Journal of Information and Computing Science
J. Info. Comput. Sci., Vol. 18, No. 1, pp. 27-50

DOI: 10.4208/JICS-2023-003
June 2023

Prediction of Properties of Anti-Breast Cancer Drugs Based on PSO-BP Neural Network and PSO-SVM

Meixian Xu[1], Yan Zheng[1,*], Yanju Li[1] and Weihao Wu[1]

[1] *College of Automobile and Traffic Engineering, Nanjing Forestry University, Nanjing 210044, China*

Abstract. The process of screening and developing new drugs through experiments is very slow and requires a lot of manpower and material resources, and the use of computer-aided prediction of the molecular properties of drugs can greatly save time and cost of drug development. Therefore, in order to enable anti-breast cancer candidate drugs to have good biological activity and ADMET properties for inhibiting ERα, the random forest classifier was first used for the collected 1 974 compounds to screen the top 20 molecular descriptors with the most significant effects on biological activity. Then a QSAR model was established using this and pIC_{50} value as characteristic data. The biological activity values of 50 new compounds were predicted via the PSO optimized BP neural network, with the model fit of 0.833 7 and the root mean square error of 0.731 5, which were more consistent with the actual values than the predicted results of the BP neural network. Subsequently, in order to improve the success rate of drug development, the ADMET classification prediction model was constructed using PSO to optimize the SVM based on the existing ADMET property data. The algorithm cross-validation CV accuracy rate reached 94.076 7%, and the prediction accuracy rates of the five index models were all above 79%. The results show that the proposed model has better prediction performance than the benchmark model, and the adopted prediction strategy is effective, which can provide reference for the discovery and development of anti-breast cancer drugs.

AMS subject classifications: 62M20, 92B20

Key words: Anti breast cancer drugs, Biological activity, ADMET properties, Particle Swarm Optimization (PSO), BP neural network, Support Vector Machines (SVM).

1 Introduction

Translated from *Journal of Nanjing University of Information Science & Technology*, 2023, 15(1): 51-65.

*Corresponding author. *Email addresses:* **xumeixian3210@163.com** (M. Xu), **ZhengYan3210@163.com** (Y. Zheng).

According to the 2018 cancer data report of the American Cancer Center, breast cancer is the most common malignant tumor in women in the world, and it seriously threatens the physical and mental health of women [1]. Breast cancer has become a worldwide health care problem, and treatment options need to be both selective and take into account the probability of effectiveness. To solve this problem, a large number of drug candidates have been studied and analyzed in the field of medicinal chemistry. The experimental results of estrogen receptor α subtype (ERα) gene deletion in mice show that ERα is considered to be an important target for the treatment of breast cancer, and compounds that can antagonize ERα activity may be candidates for the treatment of breast cancer.

Anti-breast cancer drug candidates need to have good biological activity from development to use, and their pharmacokinetic properties and safety should also meet the requirements of relevant policies and regulations. If only experimental methods are used to evaluate the biological activity, pharmacokinetic properties and safety of compounds, the time and cost will be immeasurable. Its pharmacokinetic properties and safety are collectively known as ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties. Moreover, the data obtained from experimental animals does not fully coincide with clinical data, so it cannot meet the needs of modern drug research [2]. In order to save time and cost, research institutions usually choose to combine in vitro research techniques with computer computing models to establish compound activity prediction models and screen potential active compounds. That is, by collecting a series of compounds acting on ERα and their biological activity data, and selecting a series of molecular structure descriptors as independent variables, and the biological activity value of the compound as dependent variables, a quantitative structure-activity relationship (QSAR) model of the compound is constructed, and then the model is used to predict new compound molecules with better biological activity. Or to guide the structural optimization of existing active compounds. In addition to biological activity, pharmacokinetic properties and toxicity (ADMET) are also important factors in determining the success of drug development. No matter how good the activity of a compound is, if its ADMET properties are not good, such as difficult to be absorbed by the body, or the body metabolism is too fast, or has some toxicity, then it is still difficult to become a drug, so it also needs to optimize the ADMET properties.

In the case of the rapid increase in the number of drugs, the most economical and reasonable research method is to use computer-aided artificial intelligence algorithm to predict the biological activity of drugs and the properties of ADMET. Gu et al. [3] collected a large amount of drug ADMET data from multiple public databases and proposed to use graph neural network model for virtual screening of drug development after effective data cleaning. The research results showed that the model had good predictive performance and could be generalized. Considering the accuracy and fit of shallow and deep neural networks, Xie et al. [4] chose to combine several neural networks with stacking method to predict drug molecular properties, and the fusion model had high prediction accuracy and reliability. In order to effectively predict the molecular biological activity value of drug lead compounds, Qin [5] deeply studied the learning of matrix completion algorithm in labeled ligand characteristics. Compared with deep learning, the algorithm showed stronger advantages, and the predicted optimal value was more realistic. Jia [6] used three machine learning algorithms, namely random forest, support vector machine and artificial neural network, to construct a quantitative prediction model for drug targets, and compared and analyzed the prediction results of the three algorithms, indicating that the optimal model constructed by them could objectively screen out effective molecular descriptors from the perspective of molecular vibration. Shen [7] established a QSAR prediction model of small molecule ADMET based on the classical genetic algorithm by inhaling elite warehouse strategy, and evaluated the molecular structure of compounds based on information gain, which verified that the established model could be extended to drug metabolism and toxicity assessment.

Reviewing literature [1-7], it can be seen that using artificial intelligence methods to predict the biological activity of drugs and the properties of ADMET has obviously become a research hotspot. Studies have shown that using artificial intelligence algorithms to predict the biological activity of drugs and the properties of ADMET can significantly reduce the cost of research and development, improve the probability of research and development success, and be more conducive to exploring the role of drug candidates in organisms, effectively avoiding human diseases caused by side effects and toxicity of drugs, and guiding rational drug use in clinical treatment [8]. Therefore, it is of great practical significance to use computer-aided artificial intelligence algorithm to theoretically predict the biological activity and ADMET properties of anti-breast cancer drug candidates.

In this paper, data on biological activity and ADMET properties of 1 974 compounds against ERα, a therapeutic target for breast cancer, were obtained from DrugBank drug molecule database of the University of Alberta, Canada. A quantitative prediction model was established from the perspective of compound molecular descriptors using the collected information. The IC_{50} and pIC_{50} values of new compounds were predicted by BP neural network algorithm based on particle swarm optimization. At the same time, a classification prediction model was constructed to predict five ADMET properties of compounds, namely Caco-2, CYP3A4, hERG, HOB, and MN, based on particle swarm optimization support vector machine, so as to find the molecular descriptors of compounds that can satisfy the high activity of compounds and make ADMET properties as good as possible. To accelerate the development of anti-breast cancer drug candidates.

2 Data Collection

For the breast cancer therapeutic target ERα, the bioactivity data of 1 974 compounds against ERα, 729 molecular descriptors and 5 ADMET properties were obtained from the DrugBank drug molecule database of the University of Alberta [9]. DrugBank database has unique bioinformatics and cheminformatics resources. It combines detailed drug data with comprehensive drug target information so that scientists can study drug mechanisms and explore novel drugs. The data collected in this paper include SMILES structural formula of the compound, IC_{50} and pIC_{50} values of the compound's bioactivity to ERα, 729 molecular descriptor information (independent variables), interpretation of the meaning of the molecular descriptor. The pharmacokinetic properties and toxicity of Caco-2, CYP3A4, hERG, HOB and MN were obtained by 0-1 bi-classification method.

3 Filter the main molecular descriptors

3.1 Data Preprocessing

The 729 molecular descriptors collected were observed and the data were processed to find that some descriptors of 1 974 organic compounds were all 0, for example, the molecular descriptor *n*B (boron atom number) was all 0. A large number of "0" data is not missing, but the molecular descriptor of the compound is the number "0" [10], which is of practical significance for pharmaceutical research, so it is not necessary to remove all 0 descriptor rows during data preprocessing. Therefore, the quantitative structure-activity relationship (QSAR) model can be built directly using the original data of 729 molecular descriptors of 1 974 compounds as independent variables and biological activity values as dependent variables.

In the collected data set, the bioactivity values of the compounds against ERα were expressed as IC_{50}. IC_{50} is the experimental determination value, the unit is nmol/L, and the smaller the value, the greater the biological activity and the more effective the inhibition of ERα activity. It can be seen from the tests in reference [7-10] and the special software PaDEL-Descriptor calculated by molecular descriptors that pIC_{50} value is usually obtained by IC_{50} conversion (that is, the negative logarithm of IC_{50} value), and pIC_{50} value is usually positively correlated with biological activity, that is, the larger pIC_{50} value is, the higher the biological activity is. In practical QSAR theoretical modeling, pIC_{50} value is generally adopted to represent the biological activity value. First of all, 729 molecular descriptors of 1 974 compounds need to be selected for variables, and the top 20 molecular descriptors (i.e. independent variables) that have the most significant impact on biological activity should be sorted according to the importance of each variable on biological activity. Since the collected molecular descriptor data is two-dimensional data, that is, the solubility, surface area and other information of the corresponding molecules, it is necessary to screen out several features that have the greatest impact on the results, and use them as the feature data when establishing the model. Common solution methods include principal component analysis, LASSO, random forest, etc. However, classical algorithms such as principal component analysis and LASSO will bring ambiguity when feature extraction and dimension reduction of 729 variable indicators, resulting in the loss of clarity and accuracy of the original variable meaning [11]. Therefore, the random forest (RF) algorithm was used to evaluate the importance of features and screen out the molecular descriptors that had a great influence on the activity value.

3.2 Screening molecular descriptors based on random forest

Based on Bagging algorithm, random forest generates an independent set of equally distributed training samples for each decision tree, and the voting of all decision trees will determine the final classification result. Based on the random forest model, the collected molecular descriptor data was input into the MATLAB software for calculation. The results of the *i* and *j* programs were shown in Figure 1 and Figure 2 respectively.

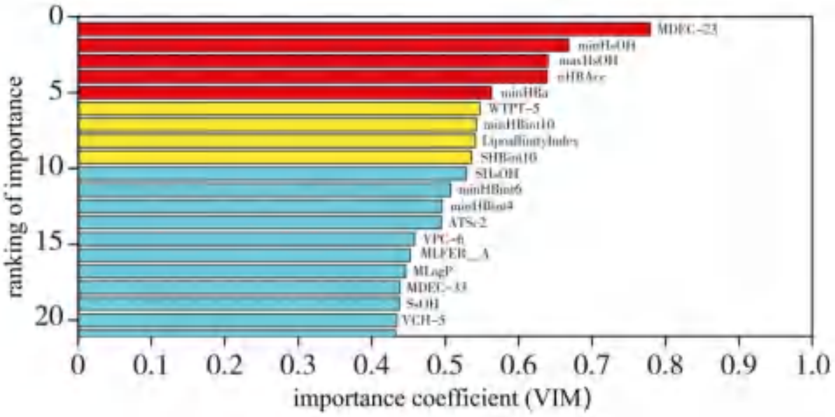


Figure 1: Relative importance of molecular descriptors (variables) in the *i*-TH experiment

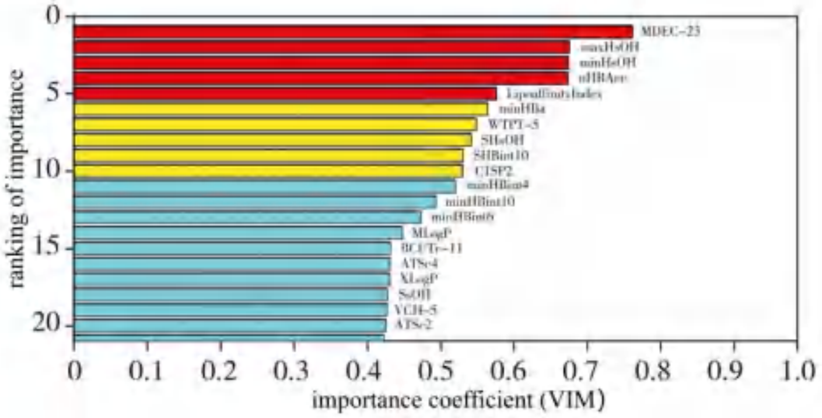


Figure 2: Relative importance of molecular descriptors (variables) in the *j*-TH experiment

Since each training is random sampling, the ranking results of the importance of molecular descriptors are different after the program is run, so 10 experiments are designed to carry out statistics on the importance of molecular descriptors. Let VIM be the importance coefficient, then VIM_j^i represents the importance coefficient of the i-TH molecular descriptor of the j-TH experiment, respectively. By counting the top 20 molecular descriptors that have appeared in 10 experiments, the average importance coefficient of the statistical molecular descriptors is calculated, which is denoted as $\overline{\text{VIM}}$. Finally, the statistical molecular descriptors were sorted according to $\overline{\text{VIM}}$, and the top 20 of the average importance coefficients were selected as the most significant molecular descriptors. The occurrence of 10 experimental molecular variable symbols is shown in Table 1. It can be seen from Table 1 that the occurrence frequency of 27 variables is sorted, and the importance coefficient of the higher the theoretical occurrence frequency is relatively large. Through statistics of the average importance coefficients of these 27 variables, the ranking of the average importance coefficients in 10 experiments can be obtained, as shown in Figure 3. According to Figure 3, these 20 molecular descriptors can be derived to describe the biological activity of the compound as best as possible.

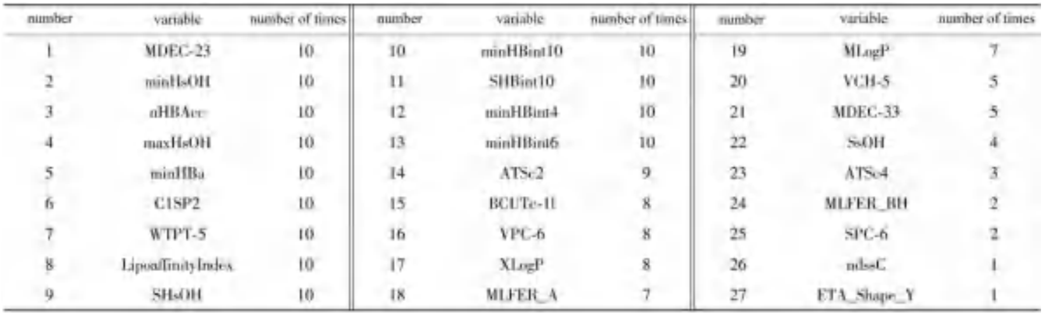
4 Predictive analyses of QSAR model optimized by BP neural network based on PSO

After dimensionality reduction of molecular descriptor data, the amount of data is greatly reduced. Since the BP neural network model has strong adaptability, generalization and fault tolerance, and can approximate any linear continuous function through data, this feature is consistent with the characteristics of the molecular descriptor data properties on drug candidates. Therefore, BP neural network can be selected for training and learning, and IC_{50} values and pIC_{50} values of 50 compounds can be predicted. In this section, the biological activity value prediction method based on BP neural network is analyzed, and the particle swarm optimization (PSO) with fast running speed and good global optimization ability is introduced to avoid the problem that traditional BP neural network is easy to fall into the local optimal solution.

4.1 BP neural network biological activity value prediction model

The neural network, which includes input layer, hidden layer and output layer, is used for training and prediction. As shown in Figure 4, the input data is set to the 20 molecular descriptors selected above, that is, the number of neuron nodes in the input layer is 20, and the number of neuron nodes in the output layer is 1[12]. The number of neuron nodes in the hidden layer can be determined according to empirical formula (1), which ranges from 4 to 14. In this section, the number of neuron nodes in the hidden layer is set as 10:

Table 1: Number of occurrences of the 27 variables in 10 experiments



number	variable	number of times	number	variable	number of times	number	variable	number of times
1	MDEC-23	10	10	minHBint10	10	19	MLogP	7
2	minHsOH	10	11	SHBint10	10	20	VCH-5	5
3	nHBAcc	10	12	minHBint4	10	21	MDEC-33	5
4	maxHsOH	10	13	minHBint6	10	22	SsOH	4
5	minHBa	10	14	ATSc2	9	23	ATSc4	3
6	C1SP2	10	15	BCUTc-1l	8	24	MLFER_BH	2
7	WTPT-5	10	16	VPC-6	8	25	SPC-6	2
8	LipoaffinityIndex	10	17	XLogP	8	26	ndssC	1
9	SHsOH	10	18	MLFER_A	7	27	ETA_Shape_Y	1

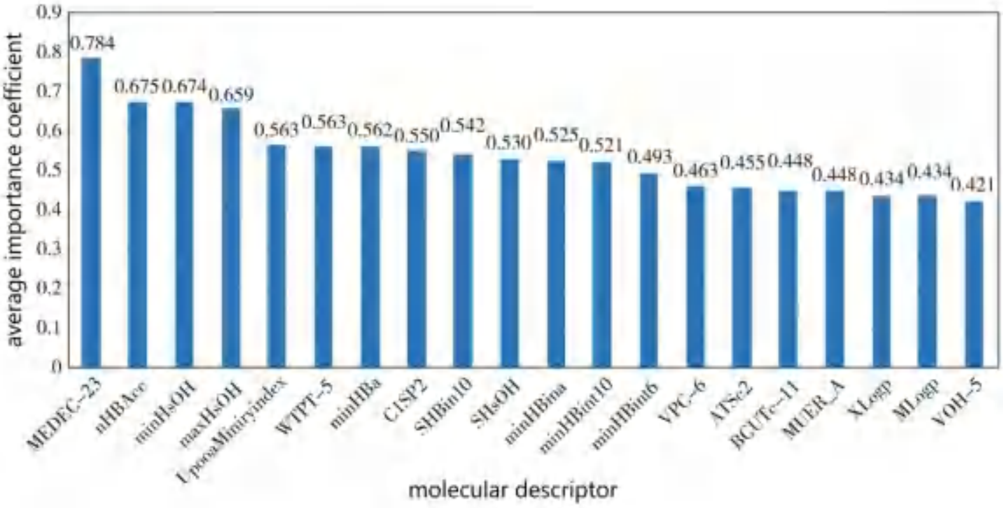


Figure 3: Average importance coefficients of 20 molecular descriptors

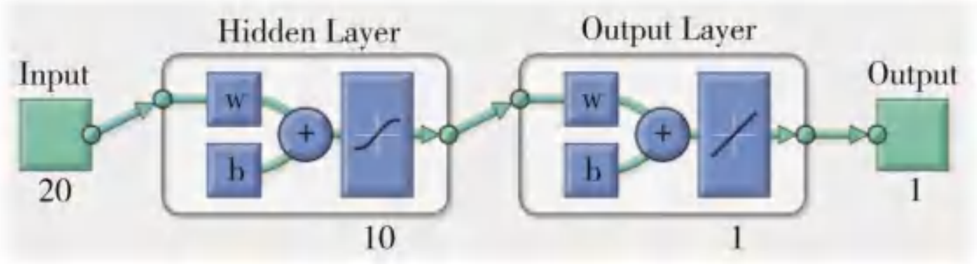


Figure 4: Three-layer BP neural network structure

$$q = \sqrt{k + l} + a, \tag{1}$$

in formula (1), q is the number of hidden layer neurons; k is the number of neurons in the input layer; l is the number of neurons in the output layer; a is a fixed constant value, ranging from 0 to 10 [13].

In the BP neural network, the activation function of the hidden layer is sigmoid, and the activation function of the output layer is relu, which is expressed by functions (2) and (3):

$$\text{sigmoid}\ (z) = \frac{1}{1+e^{-z}}, \tag{2}$$

$$\mathrm{relu}\ (z) = \begin{cases} z, z > 0, \\ 0, z \leq 0. \end{cases} \tag{3}$$

the activation function output of the j-TH neuron in layer l is represented by $S_j^{[l]}$. ω_{jk}^l represents the connection weight from k neurons in layer $(l-1)$ of the network to the j-TH neuron in layer l [14]. The activation function is denoted by σ.

The calculation formula from the input layer to the hidden layer is

$$S_j^l = \sigma\left(\sum_{p=1}^{P} \omega_{pl} x_p + b_1\right), p = 1,2,\cdots,P; l = 1,2,\cdots,L. \tag{4}$$

The calculation formula from the hidden layer to the output layer is

$$S_m = \sigma\left(\sum_{l=1}^{L} \omega_{lm} f_1\left(S_j^l\right) + b_2\right), l = 1,2,\cdots,L; m = 1,2,\cdots,M. \tag{5}$$

in formula (4) and (5), b_1 and b_2 are threshold values; ω_{pl} and ω_{lm} are connection weights; hidden layer output is $f_1(S_l)$, f_1 is relu activation function; the output result of the output layer is $f_2(S_m)$, and f_2 is the output function of the output layer.

4.2 Analysis of BP neural network solution results

Based on the traditional BP neural network model, 1 974 sample data were divided into a training set and a test set according to the ratio of 8:2. The training set was used to train the model, and then the trained model was used to verify the effect on the test set. The training regression results are shown in Figure 5.

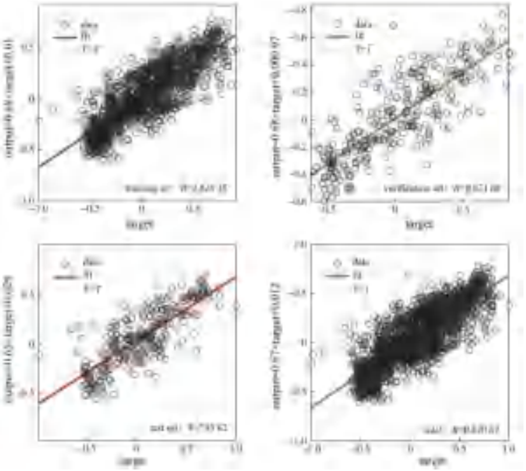


Figure 5: Regression results of BP neural network training

By observation, the fit degree of this model is 0.820 62, and the training and testing data are relatively concentrated. The error of test prediction results is shown in Figure 6 and Figure 7. As can be seen from Figure 6, the 50 test sets selected have some fluctuations in the prediction, and there are some large individual errors, but they are mainly concentrated in the range of 0.1~0.3, and the average test error is 21. 671 5%. As can be seen from Figure 7, there is an error between the pIC_{50} value predicted by 50 test sets and the actual test value. The root mean square error RMSE is 1.416 4, and the coefficient of determination R^2 is 0.466 69. It can be found that although BP neural network model prediction can predict a certain pIC_{50} value, it is not accurate, and relevant algorithms should be used to optimize the model to reduce the error. Since the particle swarm optimization algorithm (PSO) does not depend on the problem information and uses real numbers to solve it, the algorithm has strong generality [15], is easy to implement and has fast convergence speed. Therefore, based on the pursuit of small error, the BP neural network model can be optimized for prediction.

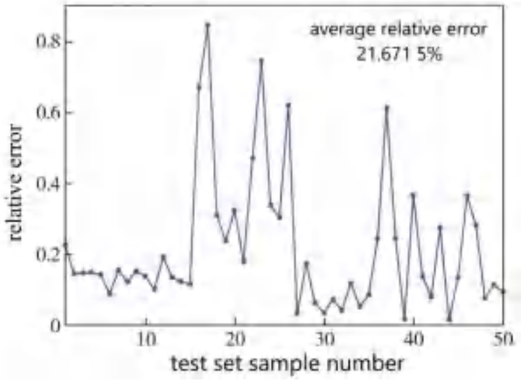


Figure 6: Relative error between predicted values and real values

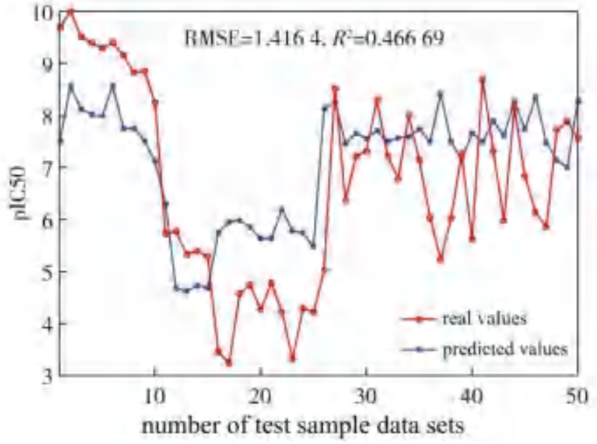


Figure 7: Comparison between predicted values and real values of test set

4.3 PSO optimized BP neural network model

BP neural network will fall into local optimal solution due to unreasonable selection of initial threshold and weight. At the same time, if a lot of training is required, it is very easy to cause overfitting, which will affect the generalization ability to a certain extent. In view of the shortcomings of BP neural network, genetic algorithm or particle swarm optimization can be considered to optimize the network. In this paper, considering that PSO algorithm adopts real number coding, which runs faster than genetic algorithm using binary coding, the mutation idea of genetic algorithm can be used to increase mutation operators and dynamically adjust learning factors to improve the shortcomings [16]. Avoid falling into local optimality and ensure population diversity. The flow of optimized BP neural network algorithm with PSO is shown in Figure 8.

When updating the speed and position of the particle, the position and speed of the particle can be adjusted according to equation (6):

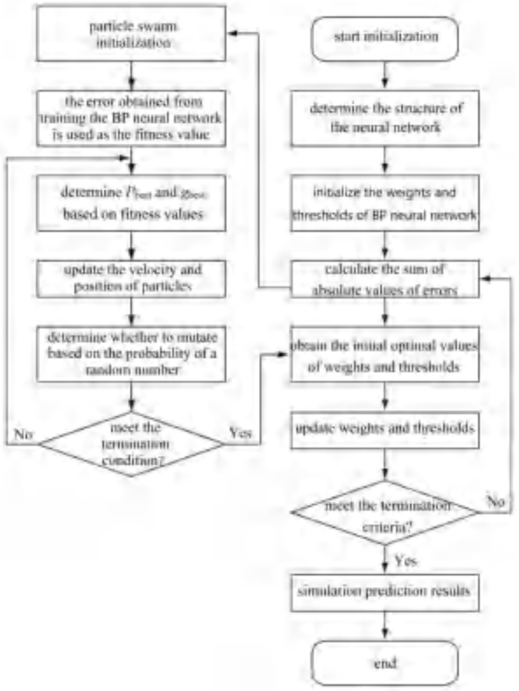

particle swarm initialization
the error obtained from training the BP neural network is used as the fitness value
determine P_{best} and g_{best} based on fitness values
update the velocity and position of particles
determine whether to mutate based on the probability of a random number
meet the termination condition?
No
Yes
start initialization
determine the structure of the neural network
initialize the weights and thresholds of BP neural network
calculate the sum of absolute values of errors
obtain the initial optimal values of weights and thresholds
update weights and thresholds
meet the termination criteria?
No
Yes
simulation prediction results
end

Figure 8: PSO optimized BP neural network algorithm flow

$$\begin{cases} V_i^{k+1} = \varepsilon V_i^k + c_1(j) \times r_1 \times \left(P_{best} - X_l^k\right) + c_2(j) \times r_2 \times \left(g_{best} - X_i^k\right), \\ X_i^{k+1} = X_i^k + V_i^{k+1}, \\ c_2(j) = c_{\max} - \left(c_{\max} - c_{\min}\right) \times \dfrac{\left(i_{t\max} - j\right)}{i_{t\max}}, \\ c_1(j) = 4 - c_2(j). \end{cases} \tag{6}$$

in formula (6), $c_1(j)$ and $c_2(j)$ represent the learning factors generated by the j-TH iteration; j represents the number of iterations; ε represents the weight coefficient; r_1 and r_2 are random functions.

4.4 Analysis of prediction results of biological activity by BP neural network optimized by PSO

A quantitative prediction model was established based on the optimized BP neural network algorithm based on PSO. The 1974 sample data were also randomly divided into 80% training set and 20% test set. The model was trained with the training set and tested with the test set. The prediction results are shown in Figure 9. The goodness of fit of the training set and the test set are 0.862 77 and 0.745 85 respectively, and the overall goodness of fit of the prediction model is 0.833 7, which is better than that of the BP neural network before optimization.

The test prediction error of BP neural network algorithm optimized by PSO is shown in Figure 10 and Figure 11. It can be seen from Figure 10 that the average relative error of the prediction of the test set sample is 9.491 3%, and the prediction accuracy has been improved, and the data of the test set is relatively concentrated. Figure 11 shows that the mean root square error RMSE is 0.731 5, and the coefficient of determination R^2 is 0.740 92. Compared with the BP neural network prediction results before optimization, RMSE decreased and R^2 increased, indicating that the biological activity value data predicted by the optimized network was closer to the real value. Through fitting degree and error analysis, it was demonstrated that the overall effect of the BP neural network model optimized by PSO was better.

The bioactivity values of 50 compounds were predicted by the quantitative prediction model of ERα bioactivity established above. In the data set, the unit of IC_{50} value is nmol/L, so the negative logarithm cannot be taken directly with the IC_{50} value, and the negative logarithm should be multiplied by 10 to the -9 power, so the relationship between IC_{50} and pIC_{50} is $\mathrm{IC}_{50} = 10^{(9-\mathrm{pIC}_{50})}$, and pIC_{50} is the conversion value of IC_{50}, and there is no unit. From this, the predicted values before and after model optimization can be obtained. However, after comparison, only IC_{50} values and corresponding pIC_{50} values predicted by the BP neural network model optimized by PSO are finally selected, as shown in Table 2.

5 Analysis of ADMET property prediction model based on PSO-optimized SVM

5.1 Compound ADMET property analysis and prediction model construction

In order for a compound to be a drug candidate, in addition to having good biological activity (that is, anti-breast cancer activity), it also needs to have good pharmacokinetic properties and safety in human beings. They are collectively referred to as the properties of ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) [17]. Among them, ADME mainly refers to the pharmacokinetic properties of the compound, describing the law of the concentration of the compound in the biological body with time, and T mainly refers to the toxic side effects that the compound may produce in the human body. No matter how good the activity of a compound is, if its ADMET properties are not good, such as being difficult to be absorbed by the body, or being metabolized too quickly in the body, or having some toxicity, it is still difficult to become a drug, so ADMET properties need to be optimized. Due to the complexity of modeling optimization, only five ADMET properties of the compound were considered in this paper, which were :1) small intestinal epithelial cell permeability (Caco-2), which can measure the absorption ability of the compound; 2) cytochrome P450 (CYP) 3A4 subtype (CYP3A4), which is the main metabolic enzyme in the human body and can measure the metabolic stability of compounds; 3) compound cardiac safety evaluation (human Ether-a-go-go Related Gene, hERG), which can measure the cardiac toxicity of the compound; 4) human Oral Bioavailability (HOB), which can measure the proportion of drug amount absorbed into the human blood circulation after entering the human body; 5) micronucleus (MN) is a method for detecting whether a compound has genotoxicity [18]. For the convenience of discussion, this paper uses the binary classification method to provide corresponding values for the properties of ADMET. For example, for Caco-2, "1" means that the small intestinal epithelial cells of the compound have good permeability, and "0" means that the small intestinal epithelial cells of the compound have poor permeability. The binary classification of the other four can be followed.

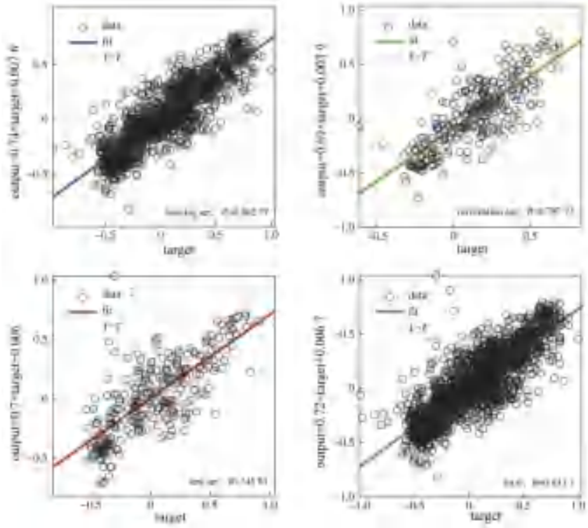


Figure 9: Regression results of the PSO optimized BP neural network training

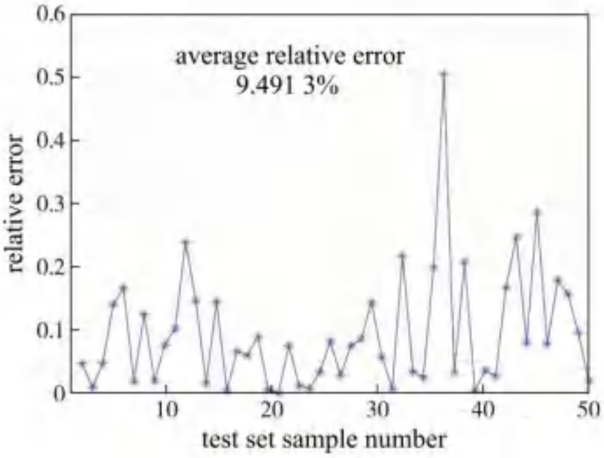


Figure 10: Relative error between predicted values and real values of the PSO optimized BP neural network

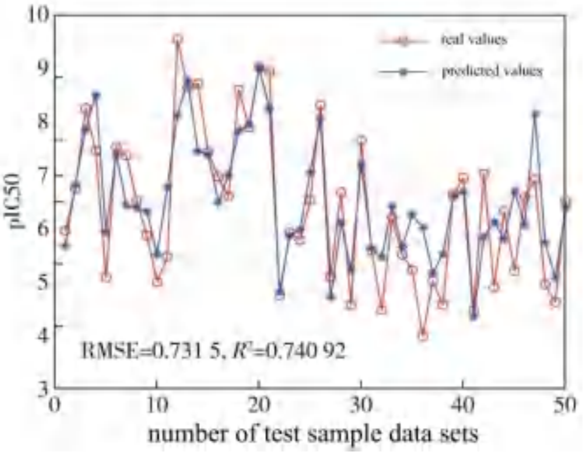
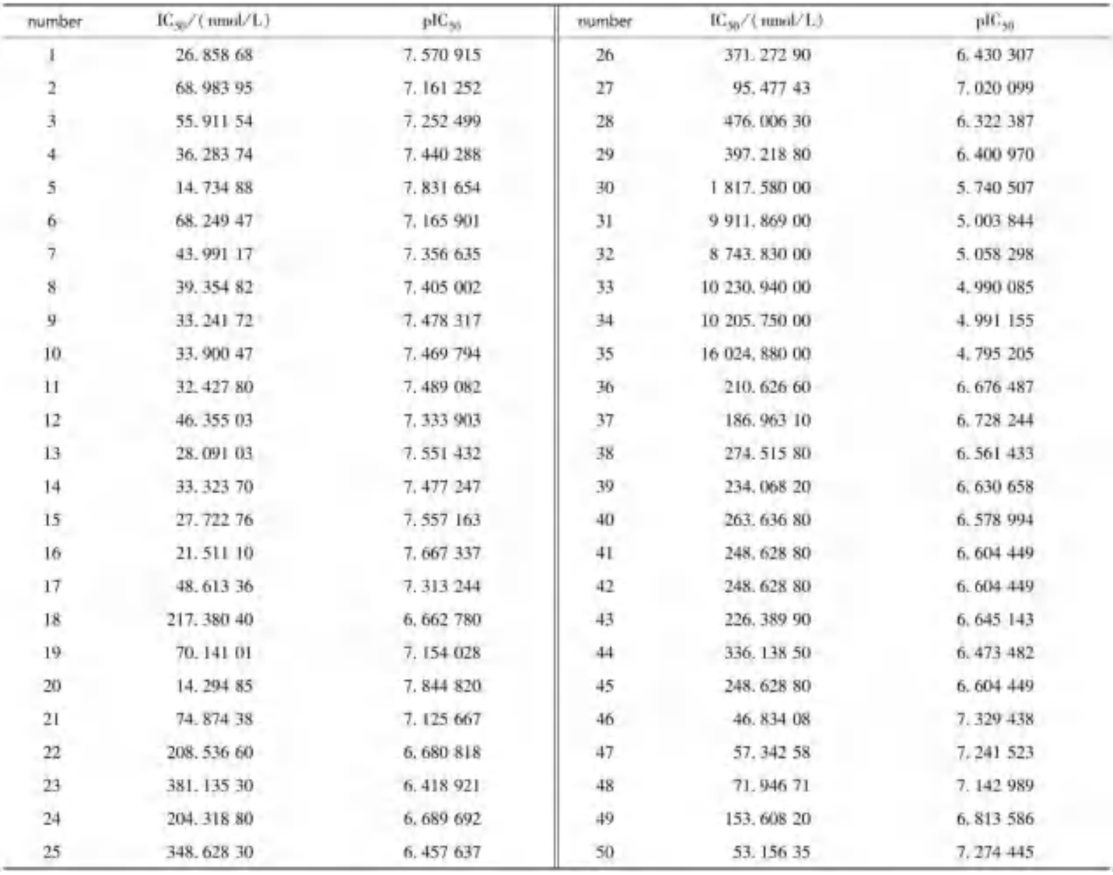


Figure 11: Comparison between predicted values and real values of the PSO optimized BP neural network on test set

Due to the limited sample size of the collected ADMET property data, and the characteristics of nonlinear and dimensional, it is easy to be affected by complex factors such as the operating environment during the collection process, which makes the data highly noisy and prone to missing and error. Therefore, when selecting the data mining algorithm for analysis and prediction, the applicability of the algorithm should be considered. Through comparison and analysis of several commonly used algorithms, it is found that naive Bayes algorithm is very sensitive to the expression form of input data, has a certain error rate in classification decision, has low training efficiency and complex operation framework, and is not suitable for ADMET property prediction of compounds. The performance of decision tree algorithm is general when dealing with data with strong feature correlation, and it is easy to overfit. The final decision function of SVM algorithm is determined only by a small number of support vectors, and the complexity of calculation depends on the number of support vectors rather than the dimension of sample space, avoiding the "dimensional disaster". Moreover, it is highly interpretable for nonlinear classification tasks, and can find crucial key samples, and the algorithm has high fitting accuracy and good robustness. It has strong applicability to predict the properties of ADMET compounds. Therefore, support vector machine is used to establish the 0-1 binary classification model of 5 ADMET properties. However, the selection of parameter kernel g and penalty parameter c will limit the further development of this method. According to the existing research, there is no good, recognized and fixed parameter selection method up to now. Generally speaking, the empirical estimation method is the most commonly used method in the research, but this method is relatively random in the selection of parameters, resulting in greater limitations. Particle swarm optimization (PSO) has significant advantages in the process of parameter optimization, and the model structure of this method is relatively simple [19]. Therefore, particle swarm optimization (PSO) will be used in this paper to optimize the parameters of support vector machine (SVM), and the operation flow of the algorithm is shown in Figure 12.

Table 2: Prediction result of IC_{50} values and corresponding pIC_{50} values



number	IC_{50}/(nmol/L)	pIC_{50}	number	IC_{50}/(nmol/L)	pIC_{50}
1	26. 858 68	7. 570 915	26	371. 272 90	6. 430 307
2	68. 983 95	7. 161 252	27	95. 477 43	7. 020 099
3	55. 911 54	7. 252 499	28	476. 006 30	6. 322 387
4	36. 283 74	7. 440 288	29	397. 218 80	6. 400 970
5	14. 734 88	7. 831 654	30	1 817. 580 00	5. 740 507
6	68. 249 47	7. 165 901	31	9 911. 869 00	5. 003 844
7	43. 991 17	7. 356 635	32	8 743. 830 00	5. 058 298
8	39. 354 82	7. 405 002	33	10 230. 940 00	4. 990 085
9	33. 241 72	7. 478 317	34	10 205. 750 00	4. 991 155
10	33. 900 47	7. 469 794	35	16 024. 880 00	4. 795 205
11	32. 427 80	7. 489 082	36	210. 626 60	6. 676 487
12	46. 355 03	7. 333 903	37	186. 963 10	6. 728 244
13	28. 091 03	7. 551 432	38	274. 515 80	6. 561 433
14	33. 323 70	7. 477 247	39	234. 068 20	6. 630 658
15	27. 722 76	7. 557 163	40	263. 636 80	6. 578 994
16	21. 511 10	7. 667 337	41	248. 628 80	6. 604 449
17	48. 613 36	7. 313 244	42	248. 628 80	6. 604 449
18	217. 380 40	6. 662 780	43	226. 389 90	6. 645 143
19	70. 141 01	7. 154 028	44	336. 138 50	6. 473 482
20	14. 294 85	7. 844 820	45	248. 628 80	6. 604 449
21	74. 874 38	7. 125 667	46	46. 834 08	7. 329 438
22	208. 536 60	6. 680 818	47	57. 342 58	7. 241 523
23	381. 135 30	6. 418 921	48	71. 946 71	7. 142 989
24	204. 318 80	6. 689 692	49	153. 608 20	6. 813 586
25	348. 628 30	6. 457 637	50	53. 156 35	7. 274 445

When the PSO-optimized SVM method is used to calculate the fitness value of each particle, the fitness function takes the mean square error (MSE), as shown in equation (7):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{7}$$

in formula (7), y_i is the actual value; $\hat{y}_i$ is the predicted value; n is the number of samples trained.

5.2 Analysis of classification prediction results based on PSO-based SVM optimization

Based on the above PSO-optimized SVM algorithm, the ADMET prediction model of the compound was constructed. The prediction analysis was carried out on the five indexes respectively, and the output variable indexes were set up successively as Caco-2, CYP3A4, hERG, HOB, MN, and then substituted into the MATLAB software for running.

5.2.1 Prediction of Caco-2 permeability of small intestinal epithelial cells of compounds

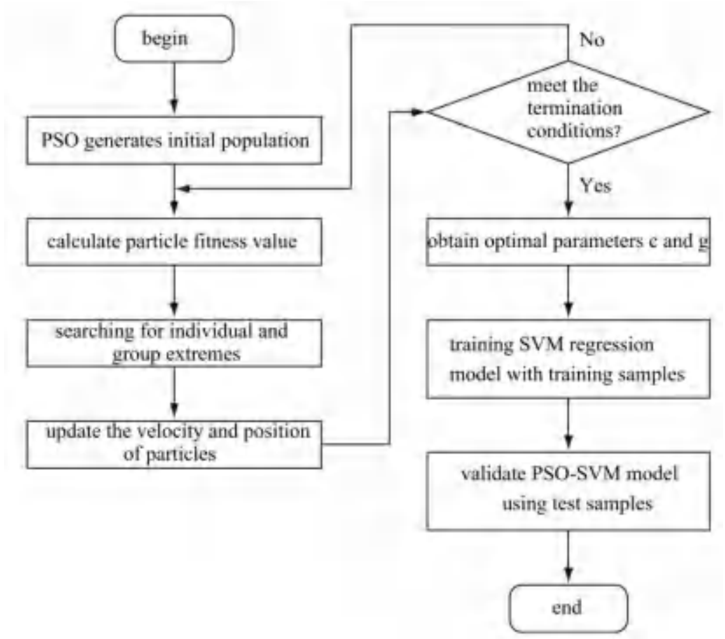

begin
PSO generates initial population
calculate particle fitness value
searching for individual and group extremes
update the velocity and position of particles
No
meet the termination conditions?
Yes
obtain optimal parameters c and g
training SVM regression model with training samples
validate PSO-SVM model using test samples
end

Figure 12: Flow chart of PSO optimized SVM algorithm

As for the prediction of Caco-2 index, Figure 13 shows the iterative process of Caco-2 optimized by PSO SVM, after which the optimized penalty parameter c= 268.757 6 and the kernel parameter g=0.001 can be obtained, and the accuracy of cross-validation CV reaches 94.076 7%, which is of good accuracy and has certain reference value for Caco-2 index prediction. Figure 14 shows the confusion matrix of 574 test data, among which the actual sample classification values and model predicted classification values of 396 compounds are "0", and the actual sample classification values and model predicted classification values of 117 compounds are "1". The accuracy of the confusion matrix is 80.7%, the recall rate is 78.0%, and the specificity is 93.4%. Figure 15 shows the comparison between actual classification and predicted classification after SVM optimization by PSO [20]. For 574 test data, the real value of Caco-2 and predicted value are mostly consistent, and the prediction accuracy is 89.372 8%.

5.2.2 Predictive analysis of CYP3A4 metabolic stability of compounds

For the prediction of index CYP3A4, Figure 16 shows its iterative process. After optimization, the penalty parameter c= 549.464 9 and the kernel parameter g=0.001, and the accuracy rate of CV in the cross-validation iteration process is 97.735 2%, which has a good accuracy. Figure 17 shows the confusion matrix of 574 test data, among which 59 compounds have the actual sample classification value and the model predicted classification value of "0", and 481 compounds have the actual sample classification value and the model predicted classification value of "1". The accuracy of the confusion matrix is 97.0%, the recall rate is 96.2%, and the specificity is 79.7%. Figure 18 shows the actual classification and predicted classification results of the test set when predicting CYP3A4 index [21]. The actual classification and predicted classification of the test set are also relatively high, with a prediction accuracy of 94.076 7%.

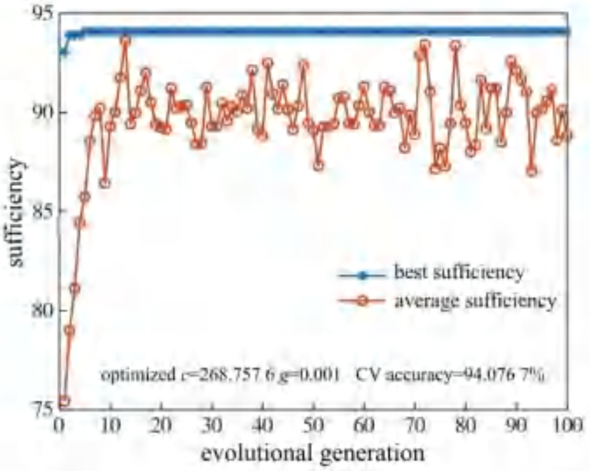


Figure 13: PSO optimizing the Caco-2 iterative process of SVM

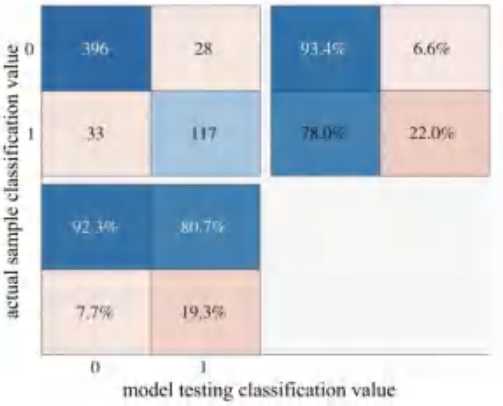


Figure 14: Confusion matrix of Caco-2 test group data

5.2.3 hERG prediction analysis of cardiotoxicity of compounds

For the prediction of indicator hERG, Figure 19 shows its iterative process. After optimization, the penalty parameter c= 891.311 9 and the kernel parameter g=0.001, and the accuracy of CV in the cross-validation iteration process is 89.198 6%, with average accuracy. Figure 20 shows the confusion matrix of 574 test data, among which 93 compounds have the actual sample classification value and the model predicted classification value of "0", and 390 compounds have the actual sample classification value and the model predicted classification value of "1". The accuracy of the confusion matrix is 84.4%, the recall rate is 95.4%, and the specificity is 56.4%. Figure 21 shows the actual classification and prediction classification results of the test set when predicting hERG index [22]. The actual classification and prediction classification of the test set are also relatively high, and the prediction accuracy is 84.146 3%.

5.2.4 HOB prediction analysis of compounds

For the prediction of indicator HOB, Figure 22 shows its iterative process. After optimization, the penalty coefficient c= 119.618 4 and the kernel parameter g=0.001, and the accuracy of CV in the cross-validation iteration process is 87.971 9%, with average accuracy. Figure 23 shows the confusion matrix of 574 test data, among which the actual sample classification values and model prediction classification values of 394 compounds are "0", and the actual sample classification values and model prediction classification values of 60 compounds are "1". The accuracy of the confusion matrix is 50%, the recall rate is 50%, and the specificity is 86.8%. Figure 24 shows the actual classification and prediction classification results of the test set when predicting HOB index [23]. The actual classification and prediction classification of the test set are also relatively high, and the prediction accuracy is 79.094 1%.

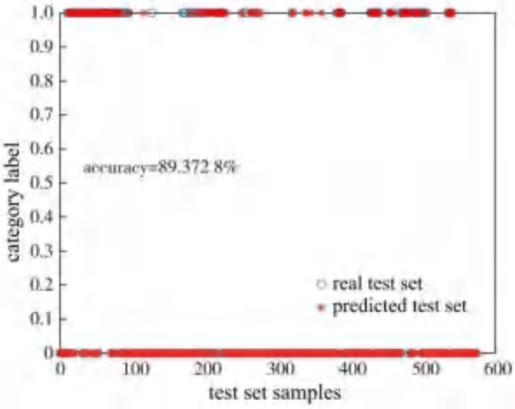


Figure 15: The actual classification and predicted classification for the test set when predicting the Caco-2 indicator

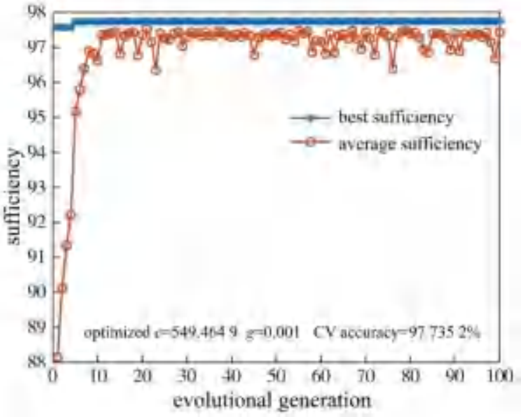


Figure 16: PSO optimizing the CYP3A4 iterative process of SVM

5.2.5 MN prediction analysis of genotoxicity of compounds

For the prediction of index MN, Figure 25 shows its iterative process. After optimization, the penalty coefficient c=63.284 6 and the kernel parameter g=0.001, and the accuracy of CV in the cross-validation iteration process is 92.508 7%, with average accuracy. Figure 26 shows the confusion matrix of 574 test data, among which 104 compounds have the actual sample classification value and the model predicted classification value of "0", and 381 compounds have the actual sample classification value and the model predicted classification value of "1". The accuracy of the confusion matrix is 86.4%, the recall rate is 92.9%, and the specificity is 63.4%. Figure 27 shows the actual classification of the test set and the predicted classification results when predicting the MN index [24]. The actual classification and prediction classification of the test set were also relatively high, with a prediction accuracy of 84.494 8%.

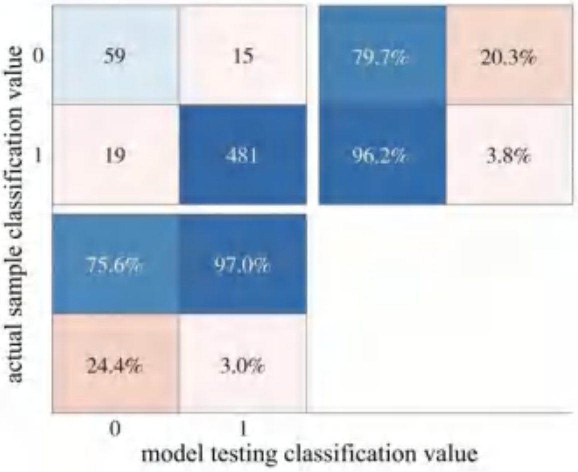


Figure 17: Confusion matrix of CYP3A4 test group data

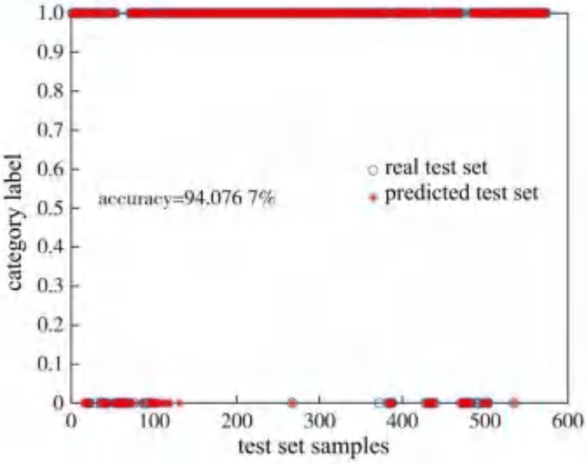


Figure 18: The actual classification and predicted classification for the test set when predicting CYP3A4 indicators

According to the classification prediction models Caco-2, CYP3A4, hERG, HOB and MN built above, due to the relatively high prediction accuracy of the models, the ADMET properties of 50 new compounds can be predicted according to the structural formula of compound molecules, so as to judge the quality of the new compounds. It provides certain reference value for the judgment of drug properties, and the prediction results are shown in Table 3.

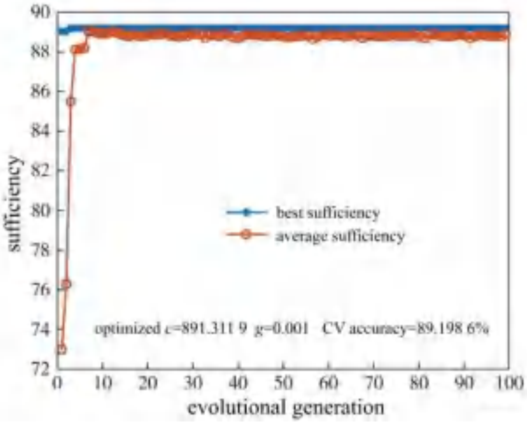


Figure 19: PSO optimizing the hERG iterative process of SVM

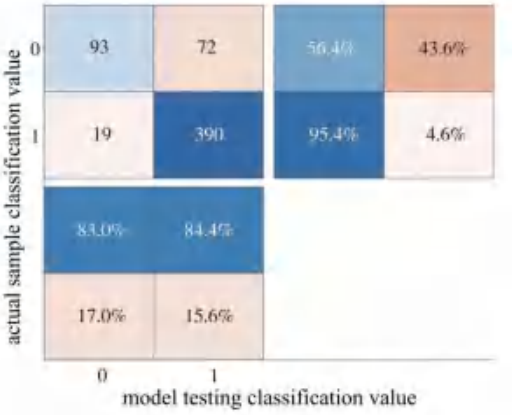


Figure 20: Confusion matrix of hERG test group data

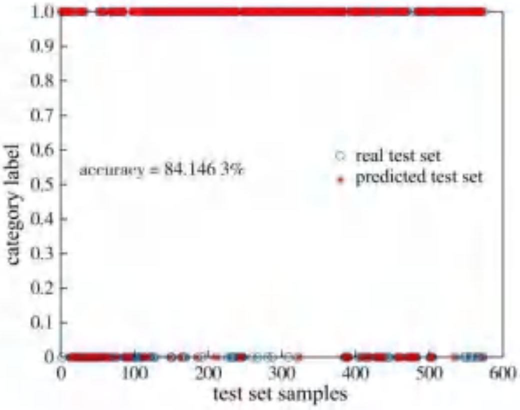


Figure 21: The actual classification and predicted classification for the test set when predicting hERG indicators

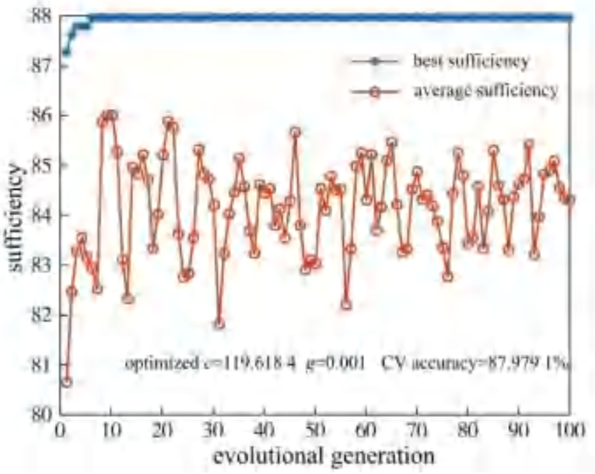


Figure 22: PSO optimizing the HOB iterative process of SVM

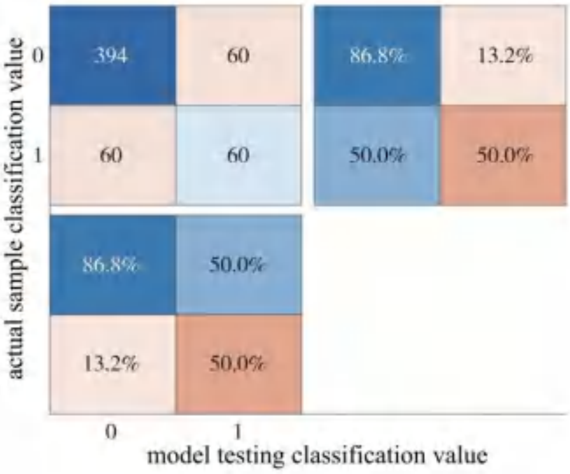


Figure 23: Confusion matrix of HOB test group data

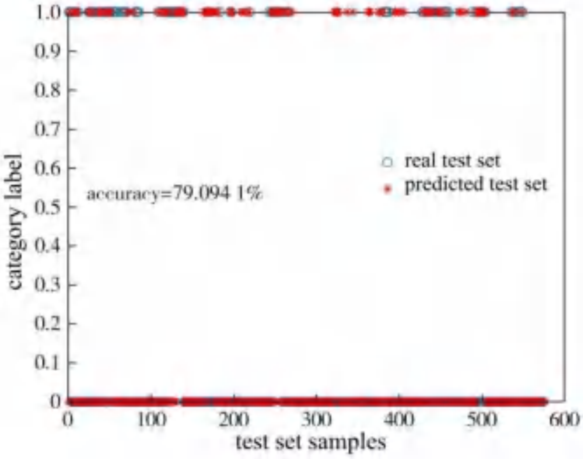


Figure 24: The actual classification and predicted classification for the test set when predicting the HOB index

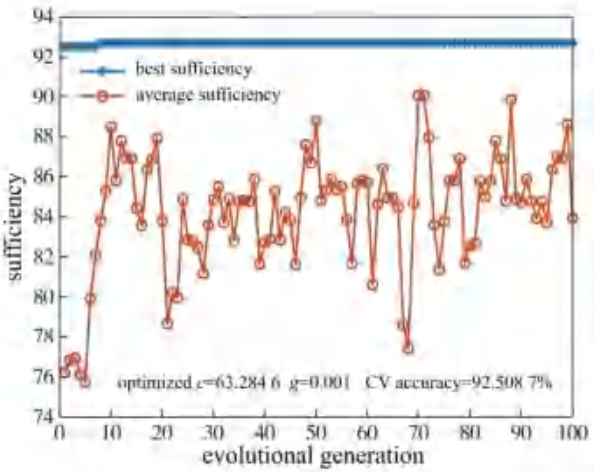


Figure 25: PSO optimizing the MN iterative process of SVM

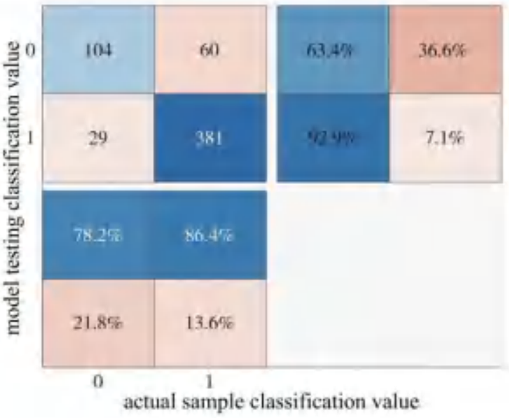


Figure 26: Confusion matrix of MN test group data

Table 3: ADMET property prediction results of 50 compounds

number	Caco-2	CYP3A4	hERG	HOB	MN	number	Caco-2	CYP3A4	hERG	HOB	MN
1	0	1	1	0	1	26	1	1	1	0	0
2	0	1	1	0	0	27	0	1	1	0	0
3	0	1	1	0	0	28	0	1	1	0	1
4	0	1	1	0	0	29	0	1	1	0	1
5	0	0	1	0	0	30	0	1	1	1	1
6	0	1	1	0	0	31	1	1	1	1	1
7	0	1	1	0	0	32	1	1	1	1	1
8	0	1	1	0	0	33	1	1	1	1	1
9	0	1	1	0	0	34	1	1	1	1	1
10	0	1	1	0	1	35	0	1	1	1	1
11	0	1	1	0	1	36	0	1	1	0	0
12	0	1	1	0	1	37	0	1	1	0	0
13	0	1	1	0	1	38	0	1	1	0	0
14	0	1	1	0	1	39	0	1	0	1	1
15	0	1	1	0	1	40	0	1	1	1	1
16	0	1	1	0	1	41	0	1	1	1	1
17	0	1	1	0	0	42	0	1	1	1	1
18	0	1	0	0	0	43	0	1	0	1	1
19	1	1	1	0	0	44	0	1	1	1	1
20	0	0	1	0	0	45	0	1	1	1	1
21	0	1	1	0	0	46	0	1	1	0	1
22	0	1	1	0	0	47	0	1	1	0	1
23	1	0	1	0	0	48	0	1	1	0	1
24	1	0	1	0	0	49	0	1	1	0	1
25	1	1	1	0	1	50	0	1	1	0	0

6 Conclusions

In order to predict the biological activity and ADMET properties of anti-breast cancer drug candidates, a computer-aided method was chosen in this paper. From the perspective of "characteristic importance analysis" of compounds, the random forest classifier was used to evaluate the importance of 1 974 compounds, so as to reorder the importance of molecular descriptors on biological activities, and select the top 20 molecular descriptors that have the most significant impact on biological activities. Secondly, particle swarm optimization BP neural network was used to construct a quantitative prediction model to obtain the IC_{50} and pIC_{50} values of 50 compounds, and the model fit was 0.833 7. Compared with the BP neural network before optimization, the RMSE value was reduced and R^2 was increased, and the predicted biological activity value after optimization was closer to the real value. Furthermore, a classification prediction model of five indexes of compound ADMET properties, Caco-2, CYP3A4, hERG, HOB and MN, was constructed by particle swarm optimization support vector machine algorithm, and cross-verified CV accuracy reached 94.076 7% through training and testing. The prediction accuracy of the five indexes was 89.372 8%, 94.067 7%, 84.146 3%, 79.094 1% and 84.494 8%, respectively, and the ADMET binary classification values of 50 compounds were obtained.

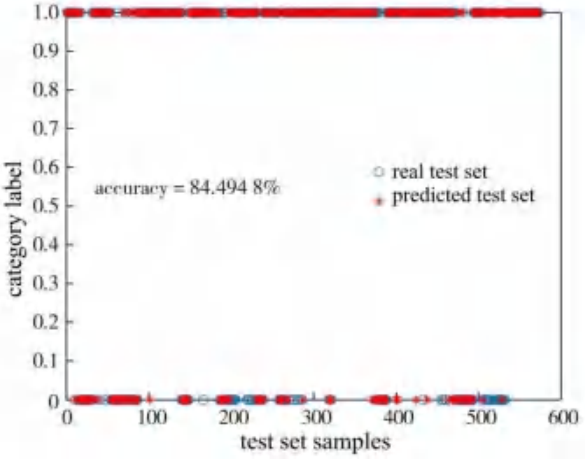


Figure 27: The actual classification and predicted classification for the test set when predicting the MN index

The results show that the prediction model constructed in this paper is better than the benchmark model, which verifies the applicability of the model. The predictive analysis of compound molecular descriptors can provide an effective reference in the development of anti-breast cancer drug candidates, and the established model can also be extended to solve other practical problems such as data analysis and prediction and multi-objective optimization. It has a certain guiding role in the prevention and treatment of human life and health research such as breast cancer, leukemia, cervical cancer or other tumor diseases [25].

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 71701099, 71501090) and Natural Science Research Project of Higher Education Institutions of Jiangsu Province (Grant No. 17KJB580008).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] H. C. S. Chan, H. B. Shan, T. Dahoun, et al., Advancing drug discovery via artificial intelligence, *Trends Pharmacol. Sci.*, 2019, 40(8): 592-604.

[2] C. Shen, J. J. Ding, Z. Wang, et al., From machine learning to deep learning: advances in scoring functions for protein-ligand docking, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2020, 10(1): e1429.

[3] Y. W. Gu, B. W. Zhang, S. Zheng, et al., Predicting drug ADMET properties based on graph attention network, *Data Anal. Knowl. Discov.*, 2021, 5(8): 76-85.

[4] L. X. Xie, F. Li, J. P. Xie, et al., Predicting drug molecular properties based on ensembling neural networks models, *Computer Sci.*, 2021, 48(9): 251-256.

[5] J. Qin, Research on matrix completion with side information for better modeling bioactivates of drug leads, *Nanjing: Nanjing University of Posts and Telecommunications*, 2020.

[6] C. M. Jia, Study on drug target recognition and activity prediction model based on molecular vibration characteristics, *Beijing: Beijing University of Chinese Medicine*, 2019.

[7] J. Shen, Development of drug ADMET theory prediction method and drug design research targeting estrogen receptor, *Shanghai: East China University of Science and Technology*, 2011.

[8] J. Wenzel, H. Matter, and F. Schmidt, Predictive multitask deep neural network models for ADME-tox properties: learning from large data sets, *J. Chem. Inf. Model.*, 2019, 59(3): 1253-1268.

[9] T. L. Lei, H. Y. Sun, Y. Kang, et al., ADMET evaluation in drug discovery. 18. Reliable prediction of chemical-induced urinary tract toxicity by boosting machine learning approaches, *Mol. Pharmaceut.*, 2017, 14(11): 3935-3953.

[10] H. Lu, and Y. Q. Zhang. Expression and significance of androgen receptor in estrogen receptor-positive breast cancer, *China J. Modern Med.*, 2021, 31(18): 55-59.

[11] B. B. Cong, and Y. S. Wang, Treatment landscape and challenges of managing the hormone receptor-positive early breast cancer, *China Oncol.*, 2021, 31(8): 689-696.

[12] Z. Q. Wu, B. Ramsundar, E. N. Feinberg, et al., MoleculeNet: a benchmark for molecular machine learning, *Chem. Sci.*, 2017, 9(2): 513-530.

[13] D. J. Yang, X. C. Yao, Z. Y. Xu, et al., Molecular docking of the chemicals of *Illicium lanceolatum* in lowering uric acid and ADMET properties, *Chin. J. Clin. Pharmacol.*, 2018, 34(23): 2750-2752, 2777.

[14] C. F. Zhang, H. T. Xie, and G. Y. Pan, Absorption, distribution, metabolism, excretion and toxicity of biologics and its application in pharmacokinetic modeling, *Acta Pharm. Sinica*, 2016, 51(8): 1202-1208.

[15] K. Mansouri, N. F. Cariello, A. Korotcov, et al., Open-source QSAR models for pKa prediction using multiple machine learning approaches, *J. Cheminformatics*, 2019, 11(1): 60.

[16] X. Chen, Studies on a few key problems of QSAR/QSPR modeling based on the OECD principles, *Changsha: Central South University*, 2013.

[17] P. A. Shar, W. Y. Tao, S. Gao, et al., Pred-binding: large-scale protein-ligand binding affinity prediction, *J. Enzyme Inhib. Med. Chem.*, 2016, 31(6): 1443-1450.

[18] M. Y. Su, H. S. Liu, H. X. Lin, et al., Machine-learning model for predicting the rate constant of protein-ligand dissociation, *Acta Phys. Chim. Sin.*, 2020, 36(1): 179-187.

[19] G. H. Liu, J. Hu, and D. J. Yu, Predicting GPCR-drug interactions with multi-view feature combination and random forest, *J. Nanjing Univ. Sci. Technol.*, 2016, 40(1): 1-9.

[20] X. Q. Li, M. Mo, F. Wu, et al., Artificial neural network models based on questionnaire survey for prediction of breast cancer risk among Chinese women in Shanghai, *Tumor*, 2018, 38(9): 883-893.

[21] Y. Q. Liu, C. Wang, and L. Zhang, Neural network based models for predicting breast cancer survivability, *Chin. J. Biomed. Eng.*, 2009, 28(2): 221-225.

[22] Q. Min, J. Liao, and T. Lu, Drug-drug interaction predicting model based on large scale drug databases, *Chin. J. Clin. Pharmacol.*, 2016, 32(11): 1034-1036.

[23] J. T. Tang, Y. Cao, J. Y. Xiao, et al., Remifentanil blood concentration forecast model based on support vector machine with particle swarm optimization, *Chin. Pharm. J.*, 2013, 48(16): 1394-1399.

[24] R. Bai, Q. Z. Teng, X. M. Yang, et al., Prediction of combinative activity of drugs and human serum albumin by using SVM and GA, *Comput. Eng. Appl.*, 2009, 45(12): 226-228, 248.

[25] X. Q. Yuan, Study on SVM prediction model of compound hepatotoxicity based on gene expression data, *Zhenjiang: Jiangsu University*, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Global Science Press and/or the editor(s). Global Science Press and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.