



Article ID 1007-1202(2023)03-0257-14

DOI <https://doi.org/10.1051/wujns/2023283257>

Machine Learning-Based Quantitative Structure-Activity Relationship and ADMET Prediction Models for ER α Activity of Anti-Breast Cancer Drug Candidates

□ XU Zonghuang

School of Information Management, Nanjing University, Nanjing 210023, Jiangsu, China

© Wuhan University 2023

Abstract: Breast cancer is presently one of the most common malignancies worldwide, with a higher fatality rate. In this study, a quantitative structure-activity relationship (QSAR) model of compound biological activity and ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties prediction model were performed using estrogen receptor alpha (ER α) antagonist information collected from compound samples. We first utilized grey relation analysis (GRA) in conjunction with the random forest (RF) algorithm to identify the top 20 molecular descriptor variables that have the greatest influence on biological activity, and then we used Spearman correlation analysis to identify 16 independent variables. Second, a QSAR model of the compound were developed based on BP neural network (BPNN), genetic algorithm optimized BP neural network (GA-BPNN), and support vector regression (SVR). The BPNN, the SVR, and the logistic regression (LR) models were then used to identify and predict the ADMET properties of substances, with the prediction impacts of each model compared and assessed. The results reveal that a SVR model was used in QSAR quantitative prediction, and in the classification prediction of ADMET properties: the SVR model predicts the Caco-2 and hERG(human Ether-a-go-go Related Gene) properties, the LR model predicts the cytochrome P450 enzyme 3A4 subtype (CYP3A4) and Micronucleus (MN) properties, and the BPNN model predicts the Human Oral Bioavailability (HOB) properties. Finally, information entropy theory is used to validate the rationality of variable screening, and sensitivity analysis of the model demonstrates that the constructed model has high accuracy and stability, which can be used as a reference for screening probable active compounds and drug discovery.

Key words: anti-breast cancer; drug discovery; quantitative structure-activity relationship (QSAR) model; ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) prediction; machine learning

CLC number: R 965.1; Q-03

0 Introduction

Breast cancer is currently one of the most common malignancies in the world with a higher mortality rate^[1].

Breast cancer development is intimately linked to estrogen receptors. According to research, estrogen receptor alpha (ER α) is present in less than 10% of normal breast epithelial cells, whereas it is expressed in 50%-80% of

Received date: 2022-11-30

Foundation item: Supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_0082)

Biography: XU Zonghuang, male, Ph.D. candidate, research direction: information science, mathematical modeling, big data analysis, machine learning and national scientific and technological intelligence. E-mail: zonghuangxu@yeah.net

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

breast carcinoma cells^[2-4]. While the experimental results of mice lacking in the ER α gene suggest that ER α does play a significant role in breast development^[5,6]. Antihormonal treatment, which reduces the level of estrogen in the body by controlling the activity of estrogen receptors, is commonly utilized in breast cancer patients with ER α expression now^[7]. ER α is seen as a key target for breast cancer treatment, and chemicals that may inhibit ER α activity may be potential medications for the treatment of breast cancer^[8,9]. Tamoxifen and raloxifene, two of the most often used medications in the therapeutic treatment of breast cancer, are ER α antagonists^[10].

To screen probable active compounds in drug development, the process of building a compound activity prediction model is now employed to save time and money. Because each compound sample contains 729 compound descriptor variables, and not all of them can offer information on the biological activity of ER α , Li *et al*^[11] proposed a technique for selecting representative features based on Pearson correlation analysis and Euclidean distance, and then used this method to screen the representative molecular descriptors in the quantitative structure-activity relationship (QSAR) model. To hunt for characteristic compound descriptors and construct quantitative prediction models for QSAR, several studies used machine learning approaches such as regression analysis^[12-14], decision tree algorithm^[15], artificial neural network^[16-18], support vector machine^[19], and so on. The specific technique is as follows: given a disease-related target (ER α), gather a series of compounds that operate on the target and their biological activity data, and then utilize a series of molecular structure descriptors as independent variables and the biological activity values of the compounds as a dependent variable in the construction of a QSAR model of the compound, which is subsequently used to forecast novel compound molecules with improved biological activity^[20-23].

To become a drug candidate, a compound must have good biological activity (anti-breast cancer activity), as well as good pharmacokinetic properties and safety in the human body, which are referred to collectively as ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties. Among them, ADME mostly relates to the compound's pharmacokinetic features, which define the law of the compound's concentration in the organism over time, and T primarily refers to the hazardous and adverse effects that the chemical may create in the human body. No matter how

active a molecule is, if its ADMET qualities are bad, for example, it is difficult to absorb by the human body, or the metabolism rate in the body is too quick, or it is poisonous. It will be difficult to become a medicine, thus ADMET properties must be adjusted. At present, machine learning techniques based on random forest^[24, 25], naive Bayes^[26], and neural network algorithms^[27, 28] are widely utilized in ADMET properties classification and prediction models, successfully advancing the drug development process^[29-32]. This paper considers the compound's five ADMET properties, which are as follows: 1) Small intestine epithelial cell permeability (Caco-2), which assesses a compound's capacity to be absorbed by the human body. 2) Cytochrome P450 enzyme 3A4 subtype (CYP3A4), which is the main metabolic enzyme in the human body and can measure the metabolic stability of the compound. 3) Compound cardiac safety evaluation (hERG), which can measure the cardiotoxicity of the compound. 4) The Human Oral Bioavailability (HOB), which measures the proportion of a medicine absorbed into the human bloodstream after it enters the human body. 5) The Micronucleus (MN) test determines if a chemical is genetically harmful.

In this study, a QSAR model of compound biological activity and a classification prediction model of ADMET properties were performed for the biological activity of anti-breast cancer drug candidates. Specifically, the main contributions of this paper are as follows:

- Grey correlation analysis (GRA) model combined with random forest (RF) method was used to determine the top 20 molecular descriptor variables that had the greatest influence on biological activity. Moreover, information entropy theory and Spearman correlation analysis were used respectively to verify the rationality of variable selection and the independence of variables. This method can improve the preparation efficiency of new drug compounds to a certain extent.

- Support vector regression (SVR) model was used to construct QSAR prediction model. The prediction effect of the model is better than that of BP neural network (BPNN) model and the BP neural network model optimized by genetic algorithm (GA-BPNN), which is of great value in the research of anti-breast cancer drugs.

- By comparing and evaluating the prediction effect of the models, different biological activities of compounds were predicted based on BPNN model, SVR model and logistic regression (LR) model, which can provide reference for screening possible active com-

pounds and drug discovery.

1 Materials and Methods

1.1 Dataset

The data set for this study was obtained from the 2021 "Huawei Cup" China Postgraduate Mathematical Contest in Modeling^[33]. There are three files in total: ER α _activity.xlsx, Molecular_Description.xlsx and ADMET.xlsx. The dataset in this study provided data on the biological activity of 1 974 compounds against ER α , 729 molecular descriptor information (independent variables), and data on five ADMET properties are supplied for the breast cancer therapy target ER α . The biological activity data of 1 974 chemicals on ER α were supplied for breast cancer therapy target ER α . These data may be found in the "ER α activity.xlsx" file's training table (training set). There are three columns in the training table. The structural formulae of 1 974 compounds are provided in the first column, which is represented by the one-dimensional linear expression SMILES (Simplified Molecular Input Line Entry System). The second column contains the biological activity value of the compound against ER α (represented by IC₅₀, it is an experimental measurement value, the unit is nmol/L. The smaller the value, the greater the biological activity, and the more effective it is to inhibit ER α activity). The third column contains the pIC₅₀ (the negative logarithm of the IC₅₀ value, which is the negative logarithm of the IC₅₀ value in the second column), which is usually positively correlated with biological activity; that is, the higher the pIC₅₀ value, the greater the biological activity. The file also includes a test table (test set) with SMILES formulae for 50 substances.

The 729 molecular descriptor information (independent variables) of the 1 974 compounds are provided in the training table (training set) in the file "Molecular_Descriptor.xlsx". The first column also contains the compound's SMILES formula, and there are 729 columns after that. Each column denotes a molecular description (an independent variable) of the molecule. The molecular descriptor of a compound is a set of parameters used to describe its structure and properties, such as physical and chemical properties, topological structure characteristics, and many more. Similarly, there is a test table in the file that lists the 729 molecular descriptors of the 50 test set compounds.

In addition to considering the compound's biologi-

cal activity, it is also vital to assess its ADMET properties. Data on the ADMET attributes of the 1 974 compounds are supplied in the training table (training set) in the file "ADMET.xlsx". The first column also reflects the compound structure's SMILES formula, and the next 5 columns correspond to the ADMET attributes of each compound, with the two-classification technique utilized to generate the relevant value, as shown below: 1) Caco-2: a value of 1 indicates that the chemical has improved small intestine epithelial cell permeability, while a value of 0 suggests that the compound has decreased intestinal epithelial cell permeability. 2) CYP3A4: A value of 1 indicates that the chemical may be metabolized by CYP3A4, whereas a value of 0 indicates that the molecule cannot be metabolized by CYP3A4. 3) hERG: a value of 1 indicates that the chemical has cardiotoxicity, whereas a value of 0 indicates that the compound does not have cardiotoxicity. 4) HOB: A value of 1 indicates that the chemical has excellent oral bioavailability, whereas a value of 0 indicates that the drug has low oral bioavailability. 5) MN: A value of 1 indicates that the chemical is genotoxic, whereas a value of 0 indicates that the compound is not genotoxic.

1.2 Data Processing

If more than 90% of the compounds have a molecular descriptor variable value of 1, this molecular descriptor variable is thought to have minimal influence and will not affect the biological activity of the ER α antagonist. We delete the molecular descriptor variables with variances less than 10 and variable values all of which are 0. The remaining 474 molecular descriptor variable data are normalized. Although the biological activity and compound molecular descriptor data of ER α are not monitored values, they are given molecular characteristics, and the data are extremely trustworthy, but the measurement units differ owing to the varied properties of the indicators. Therefore, standardizing the data is required to reduce the impact of dimensions between data components and to produce a consistent assessment standard that allows diverse indications to be compared.

The most common contemporary normalizing methods are linear scale transformation, range transformation, and zero-mean normalization. When new data is added to them, the maximum and lowest values of the range transformation technique change. It needs to be re-defined at this moment, and there are severe maximum and lowest values. It is better suited for modest amounts of data. Obviously, it is not applicable to the 1 974 com-

pounds' associated data. Because the zero-mean normalization criterion only applies to arrays that follow the normal distribution, we utilized the linear proportional transformation technique to normalize the data, such that all variable values are mapped to the range [0,1], and the original data is scaled proportionately.

1.3 Molecular Descriptor Variables Screening

Each compound sample has 729 compound descriptor variables, but not all of them can offer information on the biological activity of ER α . This is due to collinearity, which occurs when the variables include the same or similar information. We first preprocess the data of 729 molecular descriptor variables to remove dimensional impact and make various variables comparable. The original input spatial data will then be mapped to a new space via a specific transformation, thereby changing the molecular descriptor variables, considering the dimensionality reduction methods of data feature transformation such as principal component analysis, factor analysis, and singular value decomposition. To establish a rating of the significance of each variable to biological activity, we first apply the gray correlation analysis model to analyze the correlation between each molecular descriptor variable and the biological activity value of the molecule to ER α . On this basis, we apply the random forest technique of supervised machine learning to extract the amount of the contribution of various variables to better understand the relevance of variable discrimination and its influence on biological activity. Finally, the top 20 factors with the greatest influence on biological activity are identified, and the rationale of the chosen variables is assessed.

1.3.1 The GRA model

The GRA model can calculate the relevance of each molecular descriptor variable's effect on ER α biological activity^[34]. The GRA model is a method for measuring the degree of correlation between the properties of various variables that is based on grey theory. We can compare and rank the importance of the variables by using the degree of correlation between each molecular descriptor variables and the biological activity value of the compounds for ER α . The greater the correlation is, the more important the variable is for the biological activity of ER α . The key process phases are as follows:

1) Establish the comparison and reference sequences

The 474 molecular descriptor variable data sequences are regarded as comparison sequences, and the

experimentally observed value of the compound's biological activity IC₅₀ for ER α is regarded as the reference sequence.

2) Determine the grey correlation degree

The grey correlation degree between the comparison and reference series is defined as

$$r(x_0, x_i) = \frac{1}{1974} \sum_{k=1}^{1974} \zeta_i(k) \quad (1)$$

$$\zeta_i(k) = \frac{\min_s \min_t |x_0(t) - x_s(t)| + \rho \max_s \max_t |x_0(t) - x_s(t)|}{|x_0(k) - x_i(k)| + \rho \max_s \max_t |x_0(t) - x_s(t)|} \quad (2)$$

where $\zeta_i(k)$ represents the correlation coefficient of the comparison sequence x_i to the reference sequence x_0 on the k -th molecular descriptor variable. ρ is the model resolution coefficient and fulfills $\rho \in [0, 1]$, which is taken $\rho = 0.5$. $\min_s \min_t |x_0(t) - x_s(t)|$ represents the two-level minimum difference, and $\max_s \max_t |x_0(t) - x_s(t)|$ represents the two-level maximum difference.

1.3.2 RF model

We utilize the RF technique of supervised machine learning to generate the top 100 variables based on the GRA model to extract the amount of the contribution of various variables to better understand the relevance of variable discrimination and its influence on biological activity, and lastly obtain the top 20 factors with the greatest influence on biological activity. The RF model is a very flexible integrated machine learning system that uses a decision tree as the primary learner^[35]. Because the algorithm has its own feature variable screening method, it can assess the significance of each molecular descriptor variable's effect on ER α biological activity. The major steps in the procedure are as follows:

1) Determining the contribution of variables

The RF model is used to filter molecular descriptor variables that influence the relevance of ER α biological activity. The average value of each variable's contribution to each decision tree in the random forest is used to assess the relevance of the molecular descriptor variables. The relevance of a variable is established by comparing its contribution to that of other molecular descriptor variables. The Gini index (GI) is used to indicate the contribution of factors to ER α biological activity, and the formula is as follows:

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \quad (3)$$

where GI_m represents the Gini index of node m , p_{mk} rep-

resents the fraction of category k in node m , and K indicates the number of decision tree categories.

Then the Gini index score of each molecular description variable is

$$\text{VIM}_{jm}^{(\text{Gini})} = \text{GI}_m - \text{GI}_l - \text{GI}_r \quad (4)$$

where $\text{VIM}_{jm}^{(\text{Gini})}$ is the significance of the variable at node m , which indicates the change in the Gini index of the decision tree branch at node m ; GI_l, GI_r are the Gini indices of the two new nodes after the decision tree is branched.

2) Training using random forests

We undertake RF training on the molecular descriptor variables that determine the relevance of ER α biological activity in the RF model, which is primarily based on the Bagging concept^[36] and the feature subspace idea^[37]. The Bagging idea is to use the bootstrap approach to randomly generate k new sample sets from the original data set of molecular descriptor variables with replacement, and then build k related decision trees. The feature subspace concept is primarily represented in the fact that whenever a decision tree is constructed, each node splitting must randomly and equiprobably take a subset of attributes from all attributes. Then, from the attribute subset, we choose one attribute that is thought to be the best to split the node. Because each decision tree's training is independent of the others and each decision tree is of equal importance, there is no need to consider the weight of the decision tree, and the RF algorithm can be implemented directly through parallel processing, which can improve the efficiency of our model.

3) Sorting the features based on their weights

The RF algorithm's feature weights represent variations in molecular descriptors and the importance of the quantity to the biological activity of ER α . If the node of the molecular descriptor variable in the i -th decision tree is M , then the variable's relevance in the i -th decision tree is

$$\text{VIM}_{jm}^{(\text{Gini})} = \sum_{m \in M} \text{VIM}_{jm}^{(\text{Gini})} \quad (5)$$

If the RF contains n decision trees, then

$$\text{VIM}_j^{(\text{Gini})} = \sum_{i=1}^n \text{VIM}_{ij}^{(\text{Gini})} \quad (6)$$

The significance scores of 474 molecular descriptor variables are then normalized as

$$\text{VIM}_j = \frac{\text{VIM}_j}{\sum_{i=1}^{474} \text{VIM}_i} \quad (7)$$

Finally, the variables' relevance ratings are ranked in descending order, and the top 20 are chosen as the molecular descriptor variables with the greatest influence on ER α biological activity.

1.4 The QSAR Model

The QSAR prediction models primarily comprise linear and non-linear prediction models. A multiple linear regression model is used in the standard linear model. However, because the molecular descriptor variables of the compounds have interaction relationships, the variables are nonlinear and interrelated, making the traditional linear prediction models not applicable. More and more nonlinear prediction models are being used in QSAR prediction models, with neural network prediction models being the most common nonlinear prediction models presently. As a result, this paper employs the BP neural network model, the BP neural network model optimized by the genetic algorithm, and the support vector regression model to develop a QSAR quantitative prediction model for ER α biological activity. Furthermore, in this inquiry, we randomly choose 90% of the data from the 1 974 sample data set. The remaining 10% of the data are utilized for model accuracy verification and comparison to evaluate the model's rationality after training the three models. The technical route is shown in Fig. 1.

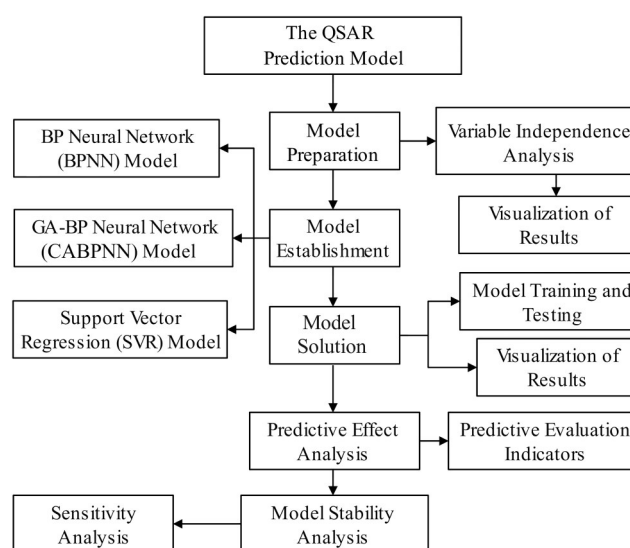


Fig. 1 The QSAR prediction model technical framework diagram

1.4.1 The BPNN model

The BPNN model is a common artificial intelligence network algorithm^[38]. It has high nonlinear map-

ping capabilities and is made up of an input layer, many hidden layers, and an output layer. Furthermore, its network structure is basic, with superior data approximation ability and high computation accuracy, making it particularly ideal for the prediction of nonlinear issues. The BPNN model's learning process may be broken into two steps: forward and back propagation. The forward propagation involves processing incoming information from the input layer via the hidden layer and passing it to the output layer. If the required output cannot be acquired in the output layer, back propagation is used, and the error signal is returned along the original connection line. The fundamental steps are as follows:

1) Determination of the hidden and output layer

The j -th neuron's input and output values in the hidden layer are as follows:

$$\begin{cases} \text{hid}_{\text{input}_j} = \sum_{i=1}^I w_{ij} x_i \\ \text{hid}_{\text{output}_j} = F(\text{hid}_{\text{input}_j}) \end{cases} \quad (8)$$

where $F(x)$ denotes the transfer function.

The output layer's input and output calculation formulae are as follows:

$$\begin{cases} \text{out}_{\text{input}_m} = \sum_{j=1}^J w_{jm} \text{hid}_{\text{output}_j} \\ \text{out}_{\text{output}_1} = F(\text{out}_{\text{input}_1}) \end{cases} \quad (9)$$

2) Determination of the training error

The n -th generation's error is:

$$e_m(n) = \text{out}_{\text{output}_1} - y_m \quad (10)$$

where y_m is the expected output.

The network's overall error is:

$$e(n) = \frac{1}{2} \sum_{m=1}^M e_m^2(n) \quad (11)$$

3) Error back propagation

The following is how the weights between the output layer and the hidden layer are corrected:

$$w_{jm}(n+1) = w_{jm}(n) + \Delta w_{jm}(n) \quad (12)$$

where $w_{jm}(n)$ is the amount of weight adjustment between the output layer and the hidden layer.

The determination of the number of nodes in each layer of the neural network is a critical technology that has a direct impact on the network simulation's performance. The BPNN model created in this article has the following structure: the input layer indicators are 16 molecular description variables chosen using SCA. According to the "Rule of Thirds" theorem, this paper defines the number of hidden layer nodes as 13 based on the

consideration of decreasing errors, enhancing network convergence effect, and lowering model calculation time, and the output layer index is a pIC_{50} variable showing the biological activity of the ER α .

1.4.2 The GA-BPNN model

Because the BPNN model belongs to the local search optimization, the prediction result may be the local optimal rather than the global optimal, the learning convergence speed is slow, and the model cannot accurately obtain the initial connection weights and thresholds during training. The genetic algorithm is combined with BPNN in this paper, and the initial weight and threshold are optimized by genetic algorithm^[39].

The GA-BPNN model is divided into two parts: genetic algorithm optimization of initial weights and thresholds, and BP neural network training and prediction. To begin with, because the BP neural network cannot precisely determine the initial connection weights and thresholds during training, it is often initialized to a random value in the range of -0.5 to 0.5 by default, which has a significant influence on the neural network's whole training process. When the genetic algorithm is applied to optimize the BP neural network's prediction: 1) The neural network topology is determined by the number of input and output parameters of the BP neural network training set data to determine the number of parameters that need to be optimized by the genetic algorithm, namely the coding length of the population individuals in the genetic algorithm. 2) Because everyone includes the connection weight and threshold value of a specific network, just the fitness function is required to determine the population's individual fitness value. 3) When estimating the pIC_{50} of ER α biological activity, our aim is to minimize the residual difference between the developed model's predicted value and the actual value. As a result, it is utilized as the genetic algorithm fitness function after sorting the norm of the model predicted value and the actual value error matrix. 4) The ideal connection weights and thresholds may then be derived using selection, crossover, and mutation operations. 5) Finally, the evolutionary algorithm-optimized connection weights and thresholds are inserted into the network's BP nerve train to obtain the prediction result.

1.4.3 The SVR model

The support vector machine (SVM) offers the advantages of rapid learning, global optimization, and a high degree of generalization capacity. Its learning outcomes out-perform other regression prediction ap-

proaches, and it has been effectively implemented in a variety of applications such as text and speech recognition, time series data prediction, and so on. The SVR model is a method for solving regression issues that employs SVM^[40]. The following are the procedures for using the SVR model to solve the regression prediction issue:

1) Mapping and the definition of a loss function

Find a nonlinear mapping Φ from the input space to the output space for the regression problem using the sample set of $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The data in the sample set is transferred to the high-dimensional space F via this mapping, and linear regression in this space F is performed using the following linear function in the feature space:

$$f(x) = t\Phi(x) + b, \Phi: \mathbb{R}^n \rightarrow F, t \in F \quad (13)$$

where b is the threshold. Because the mapping is fixed, the total of empirical risks impact t and $|t|^2$ that cause it to be flat in high-dimensional space, namely:

$$R(t) = \frac{1}{2} \|t\|^2 + \sum_{i=1}^n \varepsilon [f(x_i) - y_i] \quad (14)$$

where n is the number of samples. $\varepsilon[x]$ is the loss function, and it only computes sample points other than $[-\varepsilon, \varepsilon]$.

2) Minimizing the objective function.

To control the function's complexity, the linear regression function should be as flat as possible, even if the Euler norm of t is the smallest, and the fitting error that may exceed the accuracy is considered. The relaxation factor is introduced. According to the statistical learning theory's structural risk criteria, the SVR model determines t and b in equation (13) by minimizing the objective function, that is, the following equation is minimized:

$$R(t, \xi_i, \xi_i^*) = \frac{1}{2} \|t\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

$$\text{s.t.} \begin{cases} y_i - t\Phi(x_i) - b \leq \varepsilon + \xi_i \\ t\Phi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (15)$$

where the first term flattens the regression function and improves generalizability, and the second term is error reduction. The constant $C > 0$ is used to balance the regression function's flatness and the number of deviations bigger than the number of sample points ε .

3) Creating a Lagrangian equation

We introduce the Lagrangian multiplier $\alpha_i, \alpha_i^*, \lambda_i, \lambda_i^*$, and establish the equation:

$$L(t, b, \alpha, \xi) = \frac{1}{2} \|t\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

$$- \sum_{i=1}^n \alpha_i (\varepsilon + \xi_i - y_i + t\Phi(x_i) + b)$$

$$- \sum_{i=1}^n \alpha_i^* (\varepsilon + \xi_i^* + y_i - t\Phi(x_i) - b)$$

$$- \sum_{i=1}^n (\lambda_i \xi_i + \lambda_i^* \xi_i^*) \quad (16)$$

Allow the partial derivative of the parameters to be equal to 0 to obtain the minimum value of the equation (16), and then solve the dual optimization problem. As a result, the SVM function regression issue may be simplified to a quadratic programming problem.

We can get t represented by data points and the linear regression function by solving the quadratic programming problem:

$$f(x) = t\Phi(x_i) + b$$

$$= \sum_{i=1}^n (\alpha_i - \alpha_i^*) [\Phi(x_i)\Phi(x)] + b$$

$$= \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (17)$$

where $K(x, x_i)$ is the kernel function. Different SVM may be generated by using different types of kernel functions.

4) Selection of the kernel functions

The kernel functions that are commonly utilized include the radial basis function, polynomial function, perceptron function, linear function, and so on. Based on a priori understanding of the sample data set, we choose the radial basis function, which is appropriate for nonlinear sample data, has less parameters for the polynomial kernel function, and performs well for a variety of sample sizes. The formula is as follows:

$$K(x, x_i) = e^{-\frac{\|x - x_i\|^2}{\delta^2}} \quad (18)$$

1.5 ADMET Prediction Model

The qualities of compound ADMET are connected to its molecular features and are dependent factors. The properties of the five compounds are all classed as 0-1 binary variables. As a result, we intend to categorize and forecast compound ADMET properties using three models: the BPNN model, the SVR model, and the LR model^[41]. Five categorization prediction models of ADMET properties were determined by a comparison investigation of the prediction effects of the three models, to produce matching predictions for the 50 compounds. We develop a Logistic regression-based compound

ADMET property categorization prediction model.

An LR model can be created by knowing that $X_i = (x_1, x_2, \dots, x_r)$ is a compound molecular descriptive variable that impacts the ADMET properties of the i -type compound and Y_{ij} is the value of the j -th ADMET attribute of the i -type compound:

$$Y_{ij} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_r x_r \quad (19)$$

where α_0 is the intercept, α_r is the slope coefficient, x_r is the independent variable, that is, the compound molecular descriptor variable.

We define the probability that the compound's ADMET property is good as P , and the probability that the compound's ADMET property is poor as the value of the probability $P \in [0, 1]$. When the Caco-2 value of ADMET is 1, it suggests that the compound has superior small intestine epithelial cell permeability, indicating that the molecule has excellent properties. Similarly, a CYP3A4 score of 1 indicates that the substance has excellent properties. The compound has excellent properties when the hERG value is 0. The compound has excellent properties when the HOB value is 1. The compound has excellent properties when the MN value is 0.

We apply a Logistic transformation on P while also using Logistic P as the dependent variable to construct a linear regression equation:

$$Y_{ij} = P_{ij} = \frac{1}{1 + e^{-(a + \sum_{k=1}^r x_{ik} \cdot \alpha_k)}} \quad (20)$$

where P_{ij} represents the probability that the compound's ADMET property is excellent, a is the regression intercept, x_{ik} is the independent variable, that is, the compound molecular descriptor, which is the number of molecular descriptors of the i -th compound, and α_k represents each molecular descriptor's regression coefficient.

2 Results

2.1 The Top 20 Variables

The RF technique is used to solve the top 100 variables obtained by the GRA model, and the top 20 variables with the greatest effect on biological activity are finally achieved. The results are shown in Table 1.

2.2 Variable Independence Analysis

The top 20 variables with the greatest influence on biological activity may contain the same or comparable information, resulting in collinearity of the variables, which will be serious if there is a strong correlation between the variables. We employ Spearman's correlation

analysis (SCA) to assess the relationship between 20 variables. The SCA method is a non-parametric statistical tool created by statistician Spearman that is primarily based on the study of correlations between two variables^[42]. The greater the connection between the two variables, the bigger the Spearman correlation coefficient. The 20 molecular descriptor variables were subjected to Spearman correlation analysis, the results are shown in Fig. 2.

Figure 2 shows that the correlation between the 20 molecular descriptor variables is lighter in color, indicating that most of the variables are not linearly correlated. The more prominent correlations between a few variables are mainly 10-18 ($\zeta=0.999\ 2$), 1-14 ($\zeta=-0.921\ 9$), 12-14 ($\zeta=0.885\ 9$), 15-17 ($\zeta=0.881\ 5$), and 11-19 ($\zeta=0.821\ 5$). We excluded the four variables rated 14, 17, 18, and 19 in terms of relevance, MAXDN, minaaCH, gmax, and SP-6.

2.3 The Results of QSAR Model

In this paper, we randomly chose 90% of the data from the 1 974 set of sample data for training the three models, and the remaining 10% was utilized for model accuracy verification and comparison to evaluate the model's reasonableness. The mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (SMAPE) of the BPNN, GA-BPNN, and SVR models are used to assess the five categories of predictions. The indicators compare and examine the models' predicted impacts. The smaller the value of the five categories of prediction assessment indicators, the better the fitting effect, that is, the prediction model's prediction impact, as shown in Table 2.

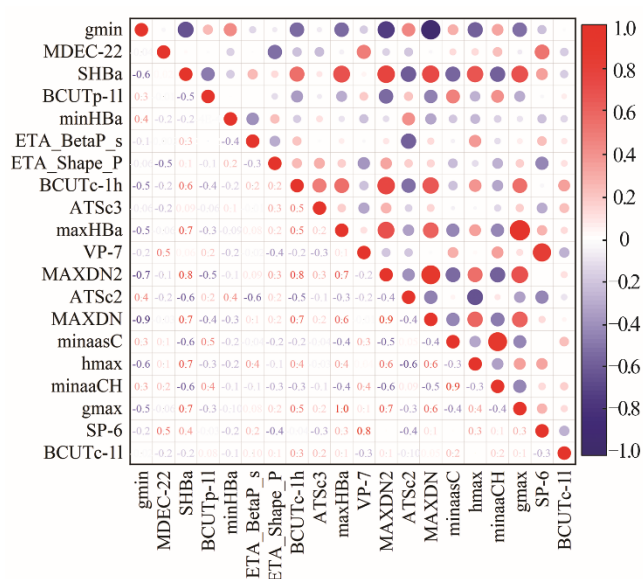
Table 2 shows that the SVR model's five types of prediction evaluation index values are all the smallest, indicating that the SVR model's prediction effect is better than the BPNN and the GA-BPNN model, so the SVR model is chosen as the QSAR quantitative prediction model.

2.4 The Results of ADMET Prediction Model

Similarly, we chose 90% of the data from the 1 974 set of sample data at random for training the three models, with the remaining 10% utilized for model accuracy verification and comparison. For the five ADMET properties, the three models were employed. Simultaneously, the four assessment indicators of the model's Accuracy

Table 1 The top 20 variables that have the greatest influence on biological activity

Molecular descriptor variables	Description	Contribution	Ranking
ATSc3	ATS autocorrelation descriptor, weighted by charges	0.020 2	1
BCUTc-1h	nlow highest partial charge weighted BCUTS	0.019 5	2
minHBa	Minimum E-States for (strong) Hydrogen Bond acceptors	0.018 7	3
ATSc2	ATS autocorrelation descriptor, weighted by charges	0.017 6	4
BCUTc-1l	nhigh lowest partial charge weighted BCUTS	0.017 5	5
BCUTp-1l	nhigh lowest polarizability weighted BCUTS	0.017 2	6
gmin	Minimum E-State	0.017 1	7
VP-7	Valence path, order 7	0.017 1	8
gmax	Maximum E-State	0.016 7	9
maxHBa	Maximum E-States for (strong) Hydrogen Bond acceptors	0.016 7	10
MDEC-22	Molecular distance edge between all secondary carbons	0.016 6	11
minaasC	Minimum atom-type E-State: :C:-	0.016 6	12
ETA_Shape_P	Shape index <i>P</i>	0.016 1	13
hmax	Maximum H E-State	0.015 2	14
MAXDN	Maximum negative intrinsic state difference in the molecule (related to the nucleophilicity of the molecule)	0.015 1	15
minaaCH	Minimum atom-type E-State: :CH:	0.015 0	16
ETA_BetaP_s	A measure of electronegative atom count of the molecule relative to molecular size	0.015 0	17
MAXDN2	Maximum negative intrinsic state difference in the molecule (related to the nucleophilicity of the molecule)	0.014 9	18
SPC-6	Simple path cluster, order 6	0.014 8	19
SP-6	Simple path, order 6	0.014 7	20

**Fig. 2** The correlation coefficient diagram of 20 molecular descriptor variables**Table 2** The prediction and evaluation results of the three models

Model	MSE	RMSE	MAE	MAPE	SMAPE
BPNN	0.006 0	0.077 6	0.073 6	1.198 6	1.199 9
GA-BPNN	0.005 8	0.075 8	0.068 1	1.067 4	1.066 7
SVR	0.003 5	0.059 0	0.053 2	0.843 2	0.843 2

Score, Precision Score, Recall Score, and F1 Score, as presented in Table 3, are statistically examined.

Table 3 indicates that using the SVR model to predict Caco-2 and hERG properties, the LR model to predict CYP3A4 and MN properties, and the BPNN model to predict HOB properties yields good prediction results. As shown in Fig. 3, we visualized the prediction results of the five models that made the best use of ADMET properties.

Table 3 The prediction and evaluation results of the three models for the ADMET properties

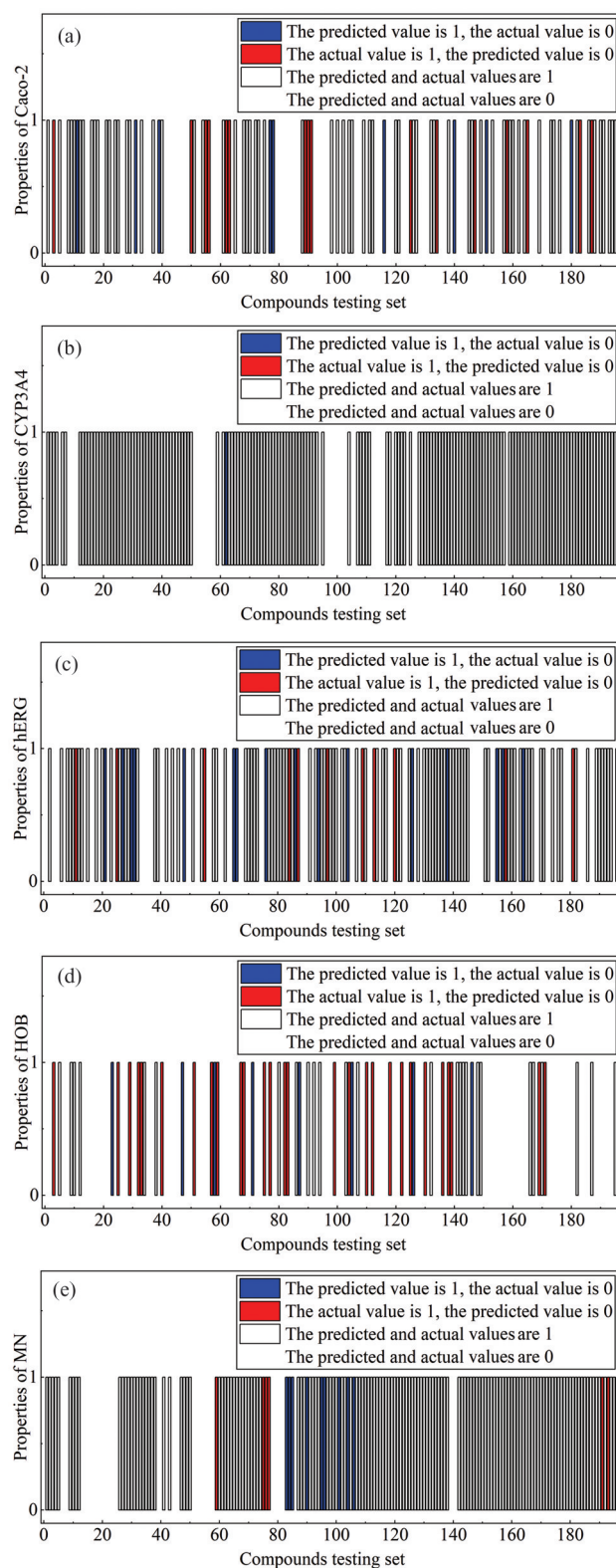
Property	Model	Accuracy Score	Precision Score	Recall Score	F1 Score
Caco-2	BPNN	0.842 6	0.805 2	0.794 9	0.800 0
	SVR	0.852 8	0.772 2	0.847 2	0.807 9
	LR	0.875 6	0.700 0	0.636 4	0.666 7
CYP3A4	BPNN	0.934 0	0.951 7	0.958 3	0.955 0
	SVR	0.918 8	0.985 8	0.908 5	0.945 6
	LR	0.956 9	0.959 3	0.989 8	0.976 6
hERG	BPNN	0.842 6	0.887 8	0.813 1	0.848 8
	SVR	0.837 6	0.850 9	0.866 1	0.858 4
	LR	0.789 3	0.894 2	0.834 4	0.863 3
HOB	BPNN	0.928 9	0.966 2	0.940 8	0.953 3
	SVR	0.857 9	0.879 0	0.938 8	0.907 9
	LR	0.878 2	0.657 1	0.389 8	0.489 4
MN	BPNN	0.842 6	0.805 2	0.794 9	0.800 0
	SVR	0.852 8	0.772 2	0.847 2	0.807 9
	LR	0.799 5	0.825 5	0.923 9	0.870 3

Figure 3 shows the classification prediction model results for five ADMET properties. If the blue and red histograms are lower, the model's accuracy is better. On the other hand, if the blue and red histograms are larger, the model's categorization prediction impact is poor. So, we established that the SVR model predicted the Caco-2 and hERG properties, the LR model predicted the CYP3A4 and MN properties, and the BPNN model predicted the HOB properties of 50 substances.

3 Discussion

3.1 Verification of Rationality

To further validate the rationality of variable selection, we employ information entropy theory to validate the top 20 factors that have the greatest influence on biological activity. Information entropy is a measure of information uncertainty in information theory. It is inspired by thermodynamic entropy, and Shannon presents a mechanism for quantifying the information delivered^[43]. The smaller the information entropy value of an indicator variable, the larger the degree of variation of

**Fig. 3** Classification and prediction model results for the five ADMET properties

- (a) Caco-2 property on SVR model; (b) CYP3A4 property on LR model; (c) hERG property on SVR model; (d) HOB property on BPNN model; (e) MN property on LR model

the variable, the greater the utility value of the information, that is, the more information that can be offered, according to the concept of entropy. On the contrary, the higher the entropy value, the lower the degree of fluctuation in the variable, the lower the utility value of the information, and the less information delivered. The information entropy cumulative weight of 20 significant variables were calculated, as shown in Fig. 4. Figure 4 shows that the cumulative information entropy weight of

the top 20 variables having the greatest influence on biological activity is 0.045 5, indicating that variable information entropy has reached 4.55% of the total information entropy of all 474 variables. The average amount of information entropy for 474 variables is $20/474=4.22\%$. It is possible to explain that the top 20 molecular descriptor variables presented in this article have a certain degree of relevance and smoothness, so confirming the rationale of the molecular descriptor variable screening.

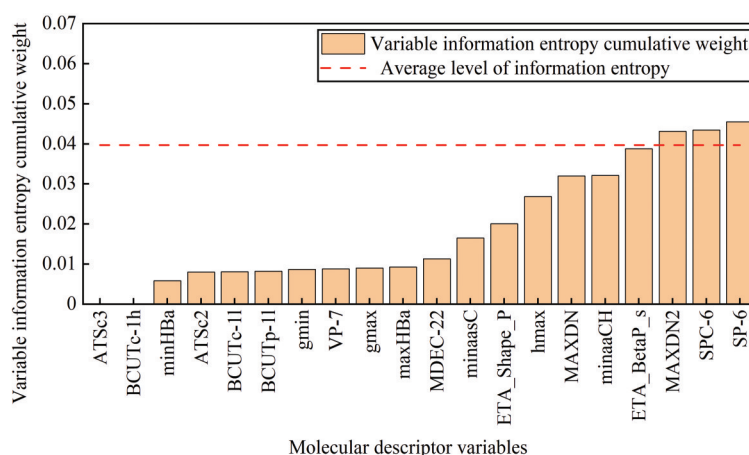


Fig. 4 Histogram of cumulative information entropy weights of molecular descriptor variables

3.2 Sensitivity Analysis

This study used the sensitivity analysis approach to quantitatively quantify the proposed QSAR model's dependency on molecular descriptor variables to validate its performance^[44]. We use the controlled variable approach in the sensitivity analysis, a case study of the compound #1, which requires us to modify only one molecular description at a time. In the range of -15% to $+15\%$, the variables are changed in 5% increments without affecting the values of the other molecular descriptor

variables, and the data after each variable change is inserted into the SVR model. The prediction of the pIC_{50} value of ER α biological activity was achieved, as shown in Fig. 5.

Figure 5 shows that, except for variations in ETA_BetaP_s and maxHBa, which generate considerable fluctuations in the projected value of ER α 's biological activity pIC_{50} , the remaining 14 molecular descriptor variables have an influence on the pIC_{50} value. The volatility of the change range is small, implying that the de-

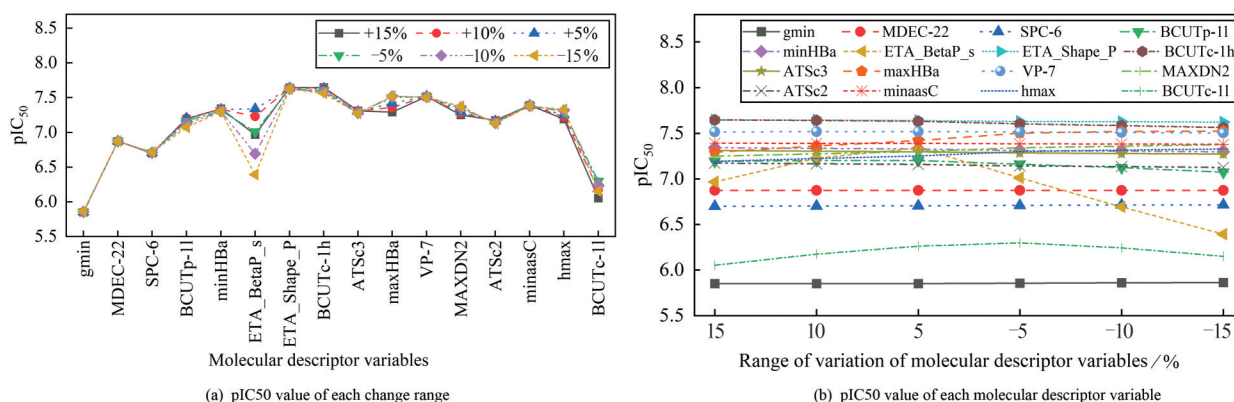


Fig. 5 Sensitivity analysis of pIC_{50} value and variable range

gree of influence is small. To some extent, it demonstrates that the established QSAR quantitative prediction model is stable. Simultaneously, if you pay attention to the changes in the two variable values of ETA_BetaP_s and maxHBa, the model's promise may be realized.

4 Conclusion

In this study, a QSAR model of compound biological activity and a classification prediction model of ADMET properties were performed for the biological activity of anti-breast cancer drug candidates.

We first utilized grey correlation analysis in conjunction with the random forest algorithm to find the top 20 molecular descriptor variables having the highest influence on biological activity, and then we used Spearman correlation analysis to discover 16 independent variables. The reasonableness of variable screening was validated using information entropy theory, and the result reveals that the cumulative information entropy weight of the top 20 variables has reached 4.55% of the total information entropy of all variables. It implies that it has a specific level of relevance and smoothness, so verifying the justification for the molecular descriptor variable screening. Second, a QSAR model of the compound were established based on BPNN, GA-BPNN, and SVR. The MSE, RMSE, MAE, MAPE, and SMAPE of the three models are used to assess the five categories of predictions.

The result shows that the SVR model's five types of prediction evaluation index values are all the smallest, and sensitivity analysis of the model demonstrates that the model has high accuracy and stability. Furthermore, The BPNN, the SVR, and the LR model were used to identify and predict the ADMET properties of substances, with the prediction impacts of each model compared and assessed. The results reveal that a SVR model was used in QSAR quantitative prediction, the SVR model predicts the Caco-2 and hERG properties, the LR model predicts the CYP3A4 and MN properties, and the BPNN model predicts the HOB characteristics in the classification prediction of ADMET properties.

The source data and codes of this study are available from GitHub at: <https://github.com/guitar-boy/source-data-and-codes>.

Conflicts of Interest

The author declares that they have no known com-

peting financial interests in this paper.

References

- [1] Li F F, Xia H S, Chen Z W. Polypeptide daintain as a new biomarker for detecting breast tumor [J]. *Wuhan University Journal of Natural Sciences*, 2008, **13**(1): 118-122.
- [2] Shull J D, Hadsell D L, *et al.* Genetic variation in sensitivity to estrogens and breast cancer risk [J]. *Mammalian Genome*, 2018, **29**(1): 24-37.
- [3] Xie G, Liu X, Zhang Y, *et al.* UTX promotes hormonally responsive breast carcinogenesis through feed-forward transcription regulation with estrogen receptor [J]. *Oncogene*, 2017, **36**(39): 5497-5511.
- [4] Lyndsay V, Rhodes, Zhu Y. Effects of SDF-1-CXCR4 signaling on microRNA expression and tumorigenesis in estrogen receptor-alpha (ER- α)-positive breast cancer cells [J]. *Experimental Cell Research*, 2011, **317**(18): 2573-2581.
- [5] Lambert C T, Lichter J B, Perry A N, *et al.* Medial amygdala ER α expression influences monogamous behaviour of male prairie voles in the field [J]. *Proceedings of the Royal Society B: Biological Sciences*, 2021, **288**(1956): 20210318.
- [6] Sutherland R, Meeson A, Lowes S. Solute transporters and malignancy: Establishing the role of uptake transporters in breast cancer and breast cancer metastasis [J]. *Cancer and Metastasis Reviews*, 2020, **39**(3): 919-932.
- [7] Chen C, Baumann W T, Xing J H, *et al.* Mathematical models of the transitions between endocrine therapy responsive and resistant states in breast cancer [J]. *Journal of the Royal Society Interface*, 2014, **11**(96): 20140206.
- [8] Katzer K, Hill J L, McIver K B, *et al.* Lipedema and the potential role of estrogen in excessive adipose tissue accumulation[J]. *International Journal of Molecular Sciences*, 2021, **22**(21): 11720.
- [9] Shyam Sundar P, Naresh P, Justin A, *et al.* Dual modulators of p53 and cyclin D in ER alpha signaling by albumin nanovectors bearing zinc chaperones for ER-positive breast cancer therapy[J]. *Mini-Reviews in Medicinal Chemistry*, 2021, **21**(7): 792-802.
- [10] Mutlu Ağardan N B, Değim Z, Yılmaz S, *et al.* Tamoxifen/raloxifene loaded liposomes for oral treatment of breast cancer[J]. *Journal of Drug Delivery Science and Technology*, 2020, **57**: 101612.
- [11] Li J S, Luo D H, Wen T T, *et al.* Representative feature selection of molecular descriptors in QSAR modeling [J]. *Journal of Molecular Structure*, 2021, **1244**(12): 131249.

- [12] Hemmateenejad B, Miri R, Elyasi M. A segmented principal component analysis-regression approach to QSAR study of peptides [J]. *Journal of Theoretical Biology*, 2012, **305**: 37-44.
- [13] Martínez M J, Razuc M, Ponzoni I. MoDeSuS: A machine learning tool for selection of molecular descriptors in QSAR studies applied to molecular informatics[J]. *BioMed Research International*, 2019, **2019**: 1-12.
- [14] Krishna J G, Khan K, Roy K. Application of QSARs in identification of mutagenicity mechanisms of nitro and amino aromatic compounds against salmonella typhimurium species [J]. *Toxicology in Vitro*, 2020, **65**: 104768.
- [15] Wang X Z, Perston B, Yang Y, *et al.* Robust QSAR model development in high-throughput catalyst discovery based on genetic parameter optimisation [J]. *Chemical Engineering Research and Design*, 2009, **87**(10): 1420-1429.
- [16] Yu Q, Deng H F, Yan H, *et al.* An accurate nonlinear QSAR model for the antitumor activities of chloroethyl nitrosoureas using neural networks [J]. *Journal of Molecular Graphics and Modelling*, 2011, **29**(6): 826-833.
- [17] Hadrup N, Frederiksen M, Wedebye E B, *et al.* Asthma-inducing potential of 28 substances in spray cleaning products—Assessed by quantitative structure activity relationship (QSAR) testing and literature review [J]. *Journal of Applied Toxicology*, 2021, **42**(3): 130-153.
- [18] Ding Q Y, Zu S P, Hou S Y, *et al.* VISAR: An interactive tool for dissecting chemical features learned by deep neural network QSAR models [J]. *Bioinformatics*, 2020, **36**(11): 3610-3612.
- [19] Chi C T, Lee M H, Weng C F, *et al.* In silico prediction of PAMPA effective permeability using a two-QSAR approach [J]. *International Journal of Molecular Sciences*, 2019, **20** (13): 3170.
- [20] Wang H, Jiang M Y, Li S J, *et al.* Design of cinnamaldehyde amino acid Schiff base compounds based on the quantitative structure-activity relationship [J]. *Royal Society Open Science*, 2017, **4**(9): 170516.
- [21] Tseng Y J, Hopfinger A J, Esposito E X. The great descriptor melting pot: Mixing descriptors for the common good of QSAR models [J]. *Journal of Computer-Aided Molecular Design*, 2012, **26**(1): 39-43.
- [22] Vu O, Mendenhall J, Altarawy D, *et al.* BCL: Mol2D—A robust atom environment descriptor for QSAR modeling and lead optimization [J]. *Journal of Computer-Aided Molecular Design*, 2019, **33**(5): 477-486.
- [23] Williams K, Bilsland E, Sparkes A, *et al.* Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases [J]. *Journal of the Royal Society, Interface*, 2015, **12**(104): 20141289.
- [24] Dong J, Wang N N, Yao Z J, *et al.* ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database [J]. *Journal of Cheminformatics*, 2018, **10**(1): 1-11.
- [25] Pires D E V, Blundell T L, Ascher D B. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures [J]. *Journal of Medicinal Chemistry*, 2015, **58**(9): 4066-4072.
- [26] Rogers D, Hahn M. Extended-connectivity fingerprints[J]. *Journal of Chemical Information and Modeling*, 2010, **50** (5): 742-754.
- [27] Teramoto R, Kato T. Transfer learning for cytochrome p450 isozyme selectivity prediction[J]. *Journal of Bioinformatics and Computational Biology*, 2011, **9**(4): 521-540.
- [28] Shi T T, Yang Y W, Huang S H, *et al.* Molecular image-based convolutional neural network for the prediction of ADMET properties [J]. *Chemometrics and Intelligent Laboratory Systems*, 2019, **194**: 103853.
- [29] Ferreira L L, Andricopulo A D. ADMET modeling approaches in drug discovery [J]. *Drug Discovery Today*, 2019, **24**(5): 1157-1165.
- [30] Jo J, Kwak B, Choi H S, *et al.* The message passing neural networks for chemical property prediction on SMILES [J]. *Methods*, 2020, **179**: 65-72.
- [31] Sorkun M C, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds [J]. *Scientific Data*, 2019, **6**(1): 143.
- [32] Shroff T, Aina K O, Maass C, *et al.* Studying metabolism with multi-organ chips: New tools for disease modelling, pharmacokinetics and pharmacodynamics [J]. *Open Biology*, 2022, **12**(3): 210333.
- [33] China Post-graduate Mathematical Contest in Modeling. "Huawei Cup" the 18th china post-graduate mathematical contest in modeling [EB/OL]. [2022-10-14]. <https://cpipc.acge.org.cn/cw/hp/4>.
- [34] Xu Q W, Xu K L, Li L, *et al.* Mine safety assessment based on basic event importance: Grey relational analysis and bow-tie model [J]. *Royal Society Open Science*, 2018, **5**(8): 180397.
- [35] Zhang H, Liu Q D, Sun X R, *et al.* Integrated bioinformatics and machine learning algorithms analyses highlight related pathways and genes associated with Alzheimer's disease [J]. *Current Bioinformatics*, 2022, **17**(3): 284-295.

- [36] Breiman L. Random forests [J]. *Machine Learning*, 2001, **45** (1): 5-32.
- [37] Breiman L. Bagging predictors [J]. *Machine Learning*, 1996, **24**(2): 123-140.
- [38] Xu Z H. Research on demand forecast of elderly beds based on multiple regression model [J]. *Journal of Shenyang University (Social Science)*, 2022, **24**(1): 52-61(Ch).
- [39] Zhang C Y, Zhang R R, Dai Z H, *et al.* Prediction model for the water jet falling point in fire extinguishing based on a GA-BP neural network [J]. *PLoS One*, 2019, **14**(9): 0221729.
- [40] Xu B B, Wang T Z, Luo K, *et al.* A fault diagnosis method based on wavelet singular entropy and SVM for VSC-HVDC converter[J]. *Wuhan University Journal of Natural Sciences*, 2020, **25**(4): 359-368.
- [41] Yadav D K, Kumar S, Saloni S, *et al.* Molecular docking, QSAR and ADMET studies of withanolide analogs against breast cancer[J]. *Drug Design, Development and Therapy*, 2017, **11**: 1859-1870.
- [42] Zhou Y, Li S J. BP neural network modeling with sensitivity analysis on monotonicity based Spearman coefficient [J]. *Chemometrics and Intelligent Laboratory Systems*, 2020, **200** (1): 103977.
- [43] Savari C, Sotudeh-Gharebagh R, Kulah G, *et al.* Detecting stability of conical spouted beds based on information entropy theory [J]. *Powder Technology*, 2019, **343**: 185-193.
- [44] Das K, Ranjith K G, Madhusudhan R K, *et al.* Sensitivity and elasticity analysis of novel corona virus transmission model: A mathematical approach [J]. *Sensors International*, 2021, **2**(1): 100088.

□