0.9	0.909	0.875	0.892	0.892	30	3	1	2
−1.13	0.92	0.9	1	0.949	0.95	25	2	1	1
−1.18	0.95	0.933	1	0.966	0.967	20	1	1	0
−1.28	0.933	0.909	1	0.953	0.955	15	1	1	0
−1.38	0.9	0.857	1	0.926	0.929	10	1	1	0
−1.49	0.8	0.667	1	0.816	0.833	5	1	1	0

[a] ACC = accuracy rate; AUC = area under the receiver operating characteristic curve; N = negative; P = positive; SE = selectivity; SP = specificity.

Analysis of privileged substructures

We used SARpy (a free tool that excavates substructures by using likelihood ratio) and IG method to identify privileged substructures of both BBB+ and BBB−.[44] For the parameter setting of SARpy, the *atom number* was set between 2 and 15 and *precision* was set to "OPTIMAL." Here, we chose only MACCS and FCFP4 fingerprints to obtain the final top substructures that has high IG values. Some representative privileged substructures are listed in Table 5, of which the IG scores were greater than 0.01 and frequency counts were greater than 25. Among these substructures, IDs 2 and 9 were calculated by using FCFP4 and others were calculated by using MACCS, and six types of BBB+ substructures and six types of BBB− substructures with their specific compounds were obtained. As shown in the table, if the propyl group was shown at the C terminal site, the molecule would be classified into the BBB+ class because this group had a high LogP value and it is easy to cross the BBB. In addition, the substructures such as IDs 10 and 11, containing more hydroxy groups, would be classified into the BBB− class. This was consistent with the common sense that the hydroxy group would increase the hydrophobicity of compounds that could not cross the BBB.

Discussion

Description of resampling methods solving the problem of imbalanced data sets

To solve the problem of imbalanced data sets, several technical strategies were tested to make positive and negative samples balanced at both the algorithmic level and the data level. At the data level, sample rescaling and resampling strategies have been used to balance data by changing the distribution of samples in different classes, including oversampling (SMOTE) and undersampling (RUS) methods.[47] At the algorithmic level, a cost-sensitive learning approach (*class weight*) has also been attempted by setting an excessive cost function to misclassification of a minority class sample.[48] In addition, an ensemble classifier combined with the resampling method such as SMOTE+ENN, which is a novel and promising route to reduce the influence of information loss or information overfit, was used for comparison. Moreover, considering the significance of both SE and SP, G-means ($=\sqrt{SP \times SE}$) was applied as the key statistical parameter.[49]

In the construction of imbalanced data set models, different *k-nearest neighbors* for resampling methods can also influence the evaluation results. For example, for both MACCS and MD features, SMOTE and SMOTE+ENN resampling methods were selected to balance the training set samples. The initial models were evaluated by using 332 test set compounds. And we performed SVM as the base estimator for training, where the fea-

Table 5. Representative substructures and their relative compounds with their possible classes identified by using IG values and frequency counts.[a]

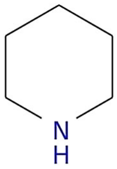
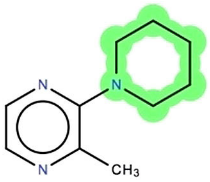
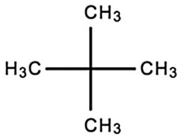
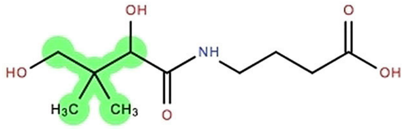
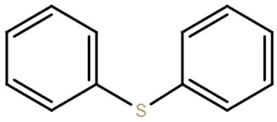
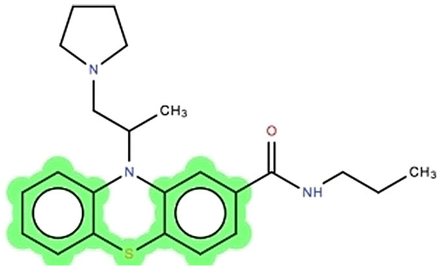
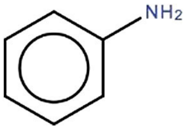
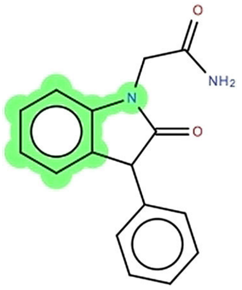
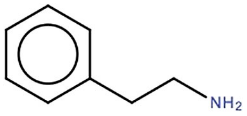
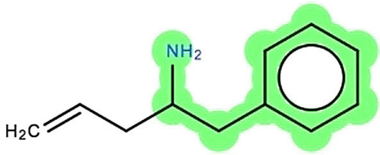
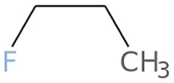
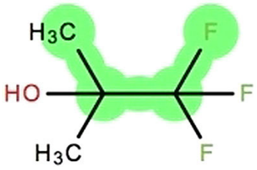
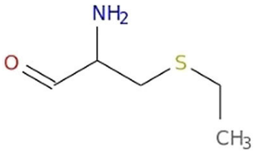
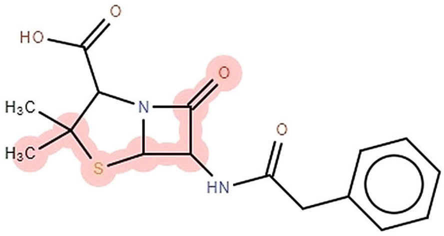
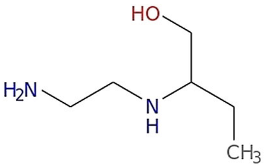
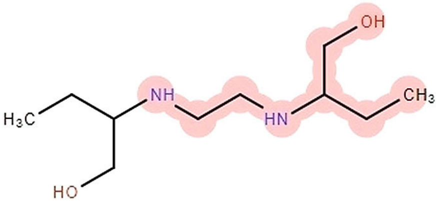
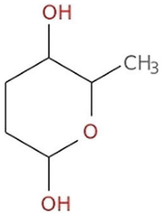
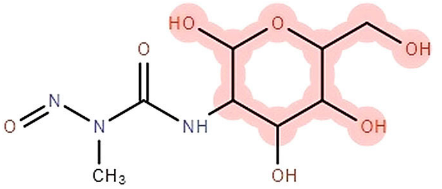
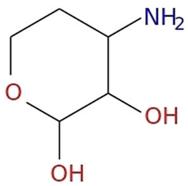
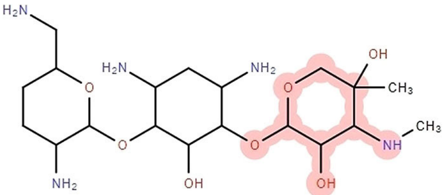
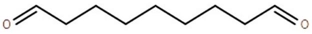
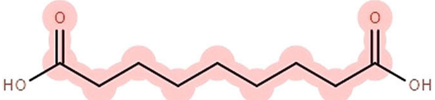
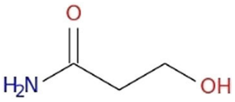
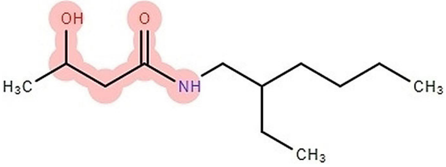
Class	ID	Substructure	Compound	IG	Frequency count
BBB +	1			0.0185	319
BBB +	2			0.0179	171
BBB +	3			0.0157	88
BBB +	4			0.0128	72
BBB +	5			0.0111	101
BBB +	6			0.0107	98
BBB−	7			0.1120	115
BBB−	8			0.0688	62

Table 5. (Continued)

Class	ID	Substructure	Compound	IG	Frequency count
BBB−	9			0.0463	52
BBB−	10			0.0315	29
BBB−	11			0.0233	35
BBB−	12			0.0166	26

[a] BBB+ = compounds that can cross the blood–brain barrier; BBB− = compounds that cannot cross the blood–brain barrier; IG = information gain.

ture selection strategy VT+RFE was adopted for MACCS feature and VT+UFE+RFE for the MD feature so that MACCS feature still contains 112 bits and the MD feature still contains 45 descriptors. As shown in Figure S3, the parameter set at 2 would achieve the best AUC and G-means values for both MACCS and MD features of the training and test set evaluation. When the *k-nearest neighbor* parameter equaled 2, SMOTE and SMOTE+ENN chose 2 samples for each training set sample and then randomly selected one of them to generate the synthetic samples from these samples. For MACCS, the SMOTE method performed better than SMOTE+ENN for AUC values in each *k-nearest neighbor* parameter. The AUC values of MACCS+SMOTE for the training set and test set were 0.960 and 0.928, respectively. But the G-means values of MACCS+SMOTE+ENN increased to 0.889, which was higher than that of MACCS+SMOTE (0.879) when $k=2$. In contrast, for MDs, in the range of *k-nearest neighbor* parameters the SMOTE method always performed better than SMOTE+ENN. By comparison of all types of statistical indicators, it was necessary to take account of the influence of *k-nearest neighbor* parameter settings when constructing robust models.

Comparison of single models with the consensus model

Our previous study demonstrated that the combination of single models to form a consensus model could have a better prediction accuracy than any single model.[50] In this study, because our resampling strategy would cause the reduction in ACC and SE values, we used the consensus model built by using MLP to take account of all the possible classification results of top single models. As shown in Table 3, the overall ACC value (0.945) of the consensus model (not adding 92 CNS+ chemicals) improved marginally compared with other single models, of which the best ACC value was 0.938. Among single models, MACCS+SVM+no resampling+class weight had the best AUC and ACC values and its SP value was lower than that of MACCS+SVM+SMOTE+class weight and MACCS+SVM+SMOTE+ENN+class weight. MD+KNN+SMOTE+ENN and MD+KNN+SMOTE had the same statistical evaluation score, which indicated that resampling methods may be not useful for the KNN algorithm. In addition, the ACC value of 2D descriptors was 0.931, which was better than that of MACCS, but models using MACCS had the highest SP values among all single models. We can see that all these weak single models performed well only in some aspects of the statistical evaluation whereas the consensus model for the external validation set (145 samples) could predict the output results of each single model better and increase its evaluation values (SE = 0.982 and SP = 0.833). Thus, it can be concluded that the consensus model would be more reliable than its component models.

Highlights of our models and comparison with previous studies

First, in the prediction models of BBB penetration, many classical models were built on the basis of MDs, and these studies were hardly concerned about the imbalanced BBB data samples or the methods used to solve this problem were not effective. The advantage of our methods is that we comprehensively considered the problem of imbalanced data samples with properly synthesizing minority samples and retained all the original samples without deletion. In addition, before synthesizing minority samples, we compared the performance of different fingerprints and chose the best fingerprint through 10-fold cross-validation. We not only were concerned about the distinguished prediction results concluded from MDs and fingerprints but also tried to integrate these results all at once. As above results had indicated, the statistical values of our models were higher than those of the models of previously reported studies and the SP values for classifying BBB− had increased significantly. The SP values of the best single model MACCS+SVM+SMOTE+ENN+class weight for the training, test, and external validation sets were 1.000, 0.873, and 0.833, respectively. Although there were biased data, the final model had a relatively good predicting ability for either BBB+ or BBB−.

Second, the consensus model was constructed and evaluated by using the external validation set, and the ACC value was 0.945, which was better than that of any other balanced models. Meanwhile, we added more CNS+ samples into the external validation set and the results suggested that the final model could predict samples more precisely (ACC = 0.966) and the predicting ability for positive samples could increase to 0.99 (SE = 0.99).

Third, we defined the SD of the consensus model, which can distinguish good predictions from bad predictions through the adjustment of parameter settings to our desired level of prediction accuracy.

Finally, using the consensus model, we made a BBB prediction of the 7179 compounds retrieved from DrugBank. The results indicated that the classification model for different drugs for various diseases had distinguished performance, which means that our models have a better capacity to predict and enrich CNS drugs as we want and for other non-CNS diseases it is easier to enrich and filter drugs of the BBB− class.

Conclusions

Six machine learning algorithms were introduced to build a series of single models and a consensus model to estimate the BBB permeability of chemicals. Imbalance learning methods and feature selection strategies were developed to solve the problem of the initial imbalanced samples of the BBB data set. Through 10-fold cross-validation for the initial training set samples, it was found that MACCS fingerprint was the best feature for model construction and outperformed other fingerprints. The comparison of different resampling methods with the evaluation based on test set samples revealed that SMOTE and SMOTE+ENN combined with the adjustment of *class weight* parameter were excellent strategies among single classifier models. Moreover, when constructing models, considering performance difference between 2D MDs and molecular fingerprints, we adopted the consensus idea and constructed a consensus model that combined the top models built by using both MACCS fingerprints and MDs. To ensure the reliability and threshold of our models, three MDs related to LogBB values were used to define the AD of our consensus model and the prediction ACC value in the best domain increased to 0.956 and the SE and SP values were 0.972 and 0.889, respectively. Meanwhile, we added 92 CNS+ chemicals to the external validation set and the ACC value increased to 0.966. These results proved that our models are impartial in predicting both positive samples and negative samples. Furthermore, the resampling strategy used in our experiment can be useful to solve the problem of biased data sets when constructing models. Last but not least, the IG method used to identify substructure patterns would be helpful in classifying BBB+ and BBB− and the substructure excavated in our study can provide significant guidance to medicinal chemists to discover potential BBB+ compounds and rapidly remove or transform compounds with poor BBB permeability. In addition, BBB models are publicly available via our web server admetSAR (http://lmmd.ecust.edu.cn/admetsar1/), which can quickly predict all kinds of small molecules with MW less than 1000 Da.

Abbreviations

ACC = accuracy rate; AD = applicability domain; ADASYN = adaptive synthetic sampling; ATC = Anatomical Therapeutic Chemical; AUC = area under the receiver operating characteristic curve; BBB = blood–brain barrier; BBB+ = compounds that can cross the blood–brain barrier; BBB− = compounds that cannot cross the blood–brain barrier; CNS = central nervous system; ENN = edited nearest neighbor; IG = information gain; KNN = *k*-nearest neighbor; LogP = lipo-hydro partition coefficient; LogPS = surface permeability product; LR = logistic regression; MD = molecular descriptor; MLP = multilayer perceptron neural network; MW = molecular weight; PCA = principal component analysis; PCC = Pearson correlation coefficient; QSAR = quantitative structure–activity relationship; RF = random forest; RFE = recursive feature elimination; RUS = random undersampling; SE = selectivity; SMOTE = Synthetic Minority Oversampling Technique; SP = specificity; SVM = support vector machine; TPSA = topological polar surface area; UFE = univariate feature selection; VT = variance threshold; WLF = weight loss function.

Acknowledgements

We gratefully acknowledge the financial support from the National Natural Science Foundation of China (grant nos. 81273438, 81373329, and 81673356).

Conflict of interest

The authors declare no conflict of interest.

Keywords: blood–brain barrier · imbalanced data · machine learning · QSAR models · resampling methods

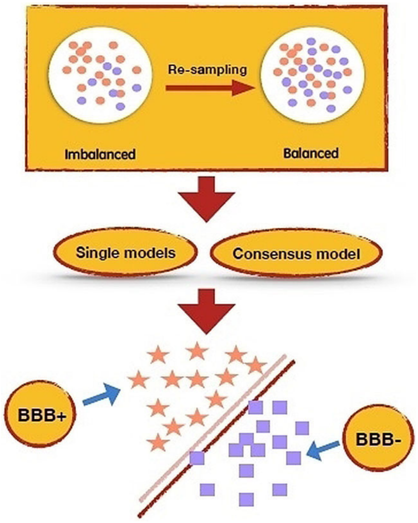
[1] H. van de Waterbeemd, E. Gifford, *Nat. Rev. Drug Discovery* **2003**, *2*, 192 – 204.
[2] D. M. Dambach, D. Misner, M. Brock, A. Fullerton, W. Proctor, J. Maher, L. Dong, K. A. Ford, D. Diaz, *Chem. Res. Toxicol.* **2016**, *29*, 452 – 472.
[3] M. T. Khan, *Curr. Drug Metab.* **2010**, *11*, 285 – 295.
[4] T. J. Hou, X. J. Xu, *J. Mol. Model.* **2002**, *8*, 337 – 349.
[5] C. Merlot, *Drug. Circ. Drug. Discovery Today* **2010**, *15*, 16 – 22.
[6] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P. W. Lee, Y. Tang, *J. Chem. Inf. Model.* **2012**, *52*, 3099 – 3105.
[7] N. Abbott, L. Ronnback, E. Hansson, *Nat. Rev. Neurosci.* **2006**, *7*, 41 – 53.
[8] P. Ballabh, A. Braun, M. Nedergaard, *Neurobiol. Dis.* **2004**, *16*, 1 – 13.
[9] L. Di, E. H. Kerns, I. F. Bezar, S. L. Petusky, Y. Huang, *Pharm. Sci.* **2009**, *98*, 1980 – 1991.
[10] T. S. Carpenter, D. A. Kirshner, E. Y. Lau, S. E. Wong, J. P. Nilmeier, F. C. Lightstone, *Biophys. J.* **2014**, *107*, 630 – 641.
[11] A. A. Toropov, A. P. Toropova, M. Beeg, M. Gobbi, M. Salmona, *J. Pharmacol. Toxicol. Methods* **2017**, *88*, 7 – 18.
[12] Y. H. Zhao, M. H. Abraham, A. Ibrahim, P. V. Fish, S. Cole, M. L. Lewis, M. J. de Groot, D. P. Reynolds, *J. Chem. Inf. Model.* **2007**, *47*, 170 – 175.
[13] I. F. Martins, A. L. Teixeira, L. Pinheiro, A. O. Falcao, *J. Chem. Inf. Model.* **2012**, *52*, 1686 – 1697.
[14] D. A. Konovalov, D. Coomans, E. Deconinck, Y. V. Heyden, *J. Chem. Inf. Model.* **2007**, *47*, 1648 – 1656.
[15] W. S. Noble, *Nat. Biotechnol.* **2006**, *24*, 1565 – 1567.
[16] L. Zhang, H. Zhu, T. I. Oprea, A. Golbraikh, A. Tropsha, *Pharm. Res.* **2008**, *25*, 1902 – 1914.
[17] A. Guerra, J. A. Páez, N. E. Campillo, *QSAR Comb. Sci.* **2008**, *27*, 586 – 594.
[18] C. Suenderhauf, F. Hammann, J. Huwyler, *Molecules* **2012**, *17*, 10429 – 10445.
[19] O. A. Raevsky, S. L. Solodova, A. A. Lagunin, V. V. Poroikov, *Biomed. Khim.* **2013**, *7*, 95 – 107.
[20] J. Shen, F. Cheng, Y. Xu, W. Li, Y. Tang, *J. Chem. Inf. Model.* **2010**, *50*, 1034 – 1041.
[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.* **2002**, *16*, 321 – 357.
[22] Q. Wang, Z. H. Luo, J. C. Huang, Y. H. Feng, Z. Liu, *Comput. Intel. Neurosci.* **2017**, 1827016.
[23] H. He, Y. Bai, E. A. Garcia, S. Li, *Neural. Networks* **2008**, 1322 – 1328.
[24] G. Lemaitre, F. Nogueira, C. K. Aridas, *J. Mach. Learn. Res.* **2017**, *18*, 1 – 5.
[25] K. Roy, P. Ambure, R. B. Aher, *Chemom. Intell. Lab. Syst.* **2017**, *162*, 44 – 54.
[26] M. Muehlbacher, G. M. Spitzer, K. R. Liedl, J. Kornhuber, *J. Comput.-Aided Mol. Des.* **2011**, *25*, 1095 – 1106.
[27] W. Wang, M. T. Kim, A. Sedykh, H. Zhu, *Pharm. Res.* **2015**, *32*, 3055 – 3065.
[28] T. J. Hou, X. J. Xu, *J. Chem. Inf. Model.* **2003**, *43*, 2137 – 2152.
[29] *Maestro*, 10.2.010; Schrödinger: New York, **2015**.
[30] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminf.* **2011**, *3*, 33.
[31] Z. Gao, Y. Chen, X. S. Cai, R. Xu, *Bioinformatics* **2016**, *33*, 901 – 908.
[32] D. S. Wishart, C. Knox, A. C. Guo, *Nucleic Acids Res.* **2006**, *34*, 668 – 672.
[33] C. W. Yap, *J. Comput. Chem.* **2011**, *32*, 1466 – 1474.
[34] S. Prasanna, R. J. Doerksen, *Curr. Med. Chem.* **2009**, *16*, 21 – 41.
[35] *Discovery Studio Modeling Environment*, Release 3.5; Accelrys Software Inc: San Diego, CA [USA], accessed May 26, **2010**.
[36] V. Khanna, S. Ranganathan, *J. Cheminf.* **2011**, *3*, 30.
[37] S. Christoph, H. Yongquan, K. Stefan, H. Oliver, L. A. Edgar, E. Willighagen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493-500.
[38] F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, *J. Mach. Learn. Res.* **2011**, *12*, 2825 – 2830.
[39] E. Gysels, P. Renevey, P. Celka, *IEEE. T. Signal. Proces.* **2005**, *85*, 2178 – 2189.
[40] P. Sedgwick, *N. Z. Med. J.* **1996**, *109*, 38.
[41] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37 – 52.
[42] R. Blagus, L. Lusa, *BMC Bioinf.* **2013**, *14*, 106.
[43] B. P. Pedersen, G. Ifrim, P. Liboriussen, K. B. Axelsen, M. G. Palmgren, P. Nissen, C. Wiuf, C. N. Pedersen, *PLoS One* **2014**, *9*, e85139.
[44] H. Du, Y. Cai, H. Yang, H. Zhang, Y. Xue, G. Liu, T. Yun, W. Li, *Chem. Res. Toxicol.* **2017**, *30*, 1209 – 1218.
[45] M. S. Pepe, T. Cai, G. Longton, *Biometrics* **2006**, *62*, 221 – 229.
[46] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, *Molecules* **2012**, *17*, 4791 – 4810.
[47] R. Blagus, L. Lusa, *BMC Bioinf.* **2013**, *14*, 64.
[48] H. He, E. A. Garcia, *IEEE. T. Knowl. Data. En.* **2009**, *21*, 1263 – 1284.
[49] S. Ertekin, J. Huang, L. Bottou, L. Giles, *ACM* **2007**, 127 – 136.
[50] F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, P. W. Lee, Y. Tang, *J. Chem. Inf. Model.* **2011**, *51*, 996 – 1011.

Manuscript received: August 8, 2018
Accepted manuscript online: August 15, 2018
Version of record online: ■■ ■■, 0000

Z. Wang, H. Yang, Z. Wu, T. Wang, W. Li, Y. Tang, G. Liu**

■■ – ■■

In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods



Beauty with brain: Data curation, feature selection, machine learning algorithms, resampling methods, consensus modeling techniques, and applicability domain analysis are used in a combinatorial manner to build classification models to estimate the blood–brain barrier permeability of chemical compounds. In addition, small molecules retrieved from DrugBank are assessed to help evaluate the best consensus model with high sensitivity and specificity values.