Computational Biology and Chemistry 87 (2020) 107303

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

Prediction of 5-hydroxytryptamine transporter inhibitors based on machine learning

Weikaixin Kong[a], Wenyu Wang[b], Jinbing An[c,*]

[a] Department of Molecular and Cellular Pharmacology, School of Pharmaceutical Sciences, Peking University, Beijing, 100191, China
[b] School of Nursing, Peking University, Beijing, 100191, China
[c] Department of Health Informatics and Management, Peking University, Beijing, 100191, China

ARTICLE INFO

Keywords:
Depression
5-Hydroxytryptamine transporter
Machine learning
Ensemble model
Inhibitor

ABSTRACT

In patients with depression, the use of 5-HT reuptake inhibitors can improve the condition. Machine learning methods can be used in ligand-based activity prediction processes. In order to predict SERT inhibitors, the SERT inhibitor data from the ChEMBL database was screened and pre-processed. Then 4 machine learning methods (LR, SVM, RF, and KNN) and 4 molecular fingerprints (CDK, Graph, MACCS, and PubChem) were used to build 16 prediction models. The top 5 models of accuracy (Q) in the cross-validation of training set were used to build three different ensemble learning models. In the test1 set, the VOT_CLF3 model had the largest SP (0.871), Q (0.869), AUC (0.919), and MCC (0.728). In the unbalanced test2 set, VOT_CLF3 had the largest SE (0.857), SP (0.867), Q (0.865) and MCC (0.639). VOT_CLF3 was recommended for the virtual screening process of SERT inhibitors. In addition, 12 molecular structural alerts that frequently appear in SERT inhibitors were found ($P < 0.05$), which provided important reference value for the design work of SERT inhibitors.

1. Introduction

Depression is a type of mood disorder characterized by a marked and persistent state of mind. It is one of the most common mental disorders. At present, the etiology and pathogenesis of depression are still unclear. It is generally believed that the onset of depression is closely related to three factors: genetic factors, biochemical factors such as norepinephrine (NE), serotonin (5-hydroxytryptamine, 5-HT) and dopamine (DA), as well as social and environmental factors (Bowen et al., 2020). Depression not only leads to a series of physical, psychological, social dysfunctions, related complications, and potentially high suicide risk, but also increases the burden on patients, families, and society, and seriously reduces the quality of life of patients. According to a World Health Organization (WHO) survey, there are approximately 322 million people with depression worldwide, accounting for 4.4% of the world's population (Smith, 2014). Studies have also shown that depression has become the world's largest disabling disease (Friedrich, 2017) and is expected to rise to the top of the world's disease burden by 2030 (Angold, 1988).

Serotonin is a highly conserved chemical signal that is widely distributed in vertebrates and invertebrates. It exists in the brain and digestive tract and plays a key role in various regulatory processes. It can stimulate the target organ and participate in a variety of processes, including mood, motivation, thinking, diet and nociception (Serretti et al., 2007). After the physiological action is applied, 5-HT is inactivated to avoid sustained excitation of the target organ and desensitization of the 5-HT receptor, which is mainly accomplished by the 5-HT transporter (5-HTT / 5-hydroxytryptamine transporter, SERT). 5-HT can make the brain feel happy, 5-HT transporteris a transmembrane transporter with high affinity for 5-HT, and 5-HT is widely present in the intestine chromaffin cell membrane, mast cells and serotonergic neurons presynaptic membrane (Chen et al., 1998). SERT is a component of the synapse at the nerve endings, located on the presynaptic membrane at the nerve endings. It re-enters the 5-HT of presynaptic neurons from the synaptic cleft, directly reducing the concentration of serotonin in the synaptic cleft (Lesch et al., 1996), preventing the occurrence of adverse reactions. SERT removes 5-HT from the synaptic cleft and affects the number and duration of postsynaptic receptor-mediated signaling, which plays a key role in the fine-tuning of overall pulse delivery (Fan et al., 2011).

If the SERT structure changes during physiological processes, the synthesis, clearance, and function of the 5-HT and 5-HT receptors will be greatly affected. These changes have important clinical implications for SERT inhibitors. The mechanism of action of selective serotonin reuptake inhibitors (SSRIs) in antidepressants is selective inhibition of the reuptake effect of 5-HT in central nervous system presynaptic

[*] Corresponding author.
E-mail address: termison@sina.cn (J. An).

https://doi.org/10.1016/j.compbiolchem.2020.107303
Received 4 March 2020; Received in revised form 29 May 2020; Accepted 4 June 2020
Available online 06 June 2020
1476-9271/

membranes, increasing the 5-HT concentration in the synaptic cleft, exciting the brain, thereby achieving an antidepressant effect. It is characterized by strong specificity and selectivity for 5-HT, limited effect on other neurotransmitters, good oral absorption and less adverse reactions (Tebartz van Elst et al., 2006). In general, studying molecular biology and pharmacological properties of SERT and exploring SSRIs with selective high affinity for SERT have brought bright prospects for the treatment of these diseases.

The method of computer-aided drug design (CADD) has accelerated the discovery of SERT inhibitors. Manepalli et al. (2011) used a pharmacophore model to find two SERT inhibitor hits (SM-10 and SM-11) with specific modes of action. Pharmacophore models can find inhibitor molecules with the same mechanism of action, but they cannot take into account inhibitors with different mechanisms of action at the same time. In addition, compared to machine learning methods, the number of inhibitors included in the pharmacophore model is small. Gabrielsen et al. (2011) used molecular docking method to find SERT inhibitors. Molecular docking method can make reasonable predictions on the effectiveness of inhibitors, but it is still only focused on a specific protein site, and predictions cannot be made for inhibitors with different mechanisms of action. In addition, the molecular docking method takes a long time in the virtual screening process of a large number of molecules. In comparison, machine learning methods based on ligand molecular characteristics (fingerprints or descriptors) have the advantages of considering multiple inhibition mechanisms at the same time, and are time-saving.

In this research, the machine learning methods were used to predict the binding ability of the ligand molecules and the SERT protein, based on the molecular fingerprints. Then three different ensemble learning models were also established and evaluated in the test1 and test2 sets. The most effective model was recommended for virtual screening of SERT inhibitors. In the process, the cost of the experiment can be reduced and the drug development cycle is shortened. This work has certain practical application value.

2. Materials and Methods

2.1. Data acquisition and processing

We obtained the training set and test1 set data from the ChEMBL database (https://www.ebi.ac.uk/chembl/). The experimental model of the selected data was Hek293 cells. The experimental method was the isotope [3H] labeling method, and all the molecular SMILES provided by the ChEMBL database were read and written by the rdkit (2019.09.3) package in Python (3.7.4) to get standard SMILES and the duplicate data was removed. Considering that the half-inhibitory concentration IC_{50} of the drugs that have significant inhibition of SERT protein published in the DrugBank (https://www.drugbank.ca/) database are all less than 500 nmol·L^{-1}, and in order to make the established model have better generalization performance, the molecules with IC_{50} smaller than 500 nmol·L^{-1} were regarded as inhibitors, and molecules with IC_{50} greater than 500 nmol·L^{-1} as non-inhibitors, which is similar to the method previously reported (Sun, 2006; Kong et al., 2020). In order to further increase the molecular diversity, the data with Ki on ChEMBL was also taken into account. After comparing the Ki and IC_{50} values in the same molecules, those molecules with Ki less than 2 000 nM were also considered inhibitors, and those with Ki greater than 2 000 nM were considered non-inhibitors. The data was randomly divided into training set and test1 set. The method proposed by Roy et al. (2015) was used to determine the applicability domain of the models, and the molecules outside the applicability domain were eliminated.

The test2 set was an external data set. The test2 set was constructed using compounds which were reported by DrugBank and not used in the training and test1 sets. Seven SERT inhibitors were obtained, and 30 compounds without SERT protein binding ability were selected as non-inhibitors of SERT protein. The test2 set was an unbalanced data set that can be used to test the robustness and generalization capabilities of the models. The molecular SMILES of the training set, the test1 set, and the test2 set are presented in supplementary materials (Supplementary Tables 1–3).

2.2. The chemical space

We calculated the ECFP4 fingerprints of the molecules, and then randomly selected 100 molecules in all data sets to find the Tanimoto correlation coefficient between each two molecules (Li et al., 2017a; Butina, 1999). We finally calculated the average Tanimoto correlation coefficient of 100 molecules. The Tanimoto correlation coefficient can effectively measure the similarity between molecules, which in turn reflects the size of the chemical space occupied by the data. In the training set, we calculated the relative molecular mass (MW), the lipid-water partition coefficient (ALogP), the hydrogen bond acceptor number (nHBAcc), the hydrogen bond donor number (nHBDon), and the number of rotatable keys (nRotB), according to the Lipinski's Rules of Five (Lipinski, 2016, 2003; Manto Chagas et al., 2018). In addition, we also calculated the complexity of a molecule (FMF), sum of the atomic polarizabilities (apol), topological polar surface area based on fragment contributions (TopoPSA), kappa shape indices(Kier), topological charge (JGT), and van der Waals volume (VABC). MW, ALogP, FMF, apol, TopoPSA, Kier, JGT and VABC were used to plot scatter plots in data with different labels. The five descriptors of Lipinski rules were plotted into a radar chart to observe the chemical space distribution of the compound molecules (Li et al., 2017b). The calculation of the these descriptors was done by the "rcdk" package (3.5.0) in the R language (3.5.3) (Guha and CharlopPowers, 2014).

2.3. Extraction of molecular features and machine learning methods

We used 4 kinds of molecular fingerprints to extract the characteristics of the molecules, namely CDK (1024 bits), MACCS (166 bits), PubChem (881 bits), and Graph (1024 bits). These fingerprints are calculated by PaDEL-Descriptor software (2.14) (Yap, 2011).

In this study, the classification models were built using the support vector machine (SVM), logistic regression (LR), random forest (RF), and k nearest neighbor (KNN). An in-depth description of these four methods can be obtained from some excellent works and research papers. Here, only the main ideas of the four methods are briefly described.

LR uses the idea of maximum likelihood estimation to minimize the loss function of the regression to obtain the unknown parameters (Dreiseitl and Ohno-Machado, 2002). Using the sigmoid function on the final result of the binary classification problem, the predicted value range is mapped to [0,1], and finally, the classification is realized.

SVM is a machine learning method based on the principle of structural risk minimization (Bouboulis et al., 2014). It maps the input variables to higher-dimensional spaces by changing the kernel function, and then uses the so-called "support vector" to find the maximum interval in the new space, and clarifies the optimal classification hyperplane. Finally, the SVM completes the classification of the input data. The SVM model of this study used a linear kernel function.

RF is a collection of multiple decision trees. The process of RF establishment is to construct multiple decision trees by randomly extracting different features and different samples, that is, multiple weak classifiers. The results of multiple weak classifiers are voted to get the final result. For out-of-bag data that has not been extracted, it can be used to test the generalization performance of the model (Svetnik et al., 2003).

KNN is based on the idea of maximum likelihood estimation and finds the distance between the new input sample point and the training sample points in the multidimensional space formed by the features. First, the specific N training sample points closest to the new sample point are selected, and then the distance between the samples and labels of each category in the N training sample points are comprehensively considered to determine the final classification result (Ruan et al., 2017).

The above models were built through the sklearn package (0.22.1), and the default parameters were used. Based on 4 feature extraction methods and 4 classification methods, 16 (4 * 4) classification models were constructed and used to predict test1 and test2 sets. In addition, three different ensemble learning models (Fang et al., 2014) (VOT_CLF1, VOT_CLF2, VOT_CLF3) have also been established (Fig. 5). In VOT_CLF1 model, the simple voting result of the labels predicted by multiple classifiers is the final classification label. In VOT_CLF2 model, considering the class probability of each model, the result of the average probability is the final classification result. In VOT_CLF3 model, the classification probabilities of single models were used as the input to build an ANN model, and the prediction result of the ANN model was regarded as the final classification result. ANN related parameters: the activation function was "relu"; the number of neurons was (200, 20, 1). A study (Bauer and Kohavi, 1999) shows that when the classification principles of multiple sub-classifiers are different and the performance is good enough, the ensemble learning model often has the performance improvement.

2.4. Prediction results evaluation

In order to evaluate different models, the following statistical indicators were used: sensitivity (SE), specificity (SP), accuracy (Q), and Matthew's correlation coefficient (MCC) (Cai et al., 2017). In addition, the receiver operating characteristic (ROC) curves were also drawn, and the corresponding area under the curve (AUC) was used to evaluate the performance of the models.

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

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

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

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

In the above formula, TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively. In this study, we first performed five-fold cross-validation on the training set. The top 5 models of accuracy (Q) in cross-validation were used to build ensemble learning models and make predictions on test1 set. The model with the largest MCC value in test1 set was considered to be the best model. The ensemble learning models were used to make predictions on test2 set.

2.5. Structural alert

We used Python's package "bioalerts" (Cortes, 2016) to find substructures related to molecular neutralization activity. This method searches and counts substructures in a molecule by setting a certain search radius. The molecular fingerprint used in the search for substructures was ECFP4. The number of a molecular substructure can be seen as obeying the hypergeometric distribution (Ahlberg et al., 2014). When the number of a molecular substructure in inhibitors is significantly higher than the number in non-inhibitors, it can be considered that this substructure plays an important role in the binding process with SERT protein. We used the training set to find structural alerts.

Table 1
The statistics of molecules in the data sets.

Data sets	Training set	Test1 set	Test2 set	Total
Non-inhibitors	926	232	30	1 188
Inhibitors	1 478	369	7	1 854
Total	2 404	601	37	3 042

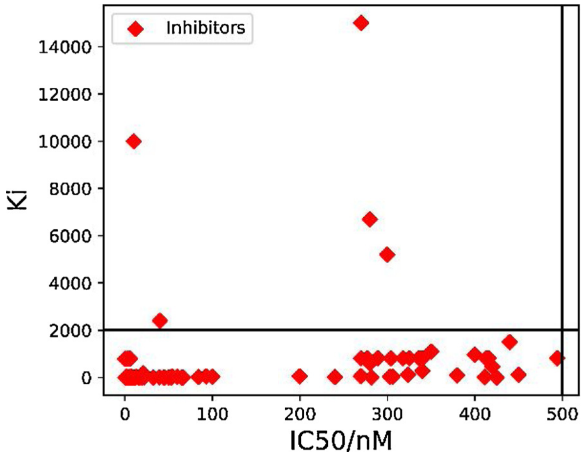


Fig. 1. Correspondence between IC_{50} and Ki in the same molecules.

3. Results and Discussion

3.1. The distribution of the chemical space

The number of molecules in the data sets were shown in Table 1. As can be seen from correspondence between IC_{50} and Ki in the same molecules (Fig. 1), inhibitors with IC_{50} less than 500 nM usually had Ki less than 2 000 nM. This shows that it was reasonable to consider the molecules with IC_{50} less than 500 nM or Ki less than 2 000 nM as inhibitors. The heat map of the Tanimoto correlation coefficient of 100 randomly selected molecules is shown in Fig. 2. The average Tanimoto correlation coefficient was 0.192, which indicated that the differences between molecules were large, and the models trained based on such data can have strong generalization ability. In the radar chart, the inside was a small value and the outside was a large value. It can be seen from Fig. 3 that the molecular distribution of the training set was broad and did not exhibit preference, whether it was a discrete variable or a continuous variable. From the scatter plots (Fig. 4), we can see that there was no difference in the distribution of inhibitors and non-

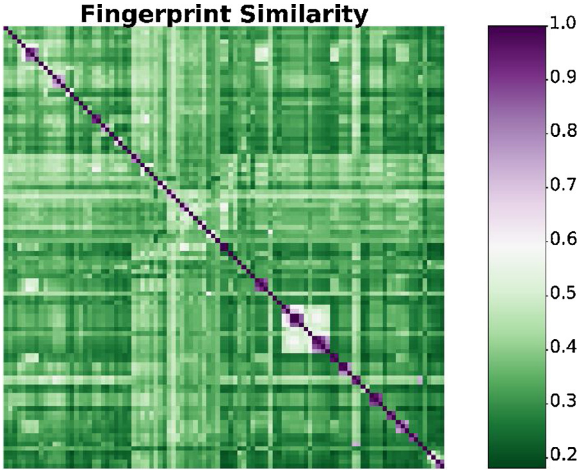


Fig. 2. Heat map of Tanimoto correlation coefficient.

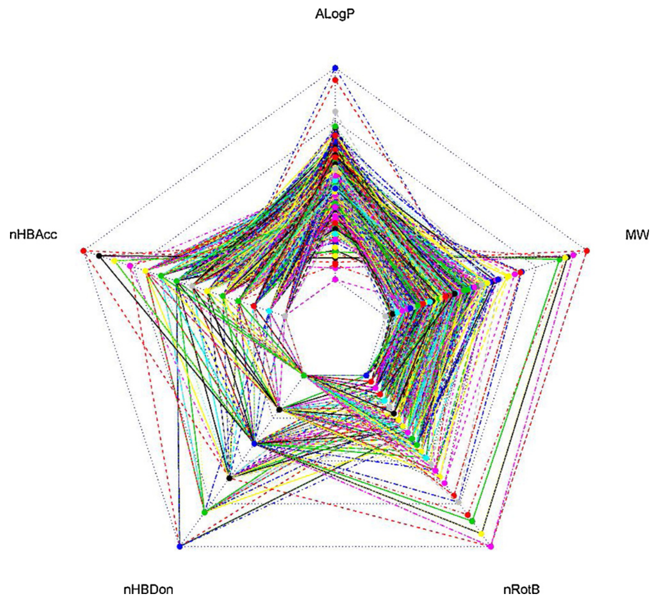


Fig. 3. Radar map of molecular properties of training data sets.

inhibitors between these descriptors. This suggests that it was difficult to successfully distinguish between the two types of molecules by simply relying on these simple chemical properties, and it was necessary to establish machine learning models to predict inhibitors of SERT.

3.2. Cross-validation result

In the different machine learning models, the training set was spilt with the ratio 1:4 five times to complete the cross-validation. The results of the cross-validation are shown in Table 2. The top five models were MACCS_SVM, PubChem_SVM, PubChem_RF, CDK_LR, MACCS_RF. These five models were used for the subsequent construction of ensemble learning models.

3.3. Test1 set results

After cross-validation, we trained the models with all the data of the training set. The five selected models were used to build the ensemble learning models (VOT_CLF1, VOT_CLF2, VOT_CLF3). The evaluation results of 5 single models and 3 ensemble learning models on test1 set are shown in Table 3. It can be seen that the ensemble learning models generally had a larger MCC value (VOT_CLF1 0.674; VOT_CLF2 0.696; VOT_CLF3 0.728). In the ROC curves of single models (Fig. 6A), CDK_LR (0.904), MACCS_RF (0.911) and PubChem_RF (0.896) have large AUC, while MACCS_SVM (0.877) and PubChem_SVM (0.826) have small AUC. This shows that the SVM model was less effective than other classification methods in predicting SERT inhibitors. Among the five single models, CDK-LR had the largest MCC value (0.677). Among all models, the VOT_CLF3 model had the largest SP (0.871), Q (0.869), AUC (0.919), and MCC (0.728). This showed that VOT_CLF3 could be regarded as the best model in the prediction of SERT inhibitors, and could be recommended for the virtual screening process of drugs. At the same time, it also indicated that the ensemble learning method applied by VOT_CLF3 had better performance than other ensemble learning methods. This provided a reference for the work of other researchers.

To ensure that the evaluation of the models in the test1 set was

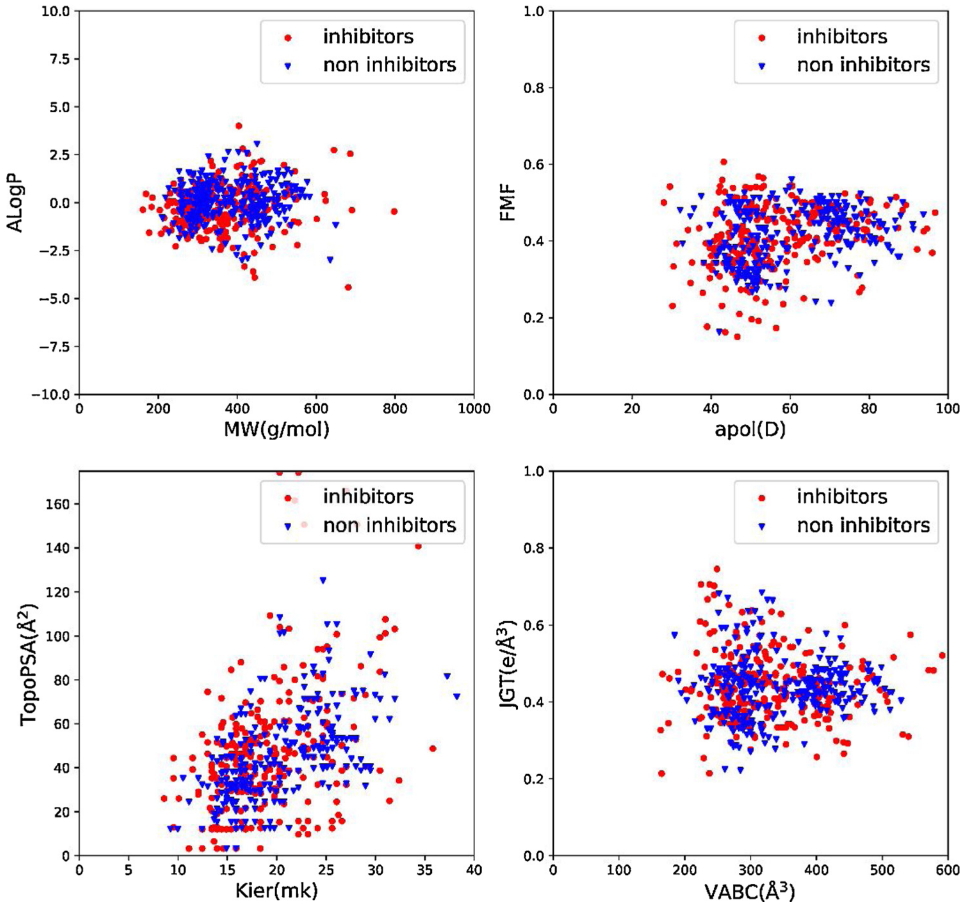


Fig. 4. Scatter plots of molecular properties in different categories.

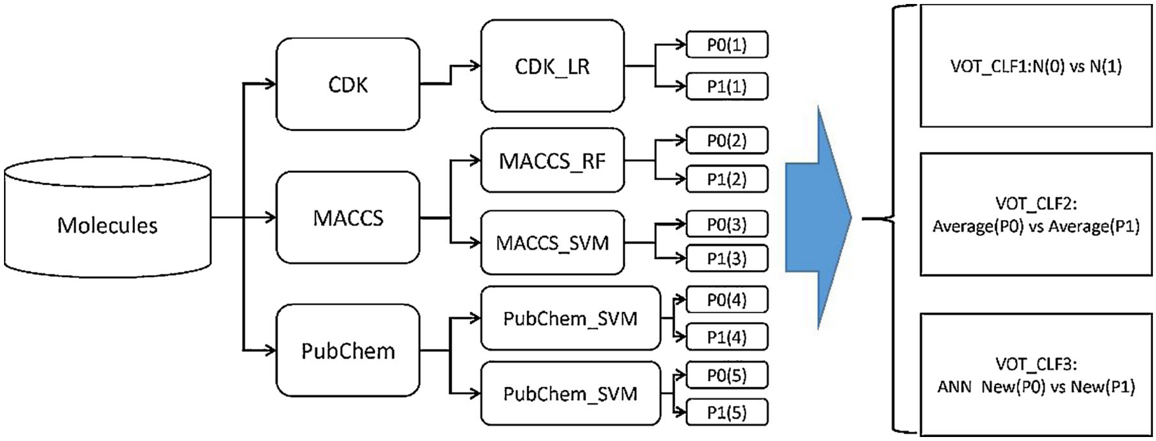


Fig. 5. Construction method of VOT_CLF models.

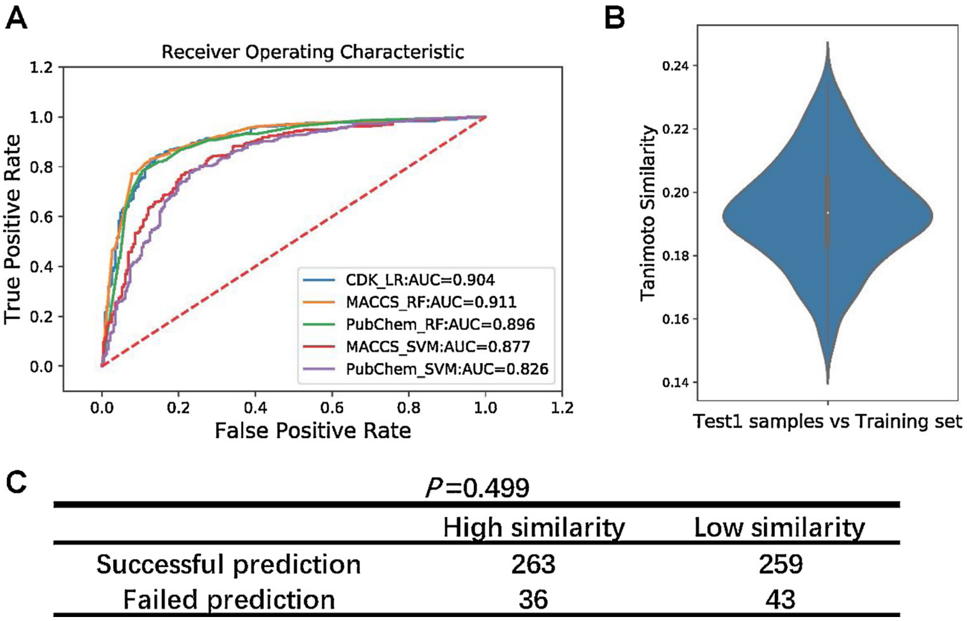
Table 2
Five-fold cross-validation of different machine learning methods in the training set.

Model	MACCS_SVM	PubChem_SVM	PubChem_RF	CDK_LR	MACCS_RF	CDK_RF	CDK_SVM	MACCS_KNN
Q ± s	0.867 ± 0.022	0.845 ± 0.019	0.845 ± 0.017	0.843 ± 0.016	0.838 ± 0.009	0.837 ± 0.021	0.834 ± 0.020	0.832 ± 0.015
Model	Graph_SVM	PubChem_KNN	MACCS_LR	Graph_RF	PubChem_LR	Graph_KNN	Graph_LR	CDK_KNN
Q ± s	0.832 ± 0.018	0.831 ± 0.017	0.830 ± 0.016	0.826 ± 0.021	0.825 ± 0.015	0.822 ± 0.027	0.821 ± 0.019	0.812 ± 0.015

Table 3
The test1 set prediction result.

Model	SE	SP	Q	AUC	MCC
CDK_LR	0.875	0.802	0.847	0.904	0.677
MACCS_RF	0.864	0.815	0.845	0.911	0.676
PubChem_RF	0.829	0.832	0.830	0.896	0.651
MACCS_SVM	0.859	0.659	0.782	0.877	0.533
PubChem_SVM	0.924	0.513	0.765	0.826	0.495
VOT_CLF1	0.900	0.763	0.847	0.903	0.674
VOT_CLF2	0.897	0.793	0.857	0.913	0.696
VOT_CLF3	0.867	0.871	0.869	0.919	0.728

credible, we analyzed the similarity of the molecules in the training set and test1 set. Tanimoto similarity (based on ECFP4) were calculated of any two molecules from test1 set and training set respectively. The distribution diagram of the average similarity of a certain molecule in the test1 set and all molecules in the training set is shown in Fig. 6B. The median of similarity was 0.1935. In order to explore whether the similarity was related to the prediction results, chi-square test was used in this process. According to the median value of similarity, the molecules in test1 set were divided into two groups: "High similarity" and "Low similarity". Chi-square test (Fig. 6C) showed that there was no significant correlation between the prediction results in test1 set and similarity with training set (P = 0.499). This showed that fingerprint-based classification models were not based on molecular similarity, but



P=0.499

	High similarity	Low similarity
Successful prediction	263	259
Failed prediction	36	43

Fig. 6. A. Receiver operating characteristic curves for five models. B. Violin diagram of the average Tanimoto similarity between the samples in the test1 set and the training set. C. Chi-square test for Tanimoto similarity (grouped by median value of similarity, 0.1935) and prediction results.

Table 4
Unbalanced test2 set prediction result by VOT_CLF models.

Model	SE	SP	Q	MCC
VOT_CLF1	0.429	0.800	0.730	0.209
VOT_CLF2	0.714	0.800	0.784	0.441
VOT_CLF3	0.857	0.867	0.865	0.639

focused on important molecular features.

3.4. Test2 set results

The ratio of SERT protein inhibitor to non-inhibitor in the test set 2 was 7:30, which was an unbalanced set, and the negative samples accounted for the majority. The relevant prediction result by VOT_CLF models are shown in Table 4. Among the three ensemble learning models, VOT_CLF3 had the largest SE (0.857), SP (0.867), Q (0.865) and MCC (0.639). The test2 set and the training set were from different databases, so VOT_CLF3 model had good generalization ability and application potential.

At present, how to find the possible active drug from the unbalance data is a major problem in machine learning to predict the combination of protein and small molecules (Farquad and Bose, 2012), the researchers tried to solve this problem, and then adopted oneshot-learning method (Altae-Tran et al., 2016). It can be seen from test1 set and test2 set that the VOT_CLF3 model had a good performance. This indicates that the VOT_CLF3model is conducive to processing unbalanced data sets characterized by molecular fingerprints

3.5. The wrongly predicted drugs

In test2 set, the VOT_CLF3 model of this study successfully predicted six molecules (Fig. 7) in 7 SERTIs and erroneously predicted one molecular (Fig. 8). The reasons for cocaine's wrong prediction may be as follows: The lack of compounds containing nitrogen bridged rings in the training samples, that is, the limited chemical space of the training samples, makes the extracted molecular fingerprints limited, and it is difficult to characterize the molecular structure of the aza bridges. The lack of fingerprints used to accurately characterize the properties of the aza bridges limits the model's predictive ability and coverage of predicted compounds. Therefore, to improve model performance, it is necessary to explore new descriptors or fingerprints, and enhance data collection to increase the number of training samples. In addition, we will try other combinations of nonlinear features that are more in line with molecular structural features to further improve the efficiency of machine learning.

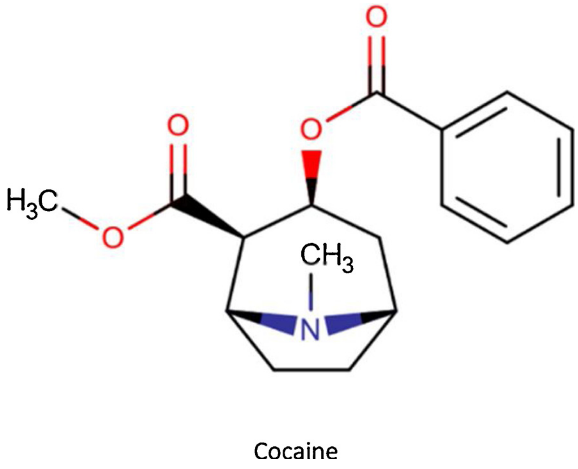


Fig. 8. The wrongly predicted molecule in the VOT_CLF3 model.

3.6. Molecular structural alerts for SERT protein

It can be seen from these structures (Table 5) that when the molecular structure tends to contain piperidine (S2-S5) or form a conjugated structure (S6, S9, S11), it has a higher binding ability to the SERT protein. At this time, the molecule is more basic and more easily forms a hydrogen bond with the amino acid residue on the SERT to function. Many inhibitors of enzymes also contain imidazole structures (Edwards, 1993), which is very similar to S12. In summary, these 12 important substructures can provide a reference for the structural design of SERT inhibitors.

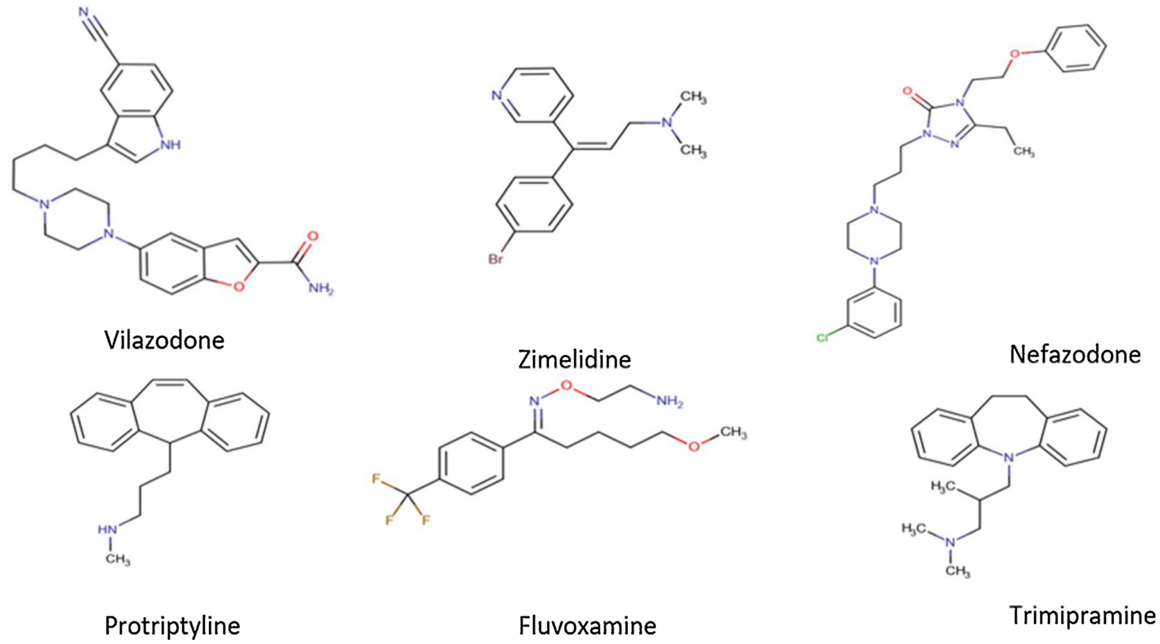
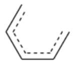
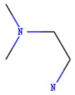
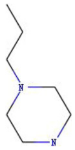
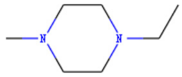
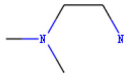
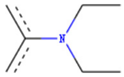
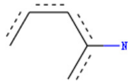
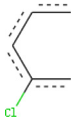
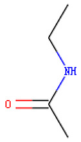
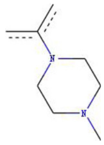
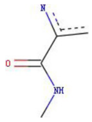
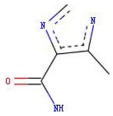


Fig. 7. The correctly predicted molecules in the VOT_CLF3 model.

Table 5
Molecular Structural Alerts for SERT Protein Binding Ability.

Substructure	Substructure ID	*P*	Compounds with substr.	Comp. with substr. active	Comp. with substr. inactive
	S1	< 0.001	278	255	23
	S2	0.0053	203	179	24
	S3	0.0041	135	122	13
	S4	< 0.001	148	135	13
	S5	< 0.001	192	176	16
	S6	< 0.001	171	162	9
	S7	< 0.001	183	172	11
	S8	< 0.001	132	126	6
	S9	< 0.001	143	136	7
	S10	< 0.001	171	162	9
	S11	< 0.001	108	107	1
	S12	< 0.001	105	105	0

4. Conclusion

In this study, we established multiple predictive models of SERT inhibitors, and based on these models, three different ensemble learning models were also established. Among them, VOT_CLF3 had the highest MCC on the test1 (0.728) and test2 (0.639) data sets. VOT_CLF3 is recommended for the virtual screening of SERT inhibitors. In addition, we obtained molecular structural alerts for SERT inhibitors to provide references for other researchers. But so far, the SERTI training data is still very limited, the sample space is small, and the molecular fingerprints are not comprehensive. However, we believe that in the near future, with the accumulation of data and the improvement of molecular characterization methods, this shortcoming will gradually be addressed.

Author Statement

All authors contributed to the work presented in this paper. Weikaixin Kong and Jinbing An designed the calculation process, wrote the manuscript, and prepared the figures. Weikaixin Kong and Wenyu Wang completed the selection of topics, the collection of background materials and data, and wrote relevant content. Weikaixin Kong completed the calculations and wrote relevant content. Jinbing An gave advice on the determination of analytical methods and the writing of the article. The manuscript was reviewed and approved by all authors.

Funding

This work was sponsored by the National Natural Science Foundation of China (No. 11771048 and No. 81871854).

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgement

This study is the project content of the "1st Big Data Workshop" (March, 2018) organized by Department of Health Informatics and Management, Peking University.

Appendix A. Supplementary Data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.compbiolchem.2020.107303.

References

Ahlberg, E., Carlsson, L., Boyer, S., 2014. Computational derivation of structural alerts from large toxicology data sets. J. Chem. Inf. Model. 54.
Altae-Tran, H., Ramsundar, B., Pappu, A., Pande, V., 2016. Low data drug discovery with one-shot learning. ACS Cent. Sci. 3.
Angold, A.S., 1988. Childhood and adolescent depression. II: research in clinical populations. Br. J. Psychiatry 153, 476–492.
Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36, 105–139.
Bouboulis, P., Theodoridis, S., Mavroforakis, C., Dalla, L., 2014. Complex support vector machines for regression and quaternary classification. IEEE Trans. Neural Netw. Learn. Syst. 26.
Bowen, D., Powers, D., Russo, J., Arao, R., LePoire, E., Sutherland, E., Ratzliff, A., 2020. Implementing collaborative care to reduce depression for rural native American/Alaska native people. BMC Health Serv. Res. 20.
Butina, D., 1999. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large data sets. J. Chem. Inf. Comput. Sci. 39, 747–750.
Cai, C., Wu, Q., Luo, Y., Ma, H., Shen, J., Zhang, Y., Yang, L., Chen, Y., Wen, Z., Wang, Q., 2017. In silico prediction of ROCK II inhibitors by different classification approaches. Mol. Divers. 21.
Chen, J., Pan, H., Rothman, T., Wade, P., Gershon, M., 1998. Guinea pig 5-HT transporter: cloning, expression, distribution, and function in intestinal sensory reception. Am. J. Physiol. 275, G433–48.
Cortes, I., 2016. Bioalerts: A python library for the derivation of structural alerts from bioactivity and toxicity data sets. J. Cheminform. 8.
Dreiseitl, S., Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Inform. 35, 352–359.
Edwards, D., 1993. Nitroimidazole drugs - Action and resistance mechanisms I. Mechanisms of action. J. Antimicrob. Chemother. 31, 9–20.
Fan, D.F., Geng-Si, X.I., Liang, K.D., Zhang, W.H., 2011. Serotonin transporter progress. Chinese Bulletin Life Sci. 23, 385–389.
Fang, J., Yang, R., Yang, S., Pang, X., Li, C., He, Y., Liu, A., Du, G.-H., 2014. Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery. Mol. Divers.
Farquad, H., Bose, I., 2012. Preprocessing unbalanced data using support vector machine. Decis. Support Syst. 53, 226–233.
Friedrich, M.J., 2017. Depression is the leading cause of disability around the world. JAMA 317, 1517.
Gabrielsen, M., Kurczab, R., Ravna, A., Kufareva, I., Abagyan, R., Chilmonczyk, Z., Bojarski, A., Sylte, I., 2011. Molecular mechanism of serotonin transporter inhibition elucidated by a new flexible docking protocol. Eur. J. Med. Chem. 47, 24–37.
Guha, R., CharlopPowers, Z., 2014. rcdk: Interface to the CDK Libraries.
Kong, Weikaixin, Tu, Xinyu, Huang, Weiran, Yang, Yang, Xie, Zhengwei, Huang, Zhuo, 2020. Prediction and Optimization of NaV1.7 Sodium Channel Inhibitors Based on Machine Learning and Simulated Annealing. Journal of Chemical Information and Modeling. https://doi.org/10.1021/acs.jcim.9b01180.
Lesch, K.P., Bengel, D., Heils, A., Sabol, S., Greenberg, B., Petri, S., Benjamin, J., Müller, C., Hamer, D., Murphy, D., 1996. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Science 274, 1527–1531.
Li, X., Zhang, Y., Chen, H., Li, H., Zhao, Y., 2017a. Insights into molecular basis of acute contact toxicity of diverse organic chemical in Honey Bee. J. Chem. Inf. Model. 57.
Li, F., Fan, D., Wang, H., Yang, H., Li, W., Tang, Y., Liu, G., 2017b. In silico prediction of pesticide aquatic toxicity with chemical category approaches. Toxicol. Res. 6.
Lipinski, C., 2003. Chris Lipinski discusses life and chemistry after the Rule of five. Drug Discov. Today 8, 12–16.
Lipinski, C., 2016. Rule of five in 2015 and beyond: target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. Adv. Drug Deliv. Rev. 101.
Manepalli, S., Geffert, L., Surratt, C., Madura, J., 2011. Discovery of novel selective serotonin reuptake inhibitors through development of a protein-based pharmacophore. J. Chem. Inf. Model. 51, 2417–2426.
Manto Chagas, C., Moss, S., Alisaraie, L., 2018. Drug metabolites and their effects on the development of adverse reactions: revisiting Lipinski's rule of five. Int. J. Pharm. 549.
Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 145.
Ruan, Y., Xue, X., Liu, H., Tan, J., Li, X., 2017. Quantum algorithm for K-Nearest neighbors classification based on the metric of hamming distance. Int. J. Theor. Phys. 56.
Serretti, A., Calati, R., Mandelli, L., Ronchi, D., 2007. Serotonin transporter gene variants and behavior: a comprehensive review. Curr. Drug Targets 7, 1659–1669.
Smith, K., 2014. Mental health: a world of depression. Nature 515, 181.
Sun, H., 2006. An accurate and interpretable bayesian classification model for prediction of hERG liability. ChemMedChem 1, 315–322.
Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., Feuston, B., 2003. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.
Tebartz van Elst, L., Ebert, D., Hesslinger, B., 2006. Depression - Augmentation or switch after initial SSRI treatment. N. Engl. J. Med. 354 2611-3; author reply 2611.
Yap, C.W., 2011. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. J. Comput. Chem. 32, 1466–1474.