ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Biological Activity Predictions of Ligands Based on Hybrid Molecular Fingerprinting and Ensemble Learning

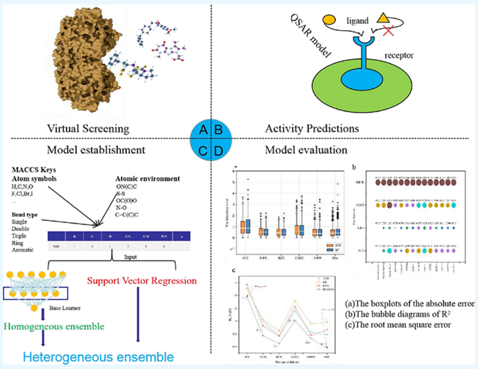
Mengshan Li,* Ming Zeng, Hang Zhang, Huijie Chen, and Lixin Guan

Cite This: *ACS Omega* 2023, 8, 5561−5570

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The biological activity predictions of ligands are an important research direction, which can improve the efficiency and success probability of drug screening. However, the traditional prediction method has the disadvantages of complex modeling and low screening efficiency. Machine learning is considered an important research direction to solve these traditional method problems in the near future. This paper proposes a machine learning model with high predictive accuracy and stable prediction ability, namely, the back propagation neural network cross-support vector regression model (BPCSVR). By comparing multiple molecular descriptors, MACCS fingerprint and ECFP6 fingerprint were selected as inputs, and the stable prediction ability of the model was improved by integrating multiple models and correcting similar samples. We used leave-one-out cross-validation on 3038 samples from six data sets. The coefficient of determination, root mean square error, and absolute error were used as the evaluation parameters. After comparing the multiclass models, the results show that the BPCSVR model has stable prediction ability in different data sets, and the prediction accuracy is higher than other comparison models.



1. INTRODUCTION

Ligand biological activity is an important parameter for receptors and ligand binding, and it is also the primary factor in the drug screening process.[1] By predicting ligand biological activity, the number of compounds to be screened can be reduced so as to improve the efficiency of the drug screening process, reduce the cost, and increase the positive rate of drug screening.[2] In summary, ligand biological activity prediction is a popular and valuable research topic in the field of drug screening.

The most common approach for the prediction of biological activity based on ligands is the quantitative structure−activity relationship (QSAR) proposed by Hansch et al.,[2,3] which is based on the principle that ligand activity is correlated with molecular structure and the activity value can be predicted by establishing a mathematical model based on the molecular structure of ligands.[4,5] The Hansch equation is the first one implemented in QSAR. This equation was formed by Hansch and Fujita, and it uses ED50 as the activity parameter and the electrical parameter., steric parameters, and hydrophobic parameters as variables for linear regression analysis.[6] Guided by the Hansch equation, 4-quinolone antibacterial drugs such as norfloxacin have been successfully designed, and it proves the validity of the Hansch equation.[7] However, the Hansch equation has many parameters that make the modeling process difficult. During the same period as when the Hansch equation was formed, Freeman-Cook et al. proposed the Free−Wilson method,[8] which is a method to quantitatively express the relationship between the chemical structure and the biological activity of drugs by mathematical formulas. The Free−Wilson method can reduce the modeling difficulty, but it is not as accurate as the Hansch equation and has not been widely used. It is generally considered that the more data characteristic parameters, the more accurate the model prediction. But the more data characteristic parameters, the more difficult it is to establish and solve the model. Due to the accumulation of a large number of experimental data, more and more researchers had been attempting to use machine learning methods to build models to predict the activity of compounds. The advantage of the machine learning method is that parameters are automatically learned according to the given data, which can reduce the difficulty of modeling and improve the efficiency of model solving. The more information contained in the data features of the machine learning model, the higher the prediction accuracy. For the activity prediction problem, we need a molecular descriptor that we can use for machine learning.

A variety of descriptors have been developed over the past few decades.[9−11] Molecular descriptors can be simply divided into two categories, namely, two-dimensional descriptors and high-dimensional descriptors. The two-dimensional molecular

Received: October 27, 2022
Accepted: December 23, 2022
Published: February 1, 2023

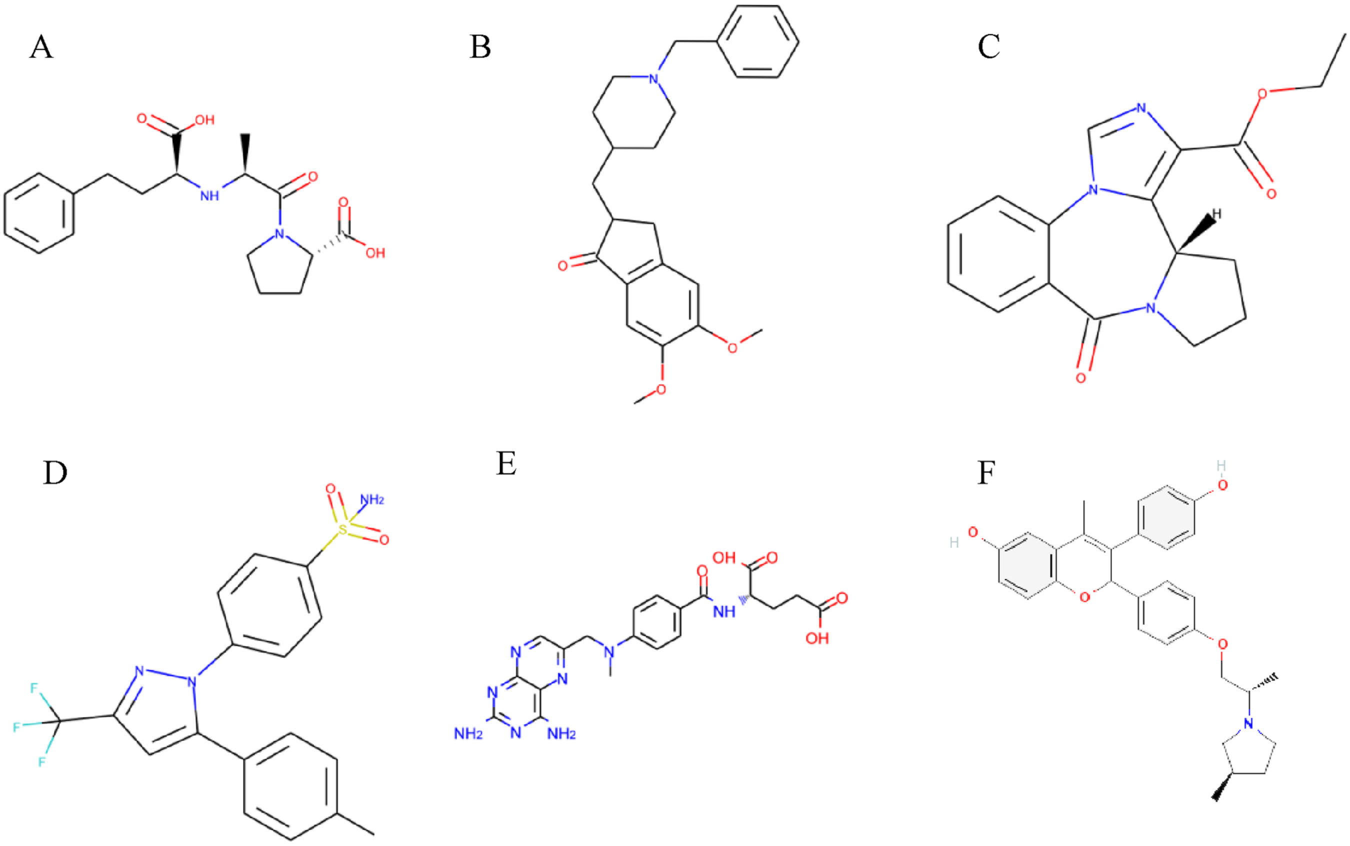


Figure 1. Representative compounds from six data sets: (A) enalaprat (ACE); (B) E2020 (AchE); (C) Ro14-5974 (BZR); (D) celecoxib (COX2); (E) methotrexate (DHFR); (F) SKPHGDXBXVGVNT-PGJZWCTDSA-N (ERα).

descriptors can be divided into four categories: (1) substructural bond-based fingerprints; (2) topology- or path-based fingerprints; (3) circular fingerprints; (4) pharmacophore fingerprints. 2D molecular fingerprints do not contain 3D structural information, greatly simplifying descriptors. The 3D structure is an important characteristic of molecules. Especially for molecules with a complex structure, 2D molecular descriptors often fail in the face of molecules with a complex 3D structure. Therefore, many researchers have proposed many new molecular descriptors. For example, algebraic topology, differential geometry, and graph theory are used to build a low-dimensional and extensible molecular descriptor, and the best prediction results are obtained in the D3R competition by these fingerprints.[12] In addition, the linear descriptors of hundreds of millions of compounds in the three databases of ChEMBL, PubChem, and ZINC are self-supervised prelearned and then fine-tuned to generate a molecular descriptor. This descriptor is novel and shows strong predictive ability in multiple tasks.[13] The generation and performance comparison of molecular descriptors is an important scientific research direction, while we will not delve into fingerprints in this article. By reading the references, we compare a variety of molecular fingerprints and find that molecular fingerprints have limited influence on the core point of this paper, and we realize that 2D fingerprints perform well enough at describing molecular information, so we use two 2D molecular fingerprints for experimental verification.[14] The two fingerprints are a substructure key-based fingerprint (MACCS) and circular fingerprint (ECFP6).

In the process of research, we found that many existing data sets have small sample amount, which leads to poor prediction results of the trained model. The small amount of data samples leads to a great difference in the prediction ability, that is, the less similar the sample, the less accurate the prediction. Also, different data sets have different experiment conditions, which lead to the same model having different predictive abilities for different data sets, that is, the model performs well on some data sets, while the prediction ability of some data sets may be extremely poor. The drug screening process will first predict a large number of unknown samples, which requires the model to be stable in predicting the activity of compounds under individual and different classes of compounds.

This paper proposes a heterogeneous integration model, the back propagation neural network cross-support vector regression (BPCSVR) model. This model has two components, which are the back propagation neural network cross (BPC) and support vector regression (SVR). Ensemble learning[15] is to build and combine multiple learners to accomplish learning tasks, which usually achieve significantly better generalization performance than a single learner, and it can improve the prediction ability and stability of the model. The bagging method of ensemble learning is to divide data sets and combine them into multiple training sets to improve the learning effect,[16] and the BPC model is based on this idea. In the training process, we screen a group of 5 to 10 samples from the training data as the validation set, and the selected samples are those with relatively high similarity between the training set and the test sample, but we should avoid selecting the five samples with the highest similarity.[17] Different validation sets correspond to different training sets, and different models can be obtained by using BP models with the same parameters. Each BP network calculates the determination coefficient and average error of the validation set, selects a model with higher determination coefficient to predict the test sample, and adds the average error to the final prediction value.[18,19]

The reason for the integration of the BPC model and SVR model is that the machine learning model has different effects on different data sets, that is, the model has unstable prediction effects on multiple different data sets. The integration of two different models can improve the overall stability of the model. To demonstrate the advantages of the BPCSVR model, the evaluation parameters RMSE, R^2, and AE and the performance are compared with those of SVR, DT, KNN, CoMFA, and RFSVR models.[20] To avoid the contingency of experimental results, the machine learning models presented in this paper all use leave-one-out cross-validation.

The contributions of this paper are as follows: (1) illustrate the feasibility of machine learning to build a QSAR model; (2) propose a machine learning model applicable to small sample data sets; (3) improve the accuracy and stability of ligand biological activity prediction based on molecular fingerprinting.

2. DATA SETS AND MODELS

2.1. Data Sets. Six data sets are studied in this paper: (1) angiotensin-converting enzyme (ACE) inhibitors, which are zinc-containing metallopeptidases;[21] (2) acetylcholinesterase (AchE) inhibitors, which can bring about memory improvement in SDAT;[16,22] (3) benzodiazepine receptors (BZRs), which are commonly used in the therapeutic treatment of anxiety, insomnia, seizure disorders, and certain kinds of spasticity;[23,24] (4) cyclooxygenase-2 (COX2), which is a highly selective and potent inhibitor;[25,26] (5) dihydrofolate reductase (DHFR) inhibitors, which can produce unacceptable toxicity to host proliferative tissues;[27−29] (6) estrogen receptor alpha (ERα) inhibitors, which are considered possible candidates for the treatment of breast cancer.[30] Representative structures of the compounds are shown in Figure 1.[31]

Data sets ACE, AchE, BZR, COX2, and DHFR have been frequently studied in references, which show the importance of these data sets. It is necessary to look at these data sets again, and past research helps us compare models. Second, the number of samples in these data sets is small and the distribution of ligand activity is unbalanced. These characteristics are convenient for us to study whether the model has stable prediction ability in small samples and unbalanced distribution data sets.

At the same time, to show whether the model can still maintain the predictive ability in the large data set, we add the ERα data set containing 1973 samples.

A total of 3038 samples were included in the six data sets, and the corresponding number and range of activity values are shown in Table 1.

The frequency distribution histograms and the kernel density function plots of the six data sets are shown in Figure 2. The abscissa of each subgraph is the activity value of the molecule, and the ordinate is the percentage of the number of molecules within the corresponding activity value range to the total number of molecules; the red line is the kernel density function curve of the data distribution. Analyzing Figure 2, the distribution of six data sets has their own characteristics, and the distribution characteristics are important factors affecting the training results. The samples of ACE are few, but the distribution is balanced, which is good for model prediction. AchE conforms to the characteristics of normal distribution, and the activity value is concentrated between 6 and 8, which is not good for model prediction. The distribution of BZR data leans to the right, which will make the model training prediction result also lean to a higher active value. The COX2 data set has a large number of samples with an activity value of 4, and the overall distribution is uneven, which will interfere with the model training; the distribution of DHFR is uneven. The activity values of data samples are concentrated in 6−8, and there are few samples at both ends. The prediction effect of this part may be poor; the distribution of ERα is also uneven, but due to the large number of samples, the overall prediction effect will not be too bad.

Table 1. Data Sets and Training Division

name	sample size	pIC50 value
ACE	114	2.14−9.94
AchE	93	4.27−9.52
BZR	145	5.3−8.85
COX2	318	4.0−9.0
DHFR	395	2.43−9.8
ERα	1973	2.45−10

2.2. Fingerprint. MACCS (Molecular Access System) predefines 166 specified structures (such as specific substructures or fragments) of the molecule, and if the molecule has a predefined feature, the corresponding feature bit is 1; otherwise, it is 0.[32] The advantage of the MACCS fingerprint is the compact structure and rich information, but the disadvantage is that it cannot define a single molecule, that is, the MACCS fingerprint of different compounds may be the same, which is not conducive to distinguishing compounds. The molecular structure of a compound contains information that determines its physical, chemical, and biological properties, so it is feasible to study the activity of a compound by molecular fingerprinting.[14] Unlike MACCS, which distinguishes compounds from their substructures, ECFPs (Extended-Connectivity Fingerprints) distinguish compounds based on the connectivity between atoms within them. They are a novel class of topological fingerprints for molecular characterization, and they are circular fingerprints with a number of useful qualities.[9] This is because this fingerprint requires setting a radius of n (the number of iterations) and then calculating each atomic environment identifier (identifier), which is similar to the connectivity in Morgan's fingerprint, except that the identifier here is ultimately determined by the environment of radius n. $n = 1$ is ECFP2, $n = 2$ is ECFP4, and so on. When n equals 0, we are just looking at the atoms themselves. The ECFP6 (1024-bit) fingerprint with the radius $n = 3$ is used in this article.

2.3. BPCSVR Model. The BPCSVR model is a heterogeneous, integrated learning method, and the component learner combines the BPC and SVR models, as shown in Figure 3. The BPC model of the component learner is a homogeneous ensemble learning model, which contains only individual learners of the same type, and its base learner is four BP models with the same training parameters and different training samples.[33]

The BP model mentioned above is built by the PyTorch framework, and the training logic is shown in Figure 4. The training process is to obtain the molecular fingerprints of compounds using software, then train the molecular fingerprint as the input of the BP neural network, and finally predict the molecular activity value.

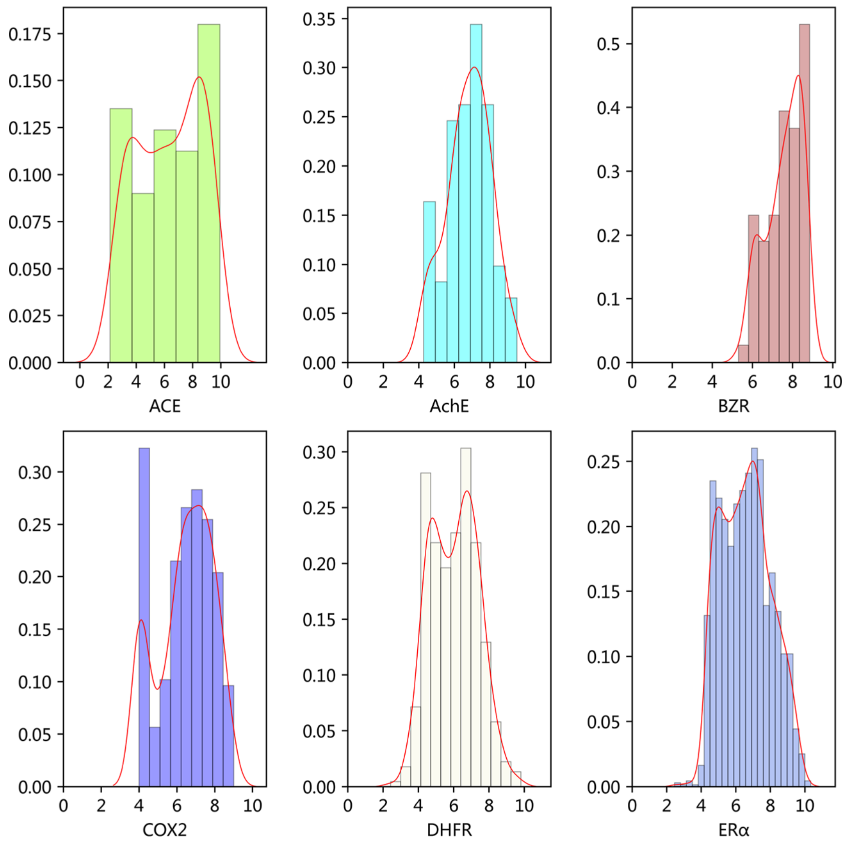


Figure 2. Frequency distribution histograms and kernel density maps corresponding to the six data sets. For each subgraph, the abscissa is the activity value of the molecule, and the ordinate is the percentage of the total.

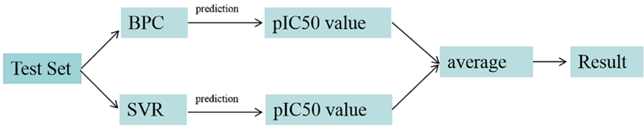


Figure 3. Integration process of the BPCSVR model.

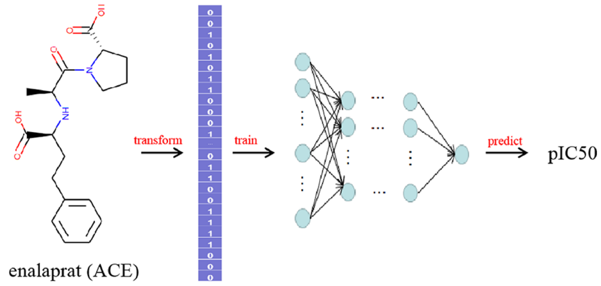


Figure 4. Schematic diagram of BP network training.

The training method of the BPC model is shown in Figure 5. Three groups of data were selected from the training data set as the validation set, and each group contained five samples. According to the molecular similarity between the training data and the test samples from high to low, the three validation sets were evenly distributed among the samples ranked from 5 to 15. Three models A, B, and C can be trained by the three partitioning methods, and which model is more reliable can be judged by calculating the coefficient of determination of the validation set corresponding to the three models. That is, the model with a high coefficient of determination is selected to predict the test samples, and then the prediction results are corrected by adding the average error of the validation set.

Another component learner of the BPCSVR model is the SVR model, which directly trains all training samples without setting a verification set or subtraining set during training. The SVR model is a common model with stable prediction ability, which is built by the SVR interface in the scikit-learn module.

2.4. Model Evaluation. In this paper, the goal was develop a model with high prediction accuracy and relatively stable prediction results. Four parameters have been used to evaluate the performance of the model: coefficient of determination (R^2), root mean square error (RMSE), absolute error (AE), and model optimization rate (MOR).

The coefficient of determination is a commonly used statistic that refers to the proportion of the variation in the dependent variable that is predictable from the independent variable, usually denoted as R^2, and it is calculated according to formula 1. It is usually used to evaluate the trustworthiness of the predicted value of the model. The closer the value is to 1, the higher the trustworthiness of the predicted value is.

RMSE is the square root of the deviation between the predicted value and the actual value divided by the number of samples N, as shown in formula 2. This parameter measures the deviation, or degree of dispersion, between the predicted value and the actual value, where smaller values imply better models.

Absolute error (AE) refers to the absolute value of the deviation between the real value and the predicted value of each sample, which can truly reflect the deviation of the predicted value from the true value, as shown in formula 3.

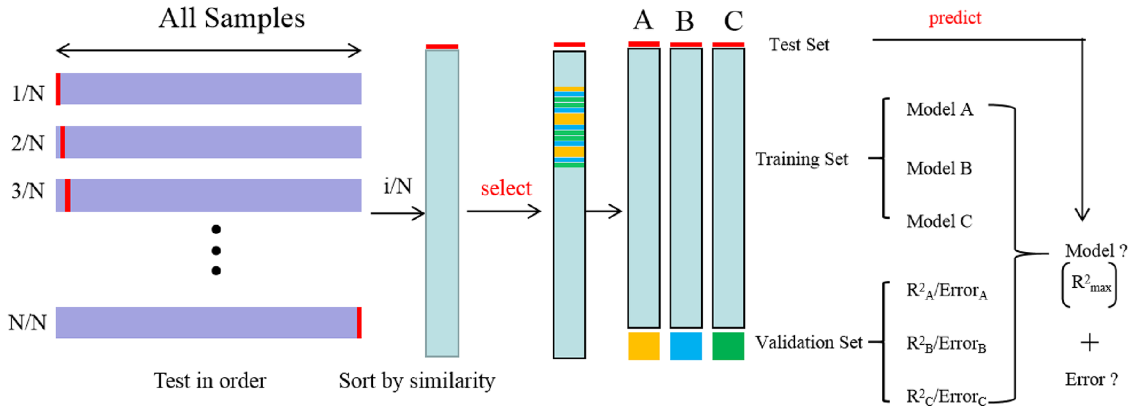


Figure 5. Schematic of the BPC model. Red is the test sample, one at a time; yellow, blue, and green represent three validation sets, each with five samples.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i^{\text{exp}} - y_i^{\text{pred}})^2}{\sum_{i=1}^{N} (y_i^{\text{exp}} - y^{\text{mean}})^2} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i^{\text{exp}} - y_i^{\text{pred}})^2} \tag{2}$$

$$\text{AE} = |y^{\text{exp}} - y^{\text{pred}}| \tag{3}$$

In the above formula, y^{exp} refers to the actual value, y^{pred} refers to the predicted value, and y^{mean} refers to the average value of the actual value.

3. RESULTS AND DISCUSSION

The BPCSVR model was built on a Windows 7 64-bit hardware environment and implemented by Python programming. The research topic of this paper is how to establish a QSAR model based on molecular fingerprinting and machine learning methods. The goal is to improve the overall prediction accuracy for small sample data sets, and to be able to stably predict values across different data sets.

3.1. Results of the Proposed Model.

The BPCSVR model adopts three strategies: (1) isomorphic integration, (2) error correction using a verification set, and (3) heterogeneous integration to improve the stability prediction ability of the model. In the process of model design, the MACCS fingerprint and ECFP fingerprint were compared, and six groups of data were tested. The effectiveness of the model was illustrated by the vertical comparison before and after model integration and the horizontal comparison between different models.

Figure 6 shows the results given in ref 13. To make the fluctuation between the data more obvious, all the data were normalized, R^2 was mapped to a range of 0.7 to 1, and the label above the circle is the real R^2 value. As can be seen from the figure, this experiment uses multiple fingerprints to train and test on four basic data sets, and the R^2 values obtained are all above 0.7. Different molecular fingerprints behave differently on the same data set, and the same fingerprint also behaves differently across different data sets, in comparison, because the data set changes. Since the influence of fingerprints on the results is less than that of the change of the data set, we choose the MACCS fingerprint and ECFP6 fingerprint for experimental comparison.

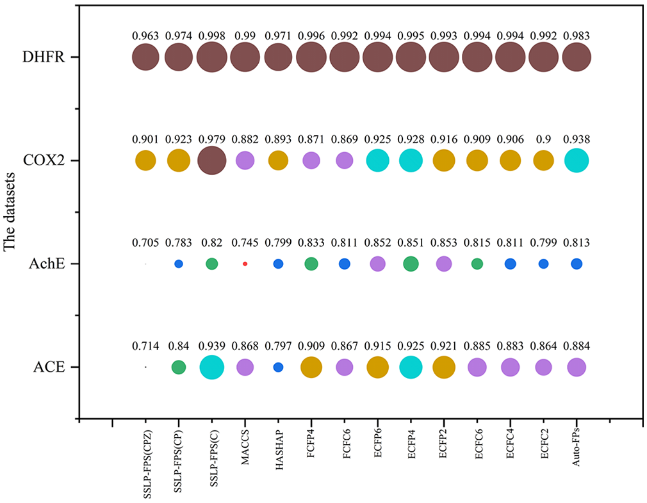


Figure 6. Bubble diagrams of R^2 for different molecular fingerprint training. R^2 is normalized, and the color and graphic size are used to distinguish R^2.

Figure 7 shows the evaluation parameters calculated by the BPC model and RF model (random forest) after training with two kinds of fingerprints. According to the figure, when the RF model and BPC model use the ECFP6 fingerprint to train and predict, R^2 is larger and RMSE is smaller. This indicates that the model trained with the ECFP6 fingerprint has a higher degree of trust, and the error between the predicted value and the actual value is smaller. Based on the above judgment, to obtain the best model, ECFP6 fingerprinting is used uniformly in the subsequent model training.

Six groups of data are predicted by using the leave-one-out cross-validation method, the AE values of the predicted values and actual values of the BP model and BPC model are calculated, and the box plot of AE is drawn as shown in Figure 8. It can be seen from the graph analysis that there are fewer outliers in the box graph of the BPC model, and the outliers are closer to 0. For the horizontal comparison of each data set, compared with the BP model, the third quartile of the BPC model in ACE, COX2, DHFR, and ERα data sets is smaller, and the first quartile in ACE, AchE, BZR, DHFR, and ERα data sets is smaller. This shows that in most data sets, the prediction error of the BPC model is smaller than that of the BP model.

From the above analysis, we can see that the BPC model can reduce the prediction error, that is, the adopted integration

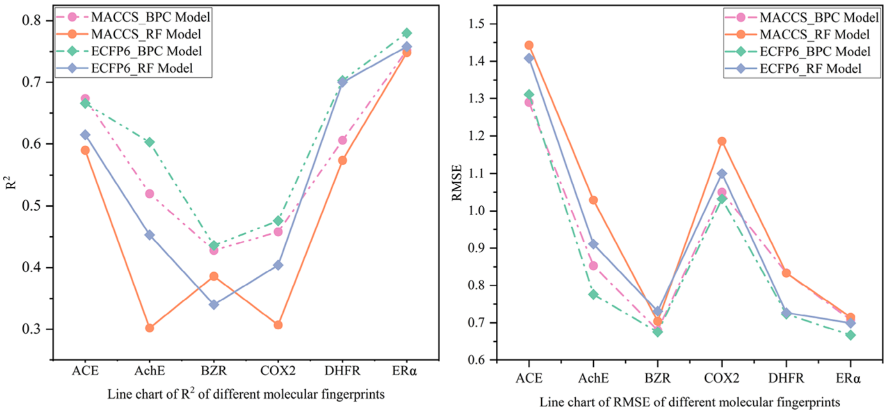


Figure 7. R^2/RMSE of two types of fingerprints in the BPC model and RF model. The evaluation parameter for the left subfigure is R^2, and that for the right subfigure is RMSE; the BPC model is represented by dotted lines, while the RF model is represented by straight lines.

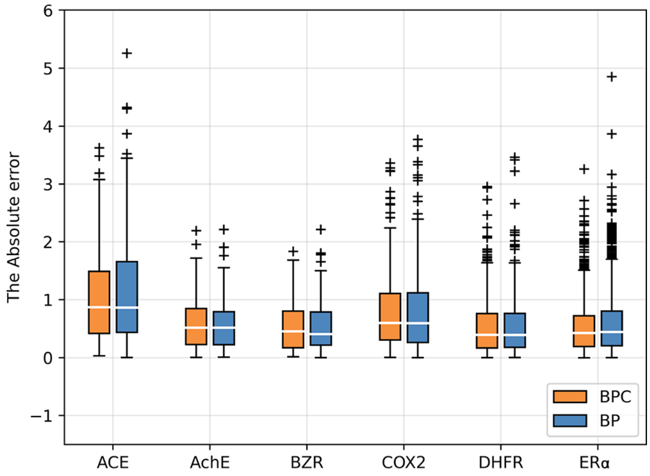


Figure 8. Box plots of the AE between predicted and actual values of the BPC model and BP model. The white line is the mean, and the black cross is an outlier.

strategy and error correction strategy can effectively reduce the prediction error.

Figure 9 shows the average values of the R^2 and RMSE obtained by the BPCSVR model and BPC model using the leave-one-out cross-validation of six data sets. It can be seen from the figure that through heterogeneous integration, the R^2 of the model becomes larger and RMSE decreases, which indicate that integrating two different models can improve the reliability and prediction accuracy of the model.

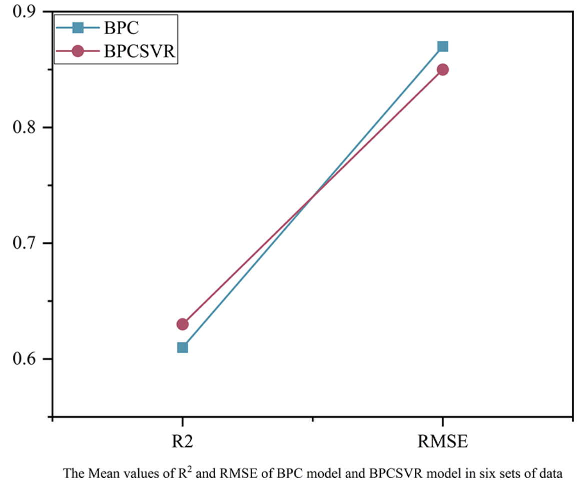


Figure 9. Mean values of R^2 and RMSE of the BPC model and BPCSVR model in six sets of data.

Table 2 shows the R^2 and RMSE of the prediction results of the BPCSVR model for the six test sets. The average of R^2 is 0.626, which indicates that the predicted value of the model has a strong correlation with the actual value; RMSE is 0.85, while the distribution range value of the sample activity is 2−10, and the deviation is only 8.5−42.5%, which means that the error is relatively small.

Through homogeneous integration and heterogeneous integration, the BPCSVR model has stronger adaptability to abnormal samples and data sets with unbalanced distribution, higher prediction accuracy, and more stable prediction effect.

3.2. Comparison of the BPCSVR Model with the Non-Integrated Model. To demonstrate the effectiveness of the BPCSVR model, several common machine learning models, including the SVR model, decision tree (DT) model, *K*-nearest neighbor (KNN) model, and CoMFA model, will be compared next. The SVR model is a kind of model that is robust to outliers and can be compared with the model proposed in this paper. Both the DT model and KNN model are traditional models, which have been applied in many fields. The CoMFA model is the most mature and widely used method in traditional QSAR models, so the effectiveness of the machine learning model can be studied by comparing with this model.

Table 3 shows the evaluation parameters R^2 and RMSE calculated by leave-one-out verification of each model. It can be seen from the table that machine learning models differ greatly in different data sets, that is, some models perform well in a data set, but the performance of another data set may be very poor. The DT model is the most significant. When the DT model is applied to the ACE data set, the decision coefficient is 0.556, while when the DT model is applied to the BZR data set, the decision coefficient is 0.112. Although the SVR model and KNN model have good performance in ACE, DHFR, and ERα data sets, they have poor performance in AchE and COX2 data sets. The only model that can maintain high prediction ability in six data sets is the BPCSVR model.

Figure 10 shows the RMSE line chart of six data sets predicted by each model. It can be seen from the figure that the RMSE of the BPCSVR model in each data set remains the minimum. Compared with the model with the largest error, the

Table 2. R^2 and RMSE Values for the Test Sets of Different Compounds

	ACE	AchE	BZR	COX2	DHFR	ER*α*	mean
R^2	0.632	0.608	0.48	0.502	0.7	0.836	0.626
RMSE	1.377	0.771	0.649	1.005	0.727	0.575	0.850

Table 3. R^2 and RMSE of Each Model

data set	evaluation	SVR	DT	KNN	CoMFA	BPCSVR
ACE	R^2	0.557	0.556	0.523	0.49	0.632
	RMSE	1.51	1.5	1.567		1.377
AchE	R^2	0.335	0.31	0.435	0.47	0.608
	RMSE	1.004	1.022	0.925		0.771
BZR	R^2	0.288	0.112	0.267	0.00	0.48
	RMSE	0.759	0.847	0.77		0.649
COX2	R^2	0.346	0.154	0.268	0.29	0.502
	RMSE	1.153	1.311	1.219		1.005
DHFR	R^2	0.612	0.476	0.655	0.59	0.7
	RMSE	0.827	0.960	0.779		0.727
ER*α*	R^2	0.695	0.526	0.633		0.836
	RMSE	0.785	0.98	0.862		0.575

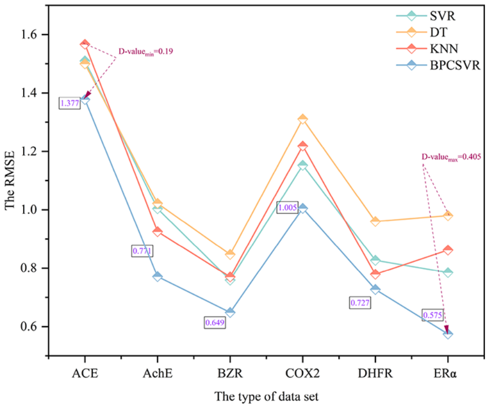


Figure 10. Line chart of RMSE comparison between the BPCSVR model and the non-integrated model.

BPCSVR model has the least and most reduction in ACE and ER*α* data sets (0.19 and 0.405, respectively). The smaller RMSE is, the smaller the deviation between the predicted value and the actual value of the model is, that is, the accuracy of the BPCSVR model is higher than that of other models.

To sum up, the BPCSVR model has higher reliability and accuracy than other non-integrated machine learning models.

3.3. Comparison between the BPCSVR Model and the Integrated RFSVR Model.

The integrated model is generally considered superior to the non-integrated model, so to compare the integrated models, the RFSVR model is additionally designed with reference to the BPCSVR model. The RFSVR model is integrated by the RF model and SVR model, that is, the average of RF prediction results and SVR model prediction results is the prediction value of RFSVR.

Figure 11 shows the R^2 and RMSE of the BPCSVR model and RFSVR model for six groups of data sets, and the left is the line chart and the right is the bar chart. From the figure, it is obvious that the R^2 of the BPCSVR model is larger, RMSE is smaller, and the RMSE of the RFSVR model fluctuates greatly among data sets. Therefore, simple stacking of models may not be able to steadily improve the prediction ability of models.

Different from the RFSVR model, which simply divides the results of two models equally, the BPCSVR model integration strategy is special in that it uses similar samples to evaluate the usefulness of the current model and uses the error of the model's prediction for similar samples to correct the deviation of test samples.

Figure 12 shows the AE distribution function of BPCSVR model and RFSVR model predictions for six groups of data. The horizontal axis is the AE value, and the vertical axis is the number of samples. From the analysis of the figure, more BPCSVR model prediction samples fall into the small error area. When the AE value is less than 1, the numbers of samples of the BPCSVR model in each data set are 70, 77, 125, 236, 345, and 1819, while the corresponding numbers of the RFSVR model are 53, 67, 122, 216, 325, and 1690, respectively. The BPCSVR model always has more samples with smaller errors between predicted values and actual values, which also show that the BPCSVR model has higher prediction accuracy.

Compared with the RFSVR model, the BPCSVR model has higher R^2, smaller RMSE, and smaller error of more sample

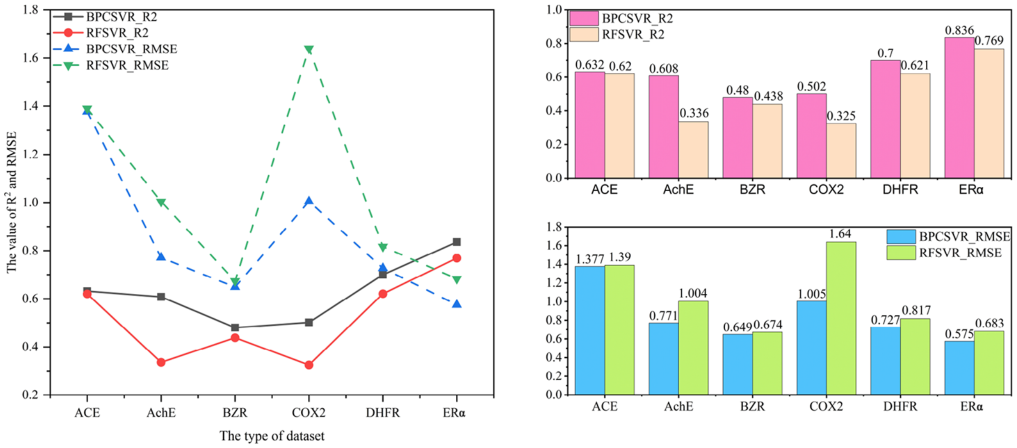


Figure 11. Figure of R^2 and RMSE of the BPCSVR model and RFSVR model on each data set.

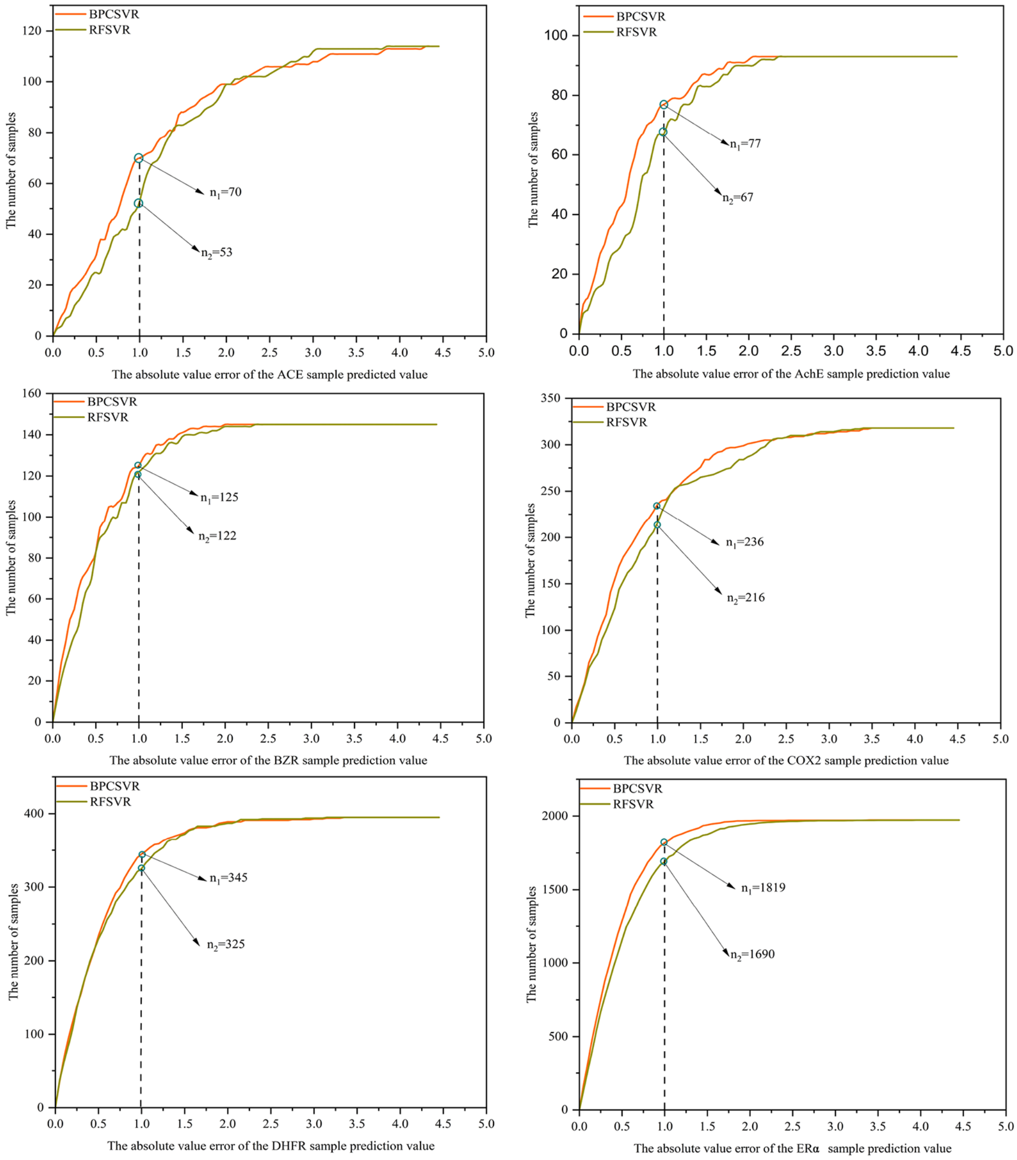


Figure 12. Distribution curve of AE predicted by the BPCSVR model and RFSVR model for six groups of data.

prediction values, which show that the BPCSVR model is more reliable and accurate than the RFSVR model.

4. CONCLUSIONS

In this paper, the BPCSVR model is proposed to train and predict the activity of six groups of ligand compounds. The experimental results show that the QSAR model based on the machine learning method is completely feasible, its prediction ability is better than the traditional COMFA model, and the model establishment is simpler. The BPCSVR model makes full use of the sample data of the small sample data set through data partitioning and integration and uses the RMSE of the validation set to correct the prediction results so as to improve the prediction accuracy of the small sample data. Through the evaluation of R^2, RMSE, AE, optimization rate, and other parameters, it can be seen that the BPCSVR model has the strongest ability and is superior to other comparison models in terms of prediction accuracy, overall prediction stability, and single prediction stability.

However, the model proposed in this paper still has some shortcomings. First of all, the prediction ability of the model still lags far behind that of the cutting-edge research. To improve the predictive ability of the model, molecular fingerprints can be improved in the next research, and the predictive ability can be improved through the improvement of molecular fingerprints. Second, compared with cutting-edge research, the experimental data sets in this paper are small, which indicates that the extensibility of the model is not fully explained. In the next study, we will add experimental data sets to illustrate the extensibility of the model.

■ ASSOCIATED CONTENT

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c06944.

Detailed description of the data source; core source code for the experiment of this article; graphic representation of MACCS fingerprints of five data sets: ACE, AchE, BZR, COX2, and DHFR (PDF)

AUTHOR INFORMATION

Corresponding Author

Mengshan Li − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi 341000, China*; orcid.org/0000-0001-7832-4185; Email: jcimsli@163.com

Authors

Ming Zeng − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi 341000, China*

Hang Zhang − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi 341000, China*

Huijie Chen − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi 341000, China*

Lixin Guan − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi 341000, China*; orcid.org/0000-0002-2554-5678

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c06944

Author Contributions

M.L. and M.Z. designed the study. M.Z., H.C., and H.Z. performed the research. M.L. and M.Z. conceived the idea. L.G. and M.Z. provided and analyzed the data. H.C. and M.Z. helped perform the analysis with constructive discussions. All authors contributed to writing and revision.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support from the National Natural Science Foundation of China (grant nos. 51663001, 52063002, 42061067, and 61741202), the Science and Technology Research Project of the Education Department of Jiangxi Province (grant no. 190745), and the Degree and Postgraduate Education and Teaching Reform Research Project of Jiangxi Province (multi thinking and "439" scientific research ability training mode for interdisciplinary academic degree postgraduate training).

REFERENCES

(1) Myint, K. Z.; Ma, C.; Wang, L.; Xie, X. Q. Fragment-similarity-based QSAR (FS-QSAR) algorithm for ligand biological activity predictions. *SAR QSAR Environ. Res.* **2011**, *22*, 385−410.

(2) Liu, X.; Zheng, W.; Jiang, X. Cell-Based Assays on Microfluidics for Drug Screening. *ACS Sensors* **2019**, *4*, 1465−1475.

(3) Xie, X.-Q.; Chen, J.-Z. Data Mining a Small Molecule Drug Screening Representative Subset from NIH PubChem. *J. Chem. Inf. Model.* **2008**, *48*, 465−475.

(4) Free, S. M.; Wilson, J. W. A MATHEMATICAL CONTRIBUTION TO STRUCTURE-ACTIVITY STUDIES. *J. Med. Chem.* **1964**, *7*, 395−399.

(5) Hansch, C.; Fujita, T. p-Ã0À Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(6) Soares, T. A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G.-W.; Merz, K. The (Re)-Evolution of Quantitative Structure−Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. *J. Chem. Inf. Model.* **2022**, *62*, 5317−5320.

(7) Myint, K. Z.; Xie, X. Q. Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *Int. J. Mol. Sci.* **2010**, *11*, 3846−3866.

(8) Freeman-Cook, K. D.; Amor, P.; Bader, S.; Buzon, L. M.; Coffey, S. B.; Corbett, J. W.; Dirico, K. J.; Doran, S. D.; Elliott, R. L.; Esler, W.; Guzman-Perez, A.; Henegar, K. E.; Houser, J. A.; Jones, C. S.; Limberakis, C.; Loomis, K.; McPherson, K.; Murdande, S.; Nelson, K. L.; Phillion, D.; Pierce, B. S.; Song, W.; Sugarman, E.; Tapley, S.; Tu, M. H.; Zhao, Z. R. Maximizing Lipophilic Efficiency: The Use of Free-Wilson Analysis in the Design of Inhibitors of Acetyl-CoA Carboxylase. *J. Med. Chem.* **2012**, *55*, 935−942.

(9) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(10) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(11) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.

(12) Nguyen, D. D.; Cang, Z.; Wei, G. W. A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4343−4367.

(13) Chen, D.; Zheng, J.; Wei, G.-W.; Pan, F. Extracting Predictive Representations from Hundreds of Millions of Molecules. *J. Phys. Chem. Lett.* **2021**, *12*, 10793−10801.

(14) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G. W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373−8390.

(15) Hu, P.; Jiao, Z.; Zhang, Z.; Wang, Q. Development of Solubility Prediction Models with Ensemble Learning. *Ind. Eng. Chem. Res.* **2021**, *60*, 11627−11635.

(16) Sugimoto, H.; Tsuchiya, Y.; Sugumi, H.; Higurashi, K.; Karibe, N.; Iimura, Y.; Sasaki, A.; Kawakami, Y.; Nakamura, T.; Araki, S. Novel piperidine derivatives. Synthesis and anti-acetylcholinesterase activity of 1-benzyl-4-[2-(N-benzoylamino)ethyl]piperidine derivatives. *J. Med. Chem.* **1990**, *33*, 1880−1887.

(17) Wu, K. A.-O.; Zhao, Z.; Wang, R.; Wei, G. A.-O. TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J. Comput. Chem.* **2018**, *39*, 1444−1454.

(18) Zhang, X.; Mao, J.; Wei, M.; Qi, Y.; Zhang, J. Z. H. HergSPred: Accurate Classification of hERG Blockers/Nonblockers with Machine-Learning Models. *J. Chem. Inf. Model.* **2022**, *62*, 1830−1839.

(19) Zhao, Z.; Qin, J.; Gou, Z.; Zhang, Y.; Yang, Y. Multi-task learning models for predicting active compounds. *J. Biomed. Inf.* **2020**, *108*, No. 103484.

(20) Caballero, J.; Saavedra, M.; Fernández, M.; González-Nilo, F. D. Quantitative Structure−Activity Relationship of Rubiscolin Analogues as δ Opioid Peptides Using Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA). *J. Agric. Food Chem.* **2007**, *55*, 8101−8104.

(21) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(22) Sugimoto, H.; Tsuchiya, Y.; Sugumi, H.; Higurashi, K.; Karibe, N.; Iimura, Y.; Sasaki, A.; Araki, S.; Yamanishi, Y.; Yamatsu, K. Synthesis and structure-activity relationships of acetylcholinesterase inhibitors: 1-benzyl-4-(2-phthalimidoethyl)piperidine and related derivatives. *J. Med. Chem.* **1992**, *35*, 4542−4548.

(23) Haefely, W. E.; Kyburz, E.; Gerecke, M.; Möhler, H. Recent advances in the molecular pharmacology of Benzodiazepine receptors and in the structure-activity relationships of their agonists and antagonists. *Adv. Drug Res.* **1985**, *14*, 165−322.

(24) Braestrup, C.; Nielsen, M. Benzodiazepine receptors. *Arzneimittelforschung* **1980**, *30*, 852−857.

(25) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-dimensional quantitative structure-activity relationships of cyclo-oxygenase-2 (COX-2) inhibitors: a comparative molecular field analysis. *J. Med. Chem.* **2001**, *44*, 3223−3230.

(26) Talley, J. J.; Brown, D. L.; Carter, J. S.; Graneto, M. J.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Rogers, R. S.; Shaffer, A. F.; Zhang, Y. Y.; Zweifel, B. S.; Seibert, K. 4-5-Methyl-3-phenylisoxazol-4-yl- benzenesulfonamide, valdecoxib: a potent and selective inhibitor of COX-2. *J. Med. Chem.* **2000**, *43*, 775−777.

(27) Gangjee, A.; Vidwans, A. P.; Vasudevan, A.; Queener, S. F.; Kisliuk, R. L.; Cody, V.; Li, R.; Galitsky, N.; Luft, J. R.; Pangborn, W. Structure-based design and synthesis of lipophilic 2,4-diamino-6-substituted quinazolines and their evaluation as inhibitors of dihydrofolate reductases and potential antitumor agents. *J. Med. Chem.* **1998**, *41*, 3426−3434.

(28) Rosowsky, A.; Mota, C. E.; Wright, J. E.; Queener, S. F. 2, 4-Diamino-5-chloroquinazoline analogues of trimetrexate and piritrexim: synthesis and antifolate activity. *J. Med. Chem.* **1994**, *37*, 4522−4528.

(29) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(30) *The 18th (2021) National Mathematical Contest in Modeling (NPMCM) test questions*. https://www.shumo.com/wiki/doku.php?id=%E7%AC%AC%E5%8D%81%E5%85%AB%E5%B1%8A_2021_%E5%85%A8%E5%9B%BD%E7%A0%94%E7%A9%B6%E7%94%9F%E6%95%B0%E5%AD%A6%E5%BB%BA%E6%A8%A1%E7%AB%9E%E8%B5%9B_npmcm_%E8%AF%95%E9%A2%98.

(31) Myint, K. Z.; Wang, L.; Tong, Q.; Xie, X. Q. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol. Pharmaceutics* **2012**, *9*, 2912−2923.

(32) *MACCS fingerprint each bit corresponding to the meaning*. https://github.com/openbabel/openbabel/blob/master/data/MACCS.txt.

(33) Zimo, Y.; Haixin, A.; Li, Z.; Guofei, R.; Yuming, W.; Qi, Z.; Hongsheng, L. Predicting the cytotoxicity of chemicals using ensemble learning methods and molecular fingerprints. *J. Appl. Toxicol.* **2019**, *39*, 1366−1377.