Oncologie

Tech Science Press

DOI: 10.32604/oncologie.2022.021256

ARTICLE

An ISSA-RF Algorithm for Prediction Model of Drug Compound Molecules Antagonizing ERα Gene Activity

Minxi Rong[1], Yong Li[1,*], Xiaoli Guo[1,*], Tao Zong[2], Zhiyuan Ma[2] and Penglei Li[2]

[1]College of Mathematics and Information Science, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

[2]College of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

[*]Corresponding Authors: Xiaoli Guo. Email: xlguo@zzuli.edu.cn; Yong Li. Email: liyong3880@163.com

Received: 05 January 2022 Accepted: 25 April 2022

ABSTRACT

Objectives: The ERα biological activity prediction model is constructed by the compound molecular data of the anti-breast cancer therapeutic target ERα and its biological activity data, which improves the screening efficiency of anti-breast cancer drug candidates and saves the time and cost of drug development. **Methods:** In this paper, Ridge model is used to screen out molecular descriptors with a high degree of influence on the biological activity of Erα and divide datasets with different numbers of the molecular descriptors by screening results. Random Forest (RF) is trained by Root Mean Square Error (RMSE) and Coefficient of determination (R^2) to determine the parameter range of RF optimized by Improved Sparrow Search Algorithm (ISSA-RF) which adds adaptive weights compared with the ordinary Sparrow Search Algorithm (SSA). Then the divided datasets were put into the ISSA-RF with defined parameter ranges to construct a regression prediction model for the biological activity of compounds on Erα, and compared with Genetic Algorithm Optimized Support Vector Machine (GA-SVM), Back Propagation Neural Network (BP), Extreme Gradient Boosting (XGBoost) for analysis and comparison. **Results:** We have tried a variety of combinations of molecular descriptors with different numbers and the above four models all achieve the best accuracy model on the dataset constructed when using 100 molecular descriptors. The ISSA-RF model proposed in this paper has a high degree of agreement between the predicted biological value of ERα and the actual value and prediction accuracy (RMSE) is 0.6876389. **Conclusions:** In the training model, ISSA-RF is proposed and it is proved that adding adaptive weights can greatly optimize the fitness accuracy of the sparrow algorithm. In the experimental part, this paper uses a variety of molecular descriptors for training, which reduces the chance of model training accuracy caused by the number of different molecular descriptors, and limits the search range of the ISSA-RF model to avoid the local optimization of the model. Secondly, the parameter optimization time is greatly reduced. In conclusion, the prediction model of drug compound molecules that antagonize ERα gene activity (ISSA-RF) proposed in this paper improves the accuracy and efficiency of anti-breast cancer drug candidates, and provides a new idea for building a quantitative structure-activity relationship model.

KEYWORDS

Anti-breast cancer drug candidates; machine learning; ridge regression; random forest; sparrow search algorithm

1 Introduction

In recent years, with the rapid development of human society, the global environment has been irreversibly destroyed and bring various diseases that have never been seen before [1]. Drug therapy is an important means to control and treat diseases. Traditional drug research and development cycle is long and the efficiency is low [2]. To reduce the cost and time of drug development, Quantitative Structure-Activity Relationship models (QSAR) are often used to construct drug compounds and target cell activity in drug discovery and development. The model is then used to predict target cell activity corresponding to new or structurally altered drug compounds. The candidate drug compound molecules are screened out according to the predicted biological activity value of the target cell, so as to achieve the purpose of computer-aided selection of the candidate drug compound.

Breast cancer [3] is a common female disease and one of the cancers with a higher mortality rate. In the early 1970s, Pietras discovered that estrogen can rapidly up-regulate the cAMP level of endometrial cells through cell membrane binding sites, and therefore speculated that there is a membrane ER. For the first time, the definition of Estrogen Receptor (ER) has been elaborated. Afterwards, estrogen was confirmed to be directly related to the malignant proliferation of breast cancer cells, and the viewpoint that breast cancer cells depend on estrogen receptors for growth was recognized. ERα is an important target for the treatment of breast cancer, if it is possible to find suitable drug candidates based on ERα activity value and molecular related factors of candidate drug compounds, it will become an effective method. In recent years, the use of machine learning [4,5] in the medical field provides an effective way for drug research and development. Machine learning can study the potential relationship between ERα activity and drug compound molecules, and build a Quantitative Structure-Activity Relationship model (QSAR) of anti-breast cancer drug candidates in order to select suitable drug compound molecules, which can not only improve the time efficiency but also provide a variety of options for the development of anti-breast cancer drugs.

2 Related Research

Machine learning is a method in which a computer automatically finds rules from the input data and can predict unknown data through such rules. It has powerful big data computing capabilities and plays an important role in data processing and data mining. In previous studies, machine learning has a strong application background in classification in the medical field. Jiang et al. [6] used the annealing algorithm and Random Forest (RF) to determine the optimal characteristics of BCRP inhibitors, and used four machine learning methods, deep learning methods, and integrated learning methods to predict BCRP inhibitors, and then evaluated the drug's effectiveness. The results showed that the Support Vector Machine(SVM) classifier showed the best classification effect, the Mathew's Correlation Coefficient (MCC) value of the test set was 0.812, and the Area Under Curve (AUC) value was 0.958. Che et al. [7] used three non-integrated machine learning algorithms, Back Propagation Neural Network (BP) and three Boosting series algorithms to predict prostate cancer. The results showed that the Decision Trees model in the non-integrated algorithm is the best with an accuracy rate of 0.933, and Extreme Gradient Boosting model (XGBoost) in the Boosting series of algorithms is the best with an accuracy rate of 0.957. Wang et al. [8] used word vector representation technology to characterize the main feature data, and then used the XGBoost model to learn the correlation between the features to identify the pathogens of food-borne diseases. The results showed that the precision rate and recall rate are 68%. Lu et al. [9] used Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Decision Trees and other methods to construct classification models for neuraminidase (NA) inhibitors and non-neuraminidase inhibitors. The results showed that the SVM algorithm gave the optimal prediction accuracy is 92.6%.

Machine learning does not have such a significant application background in prediction in the medical field and the main reason is that the feature dimension of the medical data set is large, which has a

complicated impact on the prediction results, and the prediction results cannot achieve accurate prediction. This reflects from the side that the main purpose of prediction in the medical field is to assist medical experiments. Sheridan et al. [10] mainly discussed the applicability of XGBoost in QSAR model in the paper and use Grid algorithm to optimize the model parameters. The experimental average determination coefficient (R^2) reached 0.42 which is similar to the neural network. However, the parameters range of grid optimization is defined by the author, and the range of optimization parameters has not been determined by relevant experiments. Mansouri et al. [11] provided a variety of open source QSAR models to predict the strongest acidic and strongest basic pKas of chemicals. The SVM algorithm combined with K-NN, XGBoost, and Deep Neural Network(DNN) are used to predict different open-source data sets. The optimal results show that the prediction accuracy of the deep neural network is high, and the RMSE of the optimal prediction value is 1.5, R^2 is 0.8.

The widespread use of machine learning [12–14] provides a new direction for drug research and development, and the idea of integrated learning has gradually developed in the field of machine learning. The idea of ensemble learning is to solve the shortcomings of a single model and to integrate multiple models to avoid the limitations of a single model. Random Forest(RF) is one of ensemble learning.

Since there are many training parameters for the integrated learning RF model, most scholars will train the parameters that have a greater impact on the model. Zheng et al. [15] chosed to train the n_estimators and max_depth in the coal spontaneous combustion temperature prediction model, but this did not give full play to the random selection of features by RF and the influence of the parameter max_features on the model was not considered. Most scholars believe that the choice of max_features will reduce the diversity of a single decision tree and reduce the accuracy of the RF model, so if we choose max_features, we should consider the above two aspects at the same time.

In this paper, Sparrow Search Algorithm (SSA) is combined with RF model, and adaptive weights are added to the SSA finder position update formula, and an ISSA-RF model is proposed. Before the training data was input into the model, the Ridge model was used to screen out molecular descriptors that had a greater impact on the activity of the ERα gene. Since the number of molecular descriptor inputs is uncertain, the dataset is divided by combinations of screening features with different numbers. These data are then trained using the RF model alone with RMSE and R^2 to determine the sparrow search range of max_depth, n_estimators, max_features. Then, the divided datasets were put into the ISSA-RF model to construct the QSAR model.

In order to verify the accuracy of the model, Genetic Algorithm Optimized Support Vector Machine (GA-SVM), Back Propagation Neural Network (BP), and Extreme Gradient Boosting (XGBoost) were used to construct a quantitative prediction model for the biological activity of drug compound molecules on Erα in this paper. After experimental comparison, the ISSA-RF model proposed in this paper is superior to the other three models, which can improve the efficiency of screening candidate drug molecules while ensuring the accuracy of prediction, and provides a new idea for the construction of QSAR in terms of model optimization.

This paper introduces **Principles and Methods** (*Data Source*, *Data Preprocessing*, *Basic Models*, *Model Construction* and *Model Evaluation Index*) in the third part, and displays the experimental results in the fourth part **Analysis of Results**, and the fifth part is the **Conclusion** of the article.

3 Principles and Methods

3.1 Data Source

The data in this article comes from the DrugBank drug molecule database at the University of Alberta [16]. In order to all readers to view the data, we have put the data on the github website (https://github.com/Li519445444/candidate-drug-data-source/tree/master). This data set provides:

a) The biological activity data of 1974 drug compounds on ERα and the biological activity value of the compound against ERα including IC50 and pIC50. The unit of IC50 is nM. The smaller the value, the greater the biological activity and the more effective it is to inhibit ERα activity. The pIC50 is the negative logarithm of the IC50, and this value is usually positively correlated with biological activity, that is, the larger the pIC50 value, the higher the biological activity. Generally, the pIC50 is used to indicate the biological activity.

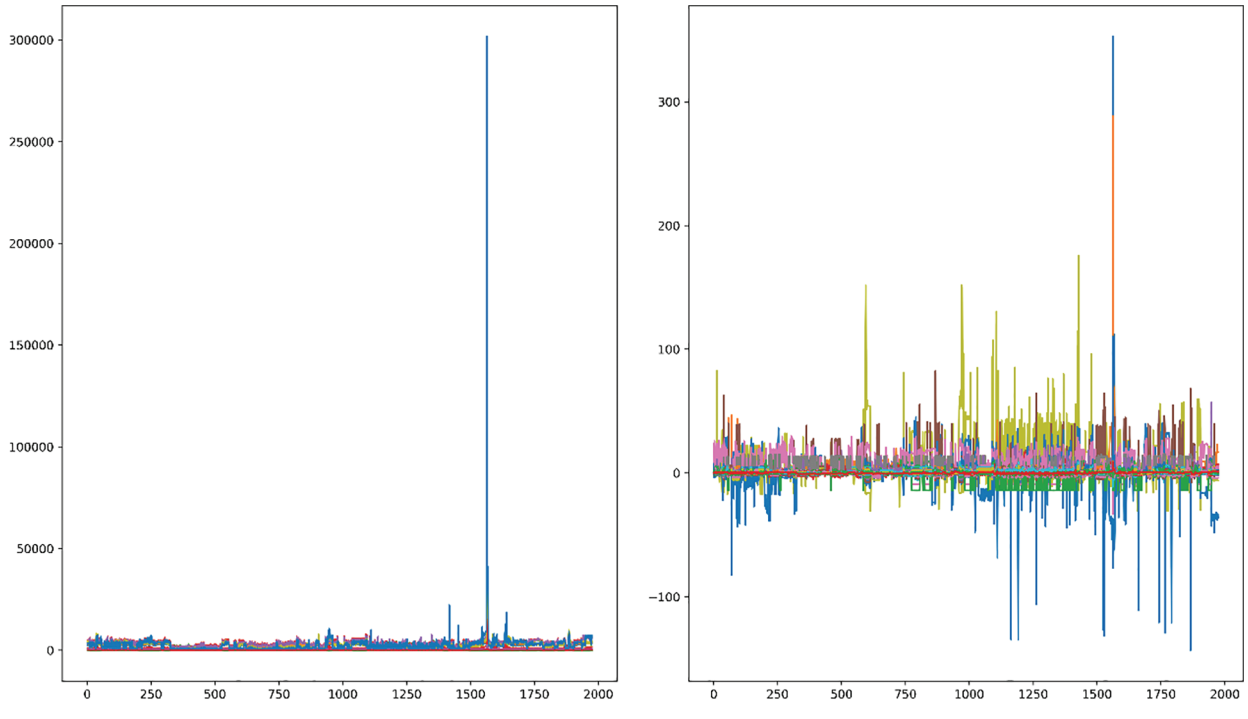
b) 729 molecular descriptor information for 1974 drug compounds. The molecular descriptor of a compound is a series of parameters used to describe the structure and properties of the compound, including physical and chemical properties (such as molecular weight, LogP, etc.), topological structure characteristics (such as the number of hydrogen bond donors, the number of hydrogen bond acceptors, etc.) and so on.

3.2 Data Preprocessing

In view of the problem of non-standard data standards, this paper adopts the following measures to solve the problem:

a) Through data observation, it is found that there are columns with all 0 values in the data. Therefore, 729 molecular descriptors whose information is all 0 are eliminated. Because these descriptors have no role in feature screening and prediction and they have no practical significance in drug development.

b) The values of drug compound molecules have a high degree of dispersion and there are abnormal values. In order to improve the accuracy of the model prediction, this paper uses the RobustScaler function to scale the features by robust statistical information to the abnormal data. Before and after data processing are shown in the Fig. 1.



(a) Before data processing (b) After data processing

Figure 1: Before and after data standardization

3.3 Basic Models

3.3.1 Ridge Model

When inputting high-dimensional features into machine learning models, there will be some features in the model that are not related to the training target or the features are redundant [17], and these redundant features not only make the prediction results of the algorithm inaccurate, but also consume computing time and computer memory. There are many excellent algorithms for the selection of data features, such as: Lasso [18], Ridge, Principal Component Analysis [19], etc. This paper adopts Ridge algorithm which is faster in calculation and better in effect.

Enumerate the expression form of the ridge regression algorithm, in Eq. (1), μ is called zero parameter.

$$\hat{\beta}^{Ridge} = (X'X + \mu I)^{-1} X'y \tag{1}$$

Take the value of the minimum penalty likelihood function as the estimated value of the regression coefficient, in Eq. (2), The penalty is $P_\lambda(|B|) = \lambda \sum_{j=1}^{P} |\beta_j|^m$, $m > 0$, λ is the adjustment parameter.

$$\hat{\beta} = \arg\min_{\beta} \left(\|Y - \beta X\|^2 + P_\lambda |\beta| \right) \tag{2}$$

When m is 2, that is the Ridge penalty item. The expression form of Ridge regression can be obtained in Eq. (3).

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta 0 - \sum_{i=1}^{P} x_{ij} \beta_j \right)^2 + \lambda \sum_{i=1}^{P} {\beta_j}^2 \right\} \tag{3}$$

In Ridge regression, the high-dimensional data X has been centered and standardized, so that the size of the standardized ridge regression coefficients can be directly compared to judge the importance of high-dimensional features. The size of the regression coefficient reflects the importance of high-dimensional features to $Y(y_1, y_2, \ldots, y_N)$ in the model. In this paper, the low-importance features are eliminated, and the high-importance input model is used for training.

3.3.2 Sparrow Search Algorithm

Sparrow Search Algorithm (SSA) [20] is a group behavior algorithm inspired by the foraging behavior and anti-predation behavior of sparrows. Individuals in the population are divided into discoverers, followers and alerters according to the division of labor. The discoverers mainly provide foraging directions and areas for the entire population. The followers follow the discoverers to forage. The alerters are responsible for monitoring the foraging area. The optimization of the model parameters is achieved through the process of updating the position of the three.

Suppose the total number of sparrow individuals is n. The dimension of the variable to be optimized is d. Then the position of the population can be expressed as:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \tag{4}$$

The discoverers are responsible for guiding the population to find food and guiding the population to a safe location. The location update formula is as follows:

$$Dx_{i,j}^{t+1} = \begin{cases} x_{i,j}^{t} \cdot \exp\left(-\dfrac{i}{\alpha \times T}\right), & R_2 < ST \\ x_{i,j}^{t} + Q \cdot L, & R_2 \geq ST \end{cases} \tag{5}$$

In Eq. (5), t represents the current iteration number, $x_{i,j}^{t}$ represents the position of the i sparrow in the j dimension at t iteration, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, d$. α is a random number between 0 and 1. T is the maximum number of iterations. R_2 is warning value between 0 and 1. ST is the preset safety threshold between 0.5 and 1. Q is a Gaussian distribution random number. L a is a matrix whose shape is $(1 * d)$ and elements are all 1.

The followers will always follow the discoverers and compete for food resources in order to obtain more food resources. When the fitness of the discoverers is low, the followers will move to other positions, The follower's position update formula is as follows:

$$Fx_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\dfrac{x_{worst}^{t} - x_{i,j}^{t}}{i^2}\right), & i > \dfrac{n}{2} \\ x_{P}^{t+1} + \left|x_{i,j}^{t} - x_{P}^{t+1}\right| \cdot A^{+} \cdot L, & i \leq \dfrac{n}{2} \end{cases} \tag{6}$$

In Eq. (6), x_{worst}^{t} represents the position of the individual with the lowest fitness in t iteration. x_{P}^{t+1} represents the position of the individual with the highest fitness in t iterations. $A^{+} = A^{T}(AA^{T})^{-1}$, A is shape of $(1 * d)$ and each element of a is randomly preset to -1 or 1.

When the population realizes the danger, alerters will quickly make an anti-predation response. The update formula of the position of the alerters is as follows:

$$x_{i,j}^{t+1} = \begin{cases} x_{best}^{t} + \beta \cdot \left|x_{i,j}^{t} - x_{best}^{t}\right|, & f_i \neq f_g \\ x_{best}^{t} + k \cdot \left(\dfrac{x_{i,j}^{t} - x_{best}^{t}}{|f_i - f_w| + \varepsilon}\right), & f_i = f_g \end{cases} \tag{7}$$

In Eq. (7), x_{best}^{t} represents the global optimal position in t iterations. β is the step size control parameter, which is a Gaussian distribution random number with mean 0 and variance 1. k is a random number between -1 and 1. f_i represents the fitness of the current individual. f_g and f_w represents the fitness of the current global best and worst individuals. ε is the smallest constant used to avoid the situation where the denominator is 0. It can be seen from this formula, $f_i \neq f_g$ means that the individual is at the periphery of the population and needs to constantly change positions to obtain higher fitness. $f_i = f_g$ means that the individual at the center of the population is aware of the danger, and will continue to approach other nearby sparrows to stay away from the danger area.

3.3.3 Random Forest Prediction Model

Random Forest (RF) [21] is a flexible and easy-to-use machine learning algorithm. It uses multiple regression trees as a basis for training and incorporates the idea of bagging. In the tree training process, random feature selection is used to reduce the correlation between sample features, thereby solving the problem of overfitting of a single decision tree model, so that the model has a better prediction effect. The basic process sees Fig. 2 below.

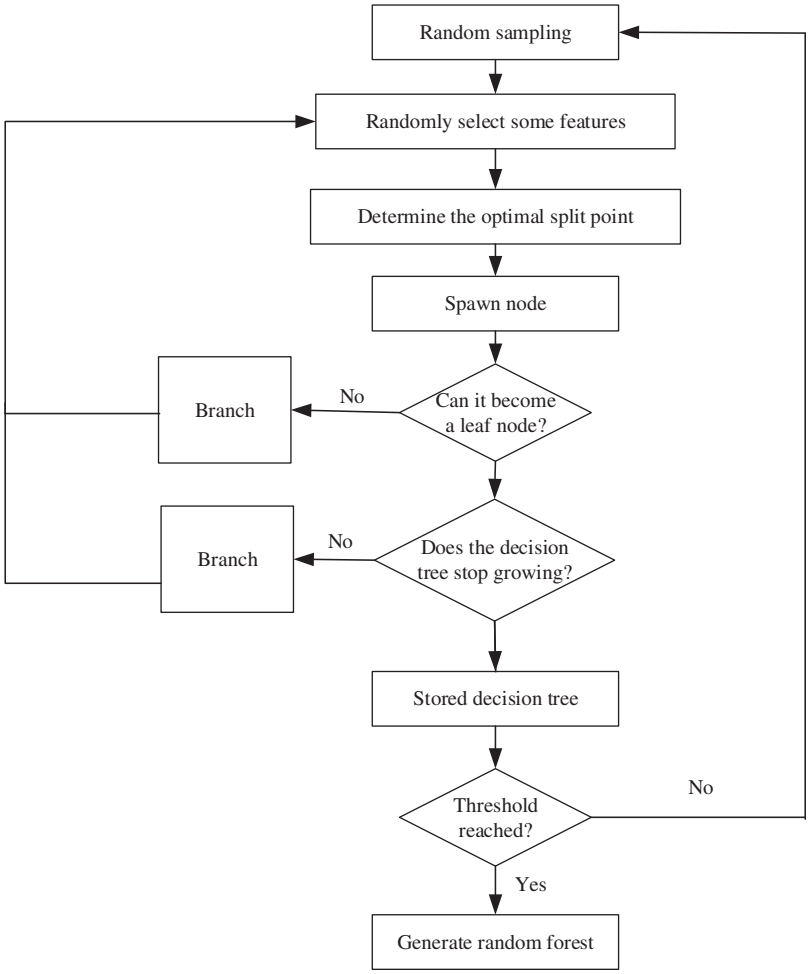


Figure 2: Random Forest algorithm flow

From the flowchart, we can see the Random Forest generation process, and the details are presented below:

a) Random replacement sampling in training samples from the original training set, repeating S times;
b) Use these S data sets as training sets to train S CART tree models;
c) If the feature dimension is M, specify a constant m, randomly select m feature subsets from M features, and each time the tree is split, select the best from these m features;
d) The generated S decision trees are formed into a random forest to ensure that each tree grows to the maximum extent;
e) For classification problems, the classification results are generated by voting by S CART trees. For regression problems, the mean value of the prediction results of S trees is used as the final prediction result.

In addition, the Random Forest incorporates the bootstrap idea when selecting samples, that is, sampling with replacement. The out-of-bag data generated by the bootstrap algorithm can be used to test the generalization ability of the model.

For the establishment of this model, this paper adopts predictive Random Forest, selects the optimal feature j and segmentation position s.

$$R_m(j,s) = \min_{j,s}[\min_{c_1} \sum (y_i - c_1)^2 + \min_{c_2} \sum (y_i - c_2)^2] \tag{8}$$

$$R_1(j,s) = \left\{x|x^{(j)} \leq s\right\}, R_2(j,s) = \left\{x|x^{(j)} > s\right\} \tag{9}$$

In Eq. (8), c_i is to divide the set sample, $m = 1, 2$. Then this paper uses the selected (j, s) to divide the area to find the corresponding output value in Eq. (10).

$$c_m = \frac{1}{N_m} \sum_{x_i \in R_m(j,s)} y_i \tag{10}$$

The smaller the c_m is, the better the split performance of the selected feature and the segmentation point is. Continue to perform the above steps on the sub-regions to generate the optimal RF model.

3.4 Model Construction

3.4.1 Molecular Descriptor Screening

Ridgecv model is trained using 5-fold cross-validation and training regression coefficients are sorted by the size. The larger the regression coefficient, the higher the influence of the molecular descriptor on the change of biological activity. The top 20 molecular descriptors are shown in the Fig. 3. In this paper, the top 20 to 100 drug compound variables with a high degree of influence are selected through the sorted characteristic regression coefficients and divide into features_num at intervals of 10 features. features_num = [20,30,40,50,60,70,80,90,100]. Divide the dataset by features_num and put the divided dataset into the model for training.

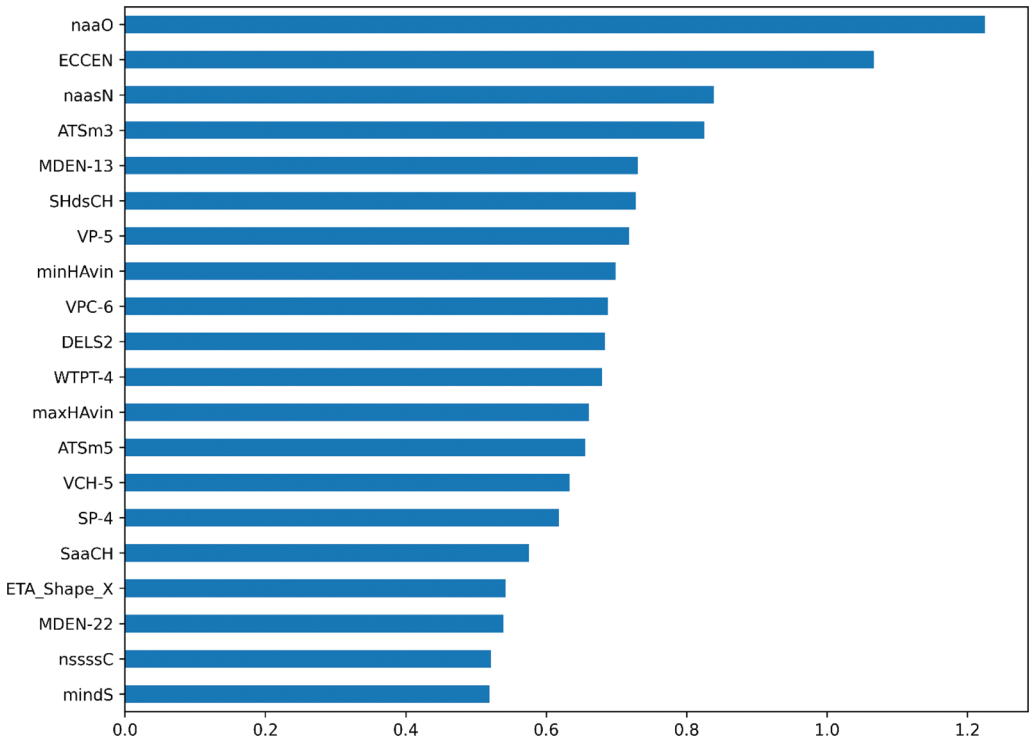


Figure 3: The importance of molecular description of each variable on biological activity is ranked in the top 20

3.4.2 ISSA-RF Model

Like other intelligence optimization algorithms, SSA has the problem of easily falling into local optimum. In the later stage of the traditional SSA algorithm iteration, the position between the three sparrows will be updated in a small range near the optimal point, which is prone to the situation that the position update in a small range is stagnant. To solve this problem, this paper proposes an Improved Sparrow Search Algorithm(ISSA). We add dynamic adaptive weights to the sparrow finder position update formula to optimize the local exploration problem of the model. The formula is as follows:

$$x_{i,j}^{t+1} = \begin{cases} w \cdot \left(x_{i,j}^{t} \cdot \exp\left(-\dfrac{i}{\alpha \cdot T} \right) \right), & R_2 < ST \\ x_{i,j}^{t} + Q \cdot L, & R_2 \geq ST \end{cases} \tag{11}$$

$$w = \sin\left(\frac{\pi \cdot t}{2 \cdot T} + \pi \right) + b \tag{12}$$

In Eq. (12), T represents the maximum number of iterations. t represents the current iteration number. b represents the bias term. The Eq. (12) curve figure is as follows.

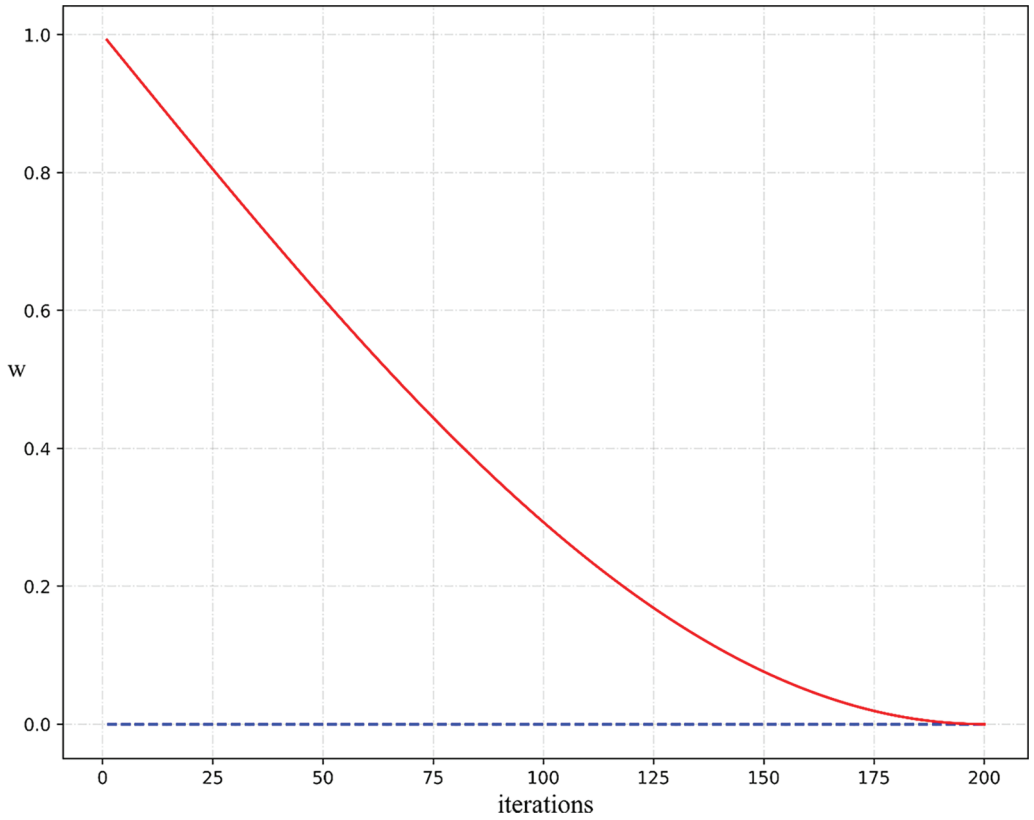


Figure 4: Eq. (12) curve

It can be seen from Fig. 4 that w is constantly changing as the number of iterations is updated. The sine function controls the range of w within the range of [−1,0], and adjusts the range of w by modifying the bias term b. This paper sets b to 1. Giving the discoverer a larger weight in the early stage of the algorithm

iteration is conducive to the global search. In the later stage of the algorithm search, w decreases slowly, and there is sufficient time for local exploration. And because w has a small decrease, it can also make a relatively large weight in the later stage of the iteration, thereby speeding up the speed of local exploration. The involvement of this weight also speeds up the overall convergence speed of the algorithm to a certain extent.

In order to illustrate the convergence effect of the ISSA algorithm proposed in this paper, this paper uses the Rosenbrock function to conduct simulation experiments. When it is a binary function, as shown in Fig. 5. The Rastrigin formula is as follows:

$$Rosenbrock = \sum_{i=1}^{N-1} 100(x_{i-1} - x_i^2)^2 + (1 - x_i)^2 \tag{13}$$

In this paper, the independent fitness convergence training of the Rosenbrock function is performed, and the results are shown in Fig. 6. Within 500 iterations, SSA reached the convergence state at 27 iterations, and the convergence fitness value precision reached 10e−7, while ISSA reached the convergence state at 41 iterations, and the convergence fitness value precision reached 10e−23. The fitness convergence accuracy is much higher than that of SSA, which shows that the improved ISSA algorithm has much higher convergence fitness accuracy than the ordinary SSA algorithm.

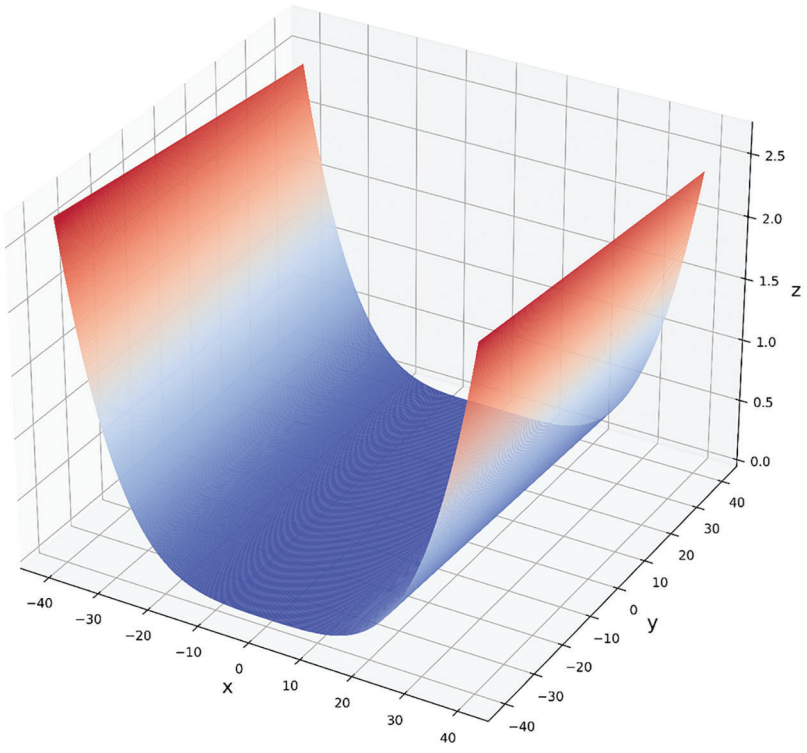


Figure 5: Rosenbrock binary function

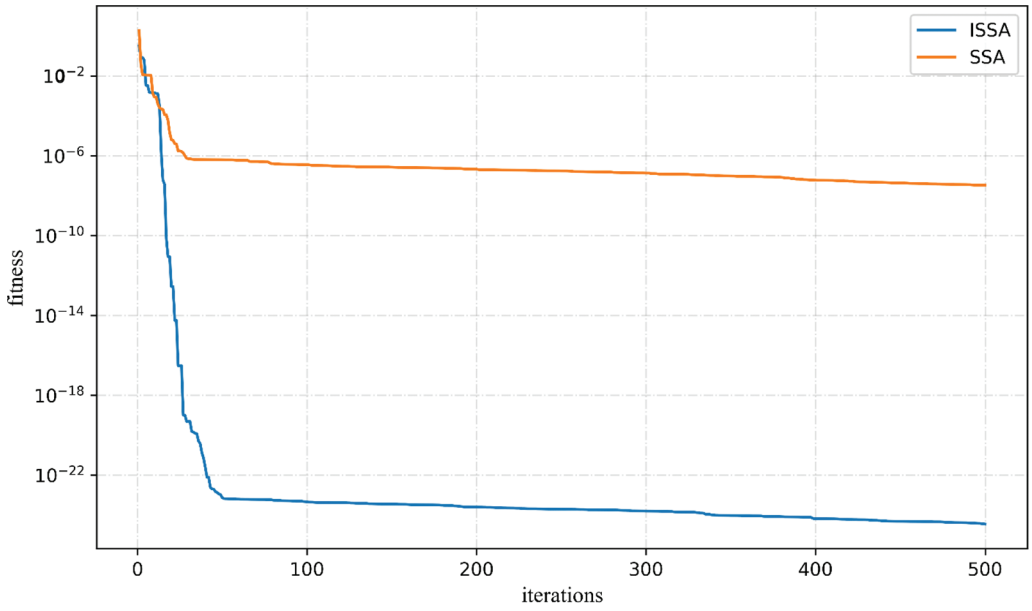


Figure 6: SSA, ISSA to Rosenbrock function convergence curve

Considering the time benefit and model accuracy of ISSA algorithm parameter optimization, too large an optimization range will lead to long model training time, and too small an optimization range will lose model accuracy, so it is necessary to limit the optimization range of parameters. Among the parameters required by the RandomForestRegressor function, n_estimators, max_depth, and max_features have a greater impact on the accuracy of the RF model. The parameters of max_depth and n_estimators are limited to the optimization range by using RF model training with MSE and R^2 as the judgment criteria. The smaller the MSE, the higher the accuracy of the prediction of biological activity. The higher the R^2, the better the model, and the stronger the interpretation of the biological activity by the molecular descriptor features of the selected compound. As shown in Fig. 7. As the number of iterations increases, MSE keeps decreasing. After 175 iterations, MSE and R^2 are almost stable with small fluctuations. This shows that when n_estimators = 175 or so, the model is stable and the accuracy reaches the highest level, so the range of n_estimators is set to [160,180]. Similarly, according to the curves shown in Figs. 7c and 7d, when max_depth = 16, the degree of fluctuation is small, and the range of max_depth is set to [10,30]. In order to satisfy the characteristics of RF model selection of features, this paper selects max_features, but considering that the number of molecular descriptors input in each training is different, the optimization range is not limited, and the ISSA algorithm directly performs the optimization operation. The parameter optimization range can be seen in Table 1.

This paper combines the ISSA with the RF algorithm to optimize the parameters of the RF model. First, the sparrow population is initialized. The the RF model is trained to calculate its fitness value. This paper sets the fitness function as the RMSE of RF model training and each iteration searches towards a position with a lower RMSE. Update the positions of discoverers, follower and alerters through the change of fitness value until the training termination condition is met. Finally, the ERα gene activity was predicted by the RF model with optimal parameters, and the model was tested by the validation set. The predicted model structures of the drug compound molecules that antagonize the activity of the ERα gene by the ISSA-RF model are shown in Fig. 8.

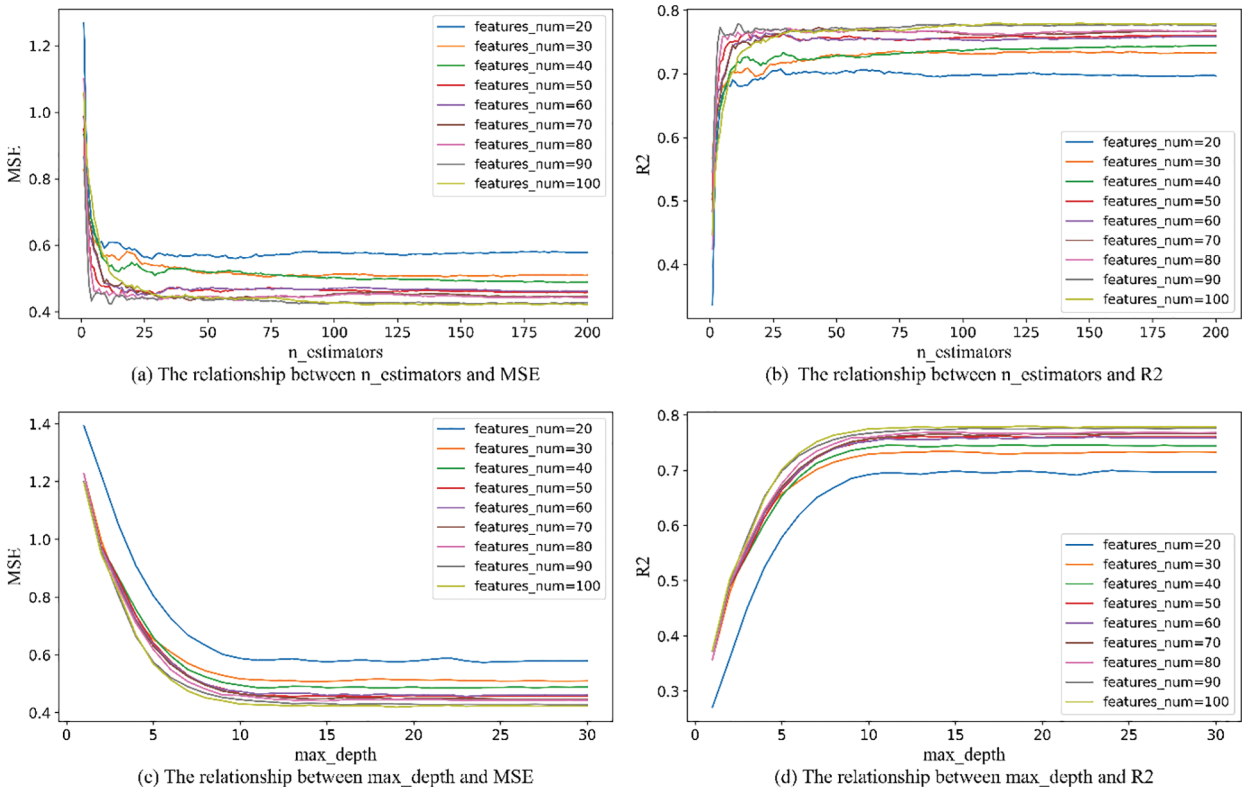


Figure 7: The relationship between the number of features_num and n_estimators and max_depth

Table 1: Parameter optimization range

Parameters	Optimization range
The number of populations	10
The maximum number of iterations	30
n_estimators	[160,180]
Max_depth	[10,30]
Max_features	[1, features_num]

3.5 Model Evaluation Index

In order to objectively evaluate the prediction effect of the established quantitative prediction model of biological activity, we introduce several model evaluation indicators to evaluate the accuracy of the model such as Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Coefficient of determination (R^2).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \tag{14}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i| \tag{15}$$

$$MAPE = \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2} \tag{17}$$

$$R^2(R2) = 1 - \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{18}$$

In the formula, y_i is the pIC50 actually given in the test set, $\hat{y}_i$ is the pIC50 predicted by the model in the test set, i is the collection position, and m is the number of samplings. When MSE, MAE, MAPE, TMSE are at a lower level, it proves that the prediction results of the model are better. The smaller the value, the higher the prediction accuracy of the established model. In addition, in Eq. (18), $\bar{y}$ is the average value of the pIC50 of the test sample. The determination coefficient R^2 takes a value between 0 and 1, The closer the value of R^2 is to 1, the better the performance of the model.

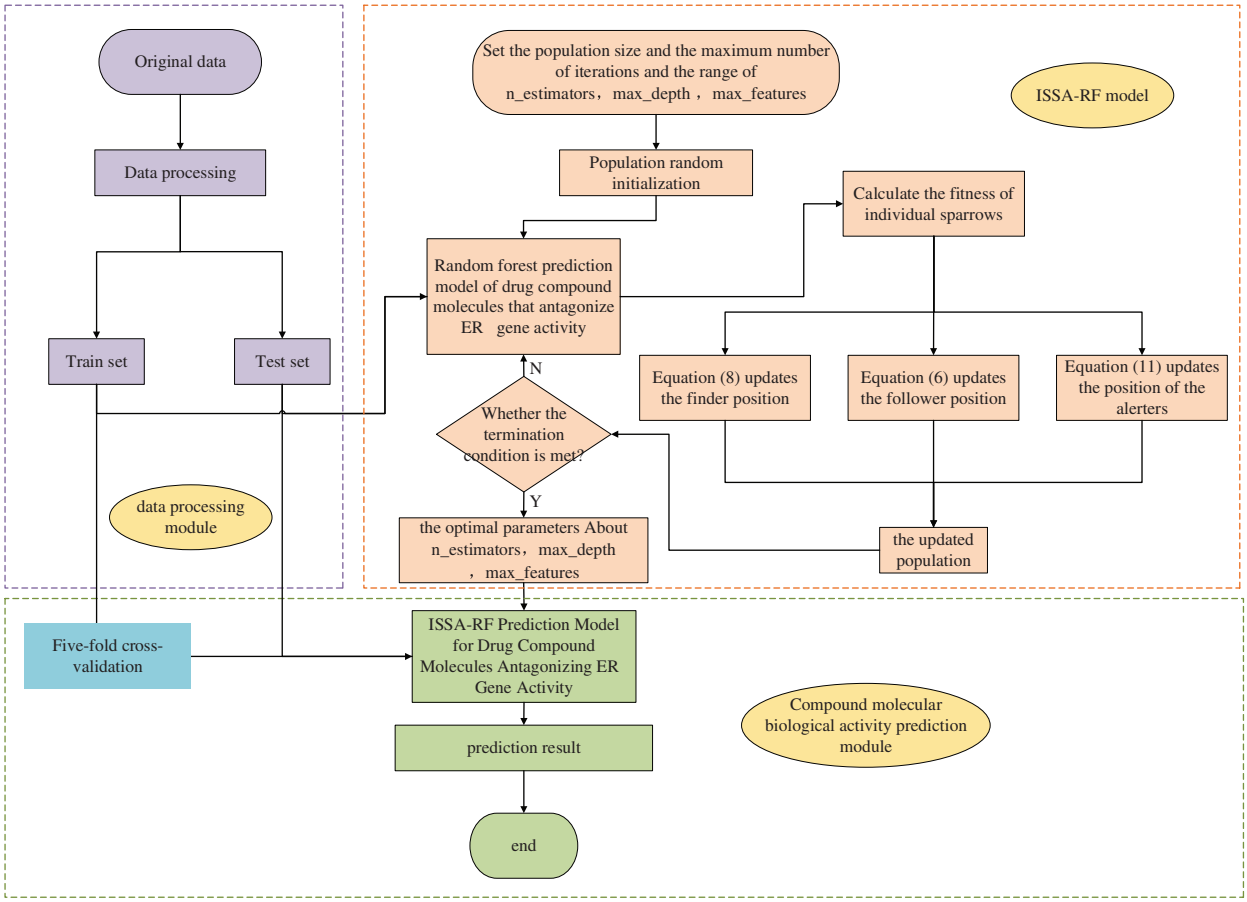


Figure 8: ISSA-RF model flowchart

4 Analysis of Results

4.1 Results of ISSA-RF Predictions

In this paper, 85% of the data is used as Train set, 15% of the data is used as Test set, the random seed parameter random_state is 0, and a 5-fold cross-validation model is used with R^2 as the target value (about 17% of the data). Dataset is divided with the selected top 20 to 100 highly influential drug compound variables, and put them into the ISSA-RF model for training, and finally predict the Test set through the RF model with the optimal parameters to obtain the final model effect. The optimal parameters can be seen in Table 2. model evaluation results can be seen in Table 3.

Table 2: Features_num corresponds to the optimal parameters for training

Features_num	n_estimators	Max_features	Max_depth
20	163	9	13
30	180	13	20
40	161	13	14
50	163	22	19
60	160	14	17
70	160	15	13
80	170	23	18
90	163	26	16
100	160	65	13

Table 3: Features_num corresponds to the evaluation index of training

Features_num	MAE	RMSE	MSE	MAPE	R^2	$Val–R^2$
20	0.5951410	0.7875591	0.6202494	0.0953065	0.6728614	0.6698001
30	0.5267269	0.7084049	0.5018375	0.0821251	0.7353155	0.7317090
40	0.5252527	0.7052240	0.4973409	0.0819899	0.7376871	0.7352390
50	0.5226297	0.7084105	0.5018454	0.0819970	0.7353113	0.7428036
60	0.5238194	0.7033877	0.4947543	0.0822189	0.7390513	0.7406314
70	0.5149397	0.6959719	0.4843770	0.0809725	0.7445247	0.7437330
80	0.5125422	0.6972933	0.4862180	0.0805919	0.7435537	0.7450803
90	0.5119813	0.6947606	0.4826924	0.0802745	0.7454132	0.7482465
100	0.4979396	0.6876389	0.4728473	0.0776072	0.7506058	0.7463062

As can be seen from Fig. 9, when features_num is increased from 20 to 30, the overall prediction effect is significantly improved. After that, the overall prediction effect is slowly improved with the increase of features_num, but the improvement effect is not obvious, which means that the number of features will increase the performance of model within a specific range. prediction accuracy. When features_num = 100, the effect is the best that RMSE is 0.6876389 and R^2 is 0.75 and $Val - R^2$ is 0.746. It can be seen from Fig. 9(d) that the Test set and cross-validation show the same line trend, and the Test set is slightly better than the cross-validation in the prediction results, which shows that the generalization ability of the ISSA-RF model is better in the antagonism Molecular prediction of drug compounds for ERα gene activity.

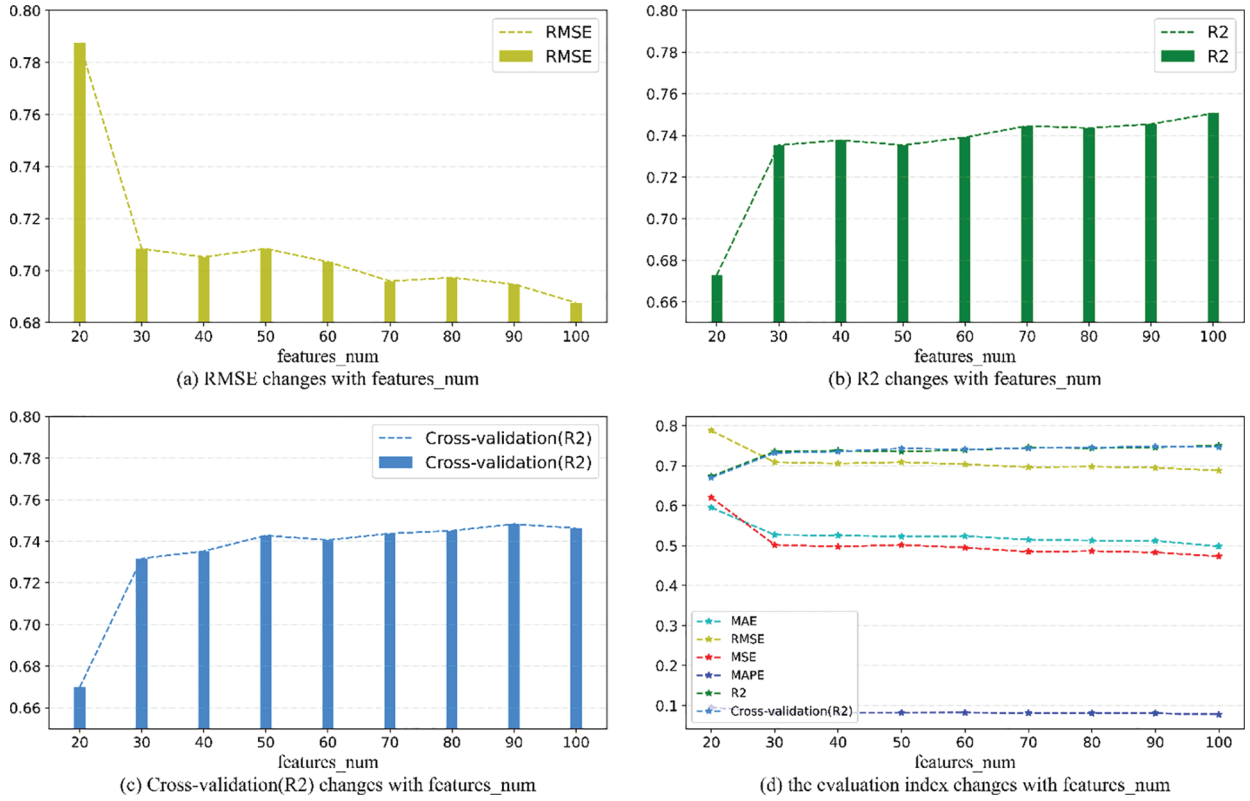


Figure 9: Evaluation metrics visualization

Table 4: features_num corresponds to the evaluation index of training

Algorithm	Features_num	MAE	RMSE	MSE	MAPE	R^2
GA_SVM	20	0.6547276	0.8719524	0.7603010	0.1048008	0.5989939
	50	0.5245315	0.7275946	0.5293939	0.0819348	0.7207813
	100	0.5165143	0.7256745	0.5266035	0.0813428	0.7222531
BP	20	0.7134478	0.9126131	0.8328628	0.1132963	0.5607226
	50	0.5876721	0.7753266	0.6011314	0.0937422	0.6829448
	100	0.5635194	0.7444502	0.5542061	0.0884355	0.7076947
XGBoost	20	0.7546805	0.9489264	0.9004613	0.1116271	0.5250690
	50	0.7423109	0.9507362	0.9038994	0.1083279	0.5232557
	100	0.7111056	0.9167204	0.8403764	0.1037813	0.5567597
ISSA_RF	20	0.5951410	0.7875591	0.6202494	0.0953065	0.6728614
	50	0.5226297	0.7084105	0.5018454	0.0819970	0.7353113
	100	0.4979396	0.6876389	0.4728473	0.0776072	0.7506058

4.2 Model Comparison

In order to verify the accuracy of the quantitative prediction model for ERα biological activity constructed by ISSA-RF model, GA_SVM, BP and XGBoost were introduced in this paper to predict the biological activity of ERα. Under the same experimental conditions, the three models are trained using the top 20, 50, and 100 molecular descriptors with a high degree of influence. The model evaluation is shown in Table 4. The comparison between the predicted results and the actual value results is shown in Fig. 10.

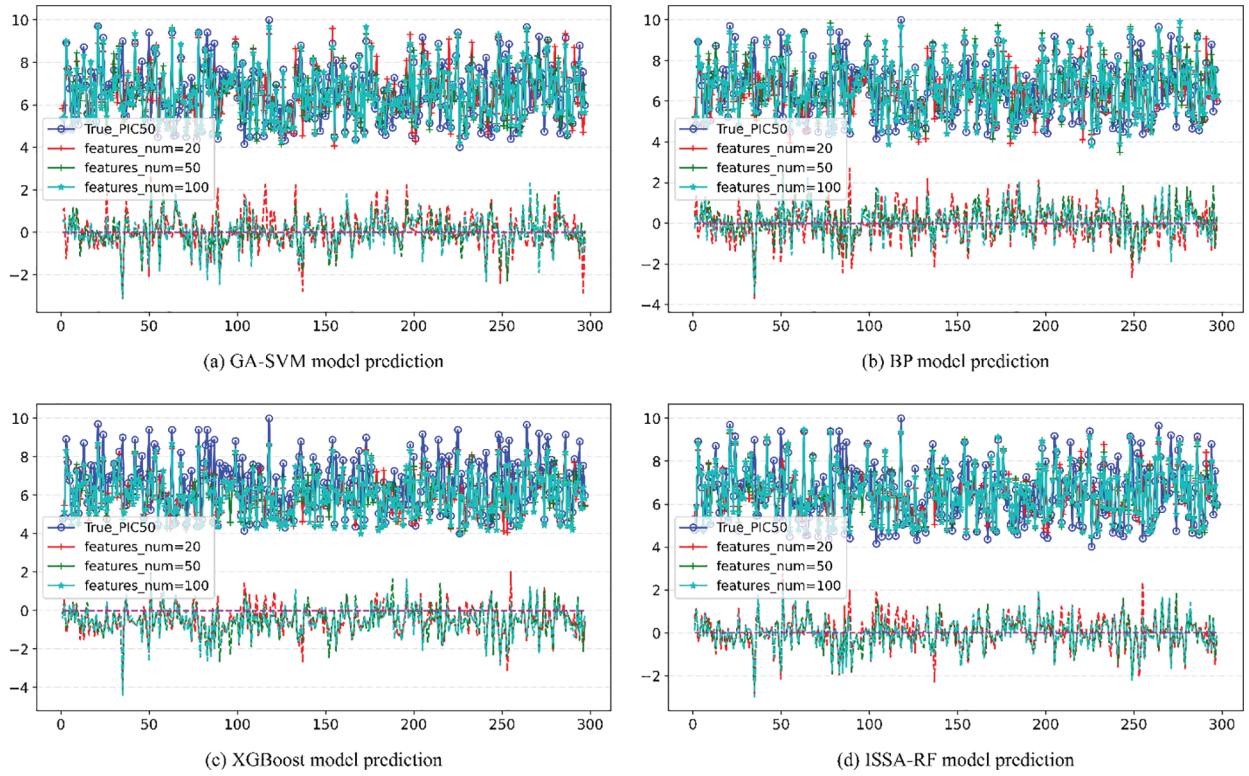


Figure 10: Comparison of pIC50 predicted value and true value of other models

The upper part of each image in Fig. 10 is the line of comparison between the predicted value and actual value of pIC50 of the model under different number of features, and the lower part is the difference between the predicted value and actual value of pIC50 of the model under different number of features Error line. According to the comparison figure, it can be seen that the predicted value of pIC50 of the model established by ISSA-RF algorithm is in good agreement with the actual value. After comparing the model evaluation results, it can be seen that different algorithm models achieve the best results when features_num = 100. The detail image when features = 100 is shown in Fig. 11. Under the same experimental conditions, the ISSA-RF model proposed in this paper R^2 is 3%, 5%, 20% higher than the other three models, and RMSE is improved by 5%, 7.6%, 24.9%. This shows that the ISSA-RF prediction model has a satisfactory effect on the quantitative prediction of the biological activity of drug compounds on ERα.

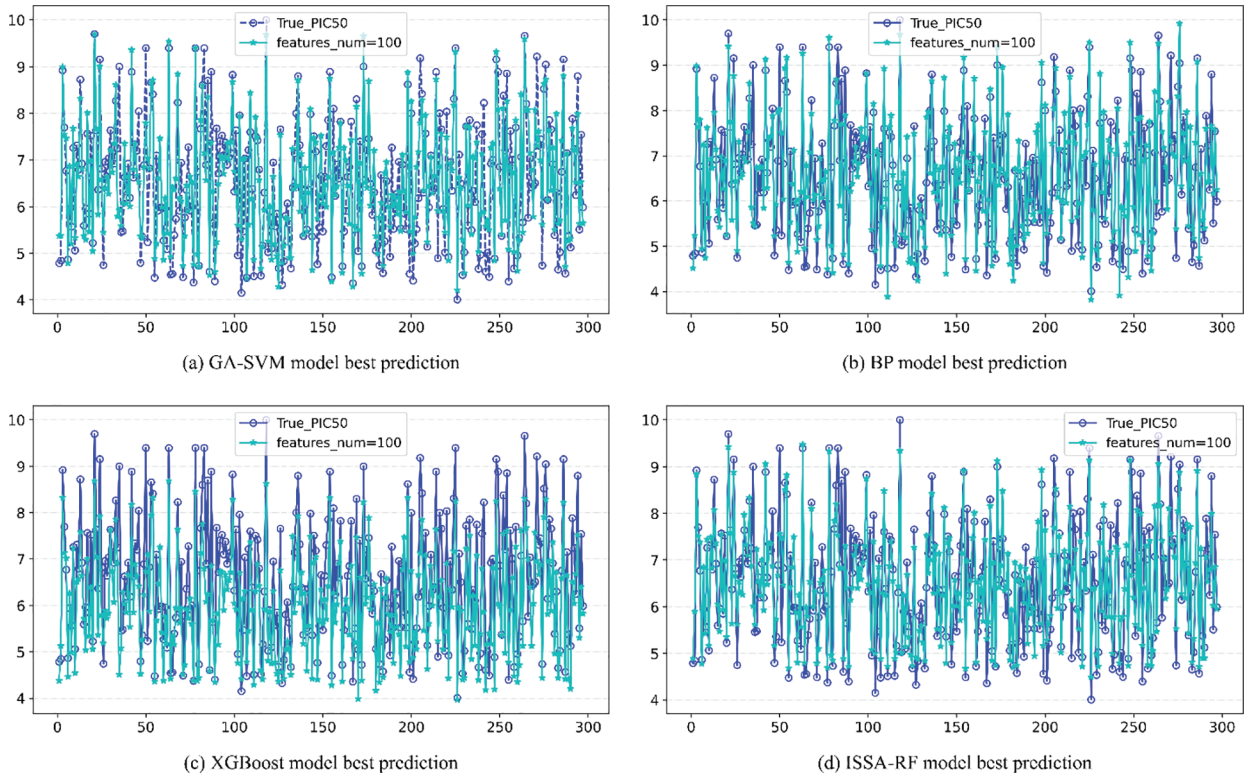


Figure 11: Comparison of the actual value and the predicted value under the optimal evaluation index

5 Conclusion

In the practice of selecting anti-breast cancer drug candidates, it is usually necessary to analyze the structure-activity relationship between compound activity data and compound molecular descriptors, and select compound molecules that satisfy biological activity as drug candidates. In this paper, a RF model optimized by the Improved Sparrow Search Algorithm (ISSA-RF) is proposed. We added an adaptive weight formula to the sparrow finder position update formula to optimize the search range and speed of sparrows in different stages, and proved that the SSA algorithm with adaptive weights has better fitness training accuracy on the Rosenbrock function than the ordinary SSA algorithm. This paper uses multi-scale molecular descriptors for model training to reduce the chance of model training accuracy caused by the number of different molecular descriptors. In addition, this paper limits the search range of ISSA-RF through RF separate training. The main purpose of this is to avoid the problem of the ISSA algorithm falling into local optimality. Secondly, this can also greatly reduce the search time of sparrows and improve the efficiency of model optimization. Finally, the prediction effect is compared with a variety of common models to verify the accuracy of the ISSA-RF model. The experimental results show that compared with other models, the ISSA-RF algorithm model proposed in this paper has a lower RMSE in the prediction of the biological activity of drug compounds on ERα, and can accurately predict the biological activity according to the molecular descriptors of the compounds, which improves the accuracy and efficiency of anti-breast cancer drug candidate screening. In addition, this model can not only be used to screen anti-breast cancer drug candidates, but also provides new ideas for constructing quantitative structure-activity relationship models of compounds.

Author Contributions: Minxi Rong, Xiaoli Guo contributed to the conception of the study; Yong Li performed the experiment and contributed significantly to manuscript preparation; Tao Zong, Zhiyuan Ma and Penglei Li helped perform the analysis with constructive discussions.

Ethics Approval and Informed Consent Statement: The datasets used in this article is a public data set from the DrugBank drug molecule database of the University of Alberta, and the data set is used as competition data in the China 2021 "Huawei Cup" Mathematical Modeling Competition, so the datasets do not involve Ethical Approval and Informed Consent Statement.

Availability of Data and Materials: The datasets used or analyzed during the current study have been posted to github website (https://github.com/Li519445444/candidate-drug-data-source/tree/master).

Acknowledgement: The authors thank the National Natural Science Foundation of China (11601491). Thanks to China's 2021 "Huawei Cup" Mathematical Modeling Competition for offering the data.

Funding Statement: This research was supported by the National Natural Science Foundation of China (11601491).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Sohn, E. (2017). Environment: Hothouse of disease. *Nature, 543(7647),* S44–S46. DOI 10.1038/543S44a.
2. Tan, L. L., Zhang, X. X., Zhou, Y. Z. (2021). Prediction of molecular biological activity based on graph convolution method of multi-characteristic fusion. *Journal of University of Electronic Science and Technology of China, 50(6),* 921–929. DOI 10.12178/1001-0548.2021158.
3. Loibl, S., Poortmans, P., Morrow, M., Denkert, C., Curigliano, G. (2021). Breast cancer. *The Lancet, 397,* 1750–1769. DOI 10.1016/S0140-6736(20)32381-3.
4. Cong, Y., Xue, Y. (2013). Quantitative structure-activity relationship study of the non-nucleoside inhibitors of HCV NS5B polymerase by machine learning methods. *Acta Physico-Chimica Sinica, 29(8),* 1639–1647. DOI 10.3866/PKU.WHXB201305171.
5. Deo, R. C. (2015). Machine learning in medicine. *Circulation, 132(20),* 1920–1930. DOI 10.1161/CIRCULATIONAHA.115.001593.
6. Jiang, D., Lei, T., Wang, Z., Shen, C., Cao, D. et al. (2020). ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning. *Journal of Cheminformatics, 12(1),* 16. DOI 10.1186/s13321-020-00421-y.
7. Che, H. X., Wang, T., Wang, W. (2021). Comparing prediction models for prostate cancer. *Data Analysis and Knowledge Discovery, 5(9),* 107–114. DOI 10.11925/infotech.2096-3467.2020.1185.
8. Wang, H., Cui, W., Guo, Y., Du, Y., Zhou, Y. (2021). Identifying pathogens of foodborne diseases with machine learning. *Data Analysis and Knowledge Discovery, 5(9),* 54–62. DOI 10.11925/infotech.2096-3467.2020.1105.
9. Lu, W., Xue, Y., Meng, Q. W. (2013). Classification prediction of inhibitors of H1N1 neuraminidase by machine learning methods. *Acta Physico-Chimica Sinica, 29(1*). DOI 10.3866/PKU.WHXB201211122.
10. Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling, 56(12),* 2353–2360. DOI 10.1021/acs.jcim.6b00591.
11. Mansouri, K., Cariello, N. F., Korotcov, A., Tkachenko, V., Grulke, C. M. et al. (2019). Open-source QSAR models for pKa prediction using multiple machine learning approaches. *Journal of Cheminformatics, 11(1),* 294. DOI 10.1186/s13321-019-0384-1.
12. Ding, L., Zhang, X. Y., Wu, D. Y., Liu, M. L. (2021). Application of an extreme learning machine network with particle swarm optimization in syndrome classification of primary liver cancer. *Journal of Integrative Medicine, 19(5),* 395–407. DOI 10.1016/j.joim.2021.08.001.

13. Zhou, C. M., Xue, Q., Liu, P. M., Duan, W., Wang, Y. et al. (2021). Construction of a predictive model of post-intubation hypotension in critically ill patients using multiple machine learning classifiers. *Journal of Clinical Anesthesia, 72,* 110279. DOI 10.1016/j.jclinane.2021.110279.
14. Luo, Y., Song, Y. L., Shang, J. L., Wang, L. (2019). Prediction of PI3K inhibitors based on naive bayesian machine learning. *Chinese Journal of New Drugs, 28(1),* 73–80.
15. Zheng, X. Z., Li, M. H., Zhang, Y. N., Jiang, P., Wang, B. Y. (2021). Research on the prediction model of coal spontaneous combustion temperature based on randem forest algorithm. *Industry and Mine Automation, 47(5),* 58–64. DOI 10.13272/j.issn.1671-251x.17700.
16. Xu, M. X., Zheng, Y., Li, Y. J., Wu, W. H. (2022). Prediction of properties of anti-breast cancer drugs based on PSO-BP neural network and PSO-SVM. *Journal of Nanjing University of Information Science & Technology,* 1–20.
17. Vilma, S. (2021). Interpretability of selected variables and performance comparison of variable selection methods in a polyethylene and polypropylene NIR classification task. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 258(8),* 119850. DOI 10.1016/J.SAA.2021.119850.
18. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58(1),* 267–288. DOI 10.1111/j.2517-6161.1996.tb02080.x.
19. Bonsignore, M., Trusso, S., de Pasquale, C., Ferlazzo, G., Allegra, A. et al. (2021). A multivariate analysis of Multiple Myeloma subtype plasma cells. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 258(9686),* 119813. DOI 10.1016/j.saa.2021.119813.
20. Xue, J. K., Shen, B. (2020). A novel swarm intelligence optimization approach: Sparrow search algorithm. *Systems Science & Control Engineering, 8(1),* 22–34. DOI 10.1080/21642583.2019.1708830.
21. Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5–32. DOI 10.1023/A:1010933404324.