frontiers | Frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 13 July 2022
doi: 10.3389/fbioe.2022.946329

MLP-Based Regression Prediction Model For Compound Bioactivity

Yongfei Qin[1], *Chao Li*[1], *Xia Shi*[1] *and Weigang Wang*[1,2]*

[1]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China, [2]Collaborative Innovation Center of Statistical Data Engineering, Technology and Application, Zhejiang Gongshang University, Hangzhou, China

The development of breast cancer is closely linked to the estrogen receptor ERα, which is also considered to be an important target for the treatment of breast cancer. Therefore, compounds that can antagonize ERα activity may be drug candidates for the treatment of breast cancer. In drug development, to save manpower and resources, potential active compounds are often screened by establishing compound activity prediction model. For the 1974 compounds collected, the top 20 molecular descriptors that significantly affected the biological activity were screened using LASSO regression models combined with 10-fold cross-validation method. Further, a regression prediction model based on the MLP fully connected neural network was constructed to predict the bioactivity values of 50 new compounds. To measure the validity of the model, the model loss term was specified as the mean squared error (MSE). The results showed that the MLP-based regression prediction model had a loss value of 0.0146 on the validation set. This model is therefore well trained and the prediction strategy used is valid. The methods developed by this paper may provide a reference for the development of anti-breast cancer drugs.

Keywords: breast cancer drug candidates, biological activity, LASSO regression, MLP, neural

OPEN ACCESS

Edited by:
Zhihua Cui,
Taiyuan University of Science and Technology, China

Reviewed by:
Zhenlong Gao,
Qufu Normal University, China
Lv Ping,
Hangzhou Normal University, China
Guangyu Yang,
Zhengzhou University, China
Sisir Nandi,
Uttarakhand Technical University, India

***Correspondence:**
Weigang Wang
wangweigang@zjgsu.edu.cn

Specialty section:
This article was submitted to Bionics and Biomimetics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *17 May 2022*
Accepted: *23 June 2022*
Published: *13 July 2022*

Citation:
Qin Y, Li C, Shi X and Wang W (2022) MLP-Based Regression Prediction Model For Compound Bioactivity. Front. Bioeng. Biotechnol. 10:946329. doi: 10.3389/fbioe.2022.946329

1 INTRODUCTION

Cancer is an important global public health problem. Breast cancer is a malignant tumor that occurs in breast epithelial tissue. Breast cancer is one of the most common and deadly cancers in the world. Through the research of Batyrova et al. (2021). (2021), with the progress of society, economic development and environmental changes, the disease burden of breast cancer is gradually increasing. In 2021, global cancer data showed that the number of new cases of breast cancer in the world was as high as 2.26 million, surpassing the 2.2 million cases of lung cancer, which replaced lung cancer as the world's leading cancer. Breast cancer causes serious damage to women's life and health. 1 in 12 women will get breast cancer. Breast cancer is one of the most common malignant tumors in women, accounting for 7%–10% of female systemic malignant tumors. In recent years, breast cancer ranks first in the number of new cancer cases and deaths among women worldwide, enough to see that breast cancer has posed a serious threat to the health of women around the world. Although global medical care has improved since the early years, the pattern of breast cancer is not promising. As a chronic disease, it is not only a serious threat to the life and health of women all over the world, but also brings a burden in all aspects. Therefore, improving the cure rate of breast cancer is of great significance to women's health.

With the rapid development of modern medical imaging technology and the deepening of biological research on breast cancer, the rate of early diagnosis of breast cancer has been increasing, which is of great significance for the early detection and timely treatment of breast cancer, as well as

improving the five-year survival rate of patients. With the research on the biology of breast cancer, in addition to the early diagnosis of breast cancer, improving the cure rate of breast cancer and selecting more effective drugs for treatment are of great significance to medical research. Chemotherapy has been used as one of the main tools in the treatment of cancer (2021) (Cheng et al., 2021). There are many drugs available for the treatment, but most of these drugs are quite toxic, which may shorten the life span of cancer patients. Therefore, there is an urgent need to predict new compounds with minimal toxicity and strong biological activity through QSAR and ligand-based design (2016, 2022) (Nandi and Bagchi, 2016; Nandi et al., 2022). Today, with the dramatic increase in the number of drugs, in order to save time and cost, the most economical and reasonable way to research is to use the computer-aided artificial intelligence algorithm to predict and analyze the biological activity of drugs through the establishment of prediction models. Drug activity prediction is performed by collecting data on a range of compounds that act on breast cancer-related targets and their biological activity. The specific approach is that, for a disease-related target, collect a series of compounds acting on the target and their bioactivity data. Then, with a series of molecular structure descriptors as independent variables and the bioactivity value of compounds as dependent variables, a prediction model is constructed to predict and guide the structural optimization of existing active compounds. Zhao et al. (2021). (2021) have shown that FBXO15 plays a critical inhibitory function in regulating breast cancer progression. Some studies have found that Estrogen Receptor Alpha (ERα) is expressed in no more than 10% of normal mammary epithelial cells, but about 50%–80% of mammary tumor cells. In mice with ERα deletion, it was found that ERα plays an important role in breast development. Currently, anti-hormone therapy is commonly used in breast cancer patients with ERα expression, which regulates estrogen receptor activity to control estrogen levels in the body. Therefore, ERα is considered an important target for the treatment of breast cancer, and compounds that can antagonize ERα activity may be candidates for the treatment of breast cancer. For example, tamoxifen and raloxifene, the classic drugs for clinical treatment of breast cancer, are ERα antagonists whereas BSC-pyrazole and MPP can fully antagonize E2 stimulation of pS2 mRNA in MCF-7 breast cancer cells.

Based on the above, in this paper, an anti-breast cancer drug candidate screening model was established. We obtained bioactivity data of 1974 compounds targeting ERα for breast cancer treatment. Next, we downscale the data. Based on the data collected, a variable screening was first performed and 20 molecular descriptors with significant effects on biological activity were selected from more than seven hundred molecular descriptors. Finally, a regression prediction model was established based on MLP fully connected neural network to predict the bioactivity values of 50 new compounds. The main contributions of this paper are as follows:

- Several commonly used dimensionality reduction algorithms are compared in high-dimensional biomedical data screening, and the LASSO algorithm is selected to reduce the dimensionality of high-dimensional biomedical data, which reduces 729 dimensions to 20 dimensions and retains more information of the original variables, And the selected 20 molecular description variables cover most description types, accounting for a suitable proportion and representative. It greatly reduces the workload in the exploration of drug molecules and thus improves the efficiency of the preparation of new drug compounds.
- A multilayer perceptron (MLP) prediction model was used to predict the biological activity of the compounds, and the model prediction results are closer to the real values than other methods. The loss value of the REGRESSION prediction model based on MLP was 0.0146 in the validation set, indicating that the model had a good prediction effect. This is of great significance for the development of anti-breast cancer drugs.

2 RELATED WORK

2.1 High-Dimensional Biomedical Data Variable Screening Methods

In recent years, with the development of biomedical testing technology, the amount of biomedical data accumulated in the research process has increased exponentially. In conventional statistical analysis, when conducting multivariate analysis, a certain ratio of the number of independent variables to the sample size is often required. However, the cumulative amount of biomedical data has resulted in a much larger number of independent variables than required, making it difficult to establish the correct association between the dependent and independent variables. In response to these problems, filtering out more critical variables from high-dimensional data has become an urgent problem for researchers.

Tibshirani et al., 1996 (1996) proposed the LASSO (Least Absolute Shrinkage and Selection Operator) estimate inspired by the penalty function. This method can compress certain constituents to zero within a reasonable choice of penalty parameters to achieve variable selection and perform parameter estimation. Then new methods of variable selection and contraction based on LASSO were gradually proposed, and other scholars also started to study and improve the LASSO method in depth. Fan and Li, 2001. (2001) proposed a penalized likelihood method called SCAD and proposed an optimization algorithm for this penalized likelihood function to address the problem that stepwise regression can be computationally costly and neglects the random errors that arise during model selection. Jolliffe et al. (2003). (2003) proposed an improved principal component method based on LASSO. Fonti and Belitser, 2017. (2017) explained and discussed how to use LASSO for feature selection. Efron et al. (2004). proposed a minimum angle regression algorithm to implement LASSO. Since the computational effort of this algorithm is the same as that of the least squares method, this has led to the widespread use of the LASSO regression method. Since the LASSO method is inconsistent with the variable selection of some cases, Zou

et al., 2006 (2006) proposed the adaptive LASSO method to overcome this problem of LASSO. Adaptive LASSO has Oracle properties and its performance remains constant under certain canonical conditions when extending it to generalized linear models. Zou et al. (2006). (2006) used LASSO to give modified principal components with sparse loadings for enhancing the interpretability of the principal components. For some strongly correlated data, such as medical data, biological data, etc., LASSO and Adaptive LASSO cannot select all highly correlated variables into the model.

Zou and Hastie, 2005. (2005) proposed the Elastic Net method by improving the LASSO method, which is a new regularization and variable selection method that can handle the complex collinearity problems in covariate. Zou and Zhang, 2009 (2009) further proposed the Adaptive Elastic-Net (AEN) method, which weights each regression coefficient of the primary penalty term so that it has the properties of both the adaptive LASSO and Elastic Net methods. Li et al., 2017. (2017) applied the Elastic Net approach to variable selection in a balanced longitudinal data model and demonstrated the compatibility and group effect properties of the approach. Lin and Yang, 2019. (2019) extended non-negative adaptive Elastic Net estimation to a high-dimensional sparse linear model and proved its Oracle property and validity under some canonical conditions with valid samples. Existing studies have demonstrated the good nature of LASSO and Elastic Net for variable screening. Yamada et al. (2014). (2014) verified the effectiveness of LASSO method through the feature selection experiment of classification and regression of thousands of features. Muthukrishnan and Rohini, 2016. (2016) explored the features of the popular regression methods, OLS regression, ridge regression and the LASSO regression. Zhang et al. (2021). (2021) used LASSO dimensionality reduction method to conduct experiments on the combination of feature sub models to obtain the best top-level feature number, thus providing support for the effective prediction of DNA binding proteins. In this paper, the original 729 variables were screened using the LASSO method according to the research objectives.

2.2 Biological Activity Prediction Model

Machine learning algorithms are widely used in biological activity prediction. Scholar Jia et al., 2019 (2019) used the RF algorithm combined with ten-fold cross-validation to build a quantitative prediction model for drug targets. The MSE of the RF model constructed on the EC50 validation set and test set were both less than 0.09, and the R2 was greater than 0.96; in the KD data set, the MSE is less than 0.12, and the R2 is greater than 0.94. Lv and Xue, 2011. (2011) tested the classification prediction model of SVM in hepatitis virus NS5B protease inhibitor and non-inhibitor, which had high model calculation efficiency and prediction accuracy. He and Zhu, 2020. (2020) modeled the drug target prediction problem with reference to the recommendation system, and made improvements to the traditional algorithm, by adding drugs and targets double regularization to improve the accuracy of the model. Based on different application scenarios, the advantages of various algorithms are different. The K-Nearest Neighbor algorithm performs well in the classification of known Hepatitis C Virus NS5B protease inhibitors and non-inhibitors. Random forest is suitable for quantitative prediction of drug EC50 and CD value. Liu et al. (2013). (2013) proved Support Vector Machine is good at processing the prediction of drug-target BuChE inhibitors and non-inhibitors. Collaborative filtering can solve the problem of sparse matrix by the mixing weighted drug-target association matrix, so it is suitable for the recommendation of "drug-target" and the prediction of "drug-target" interaction (2017) (Zhang, 2017). The artificial neural network is suitable for predicting the inhibitory activity of drug molecules against p38R MAPase (2019) (Abdolmaleki and Ghasemi, 2019).

The neural network model originated from the mathematical model of MP neurons proposed by McCulloch and Pitts in 1943. MLP model is a popular and practical one among many neural network models. Pinkus et al., 1999. (1999) discussed various approximate theoretical problems in the MLP model of neural networks. Then, Rosenblatt et al., 1958 (2017) proposed a single-layer perceptron model, but it could only distinguish basic shapes such as triangles and squares. In 1986, the second-generation neural network was proposed (2016) (Jiao et al., 2016), which introduced the Sigmoid function as the activation function, and used multiple hidden layers instead of the original single fixed feature layer to solve the nonlinear classification problem. The back-propagation algorithm was used to train the model, and the back-propagation algorithm and its improvement still play an important role in model training so far. Multilayer perceptron (MLP), also known as the feedforward neural networks, has been widely used in research. Gao et al. (2019). (2019) constructed a multilayer neural network model to accurately assess the risk of non-contact sports injuries in rugby. Xiao et al. (2021). (2021) constructed and verified the prediction model of occupational coal worker's pneumoconiosis (CWP) incidence based on multilayer perceptron neural network, and discussed its application value in the prediction of CWP incidence. Generally speaking, MLP falls into two categories: supervised and unsupervised (2021) (Dua et al., 2021). The trainer is responsible for training the MLP. It has the highest performance for new input datasets. The research in this paper is to predict the activity of new compound molecules with better biological activity, so it is suitable to use artificial neural network for prediction. In order to further optimize the effect of improving prediction, MLP neural network is applied.

3 METHODS

3.1 The LASSO Algorithm

LASSO (Least Absolute Shrinkage and Selection Operator) is a variable selection method proposed by the statistician Tibshirani. The basic idea of which is to introduce a penalty factor to the ordinary least squares estimation to penalise or constrain the estimator β by the L_1 norm. For a data set with n predictor-

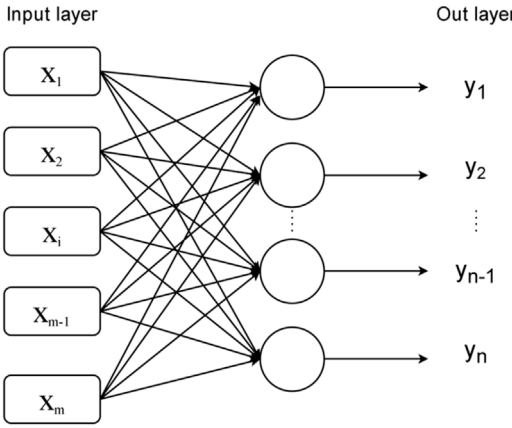

FIGURE 1 | Perceptron model.

response variable pairs $\{(x_i, y_i)\}_{(i=1)}^{n}$, LASSO seeks an estimate $\hat{\beta}$ that fits the data better by minimizing the RSS $(\hat{\beta})$.

$$RSS(\hat{\beta}) = \underbrace{\arg\min}_{\hat{\beta}} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \quad (1)$$

Where $\lambda \geq 0$ is a penalized parameter, $\lambda\sum_{j=1}^{p}|\beta_j|$ is a compression penalty.

$$\hat{\beta} = \arg\min\{\|y - x\beta\|_2^2\} \qquad s.t.\|\beta\| \leq t \quad (2)$$

$t \geq 0$ is a turning parameter, which controls the intensity of compression. If the parameter obtained by ordinary least squares estimation is denoted as $\hat{\beta}^0$, then LASSO achieves compression, as long as $t < \sum_{j=1}^{p}\|\hat{\beta}_j^0\|$. And for some models with small absolute values, the coefficients can be compressed to zero. Thus, the inequality $\|\beta\| \leq t$ can effectively constrain the parameter space and allow the final model to be well interpreted.

3.2 The Perceptron Model

The single-layer perceptron model was the first and simplest neural network model to be proposed.

As shown in the **Figure 1**, it has two layers, the input layer and the output layer, which are each composed of multiple ANs and are interconnected.

In terms of the operation principle, the single-layer perceptron first receives input data $(x_1, x_2, ..., x_m)$ at the input layer, and then accumulates the input data with their respective weights, That is, for the Kth output layer neuron, the total amount of information it receives is $\sum_{i=1}^{n} w_{ik}x_i, k = 1, 2, ...n$. So the output of the Kth output layer neuron is:

$$y_k = f\left(\sum_{i=1}^{m} w_{ik}x_i - \theta_k\right) \quad (3)$$

The model of a single-layer perceptron actually constructs some hyperplanes to separate different data sets in a high-dimensional space. So it can handle linearly separable problems better, but it is helpless for linearly indistinguishable data sets. In order to further improve the learning ability of neural networks to solve more practical problems, the multilayer perceptron model, which evolved from the single-layer perceptron, was naturally created.

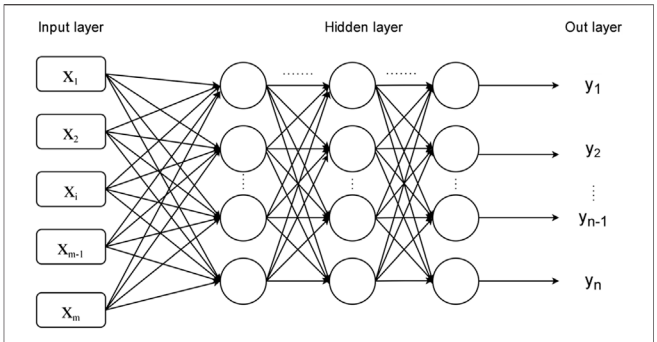

FIGURE 2 | Multilayer perceptron model.

As shown in the **Figure 2**, in addition to the input layer and output layer, the multilayer perceptron model also includes a hidden layer, which is also composed of multiple layers of artificial neurons, and each layer is fully connected to each other. In 1991, Kurt Hornik proposed the Universal Approximation Theorem, which states that if the hidden layer can consist of any number of artificial neurons, then a feedforward neural network containing a hidden layer can approximate any continuous function in the real number range. This theorem actually tells us that a multilayer perceptron neural network model with hidden layers can solve more problems than a single layer perceptron model. Due to the large number of data indicators and high dimensions involved in this paper, in order to make the prediction model in this paper have higher prediction accuracy, the multilayer perceptron model is chosen to construct the prediction model.

4 EXPERIMENTS AND RESULTS

4.1 Data Set and Data Preprocessing

For the breast cancer therapeutic target ERα, the required data were obtained from the University of Alberta's DrugBank database of drug molecules, which is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information for the study of drug mechanisms and even the exploration of novel drugs. The data used in this paper contains a total of 1974 compounds and the corresponding bioactivity values for ERα, $IC_{50,}$ and pIC_{50}. And the properties of each compound are co-represented by 729 molecular descriptors, which will also be added as independent variables in the Quantitative Structure-Activity Relationship (QSAR) model. In particular, pIC_{50} is the negative logarithm of the IC50 value, which usually has a positive correlation with biological activity, i.e., a higher pIC_{50} value indicates higher biological activity. The pIC_{50} is generally used to represent biological activity values in practical QSAR modelling.

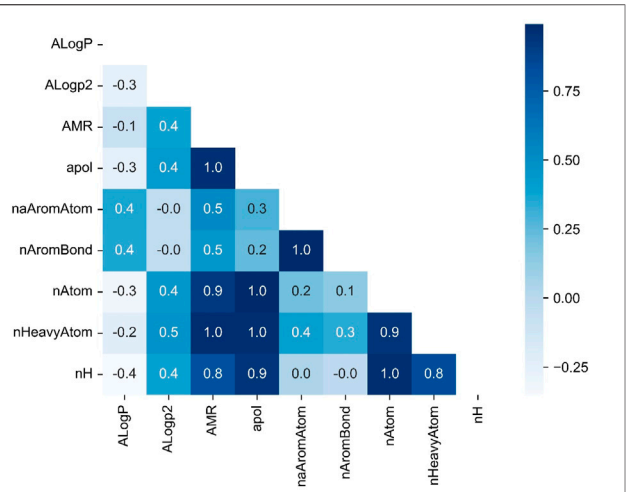

FIGURE 3 | Diagonal matrix plot of correlation coefficients (10 variables).

As the data collected for the molecular descriptors are two-dimensional, i.e., information corresponding to the solubility and surface area of the molecule. Therefore, certain characteristics that have the greatest impact on the results need to be filtered out as feature data when building the model. Based on the data collected, the importance of bioactivity values of compounds was ranked according to molecular descriptors (independent variables) to achieve the purpose of variable selection. This means that we need to find a metric to compare the importance of the variables cross-sectionally.

Due to the large number of variables, in order to explore whether there is multicollinearity among 729 variables, this paper analyzed the correlation between molecular descriptor variables by solving the correlation coefficient matrix. After arriving at a judgement based on the correlation matrix, an appropriate regression model will be built to describe the importance of the variables, thus enabling the screening of the molecular descriptors.

The first step was data pre-processing, reading the molecular descriptor data as the independent variable and the pIC50 bioactivity value data of ERα as the dependent variable. And the results of data missing judgment indicated that there were no data set missing, and the data can be directly normalized by Z-score.

In the second step, the Pearson correlation coefficient ρ was used to calculate the correlation coefficients between the respective variables. And the correlations between the molecular descriptor variables will be analyzed using the correlation coefficient matrix.

$$\rho_{XY} = \frac{COV(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (4)$$

Where $Cov(X,Y)$ is the covariance of X and Y, and $D(X)$, $D(Y)$ denote the variance of X, Y respectively. The correlation coefficients between 729 variables can be calculated according to this formula.

The diagonal matrix of the correlation coefficients between 10 arbitrarily selected variables is shown in the **Figure 3**. From the figure above, it can be obtained that there is a close correlation between some of the molecular descriptor variables. For example, the correlation coefficients between nAtom and AMR, apol reached 0.9 and 1.0 respectively, indicating the existence of multicollinearity when using this data set for regression analysis.

4.2 Filtering Molecular Descriptors Based on the LASSO Algorithm

4.2.1 LASSO + 10-Fold Cross-Validation

The LASSO algorithm is used to obtain a more refined model by constructing a penalty function that compresses the coefficients of less relevant variables to zero and removes these characteristic variables. The LASSO algorithm performs variable screening while fitting a generalized linear model. This can effectively avoid the over-fitting problem caused by too many variables and also help to solve the multicollinearity problem in the correlation analysis results of the previous step. Based on the LASSO algorithm, the molecular descriptor data were entered into the R software and the number of calculations was set to 1,000. The model stopped at 919 runs and the algorithm converged to the optimal solution, which contained 345 non-zero coefficients, i.e., 384 variables were initially eliminated, corresponding to a regularization parameter λ value of 0.00011. The **Figure 4** shows the evolution of the model coefficients.

This paper continued to use k-fold cross-validation method in conjunction with the LASSO regression algorithm to increase the accuracy of the model fit. In the cross-validation stage of the model, the 10-fold cross-validation was performed by setting random number seeds. The mean squared error (MSE) was used as the target parameter to minimize the loss, so that the regularization parameter λ that minimize the loss could be filtered.

The vertical coordinate of the **Figure 5** shows the mean squared error of the model, which changes as the penalty increases. And the horizontal coordinate at the top of the graph shows the number of independent variables. The first dashed line shows the logarithm of λ when the mean squared error is smallest, and the second dashed line shows the logarithm of λ when it is one standard error away from the minimum mean squared error. The two values are printed in the **Table 1**.

In general, the latter is chosen as the optimal value of λ. The model given by 1se has a better performance and has fewer independent variables. Since the objective of this section is to filter the independent variables, the 1se value is also chosen as the optimal solution. Therefore, the optimal λ value of the 10-fold cross-validated LASSO regression model was 0.003225, and the corresponding regression coefficients of each variable under this value were further obtained.

A total of 163 variables were retained in the 10-fold cross-validated LASSO regression model, which further eliminated 182 variables compared to the non-cross-validated model. The absolute values of the regression coefficients of the optimal model were then taken to reflect the importance of the

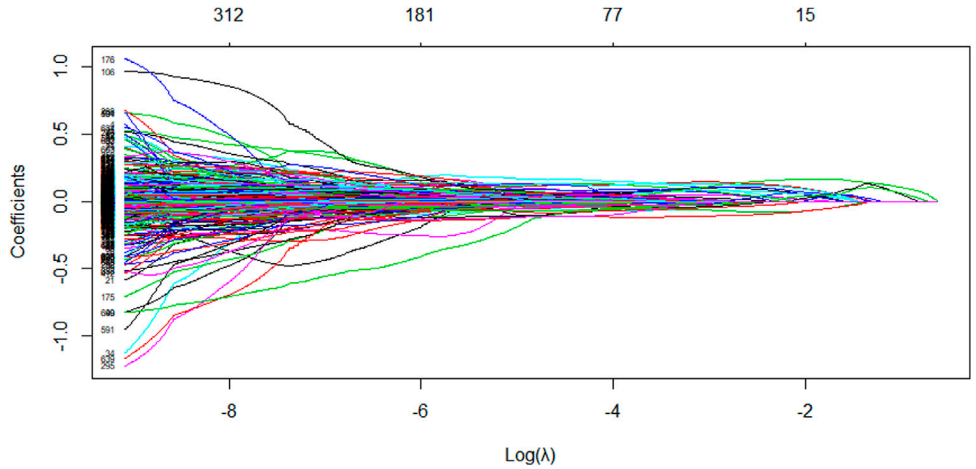

FIGURE 4 | Plot of the variation of the L1 parity against the regression coefficient.

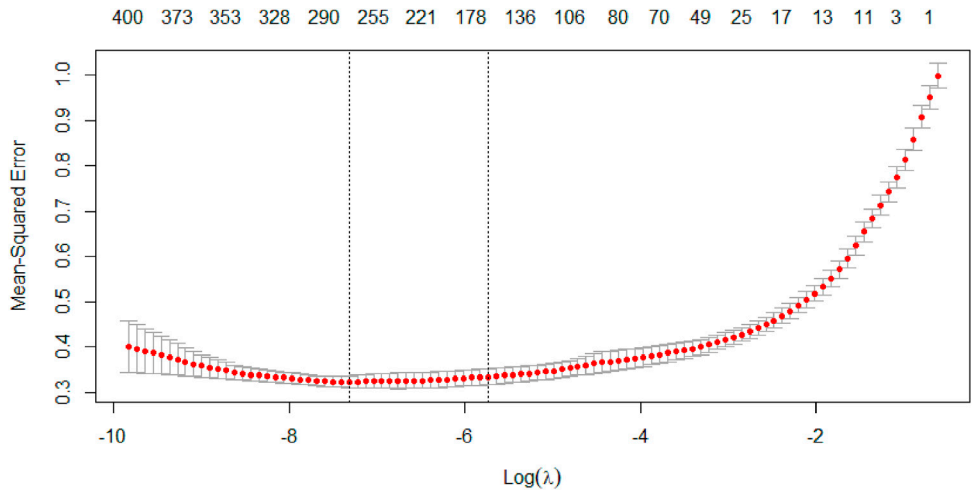

FIGURE 5 | MSE with *Log* (λ) curve.

TABLE 1 | Cross-validating lambda values.

Measure: Mean Squared error					
Lambda		**Index**	**Measure**	**SE**	**Nonzero**
Min	0.000663	73	0.3232	0.01272	281
1se	0.003225	56	0.3344	0.01698	163

variables on the pIC_{50} bioactivity values. The **Table 2** shows the top 20 molecular descriptors with the most significant effect on pIC_{50} bioactivity values.

4.2.2 Evaluation of Screening Results

This section evaluates the performance of the model by using two accuracy evaluation indicators, explainable deviation (%Dev) and mean squared error (MSE).

The **Figure 6**, with the number of independent variables indicated by the horizontal coordinate, shows the relationship between the explainable deviation and the variable coefficients for a 10-fold cross-validated LASSO regression model. The explainable deviation means the amount of sample information contained in the model, and the larger the value, the better. For example, with the number of independent variables exceeding 37, the model was able to explain 60% of the sample information. With the aforementioned λ value equal to 0.003225, the explainable deviation value of this model was 78.96, which indicates that the model has good model properties, and is able to explain 78.96% of the sample information. At the same time, the mean squared error of the model was very close to the minimum mean squared error, which can be concluded that the 10-fold cross-validated LASSO regression model constructed in this paper worked well and can achieve the purpose of screening the variables.

TABLE 2 | Molecular descriptors with significant effect on pIC_{50} bioactivity values (TOP20).

Variable	Coefficient	Abs_coeff
nHBAcc	−0.34921	0.349,211
SP-7	−0.25978	0.259,782
C3SP2	0.198,333	0.198,333
ndO	0.16484	0.16484
nsssCH	0.163,161	0.163,161
ECCEN	0.145,899	0.145,899
MLFER_A	0.140,239	0.140,239
mindO	−0.13936	0.139,355
ATSm5	0.138,853	0.138,853
minHAvin	0.137,933	0.137,933
MDEC-34	−0.12988	0.129,875
BCUTp-1h	0.129,423	0.129,423
SsF	0.123,461	0.123,461
mindsCH	−0.12217	0.122,166
C1SP3	0.117,622	0.117,622
maxHsOH	0.11567	0.11567
nHBint7	−0.11027	0.110,271
MDEN-22	0.099217	0.099217
ATSp5	−0.09359	0.093587
SHBint8	0.091518	0.091518

To further verify the rationality of the variables selected in the above model, the relationship between the corresponding type of the selected molecular descriptors and the corresponding type of all molecular descriptors is statistically shown in **Figure 7**.

In **Figure 7**, the blue one represents the categorical statistics of 729 variables and the orange one represents the categorical statistics of 20 variables screened by LASSO regression. Through searching, 46 types 727 variables were determined among 729 molecular descriptive variables. Because 50% of the descriptor types contain only one molecular descriptor, they are not shown in the graph. The category "Atom type electrotopological state" contains the largest number of descriptors, 488 in total. It is clear from the graph that the top 20 molecular descriptor variables screened by the LASSO regression cover most of the descriptor types and the proportion is appropriate, which can be considered that the variable screening is representative.

4.3 Quantitative Structure-Activity Relationship Model Prediction Analysis Based on MLP Neural Network

After dimensionality reduction of the molecular descriptor data, the top 20 molecular descriptor variables with significant influence on the bioactivity values were obtained. The information contained in these variables was then learned to find further relationships between compounds and ERα bioactivity values. In this section, the MLP neural network was trained to learn and the model relationships were used to predict the IC_{50} and pIC_{50} values of 50 new compounds.

4.3.1 Development of an MLP Neural Network Prediction Model for Biological Activity Values

MLP generalises the perceptron to some extent and its weakness of lacking the ability to recognise linear indivisible data is well overcome by MLP neural networks. In machine learning, there is generalized linear model algorithm, which can be formulated as:

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p \tag{5}$$

The MLP could be considered as a multi-layer processing algorithm for the generalized linear model. The linear weighted combination of the above formula needs to be carried out on each layer, and the final desired predicted value y can be obtained after n layers. Therefore, the MLP can be understood as a multi-layer processing model for the generalized linear model, which can be used for classification and regression prediction processing in the machine domain. We selected its function of regression prediction to make regression prediction of pIC_{50} value of the new compounds.

The **Figure 8** shows the network structure of the model constructed in this paper. The first layer is an input layer, the

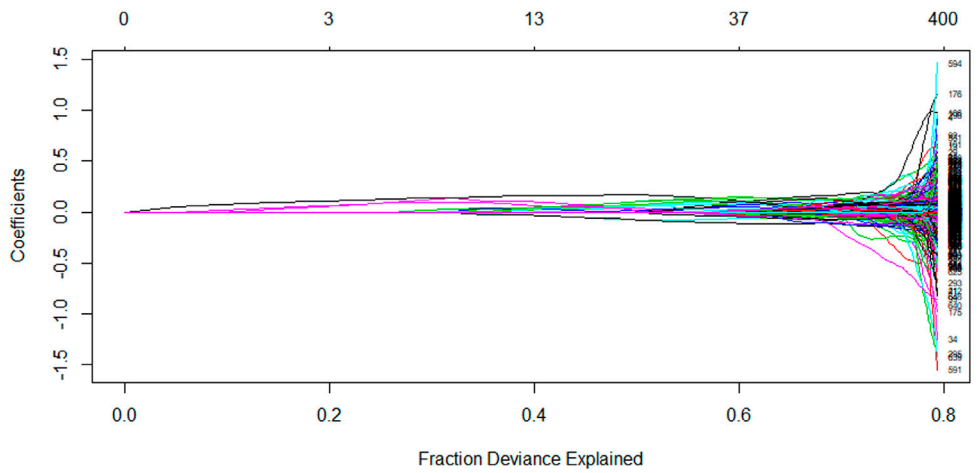

FIGURE 6 | Plot of model explainable deviation against variable coefficients.

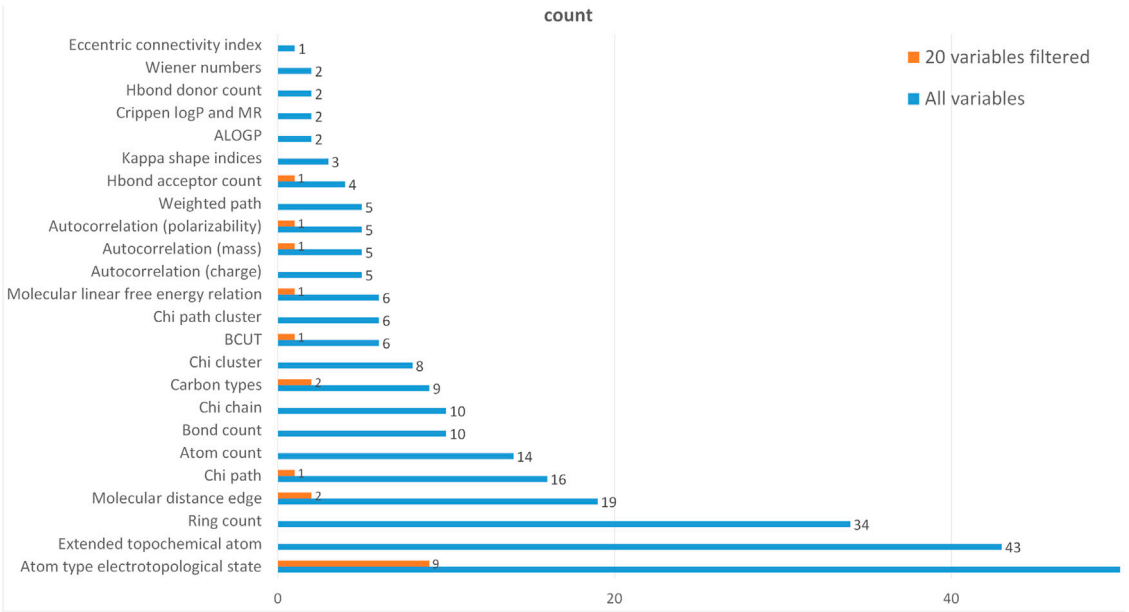

FIGURE 7 | Filter variable classification statistics.

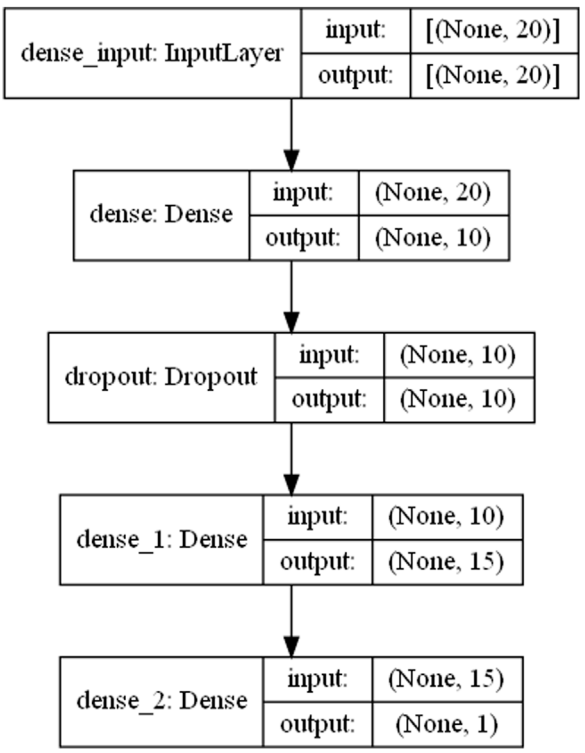

FIGURE 8 | Network structure diagram.

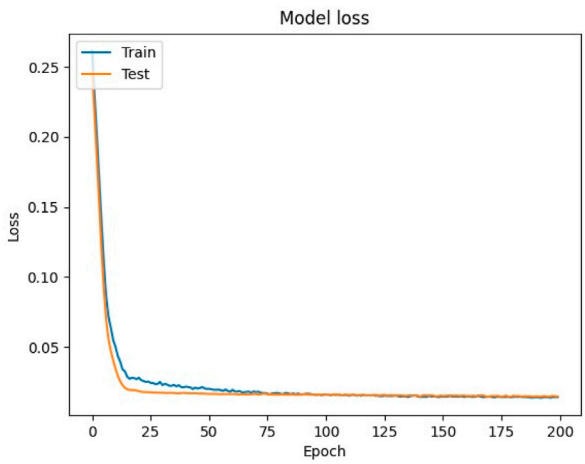

FIGURE 9 | Loss function diagram.

second layer is a fully connected layer with an excitation function of relu, and the third layer is a discard layer, which discards the neuron links with the fourth layer with a certain probability to prevent overfitting. In this model, the probability of discarding neuronal links was set to 0.2. The fourth layer is also a fully connected layer. The last layer is the output layer. And the linear excitation function was selected as the excitation function.

4.3.2 Analysis of MLP Neural Network Results

Based on the MLP neural network model, the 1974 sample data were divided into a training set and a validation set, in the ratio of 7:3, and the model was trained with the training set data and then validated on the validation set. In the setting of the neural network model parameters, for the fitting of the model loss, the mean squared error (MSE) was used as the loss function value

TABLE 3 | Table of predicted results.

Index	IC50_nM	pIC50	Index	IC50_nM	pIC50
1	625.279,085	6.203,926	26	7,550.487,226	5.122,025
2	1,514.918,949	5.819,611	27	2,975.551,524	5.526,433
3	1,230.488,010	5.909,923	28	1,056.888,664	5.975,971
4	80.848,224	7.092329	29	154.518,702	6.811,019
5	48.221,256	7.316,762	30	242.322,471	6.615,606
6	39.612,235	7.402,171	31	3,283.679,462	5.483,639
7	2.341,306	8.630,542	32	8,099.091247	5.091564
8	23.148,631	7.635,475	33	1947.127,430	5.710,606
9	47.355,375	7.324,631	34	8,526.529,200	5.069228
10	34.096260	7.467,293	35	8,391.913,980	5.076139
11	31.653,379	7.499,580	36	472.583,992	6.325,521
12	32.697,205	7.485,489	37	466.080905	6.331,539
13	25.456,401	7.594,203	38	4,262.275,525	5.370,359
14	23.424,709	7.630,326	39	458.985,483	6.338,201
15	9.698,761	8.013284	40	441.836,676	6.354,738
16	9.306,946	8.031193	41	426.397,463	6.370,185
17	26.925,417	7.569,838	42	429.814,207	6.366,719
18	36.208,316	7.441,192	43	662.573,029	6.178,766
19	157.062175	6.803,928	44	145.994,719	6.835,663
20	1953.101,219	5.709,275	45	426.397,463	6.370,185
21	26.404,043	7.578,329	46	5,860.424,608	5.232,071
22	2,333.532,620	5.631,986	47	5,528.679,594	5.257,379
23	894.967,968	6.048193	48	5,671.119,197	5.246,331
24	1,616.214,377	5.791,501	49	4,657.397,312	5.331,857
25	8,979.197,446	5.046763	50	996.189,669	6.001658

(loss) to measure the effectiveness. The number of epoch iterations was set to 200, and the size of the batch data used for gradient descent each time was also set to 200 for model training.

By visualizing the training history, it can be seen in **Figure 9** that the loss function values of the training and validation sets show a clear downward trend. The loss of both sets shows a sharp decreasing trend when the epoch was trained to around 10. Then, the loss of the training set shows a very small oscillation when the epoch was trained to between 20 and 50, while the validation set still keeps decreasing smoothly. The final loss value of the validation set was 0.0146, indicating that the model has a good training effect. At the same time, the loss function curve of the validation set is always below the loss function curve of the training set, indicating that the model did not show any overfitting phenomenon. So it can be considered that the model has good regression effect and can be used for the prediction of the test set.

Through the quantitative prediction model of ERα bioactivity constructed above, the bioactivity values of 50 new compounds were predicted. The results of the IC_{50} and pIC_{50} values are shown in **Table 3**, where the IC_{50} values were further solved from the negative logarithmic relationship between the two.

5 CONCLUSION

To address the problem of compound screening and bioactivity prediction in the anti-breast cancer drug candidates development process, this paper used a computer-aided approach combined with a neural network algorithm to conduct the analysis. Based on the molecular descriptor information and bioactivity values of compounds, and etc., correlation analysis was used to gain a preliminary understanding of the linkage between the independent variables, and the results indicated the existence of multicollinearity between them. Therefore, the LASSO regression algorithm was used to assess the feature importance of 1974 compounds. The features were further screened in combination with cross-validation method, so as to determine the importance of molecular descriptors on biological activity. At the end, the top 20 molecular descriptors with the most significant impact on biological activity were screened. It can be seen from the variation curve of explainable deviation and mean squared error of the model, as well as the statistical graph of the type of variables, that the established variable screening model is effective and the results are well representative. The MLP neural network was then used for learning. The data were divided into training and validation sets to construct a quantitative prediction model of the compound molecular descriptor variables and the bioactivity value pIC_{50}. The final validation set loss function value was 0.0146.

The results showed that the molecular descriptors screening model and bioactivity value prediction model were reasonable, and the applicability of the model was verified. The predictive analysis of the compound molecular descriptors could provide an effective reference for the development of anti-breast cancer drug candidates.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YQ and WW contributed to conception and design of the study. CL and XS organized the database. YQ,CL, and XS performed the statistical analysis. YQ wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This project is supported by the Natural Science Foundation of Zhejiang Province (LY19A010004), the National Natural Science Foundation of China(11971432), Collaborative Innovation Center of Statistical Data Engineering Technology and amp; Application.

REFERENCES

Abdolmaleki, A., and Ghasemi, J. B. (2019). Inhibition Activity Prediction for a Dataset of Candidates' Drug by Combining Fuzzy Logic with MLR/ANN QSAR Models. *Chem. Biol. Drug Des.* 93 (6), 1139–1157. doi:10.1111/cbdd.13511

Batyrova, G., Umarova, G., Kononets, V., Salmagambetova, G., Zinalieva, A., and Saparbayev, S. (2021). Air Pollution Emissions Are Associated with Incidence and Prevalence of Breast Cancer in the Aktobe Region of Western Kazakhstan [J]. *Georgian Med. News* 321, 135–140.

Cheng, S., Qu, B., Qiu, X., Li, N., Wang, X., and Hao, J. (2021). Efficacy and Safety of Kanglaite Injection Combined with Chemotherapy for Women Breast Cancer. *Medicine* 100 (22), e26245. doi:10.1097/md.0000000000026245

Dua, N., Singh, S., and Semwal, V. (2021). Multi-input CNN-GRU Based Human Activity Recognition Using Wearable Sensors[J]. *Computing* 103, 1461–1478. doi:10.1007/s00607-021-00928-8

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression [J]. *Ann. Statistics* 32 (2), 407–499. doi:10.1214/009053604000000067

Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* 96 (456), 1348–1360. doi:10.1198/016214501753382273

Fonti, V., and Belitser, E. (2017). Feature Selection Using Lasso[J]. *VU Amsterdam Res. Pap. Bus. Anal.* 30, 1–25.

Gao, X. L., Shi, Y. J., and Yang, H. J. (2019). Prediction of Injury Risk of Chinese Rugby Players by Multilayer Perceptron Neural Network Model [C]. Abstracts of the 11th National Sports Science Congress, Nanjing, 5797–5799.

He, Y. Q., and Zhu, X. J. (2020). Deep Collaborative Filtering Algorithm for Drug-Target Interaction Ptrediction. *Comput. Electr. Eng.* 41 (08), 2195–2200. doi:10.16208/j.issn1000-7024.2020.08.017

Jia, C. M. (2019). *Research on Drug Target Recognition and Activity Prediction Model Based on Molecular Vibration Characteristics[D]*. Beijing: Beijing University of Chinese Medicine.

Jiao, L., Yang, S., Liu, F., Wang, S., and Feng, Z. (2016). Neural Networks in the Past 70 years: Review and Prospect[J]. *Chin. J. Comput.* 39 (08), 1697–1716. doi:10.11897/SP.J.1016.2016.01697

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *J. Comput. Graph. Statistics* 12 (3), 531–547. doi:10.1198/1061860032148

Li, H. (2017). Variable Selection via Elastic Net Method for Variable for Variable Selection of Balanced Longitudinal Data Model[J]. *J. Taishan Univ.* 39 (3), 5–10. doi:10.1002/sim.9417

Lin, Y., and Yang, J. (2019). Nonnegative Estimation and Variable Selection via Adaptive Elastic-Net for High-Dimensional Data[J]. *Commun. Statistics Simul. Comput.* 50(12), 1–17. doi:10.1080/03610918.2019.1642484

Liu, A., Fang, J., and Yang, R. (2013). Application of Support Vector Machine (SVM) and Bayesian Data Mining Techniques in Drug Discovery[J]. *Chin. Pharmacol. Commun.* 30 (3), 65.

Lv, W., and Xue, Y. (2011). Prediction of Hepatitis C Virus Non-structural Proteins 5B Polymerase Inhibitors Using Machine Learning Methods[J]. *Acta. Phys. -Chim. Sin.* 27 (06), 1407–1416. doi:10.3866/PKU.WHXB20110608

Muthukrishnan, R., and Rohini, R. (2016). LASSO: A Feature Selection Technique in Predictive Modeling for Machine learning[C]. 2016 IEEE international conference on advances in computer applications (ICACA), 1 October 2016, India. IEEE, 18–20. doi:10.1109/ICACA.2016.7887916

Nandi, S., and Bagchi, M. C. (2016). EGFr, FGFr and PDGFr: Emerging Targets for Anticancer Drug Design[J]. *J. Cancer Res. Updat.* 5 (3), 99–108. doi:10.6000/1929-2279.2016.05.03.3

Nandi, S., Rishita, D., Asmita, S., Aaruni, S., and Anil Kumar, S. (2022). Natural Sourced Inhibitors of EGFR, PDGFR, FGFR and VEGFRMediated Signaling Pathways as Potential Anticancer Agents[J]. *Curr. Med. Chem.* 29 (2), 212–234. doi:10.2174/0929867328666210303101345

Pinkus, A. (1999). Approximation Theory of the MLP Model in Neural Networks. *Acta Numer.* 8, 143–195. doi:10.1017/s0962492900002919

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* 65, 386–408. doi:10.1037/h0042519

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO[J]. *J. R. Stat. Soc. Ser. B(Methodological)* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Xiao, S. Y., Gao, J., Sun, Z. Q., Wang, P., Zhang, X., Wang, J. L., et al. (2021). Prediction of Occupational Coal Worker's Pneumoconiosis by Multilayer Perceptron Neural Network Model[J]. *Chin. Occup. Med.* 48 (01), 19–25. doi:10.11763/j.issn.2095-2619.2021.01.004

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Comput.* 26 (1), 185–207. doi:10.1162/neco_a_00537

Zhang, S., Zhu, F., Yu, Q., and Zhu, X. (2021). Identifying DNA-Binding Proteins Based on Multi-Features and LASSO Feature Selection. *Biopolymers* 112 (2), e23419. doi:10.1002/bip.23419

Zhang, X. (2017). *Research and Application of Collaborative Filtering Algorithm in Drug Relocation[D]. Diss*. Shanghai: Donghua University.

Zhao, Y., Shim, N., Cui, Y.-H., Kang, J.-H., Yoo, K.-C., Kim, S., et al. (2021). FBXO15 Plays a Critical Suppressive Functional Role in Regulation of Breast Cancer Progression. *Sig Transduct. Target Ther.* 6 (1), 211. doi:10.1038/s41392-021-00605-4

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net[J]. *J. R. Stat. Soc.* 67, 768. doi:10.1111/j.1467-9868.2005.00527.x

Zou, H., and Zhang, H. H. (2009). ON THE ADAPTIVE ELASTIC-NET WITH A DIVERGING NUMBER OF PARAMETERS. *Ann. Stat.* 37 (4), 1733–1751. doi:10.1214/08-AOS625

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. *J. Comput. Graph. Statistics* 15 (2), 265–286. doi:10.1198/106186006x113430

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *J. Am. Stat. Assoc.* 101 (476), 1418–1429. doi:10.1198/016214506000000735

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Qin, Li, Shi and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.