

<https://helda.helsinki.fi>

Multiple machine learning methods aided virtual screening of Na(V)1.5 inhibitors

Kong, Weikaixin

2023

Kong , W , Huang , W , Peng , C , Zhang , B , Duan , G , Ma , W & Huang , Z 2023 , ' Multiple machine learning methods aided virtual screening of Na(V)1.5 inhibitors ' , Journal of Cellular and Molecular Medicine , vol. 27 , no. 2 , pp. 266-276 . <https://doi.org/10.1111/jcmm.17652>

<http://hdl.handle.net/10138/355851>
<https://doi.org/10.1111/jcmm.17652>

cc_by
publishedVersion


Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Multiple machine learning methods aided virtual screening of Na_v1.5 inhibitors

Weikaixin Kong^{1,2,3}  | Weiran Huang¹ | Chao Peng¹ | Bowen Zhang⁴ | Guifang Duan¹ | Weining Ma⁵ | Zhuo Huang^{1,6}

¹Department of Molecular and Cellular Pharmacology, School of Pharmaceutical Sciences, Peking University Health Science Center, Beijing, China

²Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

³Institute Sanqu Technology (Hangzhou) Co., Ltd., Hangzhou, China

⁴ComMedX (Computational Medicine Beijing Co., Ltd.), Beijing, China

⁵Department of Neurology, Shengjing Hospital affiliated to China Medical University, Shenyang, China

⁶State Key Laboratory of Natural and Biomimetic Drugs, Department of Molecular and Cellular Pharmacology, School of Pharmaceutical Sciences, Peking University Health Science Center, Beijing, China

Correspondence

Weining Ma, Department of Neurology, Shengjing Hospital affiliated to China Medical University, Shenyang, China.
Email: mawein1985@163.com

Zhuo Huang, Department of Molecular and Cellular Pharmacology, School of Pharmaceutical Sciences, Peking University Health Science Center, Beijing, China.
Email: huangz@hsc.pku.edu.cn

Funding information

by Chinese National Programs for Brain Science and Brain-like intelligence technology, Grant/Award Number: 2021ZD0202102; National Natural Science Foundation of China, Grant/Award Number: 31871083, 32000674 and 81371432

Abstract

Na_v1.5 sodium channels contribute to the generation of the rapid upstroke of the myocardial action potential and thereby play a central role in the excitability of myocardial cells. At present, the patch clamp method is the gold standard for ion channel inhibitor screening. However, this method has disadvantages such as high technical difficulty, high cost and low speed. In this study, novel machine learning models to screen chemical blockers were developed to overcome the above shortage. The data from the ChEMBL Database were employed to establish the machine learning models. Firstly, six molecular fingerprints together with five machine learning algorithms were used to develop 30 classification models to predict effective inhibitors. A validation and a test set were used to evaluate the performance of the models. Subsequently, the privileged substructures tightly associated with the inhibition of the Na_v1.5 ion channel were extracted using the bioalerts Python package. In the validation set, the RF-Graph model performed best. Similarly, RF-Graph produced the best result in the test set in which the Prediction Accuracy (Q) was 0.9309 and Matthew's correlation coefficient was 0.8627, further indicating the model had high classification ability. The results of the privileged substructures indicated Sulfa structures and fragments with large Steric hindrance tend to block Na_v1.5. In the unsupervised learning task of identifying sulfa drugs, MACCS and Graph fingerprints had good results. In summary, effective machine learning models have been constructed which help to screen potential inhibitors of the Na_v1.5 ion channel and key privileged substructures with high affinity were also extracted.

KEYWORDS

chemical inhibitors, machine learning, Na_v1.5, privileged substructures

Weikaixin Kong, Weiran Huang and Chao Peng contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of Cellular and Molecular Medicine* published by Foundation for Cellular and Molecular Medicine and John Wiley & Sons Ltd.

1 | INTRODUCTION

Voltage-gated sodium channel subtype 1.5 (Na_v1.5) is the major cardiac voltage-gated sodium ion channel, which plays a vital role in the generation of the cardiac action potential and in the propagation of the electrical impulses in the heart.¹ The role of Na_v1.5 in the aetiology of numerous cardiac anomalies strongly suggests the proper regulation of the channel is critical for normal heart function. Also, the pivotal function of Na_v1.5 in normal heart operation has been discovered by researching the genetic mutation of SCN5A located on chromosome 3p21 which encodes Na_v1.5. Tan et al.² found Na_v1.5 is linked to congenital and drug-acquired Long QT Syndrome (LQTS), Brugada Syndrome (BS), conduction disorders and sudden infant death syndrome. Other research has also indicated that this gene is also associated with disorder-ventricular arrhythmia and dilatative cardiomyopathy.³

Cardiac toxicity of drugs has always been forefront for drug administration, and many non-cardiac drugs, especially psychotropic drugs, can introduce ventricular fibrillation and syncope and sudden cardiac arrest (SCA) by reducing cardiac excitability through Na_v1.5.⁴

At present, the patch-clamp electrophysiological method is still the gold standard for ion channel drug screening. However, this method has disadvantages such as high technical difficulty, high cost and low speed. In this case, drug virtual screening based on computational methods can help to find drugs with higher specificity and bioactivity along with a higher speed and less consumption. Machine learning (ML) has become very popular recently, due to increased data availability and algorithmic methods, and has been employed in drug design and screening. ML approaches provide a set of tools that can use abundant, high-quality data to solve discovery and decision-making for well-specified questions. ML models are based on existing data to do predictions which can accelerate the new drug discovery process, which have been applied in all stages of drug discovery.⁵⁻⁸ Examples include the identification of prognostic biomarkers,^{9,10} drug repurposing,¹¹⁻¹⁴ and analysis of side effect.¹⁵ With the development of high-throughput screening technology, countless meaningful experimental data are being produced to the benefit of future work using computer-dependent drug design.

In this study, building ML classification models are established based on molecular features to predict chemicals that have a high affinity of Na_v1.5. A comparison with the graph convolutional neural network method is also made to find the most effective prediction method.

2 | MATERIALS AND METHODS

2.1 | Data preparation

In this study, the inhibition data were acquired from the ChEMBL Database (ChEMBL ID:CHEMBL1980). The download date was 3 October 2020. The Homo sapiens Na_v1.5 ion channel was used as the target and IC₅₀ as the experimental method. The data preparation

process was shown in Figure S1. After deleting duplicate molecules and only keeping experimentally verified molecules, there were 1957 diverse compounds left which were encoded into a standard simplified molecular-input line-entry system (SMILES).

Then, according to other researchers' previous work,¹⁶⁻¹⁸ 30,000nM was used as the threshold to divide molecules into positive molecules and negative molecules. The molecules with IC₅₀ values less than 30,000nM were tagged to the label "1" which represented the molecules that were able to inhibit Na_v1.5. In contrast, the negative molecules were tagged to the label "0." In this step, we obtained 1785 positive molecules and 172 negative molecules. To exclude the polymers and make sure selected molecules were drug-like, the compounds with atom numbers over 120 were deleted, and then only molecules confirmed by Lipinski's rule of 5 were left. After the above process, the positive molecules were reduced to 1558 molecules and the negative molecules were reduced to 96 molecules.

In general, Na_v1.5 inhibitors reported in the same research paper often have very similar structures. If these molecules appear in both the training set and test set, a "data leakage" problem would arise. As shown in Figure 1A, two molecules reported in the same article were highly similar and the Tanimoto similarity (using the ECFP4 fingerprint to obtain the similarity) of them was 0.9590.¹⁹ Hence, hierarchical clustering was performed on the selected positive molecules according to the inter-group Tanimoto distance (1-Tanimoto similarity). In clustering, we used the hierarchical clustering method. Three functions from the RDKit package (<http://www.rdkit.org/>) in python were employed in this process including BulkTanimotoSimilarity, ForwardSDMolSupplier and GetMorganFingerprintAsBitVect. Then undersampling was performed by setting a certain cutoff value in the clustering tree to reduce molecular similarity among positive molecules. To make sure the number of left positive molecules was not too small to build machine learning models, the cutoff value was set from 0.1 to 0.6, and then a suitable value was chosen.

In the drug screening process, to distinguish effective inhibitors from thousands of negative molecules, the negative molecules would have wider chemical space. We try to consider this point in the model training and evaluation processes. So there were some molecules randomly extracted from the ChEMBL database were also included as another part of negative molecules. The "chembl_webresource_client" package (https://github.com/chembl/chembl_webresource_client) in Python was used to implement this process. Then these randomly extracted molecules were processed similarly as mentioned above¹: the compounds with atom number over 120 were deleted and² only molecules confirmed by Lipinski's rule of 5 were left.

Following the previous process, a positive dataset containing 364 molecules and a negative dataset containing 400 (96+304) molecules were obtained. Then, the similarity of negative molecules were visualization by Tanimoto similarity and heatmap. The positive and negative datasets were then divided into a training set, a validation set and a test set using the proportions of 2:1:1, as shown in Table 1.

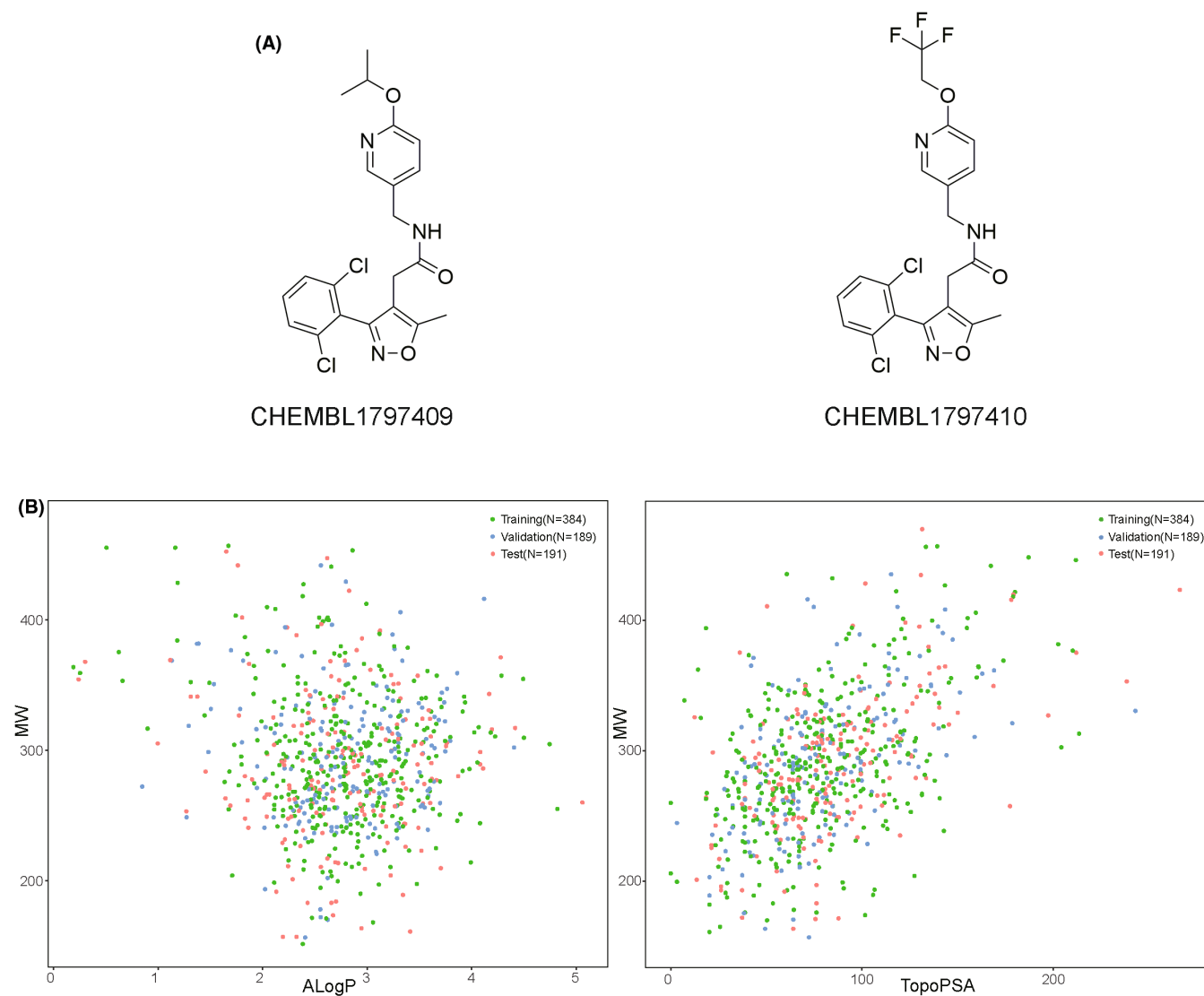


FIGURE 1 (A) Two similar molecules from the same study. (B) Chemical diversity analysis of training, test and validation datasets. The y-axis is MW, where the x-axis on the left is ALogP and the x-axis on the right is TPSA. Green, blue and red points represent the samples in the training set, validation set and test set, respectively.

TABLE 1 Number of molecules in the training set, validation set and test set.

	Positive molecules	Negative molecules	Total
Training set	184	200	384
Validation Set	89	100	189
Test set	91	100	191
Total	364	400	764

2.2 | Sample distribution and similarity test

The distribution of samples would influence the quality of classification models. To test the distribution of the datasets, three

descriptors of these molecules including molecular weight (MW), GhoseCrippen LogKow (ALogP) and Topological polar surface area based on fragment contributions (TPSA) were plotted as scatter plots to show the distribution. The AlogP value represents the partition coefficient between octanol and water, which is crucial for the hydrophobicity of the molecule. It is based on the Ghose-Crippen method,²⁰ which is calculated from a regression equation based on the hydrophobicity contributions of 120 atom types, including bonding of H, C, N, O, S and halogens. Further, molecular Tanimoto similarity characterized by ECFP4 fingerprints was employed to test the similarity of the samples. The heat maps describing the overall similarity of 100 molecules randomly selected from the training set were used to visualize molecular similarity. The average similarity was also calculated to evaluate the similarity of these molecules.

2.3 | Extraction of molecular features

Firstly, six molecular descriptors were generated using PaDEL-Descriptor to perform a simple comparison of positive and negative molecules. These six descriptors included molecule weight (MW), atomic polarizability (apol), the logarithm of 1-octanol/water partition coefficient (ALogP), the number of hydrogen bond donors (nHBDon), the number of rotatable bonds (nRot) and the number of hydrogen bond acceptors (nHBAcc),^{21–23} five of which were Lipinski's rule of five. A *T*-test was then used to evaluate the results. After comparison, to construct effective classifiers, six kinds of molecular fingerprints were used, which were generated by PaDEL-Descriptor software. These six fingerprints included CDK fingerprints (CDK, 1024 bits), Estate fingerprints (Est, 79 bits), Extended fingerprints (Ext, 1024 bits), Graph only fingerprints (GO, 1024 bits), MACCS fingerprints (MACCS, 166 bits) and PubChem fingerprints (Pub, 881 bits).

2.4 | Establishment of classification models

Five machine learning algorithms implemented by “sklearn” Python package (<http://www.scikit-learn.org/>) were used to construct machine learning models, namely logistic regression (LR), support vector machine (SVM), naive Bayes (NB), Multilayer Perceptron (MLP) and random forest (RF).²⁴ The 5-fold cross validation (CV) and grid search were used to find the best parameters of these classifiers in the training set.

Logistic regression (LR) is an algorithm which uses least squares for developing a linear model describing a response from an explanatory variable(s).²⁵ In this case, a generalized linear model is established using the Sigmoid function as the connection function.

Support vector machine (SVM) aims to find a hyperplane in the multi-dimension vector space in which each dimension represents a feature to classify two classes. In multi-dimensional problems, it uses kernel functions to map data to a feature space in which a linear separator can be found.²⁶ In this study, different molecular fingerprints consist of multi-dimensional features where positive dataset and negative dataset are distributed in different areas divided by an unknown hyperplane.

Naive Bayes (NB) is a simple approach using Bayes' theory to find the best classification method.²⁷ Bayes' theory aims to make the optimal decision by generating the posterior class probability of a test data on the basis of class conditional density estimation and class prior probability.²⁸

Multilayer perceptron (MLP) is a kind of machine learning method which has a multi-layered neuron structure. This model is suitable for nonlinear fitting, because of a lot of parameters. Many research proved that MLP methods are suitable for drug-target interaction prediction.^{29,30} When the number of training data is big, the deep learning models based on multilayer perceptron could always have good prediction results.³¹

Random forest (RF) is an ensemble method consisting of many individual decisions. The final predicted label depends on the vote of

each decision tree.³² The decision tree method is a commonly used data mining method to build classification models.³³

Graph convolutional artificial neural networks (ANN) can treat the structure of molecules as a network, transform molecules into structure matrices and feature matrices and perform feature transfer and model training on the molecular structure. Related research shows that the graph convolution method has achieved good results on very large sample volume molecular property prediction tasks, but it is not in general as good as traditional ML methods such as SVM on certain specific data sets.³⁴ The GraphConvModel function of the “DeepChem” package (<https://www.deepchem.io/>) in Python was used to establish a graph convolutional ANN model for comparison with other classification models.³⁵ The setting of parameters are batch_size = 10, mode = ‘classification’, nb_epoch = 10.

2.5 | Model evaluation

The validation set and test set were used to evaluate the performance of these different models. Prediction accuracy (Q), sensitivity (SE), specificity (SP) and Matthew's correlation coefficient (MCC) were employed to evaluate the classification models, as shown below.³⁶ Among them, the MCC evaluation index is a common classifier evaluation index calculated based on the confusion matrix. MCC value was used as a determining evaluation index³⁶ in our research. TP, TN, FP and FN were parameters which represent the number of true positives, the number of true negatives, the number of false positives and the number of false negatives, respectively. Further, the receiver operating curves (ROC) was plotted which described the relation of FPR (False Positive Rate) and TPR (True Positive Rate) in different models and the PR curves which described the Precision and TPR were also plotted. To evaluate the models, the area under the curve (AUC) of both curves was used.

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

$$SE = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

we want to find models which have good prediction performance both in the validation set and test set. Then, the top 10 models in the validation set were selected based on the MCC to do prediction in the test set. In addition, the top 5 models in the validation set were used to build an ensemble learning model to improve the prediction performance.³⁷ The prediction result of the ensemble learning model

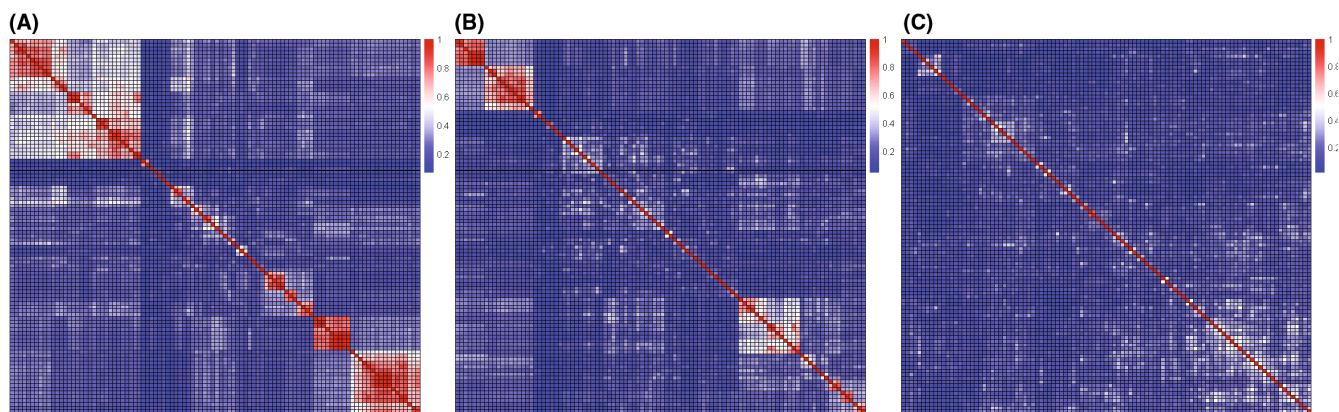


FIGURE 2 The Tanimoto similarity heatmap of 100 positive and 100 negative molecules. (A) Molecular similarity result of positive molecules before excluding similar molecules. The average similarity of all positive molecules is 0.301. (B) Molecular similarity result after excluding similar molecules. The average similarity of all positive molecules is 0.238. (C) Molecular similarity result of negative molecules. The average similarity of all negative molecules is 0.204.

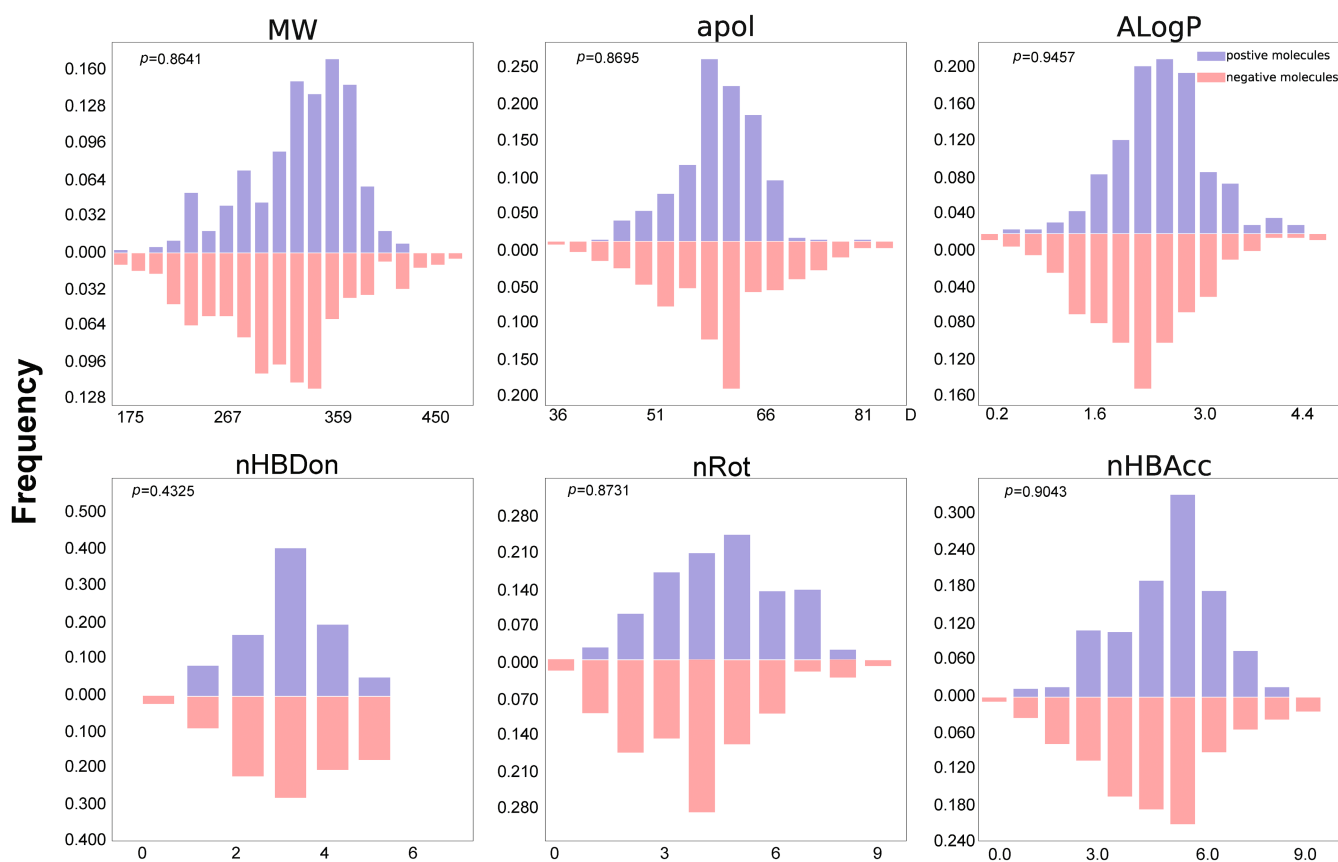


FIGURE 3 The distribution of positive blockers and negative compounds in six descriptors. The molecular descriptors are MW, apol, ALogP, nHBDon, nRot and nHBacc. Blue bars represent positive molecules and red bars represent negative molecules.

was simple majority voting of the top 5 classification models. In other words, the final classification of a new sample would depend on the number of “1” labels in the five basic classifiers. For example, if there were 3/5 basic learners predicting “1” label for the sample, the final predicted label would be “1.”

2.6 | Privileged substructure analysis

The privileged substructures (alert substructure) were the potential fragments of chemicals that may bind to target proteins and operate functions. Molecules containing such fragments are of

TABLE 2 Top ten classification models in the test set.

	Q	SE	SP	AUC*	MCC
RF_Graph	0.931	0.902	0.958	0.947	0.863
LR_CDK	0.931	0.913	0.948	0.957	0.862
RF_CDKextended	0.926	0.88	0.969	0.956	0.854
RF_PubChem	0.926	0.88	0.969	0.951	0.854
RF_CDK	0.92	0.87	0.969	0.96	0.844
LR_CDKextended	0.92	0.935	0.906	0.953	0.841
LR_PubChem	0.915	0.902	0.927	0.944	0.83
RF_Estate	0.91	0.891	0.927	0.942	0.819
LR_Graph	0.894	0.891	0.896	0.93	0.787
RF_MACCS	0.862	0.837	0.885	0.948	0.724

TABLE 3 Evaluation results of DeepChem model and ensemble models.

id	Q	SE	SP	AUC*	MCC
deepchem_validation	0.784	0.731	0.832	0.854	0.570
deepchem_test	0.782	0.630	0.927	0.837	0.586
ensemble_validation	0.937	0.914	0.959	-	0.854
ensemble_test	0.928	0.891	0.969	-	0.874

particular interest to researchers. Hence, the “bioalerts” package (<https://github.com/isidro/bioalerts>) was used to extract the alert structure of the Na_v1.5 ion channel. All datasets in this study were utilized to analyse privileged substructures. The fingerprint used in this method was ECFP4. Positive molecules and negative molecules were counted by setting searching radius (radi = 2, 3 and 4). The probability for a substructure to be a structural alert was derived from the probability density function of the binomial distribution in the positive and negative groups. These were used to calculate a P value³⁸ which indicated the level of significance when considering a given substructure as a structural alert. Only when the possibility of a certain substructure occurring in positive molecules was significantly larger than that occurring in negative molecules, the substructure would be then recognized as an alert substructure.

3 | RESULTS

3.1 | Data set preparation and analysis

As for positive molecules, we obtained 2145 molecules with IC50 values from the ChEMBL database (Figure S1A). Firstly, we deleted duplicate molecules and only keep the molecules which were verified experimentally. In this step, 1957 molecules left (Figure S1A). Then, we kept the molecules with IC50 less than 30,000nM and 1758 molecules left. In this step, there are 172 molecules with IC50 greater than 30,000nM which were regarded as part of negative molecules (Figure S1B). In 1758 positive molecules, there are 1558 molecules with N(atom) < 120 and they meet Linpiniski's rule of five

(Figure S1A). So these 1558 molecules were used to do clustering and undersampling to reduce molecular similarity. After undersampling, only 364 molecules were left.

As for negative molecules, the first part is 172 molecules as mentioned above. In these molecules, 96 molecules meet Linpiniski's rule of five and N(atom) < 120. In addition, we also extract molecules from ChEMBL randomly to extend the negative molecule set. We randomly extracted 304 molecules which met Linpiniski's rule of five and N(atom) < 120. These 304 molecules were another part of the negative molecules. In total, we got 400 (96 + 304) molecules.

These positive and negative molecules were then divided into a training set, a validation set and a test set in the proportions of 2:1:1 as shown in Table 1. Then MW, ALogP and TPSA were used as the main representation of the chemistry space to show the diversity of samples. As shown in Figure 1B, these three datasets are almost spread uniformly in the same chemistry space which proved the rationality of the sampling process.

3.2 | Result of molecular similarity test

When evaluating molecular similarity, Tanimoto similarity based on ECFP4 fingerprint was used. We used the hierarchical clustering method to perform undersampling to reduce molecular similarity and set the method parameter as “average”: the between-group distance (Tanimoto distance = 1 – Tanimoto similarity) is equal to the average distance between the two group objects. The cutoff value in the clustering tree was set to a range from 0.1 to 0.8, as shown in Figure S2. When the cutoff value increases, the number of left positive molecules decreased. And when the cutoff value changed from 0.4 to 0.5, the decreasing speed of positive molecules reached the maximum. To avoid too little molecules left, 0.4 was chosen as the final cutoff value. And after clustering and undersampling, there were 364 positive molecules left (Table 1 and Figure S1). Figure 2 showed the similarity of 100 randomly selected positive or negative molecules. The average similarity value of the whole positive molecule set before undersampling samples is 0.301 (Figure 2A); the average similarity value after undersampling is 0.238 (Figure 2B) and the average similarity of the whole negative molecule set was only

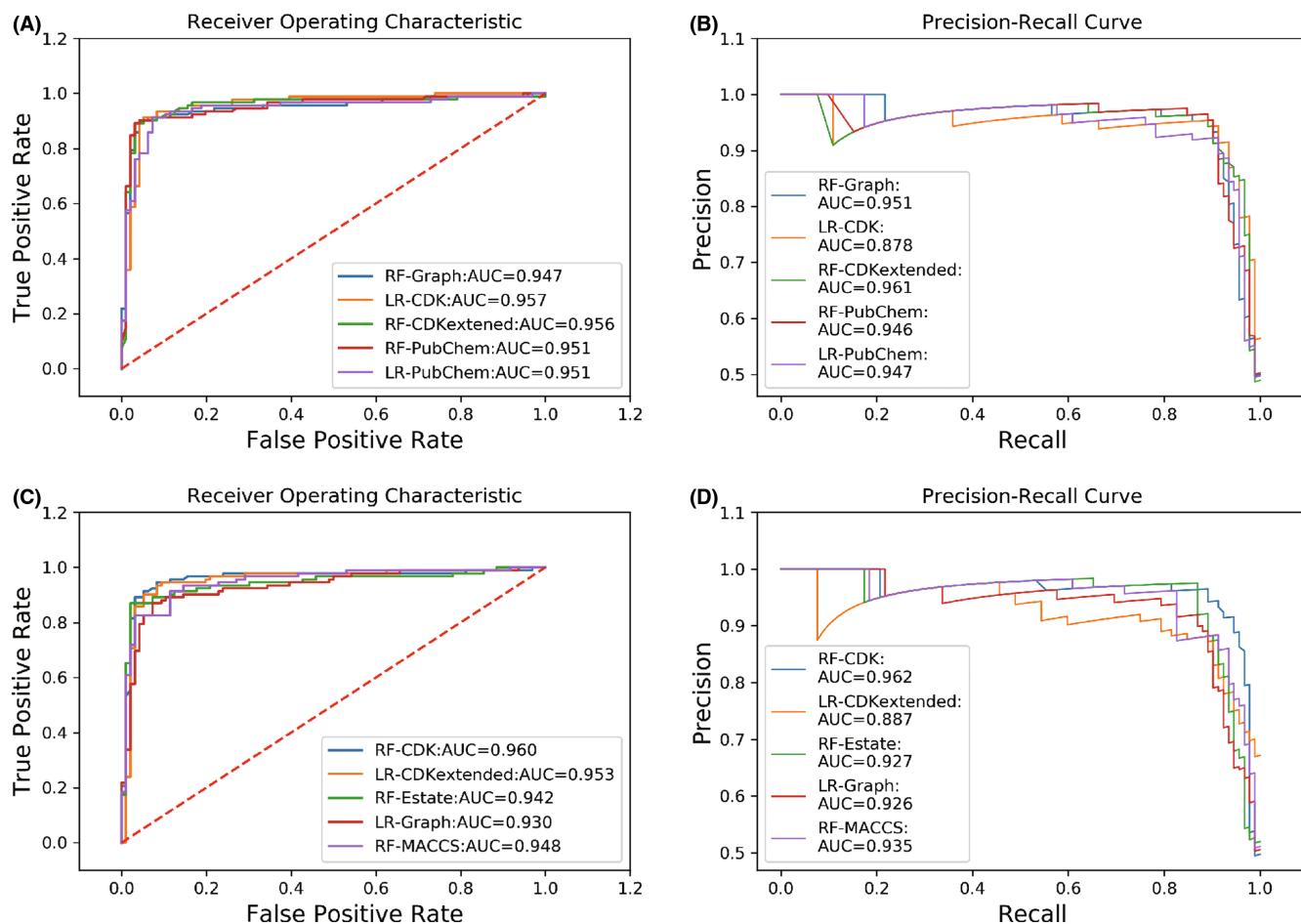


FIGURE 4 The ROC and PR curve of the top ten models in the test set. (A) The ROC curves of the top five models with biggest MCC and their AUC in the test set. (B) The PR curves of the top five models with biggest MCC and their AUC in the test set. (C) The ROC curves of the top 6–10 models and their AUC in the test set. (D) The PR curves of the top 6–10 models in the top ten models and their AUC in the test set.

0.204 (Figure 2C). And compared with other research results,^{39–41} the molecule similarity here (0.238 for positive molecules and 0.204 for negative molecules) has lower values. The above results showed that clustering and undersampling can effectively reduce molecular similarity in the positive molecules and there is no need to do clustering and undersampling in the negative molecules. Building models based on data which have wider chemical space can make the models have wider application range.

3.3 | Distribution analysis of molecular descriptors

Six molecular descriptors were used to analyse the distribution of 6 descriptors between positive and negative molecules using t-test. The *p* Values in MW, apol, AlogP, nHBDon, nRot, and nHBAcc are 0.8641, 0.8695, 0.9457, 0.4325, 0.8731 and 0.9043, respectively, which proved that positive and negative molecules have a similar distribution in these descriptors and cannot be distinguished only by some simple features, as shown in Figure 3. Hence, it is necessary to build the proposed classification models based on molecular fingerprints.

3.4 | Evaluation of classification models

The result of cross validation was shown in Table S1. Then, the best model in every machine learning method was used to establish models using the whole training set. The performance of the developed models on the training and validation sets is shown in Table S2 and Table S3 and was ranked by MCC. The best model in the training set is RF_Graph (MCC = 0.887, AUC = 0.966, Table S2), while the MCC values of RF_Graph in the validation set and test set were 0.885 (Table S3) and 0.863 (Table 2), which were not significantly less than 0.887. So in this condition, we did not meet overfitting problem in this task. In the validation set, the top ten models based on MCC value for the validation dataset are RF-Graph, LR-CDK, RF-CDKextended, RF-PubChem, LR-PubChem, RF-CDK, LR-CDKextended, RF-Estate and LR-Graph, RF-MACCS. Then, the top ten models were chosen and evaluated in the test set (Table 2). The RF-Graph model had the best result for the test dataset: Q was 0.931, SE was 0.902, SP was 0.958, AUC of the ROC curve was 0.947 and MCC was 0.863. In both the test set and the validation set, the best model was RF-Graph, which indicated that this model could be an excellent classifier for this study. The MCC values of the DeepChem model on the

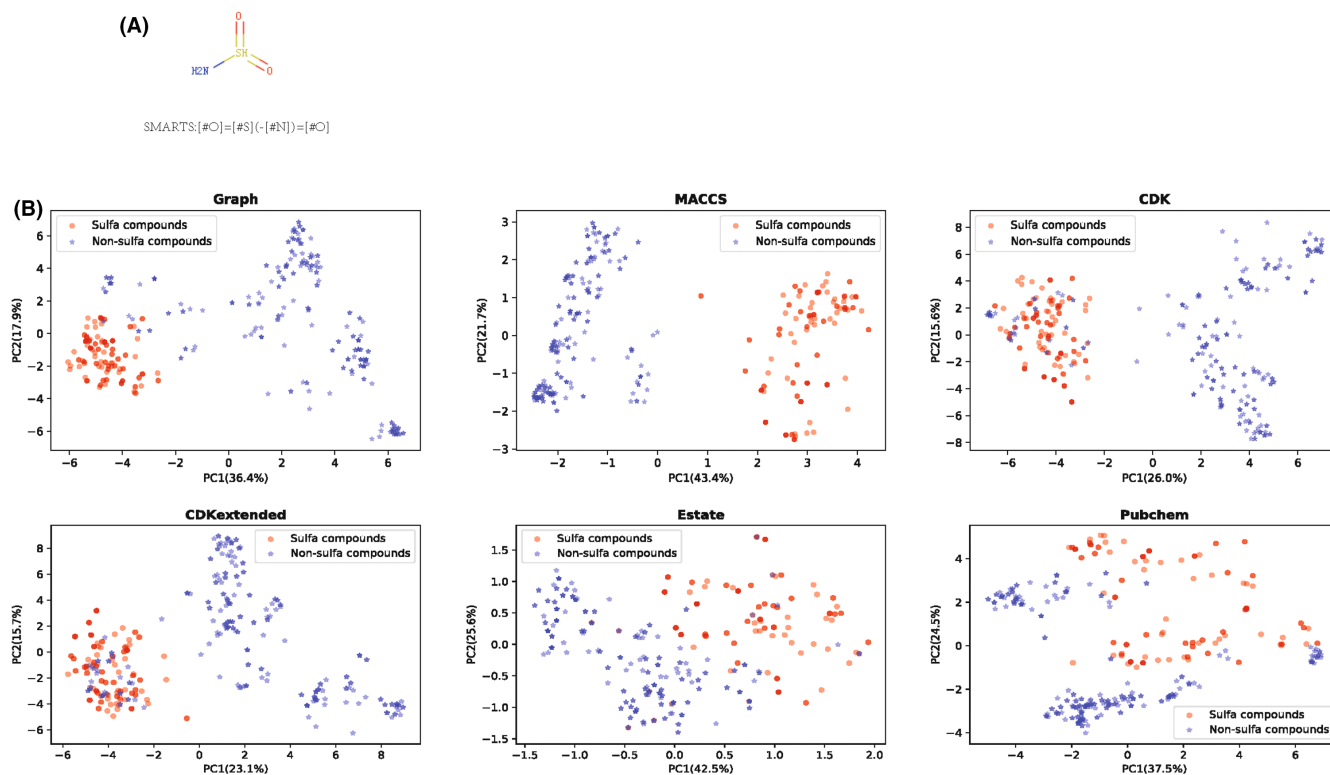


FIGURE 5 (A) The characteristic structure of sulfa drugs and its SMARTS code. (B) Principal component analysis results of positive molecules. The red dots are sulfa drugs and the blue dots are non-sulfa drugs. The x-axis is the first principal component (variance contribution rate) and the y-axis is the second principal component (variance contribution rate).

validation dataset and the test dataset were 0.570 and 0.586, respectively (Table 3). The ensemble model has an MCC value of 0.854 in the validation set and 0.874 in the test set (Table 3). Compared to the results of the RF-Graph, there is no obvious improvement in the ensemble model. In addition, the ROC curves and PR curves of the top ten models in the test set are shown in Figure 4. It can be seen that the gap between the ROC curves and the PR curves is not obvious for the top 10 models.

3.5 | Analysis of privileged substructures

The fragments with $p < 0.05$ would have a high frequency to bind to the target protein, and they were considered as an alert substructure. The privileged substructures were sorted according to the P value from small to large. As shown in Table S4, these privileged substructures included some typical fragments of sulfonamides and some fragments which have a large structural Steric hindrance. They are more common in $\text{Na}_v1.5$ blockers, indicating that chemicals that contain them may have a high possibility to inhibit $\text{Na}_v1.5$. Previous research has partially revealed the reason they can inhibit the $\text{Na}_v1.5$ effectively. There are two possible mechanisms. One is that some structure with a large Steric hindrance blocks the pore physically where sodium ions pass through.⁴² This theory fits fragments S5, S8, S9 and S6. The other theory is that some chemicals change the conformation of the channel by binding with some peptide residue

through van der Waals interactions and salt bridge, which fits some chemicals with sulfa fragments such as S1, S3, S6 and S7.⁴³

We group positive molecules based on the alert substructure. We converted all positive molecules from SMILES to SMARTS to find molecules containing the characteristic structure of sulfa drugs. The characteristic structure of sulfa drugs and their SMARTS are shown in Figure 5A. In the end, we obtained 235 sulfa drugs and 129 non-sulfa drugs from 364 positive molecules. We used six kinds of fingerprints to do the principal component analysis of positive molecules, and the relevant results are shown in Figure 5B. It can be seen that Graph and MACCS molecular fingerprints can better distinguish sulfonamides and non-sulfonamides in this unsupervised learning task. Among them, MACCS has obvious classification boundaries. Other molecular fingerprints have found no clear boundaries for these two kinds of drugs.

4 | DISCUSSION

As a critical molecule in the regulation of cardiac electrophysiology, $\text{Na}_v1.5$ has been a focal point in related research. As previously discussed, researchers have paid great attention to the potential cardiac risk caused by some $\text{Na}_v1.5$ blockers to instruct on rational drug use. In addition, some researchers concentrate on the therapeutic effect on cardiac arrhythmias, the balance between therapeutic and adverse effects being the important issue.⁴⁴ Many $\text{Na}_v1.5$ blockades

have been found to have the antiarrhythmic effect, such as lidocaine and phenytoin.^{45,46} These induce the excitability of cardiomyocytes by blocking $\text{Na}_v1.5$ to relieve arrhythmia. So $\text{Na}_v1.5$ is a key target in arrhythmia.

At present, there are several methods for screening $\text{Na}_v1.5$ ion channel drugs, such as the fluorescence resonance energy transfer method, patch-clamp electrophysiological method and fluorescence imaging plate reader.⁴⁷⁻⁴⁹ The patch-clamp electrophysiological method is still the gold standard for ion channel drug screening. However, due to the limitations of experimental equipment and the lack of professionals, screening ion channel drugs using electrophysiological methods often requires significant time and resources. Hence, we need to use cheminformatics methods to accelerate this process. In this study, the RF-Graph model had the best performance. This model will greatly reduce experimental time and cost. In addition, researchers hope to find specific $\text{Na}_v1.5$ inhibitors, which often can have huge application potential. To achieve this goal, multiple models can be constructed of different sodium ion channels ($\text{Na}_v1.7$, $\text{Na}_v1.6$, etc.). When these models are used to screen compounds at the same time, $\text{Na}_v1.5$ -specific inhibitors can be obtained.

Compared with other studies, it can be seen that the RF-Graph model has achieved higher MCC and AUC values, which may be related to the strict data processing and undersampling process.^{50,51} All the top 10 models in the test set are LR and RF based (Table 2). This shows that LR and RF are suitable for the rapid construction of classification models with small sample sizes. In the prediction problem of $\text{Na}_v1.5$ inhibitors, the RF-Graph model is better than the graph convolutional ANN model (Table 3), which is consistent with the research results of Korolev et al.³⁴ Previous research^{52,53} have shown that GCNNs are not effective at long range information propagation. However, we have not explored this aspect in our current study.

There are still areas for improvement for further work. All the data used originates from the ChEMBL database, with no data set from other sources. After screening the original data, only 364 positive molecules were obtained. The lack of experimental data also greatly limits the applicability of the model. In addition, there is scope for trying different molecular feature extraction methods and ML methods to make the predicted results more reliable. Therefore, combining ML methods to predict $\text{Na}_v1.5$ inhibitors with experimental high-throughput screening methods is planned for a future study.

5 | CONCLUSION

Based on molecular fingerprinting and machine learning methods, 30 classification models were developed and implemented which predict the binding capacity with $\text{Na}_v1.5$ protein. In all cases, the most suitable model obtained for the test set was RF-Graph, of which the Q, SE, SP, AUC and MCC values were 0.9309, 0.9022, 0.9853, 0.9473 and 0.8627. These results are significantly better than the classification model based on graph neural networks. We also extracted 10 kinds of alert substructures which were firmly related to the affinity

of inhibition to $\text{Na}_v1.5$. In the unsupervised learning task of identifying sulfa drugs, MACCS and Graph fingerprints have good results. To conclude, the model established in this research can effectively shorten the development time and cost of $\text{Na}_v1.5$ inhibitors and provide guidance for related experimental work.

AUTHOR CONTRIBUTIONS

Weikaixin Kong: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); writing – original draft (equal). **Weiran Huang:** Conceptualization (equal); data curation (equal); formal analysis (equal); methodology (equal). **Chao Peng:** Conceptualization (equal); data curation (equal); formal analysis (equal). **Zhuo Huang:** Conceptualization (equal); funding acquisition (equal); project administration (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Bowen Zhang:** Methodology (equal); writing – original draft (equal). **Guifang Duan:** Writing – original draft (equal); writing – review and editing (equal). **Weining Ma:** Methodology (equal); writing – review and editing (equal).

FUNDING INFORMATION

This work was supported by Chinese National Programs for Brain Science and Brain-like intelligence technology No.2021ZD0202102 to Z.H; National Natural Science Foundation of China Grant Nos. 31871083 and 81371432 to Z.H. and 32000674 to GFD.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

DATA AVAILABILITY STATEMENT

All data can be found in public databases.

ORCID

Weikaixin Kong  <https://orcid.org/0000-0002-1070-7136>

REFERENCES

- Abriel H, Kass RS. Regulation of the voltage-gated cardiac sodium channel $\text{Nav}1.5$ by interacting proteins. *Trends Cardiovasc Med*. 2005;15:35-40. doi:10.1016/j.tcm.2005.01.001
- Tan HL, Bezzina CR, Smits JP, Verkerk AO, Wilde AA. Genetic control of sodium channel function. *Cardiovasc Res*. 2003;57:961-973. doi:10.1016/s0008-6363
- McNair WP, Ku L, Taylor MRG, et al. SCN5A mutation associated with dilated cardiomyopathy, conduction disorder, and arrhythmia. *Circulation*. 2004;110:2163-2167. doi:10.1161/01.Cir.0000144458.58660.Bb
- Bardai A, Amin AS, Blom MT, et al. Sudden cardiac arrest associated with use of a non-cardiac drug that reduces cardiac excitability: evidence from bench, bedside, and community. *Eur Heart J*. 2013;34:1506-1516. doi:10.1093/eurheartj/ehs054
- Pramanik S, Roy K. Modeling bioconcentration factor (BCF) using mechanistically interpretable descriptors computed from open source tool "PaDEL-descriptor". *Environ Sci Pollut Res Int*. 2014;21:2955-2965. doi:10.1007/s11356-013-2247-z
- Maryam L, Usmani SS, Raghava GPS. Computational resources in the management of antibiotic resistance: speeding up drug discovery. *Drug Discov Today*. 2021;26:2138-2151. doi:10.1016/j.drudis.2021.04.016

7. Usmani SS, Bhalla S, Raghava GPS. Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. *Front Pharmacol*. 2018;9:954. doi:10.3389/fphar.2018.00954
8. Pillai N, Dasgupta A, Sudsakorn S, Fretland J, Mavroudis PD. Machine learning guided early drug discovery of small molecules. *Drug Discov Today*. 2022;27:2209-2215. doi:10.1016/j.drudis.2022.03.017
9. Kong W, Gao M, Jin Y, Huang W, Huang Z, Xie Z. Prognostic model of patients with liver cancer based on tumor stem cell content and immune process. *Aging (Albany NY)*. 2020;12:16555-16578. doi:10.18632/aging.103832
10. Zhu J, Kong W, Xie Z. Expression and prognostic characteristics of ferroptosis-related genes in colon cancer. *Int J Mol Sci*. 2021;22:5652. doi:10.3390/ijms22115652
11. Rank L, Puhl AC, Havener TM, et al. Multiple approaches to repurposing drugs for neuroblastoma. *Bioorg Med Chem*. 2022;73:117043. doi:10.1016/j.bmc.2022.117043
12. Datta A, Matlock MK, le Dang N, et al. 'Black Box' to 'Conversational' machine learning: ondansetron reduces risk of hospital-acquired venous thromboembolism. *IEEE J Biomed Health Inform*. 2021;25:2204-2214. doi:10.1109/jbhi.2020.3033405
13. Vanhaelen Q, Mamoshina P, Aliper AM, et al. Design of efficient computational workflows for in silico drug repurposing. *Drug Discov Today*. 2017;22:210-222. doi:10.1016/j.drudis.2016.09.019
14. Gu Y, Zheng S, Yin Q, Jiang R, Li J. REDDA: integrating multiple biological relations to heterogeneous graph neural network for drug-disease association prediction. *Comput Biol Med*. 2022;150:106127. doi:10.1016/j.combiomed.2022.106127
15. Lukashina N, Kartysheva E, Spiuth O, Virko E, Shpilman A. SimVec: predicting polypharmacy side effects for new drugs. *J Chem*. 2022;14:49. doi:10.1186/s13321-022-00632-5
16. Mujtaba MG, Gerner P, Wang GK. Local anesthetic properties of prenylamine. *Anesthesiology*. 2001;95:1198-1204. doi:10.1097/0000542-200111000-00025
17. la DS, Peterson EA, Bode C, et al. The discovery of benzoxazine sulfonamide inhibitors of Na(V)1.7: tools that bridge efficacy and target engagement. *Bioorg Med Chem Lett*. 2017;27:3477-3485. doi:10.1016/j.bmcl.2017.05.070
18. Mirams GR, Cui Y, Sher A, et al. Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk. *Cardiovasc Res*. 2011;91:53-61. doi:10.1093/cvr/cvr044
19. Macsari I, Sandberg L, Besidski Y, et al. Phenyl isoxazole voltage-gated sodium channel blockers: structure and activity relationship. *Bioorg Med Chem Lett*. 2011;21:3871-3876. doi:10.1016/j.bmcl.2011.05.041
20. Ghose AK, Crippen GM. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J Chem Inf Comput Sci*. 1987;27:21-35. doi:10.1021/ci00053a005
21. Cheng F, Yu Y, Zhou Y, et al. Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *J Chem Inf Model*. 2011;51:2482-2495. doi:10.1021/ci200317s
22. Zhao YH, Abraham MH, Zissimos AM. Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. *J Org Chem*. 2003;68:7368-7373. doi:10.1021/jo034808o
23. Xue L, Godden JW, Bajorath J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J Chem Inf Comput Sci*. 2000;40:1227-1234. doi:10.1021/ci000327j
24. Pedregosa F. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
25. Davis LJ, Offord KP. Logistic regression. *J Pers Assess*. 1997;68:497-507.
26. Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning*. 1995;20:273-297.
27. Plewczynski D, Spieser SAH, Koch U. Assessing different classification methods for virtual screening. *J Chem Inf Model*. 2006;46:1098-1106.
28. Poli G, Galati S, Martinelli A, Supuran CT, Tuccinardi T. Development of a cheminformatics platform for selectivity analyses of carbonic anhydrase inhibitors. *J Enzyme Inhib Med Chem*. 2020;35:365-371.
29. Ye Q, Zhang X, Lin X. Drug-target interaction prediction via graph auto-encoder and multi-subspace deep neural networks. *IEEE/ACM Trans Comput Biol Bioinform*. 2022;PP:1-12. doi:10.1109/tcbb.2022.3206907
30. Lazarczyk M, Duda K, Mickael ME, et al. Adera2.0: a drug repurposing workflow for neuroimmunological investigations using neural networks. *Molecules*. 2022;27:6453. doi:10.3390/molecules27196453
31. Wang L, Yu Z, Wang S, Guo Z, Sun Q, Lai L. Discovery of novel SARS-CoV-2 3CL protease covalent inhibitors using deep learning-based screen. *Eur J Med Chem*. 2022;244:114803. doi:10.1016/j.ejmech.2022.114803
32. Breiman L. Random Forests. *Machine Learning*. 2001;45:5-32.
33. Song Y-Y, Ying LU. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27:130.
34. Korolev V, Mitrofanov A, Korotcov A, Tkachenko V. Graph convolutional neural networks as "general-purpose" property predictors: the universality and limits of applicability. *J Chem Inf Model*. 2019;60:22-28.
35. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23:1241-1250.
36. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLoS One*. 2017;12:e0177678.
37. Polikar R. *In Ensemble Machine Learning 1-34*. Springer; 2012.
38. Cortes-Ciriano I. Bioalerts: a python library for the derivation of structural alerts from bioactivity and toxicity data sets. *J Chem*. 2016;8:13. doi:10.1186/s13321-016-0125-7
39. Li X, Zhang Y, Li H, Zhao Y. Modeling of the hERG K⁺ channel blockage using online chemical database and modeling environment (OCHEM). *Molecular Informatics*. 2017;36:1700074.
40. Zhang C, Cheng F, Li W, Liu G, Lee PW, Tang Y. In silico prediction of drug induced liver toxicity using substructure pattern recognition method. *Mol Inform*. 2016;35:136-144. doi:10.1002/minf.201500055
41. Kong W, Tu X, Huang W, Yang Y, Xie Z, Huang Z. Prediction and optimization of Na(V)1.7 Sodium channel inhibitors based on machine learning and simulated annealing. *J Chem Inf Model*. 2020;60:2739-2753. doi:10.1021/acs.jcim.9b01180
42. Narahashi T. Tetrodotoxin: a brief history. *Proc Jpn Acad Ser B Phys Biol Sci*. 2008;84:147-154. doi:10.2183/pjab.84.147
43. Baden DG, Bourdelais AJ, Jacocks H, Michelliza S, Naar J. Natural and derivative brevetoxins: historical background, multiplicity, and effects. *Environ Health Perspect*. 2005;113:621-625. doi:10.1289/ehp.7499
44. Clarkson CW, Hondeghem LM. Mechanism for bupivacaine depression of cardiac conduction: fast block of sodium channels during the action potential with slow recovery from block during diastole. *Anesthesiology*. 1985;62:396-405.
45. Clarkson CW, Follmer CH, ten Eick RE, Hondeghem LM, Yeh JZ. Evidence for two components of sodium channel block by lidocaine in isolated cardiac myocytes. *Circ Res*. 1988;63:869-878. doi:10.1161/01.res.63.5.869
46. Xu YQ, Pickoff AS, Clarkson CW. Evidence for developmental changes in sodium channel inactivation gating and sodium channel block by phenytoin in rat cardiac myocytes. *Circ Res*. 1991;69:644-656. doi:10.1161/01.res.69.3.644

47. Davis GC, Kong Y, Paige M, et al. Asymmetric synthesis and evaluation of a hydroxyphenylamide voltage-gated sodium channel blocker in human prostate cancer xenografts. *Bioorg Med Chem*. 2012;20:2180-2188. doi:10.1016/j.bmc.2011.08.061
48. Macsari I, Besidski Y, Csornyik G, et al. 3-Oxoisoindoline-1-carboxamides: potent, state-dependent blockers of voltage-gated sodium channel Na(V)1.7 with efficacy in rat pain models. *J Med Chem*. 2012;55:6866-6880. doi:10.1021/jm300623u
49. Yang SW, Ho GD, Tulshian D, et al. Bioavailable pyrrolo-benzo-1,4-diazines as Na(v)1.7 sodium channel blockers for the treatment of pain. *Bioorg Med Chem Lett*. 2014;24:4958-4962. doi:10.1016/j.bmcl.2014.09.038
50. Li X, Chen Y, Song X, Zhang Y, Li H, Zhao Y. The development and application of in silico models for drug induced liver injury. *RSC Adv*. 2018;8:8101-8111. doi:10.1039/c7ra12957b
51. Kong W, Wang W, An J. Prediction of 5-hydroxytryptamine transporter inhibitors based on machine learning. *Comput Biol Chem*. 2020;87:107303. doi:10.1016/j.compbiolchem.2020.107303
52. Matlock MK, Datta A, Dang NL, Jiang K, Swamidass SJ. Deep learning long-range information in undirected graphs with wave networks. *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019;1-8. doi:10.1109/IJCNN.2019.8852455
53. Wu Z. Representing long-range context for graph neural networks with global attention. *Adv Neural Inf Process Syst*. 2021;34:13266-13279.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kong W, Huang W, Peng C, et al. Multiple machine learning methods aided virtual screening of Na_v1.5 inhibitors. *J Cell Mol Med*. 2023;27:266-276. doi:10.1111/jcmm.17652