WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n2.19
Journal homepage: https://wepub.org/index.php/IJCSIT/index

# Optimized Modeling of Anti-Breast Cancer Drug Candidates

Xiatian Sun *

College of QingDao University, QingDao, China

## ABSTRACT

Breast cancer is often referred to as the "Pink Killer", and its incidence rate ranks first among female malignant tumors. It has been found that the expression of estrogen receptors alpha (ERα) plays a very important role in breast lesions, and compounds that can antagonize the activity of ERα may be candidates for the treatment of breast cancer. In this paper, it is of practical significance to construct a quantitative structure-activity relationship (QSAR) model of compounds to screen potential compounds that can antagonize the activity of ERα by using a machine learning approach, and to construct a classification prediction model of ADMET properties to predict the pharmacokinetic properties and safety of compounds in the human body.

## KEYWORDS

Machine learning; Random forest; LightGB regression; SVM; Feature selection

## 1. INTRODUCTION

### 1.1. Background of the Study

The 2020 Global Cancer Statistics report explains that cancer is currently the first or second leading cause of death for populations in 112 countries. 19,292,789 new cases of cancer are projected globally in 2020, with a markedly elevated trend [1]. And in China, the incidence of cancer remains high globally due to the rising proportion of an aging population. As early as 2012, China ranked first in the world in terms of new cancer cases. Among all cancers, breast cancer is the cancer with the highest incidence rate and the highest number of women in China. And with the improvement of early diagnosis and treatment of cancer, the 5-10 year survival rate of breast cancer patients has been gradually improved [2].

It has been shown by research and experimental results that estrogen receptor α subtype plays a nearly decisive role in the development of mammary glands, and there is a significant correlation between breast cancer ultrasound characteristics and ERα positive expression. At present, anti-hormonal therapy has a good effect on ERα-expressing breast cancer patients. Therefore, we expect to find a compound that can effectively antagonize the activity of Erα as a candidate for the treatment of breast cancer.

The screening of compounds to become drug candidates is more complicated. Usually, a series of compounds acting on Erα and their bioactivity data are collected, and some structural descriptors of these compounds are used as independent variables, while the physiological activity of the compounds is used as the dependent variable, in order to construct the Quantitative Structure-Activity Relationship (QSAR) model of the compounds. The QSAR model can not only predict various physicochemical properties of compounds, but also explore and determine the structural factors that determine various properties of compounds, so as to understand the influence of microstructure on various macroscopic properties at the molecular level, which can play a certain role in guiding the

design of molecules. It can be used as a guide for molecular design. Therefore, the results of QSAR studies have been increasingly applied to the design of drug molecules, chemical process design, and evaluation of the environmental behavior of organic compounds and practical applications [3].

### 1.2. Problem Introduction

Based on the above background, this paper addresses and examines the following tasks:

**Task1:** Based on the data provided in the annexes "Molecular_Descriptor.xlsx" and "ERα_activity.xlsx", select the variables for 729 molecular descriptors of 1974 compounds according to their significance in terms of their impact on biological activity. 729 molecular descriptors of 1974 compounds were selected as variables, and 20 molecular descriptors (i.e., variables) with the most significant influence on biological activity were selected, and the molecular descriptor selection process and its rationality were explained in detail.

**Task2:** In conjunction with Problem 1, select no more than 20 molecular descriptor variables and describe the process of constructing a quantitative prediction model of a compound's biological activity against ERα. Then use the constructed model to predict the $IC_{50}$ values and the corresponding $pIC_{50}$ values of the 50 compounds in the TEST table.

## 2. PROBLEM ANALYSIS

### 2.1. Analysis and Thoughts on Question One

The title gives 729 molecular descriptors for 1974 compounds, and due to the large number of molecular descriptors for compounds, there is a high risk of overfitting if the sample/parameter ratio is less than 5, from classical statistical considerations. Parameter compression should be taken and various statistical and optimization algorithms should be applied to select the features [4].

For Problem 1, this paper sequentially uses filtering methods (variance filtering and mutual information method) to eliminate unimportant features and tree model based embedding method to select important features [5]. Firstly, the variance filtering method is used to eliminate the features with variance 0, leaving 504 features, and then the mutual information method of relevance filtering is used to filter out the features that are completely irrelevant to the label, further filtering the features to 474, after that, in order to filter the important features, the sample data are analyzed by the tree model-based feature selection method in machine learning algorithms, which results in the top 20 highest importance of influencing the biological activity of the molecular descriptors [6].

### 2.2. Analysis and Ideas for Question 2

For the 20 important molecular descriptors screened in Problem 1, this paper further screens the important molecular descriptors from two aspects: the first step is to view and compare the distribution of molecular descriptors in the training set and the test set by drawing the KDE distribution graphs, and in order to avoid the deterioration of the generalization ability of the constructed model, the molecular descriptors with inconsistent distributions in the training set and the test set are deleted [7]. In the second step, considering that some molecular descriptors may be correlated with each other, in order to ensure the independence of the selected molecular descriptors, this paper visualizes the molecular descriptors using heat map and Pearson diagram, and further deletes the molecular descriptors with strong mutual correlation according to the visualization results. After that, four regression models (LightGBM regression model, random forest regression model, multiple linear regression model, KNN regression model) were constructed according to the train table in the file "ERα_activity.xlsx", and LightGBM regression model, which had the best effect, was chosen to predict the "ERα_activity.xlsx". The regression model was used to predict the IC50 values

of the chemicals in the test table in the file "ERα_activity.xlsx", and the prediction accuracies of the four regression models were calculated for different numbers of molecular descriptors, which were measured by the mean squared error (MSE). The mean squared error (MSE) was used to measure the prediction effect of the regression models [8].

# 3. MODELING AND SOLVING

## 3.1. Solving Problem 1

### 3.1.1. Flowchart of the solution idea for Problem 1

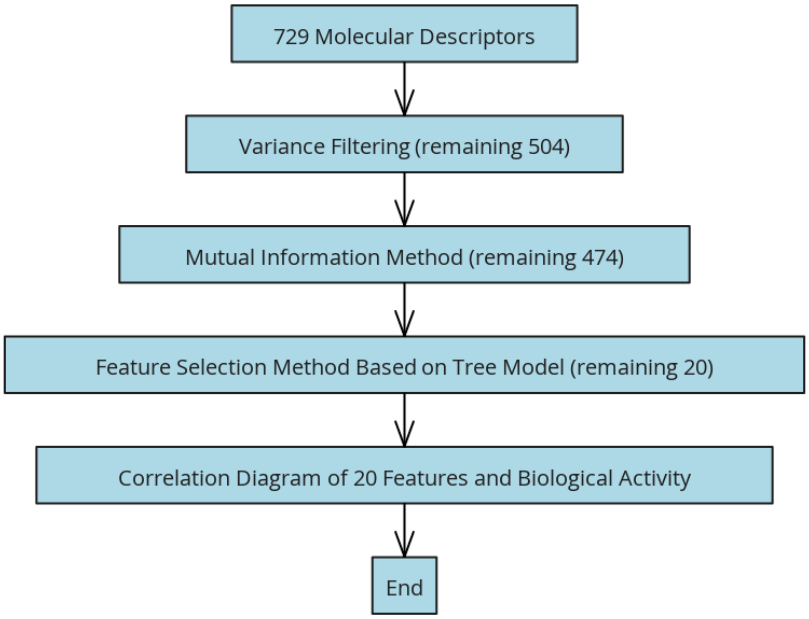


**Figure 1.** Flowchart of Problem 1 Ideas

### 3.1.2. Problem 1 solution process and results

For Problem 1, we firstly used variance filtering to eliminate the features with variance 0, leaving 504 features, and then used the mutual information method of relevance filtering to filter out the features that are completely unrelated to the label, and further filtered the features down to 474, and after that, in order to filter the important features, the sample data were analyzed using the tree model-based feature selection method in machine learning algorithms, and the importance of influencing the bioactivity was derived The 20 molecular descriptors with the highest importance affecting the biological activity, and finally the selected 20 molecular descriptors were visualized and analyzed with the biological activity.The 20 important molecular descriptors selected are shown in Table1:

**Table 1.** 20 important molecular descriptors

| | | | | |
|---|---|---|---|---|
| AlogP | ALogp2 | ATSc4 | ATSc5 | BCUTc-1l |
| BCUTc-1h | SHBint10 | SHBd | SsOH | minHsOH |
| minssCH2 | maxsOH | LipoaffinityIndex | ETA_Shape_Y | ETA_BetaP_s |
| MDEC-33 | MLFER_A | MLFER_BH | WTPT-5 | XLogP |

### 3.1.3. Plot of linear regression relationships

The linear regression relationship plot is mainly used to analyze the linear relationship between the variables and the labels, from which it can be seen the relationship between the selected 20 important molecular descriptors and $pIC_{50}$.

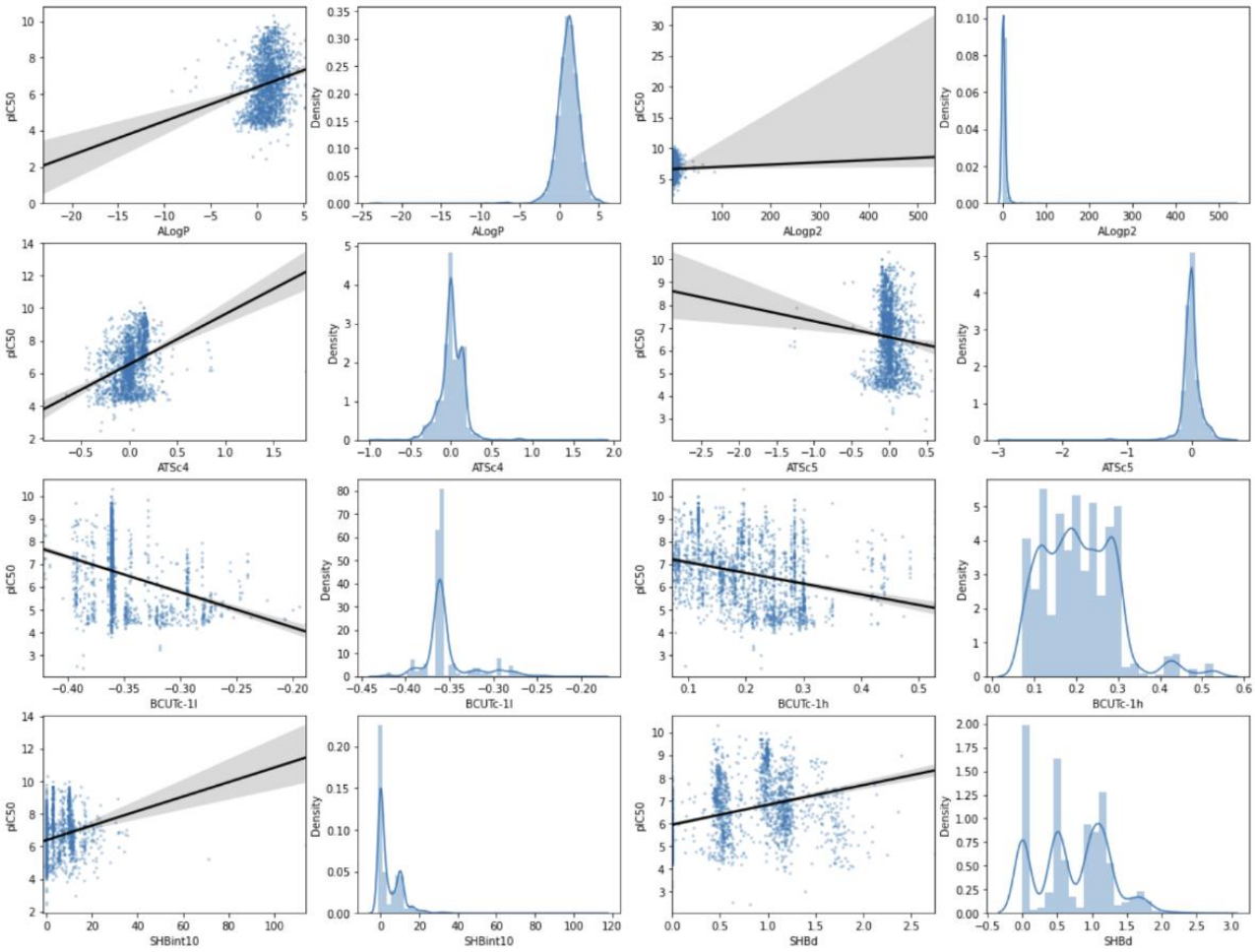


**Figure 2.** Plot of linear regression relationships between molecular descriptors (a)

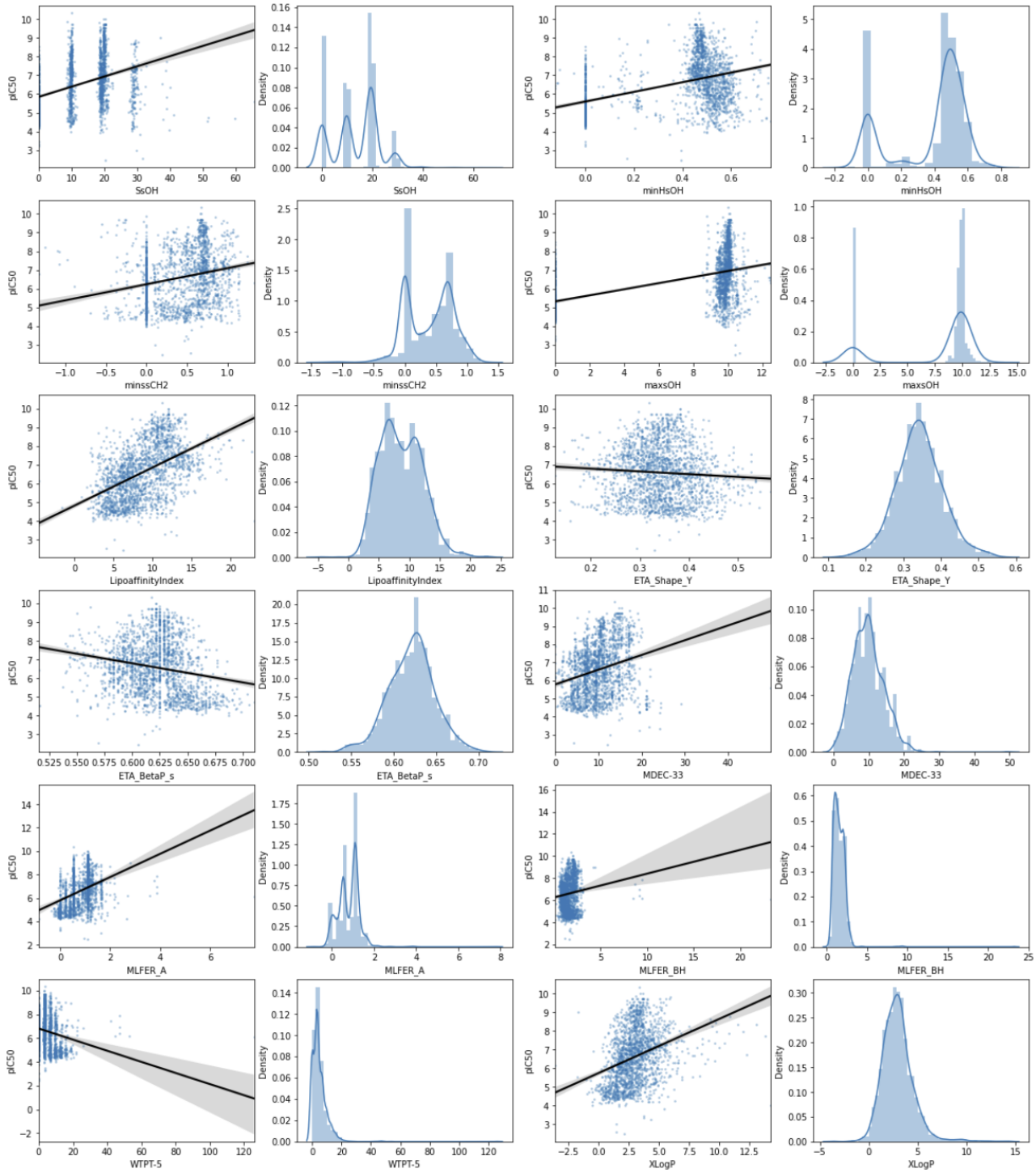


**Figure 3.** Plot of linear regression relationships between molecular descriptors (b)

From Figure 3.1.3, it can be found that LipoaffinityIndex, maxs0H, minHsOH, and MLFER_A have a strong positive linear correlation with the label $pIC_{50}$; and BCUTc-1l, BCUTc-1h, and ETA_BetaP_s have a strong negative linear correlation with the label $pIC_{50}$. The above results show that the method of screening features used in this paper is correct and feasible. method of screening features is correct and feasible [9].

## 3.2. Solving Problem 2

### 3.2.1. Flowchart of Solution Ideas for Problem 2

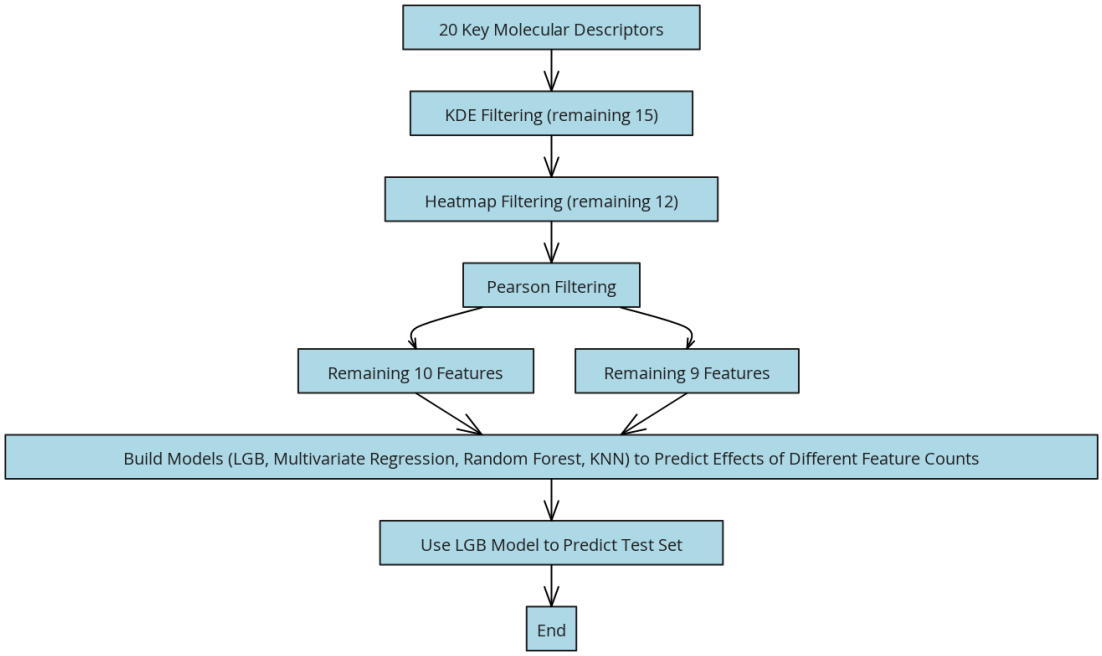


**Figure 4.** Flowchart of Problem 2 Ideas

### 3.2.2. Solution process and results of Problem 2

For modeling the 20 important molecular descriptors screened out in Problem 1, four regression models were constructed: LightGBM regression model, random forest regression model, multivariate linear regression model, and KNN regression model; and as few molecular descriptors as possible were selected under the condition of ensuring that there was not too much loss of the model's accuracy. The first step is to view and compare the distribution of molecular descriptors in the training set and the test set by plotting the KDE distribution, and to avoid the generalization ability of the model deteriorating, the molecular descriptors with inconsistent distributions in the training set and the test set are deleted [10]. Compare the KDE distribution of 20 molecular descriptors in the training and test sets:

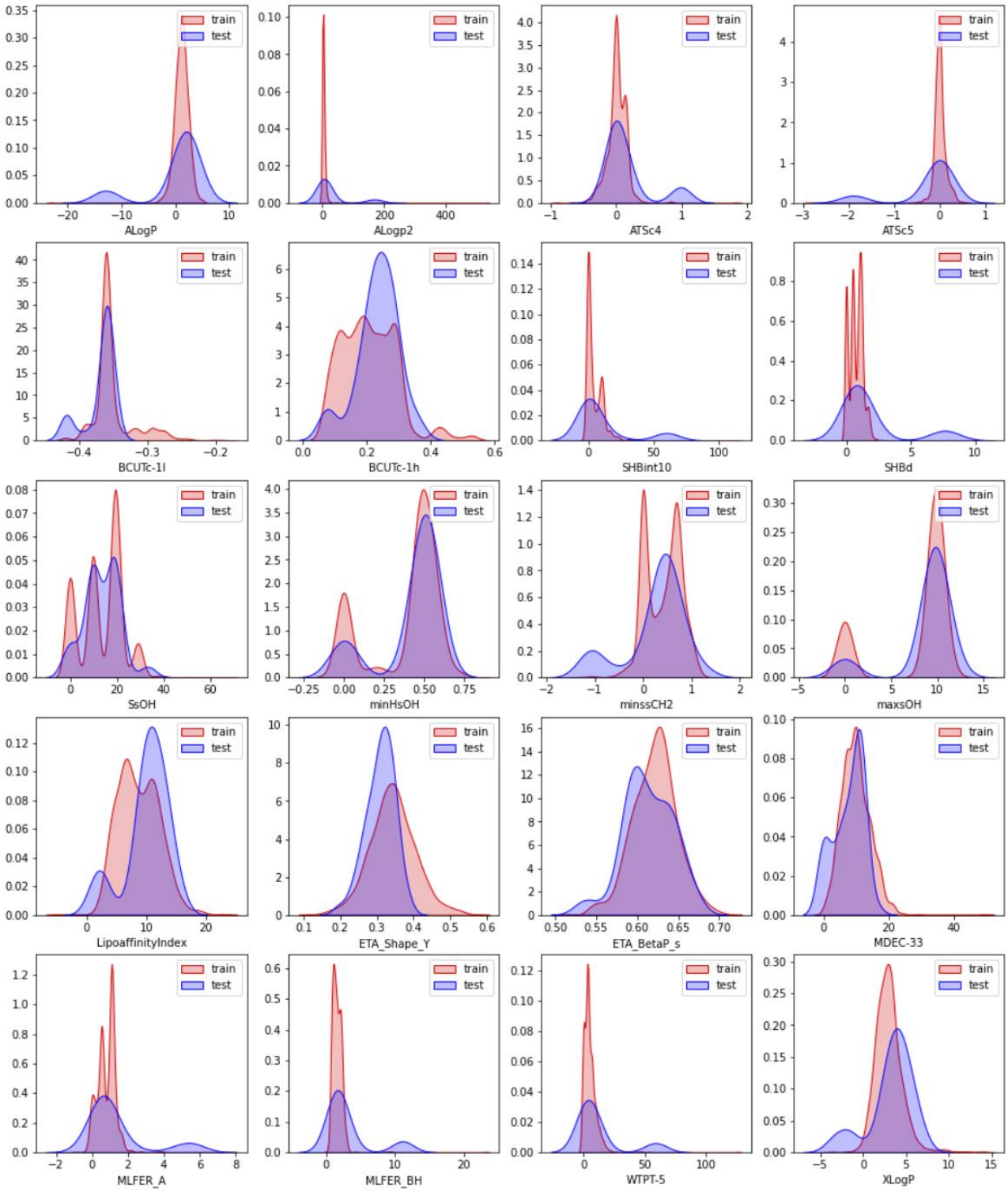


**Figure 5.** KDE distribution of the 20 molecular descriptors in the training and test sets

It can be found that five molecular descriptors: ALogp, ATSc5, SHBd, MLFER_BH and SsOH have inconsistent distributions in the training set and the test set, which will lead to poorer generalization of the model, and it is necessary to remove these five molecular descriptors [11].

The predictive evaluation (MSE) of the four models at this point is shown in Table2.

**Table 2.** Representation of the 20 and 15 molecular descriptors in the model

| Number of Features | LGB Regression | Random Forest Regression | Multiple Linear Regression | KNN Regression |
|---|---|---|---|---|
| 20.0 | 0.488 | 0.593 | 0.963 | 0.742 |
| 15.0 | 0.496 | 0.584 | 1.003 | 0.795 |

In the second step, considering that some molecular descriptors may be correlated with each other, in order to ensure the independence of the selected molecular descriptors, we will use the heat map to view the two-by-two correlation between the molecular descriptors for the remaining 15 molecular descriptors, and further delete the molecular descriptors with strong correlation according to the visualization results:

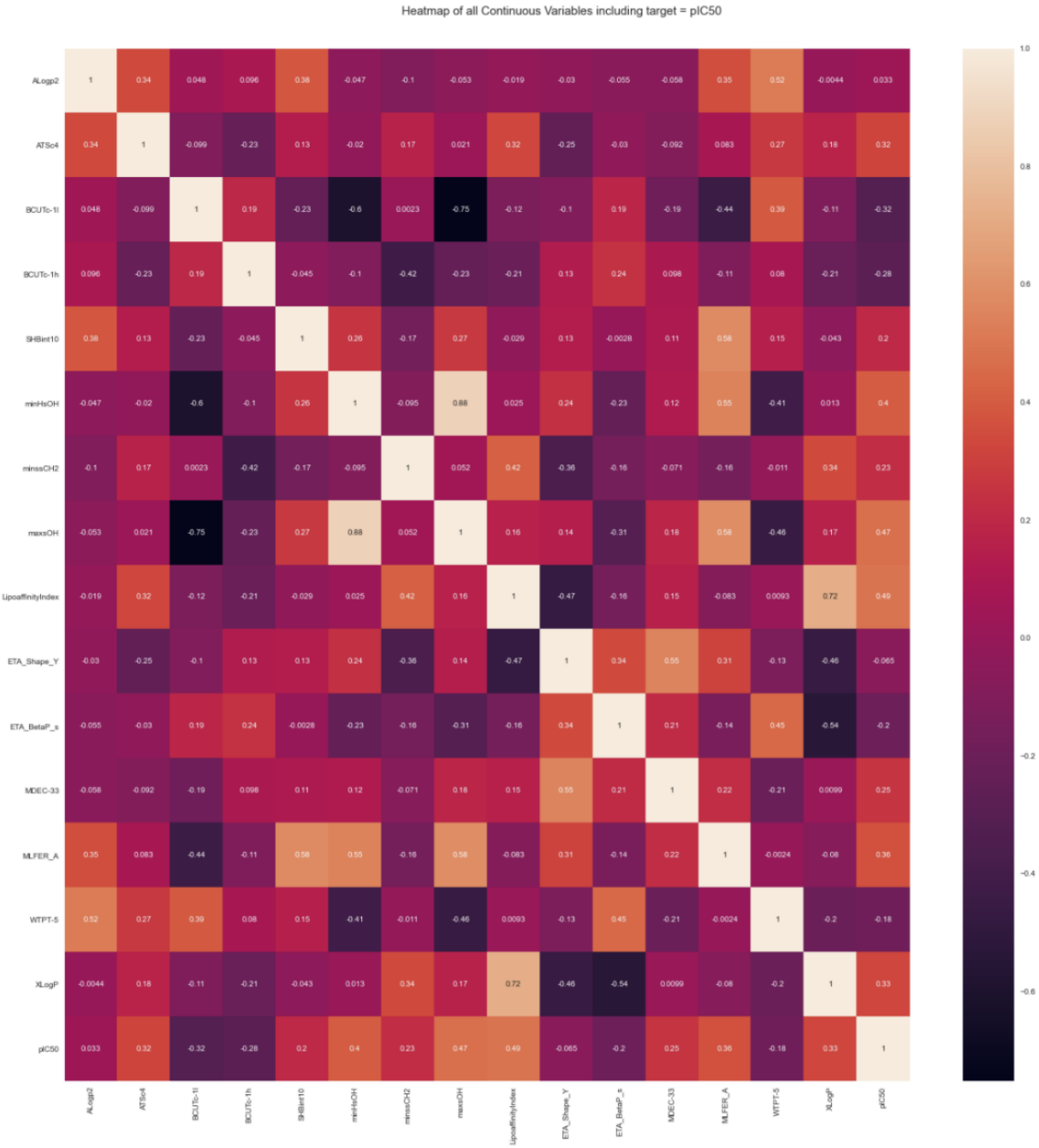


**Figure 6.** Heat map for 15 molecular descriptors

It can be seen from the heat map: the darker the color, the stronger the correlation of molecular descriptor pairs, we selected six groups of molecular descriptors with strong correlation for linear correlation visualization, and the results are as follows:

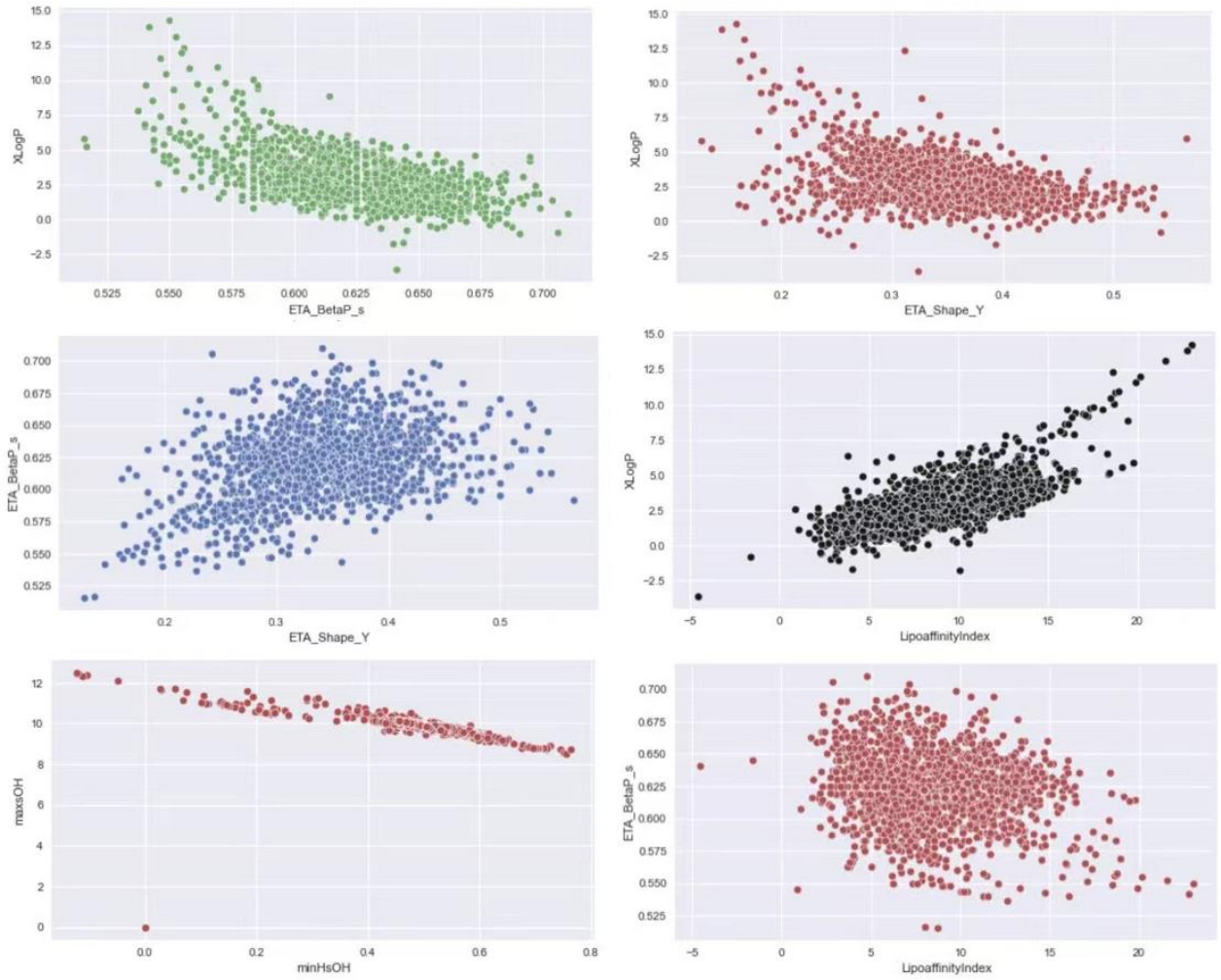


**Figure 7.** Visualization of linear correlation for 6 groups of molecular descriptors

From the linear correlation diagram, it can be seen that: the three molecular descriptors LipoaffinityIndex, ETA_Shape_Y, ETA_BetaP_s are strongly co-correlated with other features, so they are deleted, and at this time, the remaining 12 molecular descriptors in the four regression models are as follows:

**Table 3.** Representation of 20, 15, and 12 molecular descriptors in the model

| Number of Features | LGB Regression | Random Forest Regression | Multiple Linear Regression | KNN Regression |
|---|---|---|---|---|
| 20.0 | 0.488 | 0.593 | 0.963 | 0.742 |
| 15.0 | 0.496 | 0.584 | 1.003 | 0.795 |
| 12.0 | 0.525 | 0.595 | 1.105 | 0.891 |

In the third step, a Pearson plot was used to further view the direction and strength of the linear relationship between the remaining 12 molecular descriptors.

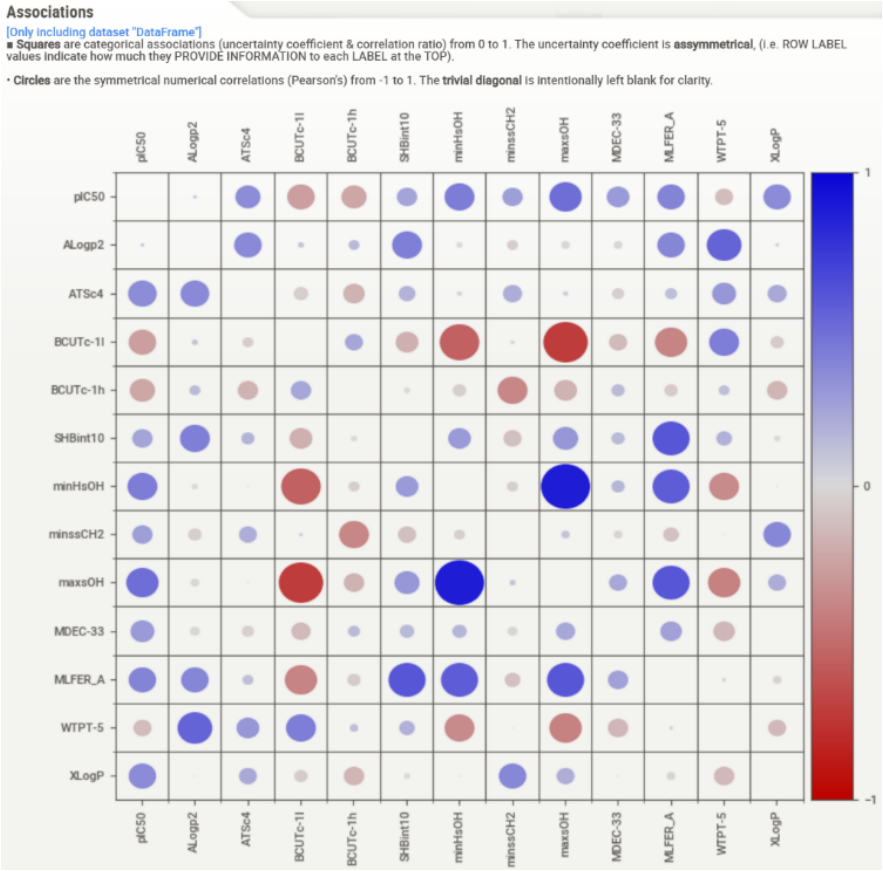


**Figure 8.** Pearson diagram for the remaining 12 molecular descriptors

The Pearson plot shows that the three molecular descriptors maxsOH, ALogp2, and MLFER_A are strongly correlated with other molecular descriptors, and at this point, taking into account the accuracy of the model, we compared the four regression models after deleting the two (maxsOH, ALogp2) and three (maxsOH, ALogp2, and MLFER_A) molecular descriptors, respectively. model performance, using the mean square error MSE as an evaluation criterion, and the results are shown in Table4:

**Table 4.** Summary of the performance of different number of molecular descriptors in the model

| Number of Features | LGB Regression | Random Forest Regression | Multiple Linear Regression | KNN Regression |
|---|---|---|---|---|
| 20.0 | 0.488 | 0.593 | 0.963 | 0.742 |
| 15.0 | 0.496 | 0.584 | 1.003 | 0.795 |
| 12.0 | 0.525 | 0.595 | 1.105 | 0.891 |
| 10.0 | 0.513 | 0.598 | 1.122 | 0.965 |
| 9.0 | 0.548 | 0.618 | 1.133 | 0.984 |

In order to ensure the accuracy of the model, we finally chose the LightGB regression model, which retains 10 molecular descriptors, to predict the test data in the file "ERα_activity.xlsx", and the prediction results are shown in the Appendix.

# 4. MODEL EVALUATION AND DISSEMINATION

## 4.1. Evaluation of the Model

### 4.1.1. Advantages of the model

(1) In model selection, the prediction effect of multiple models was considered, and the optimal model was selected to make the model more representative.

(2) For the complex relationship between important molecular descriptors and pIC50, we use the feature selection method based on the tree model, and for the selected molecular descriptors, we make the model prediction with high accuracy and excellent robustness.

(3) Use LightGB regression model with high accuracy and fast running speed.

### 4.1.2. Disadvantages of the model

(1) The models used are algorithms from machine learning libraries, which are prone to black boxes: it's hard to know what's going on.

(2) The model may not produce good classification for small or low-dimensional data (data with fewer features).

## 4.2. Extension of the Model

The model constructed in this question is to predict the value of pIC50 under the main molecular descriptors, and it can also predict other data related to pIC50, i.e., the model in this paper is of some generalized relevance.


# ACKNOWLEDGEMENTS

Thank you to my teammates, Shou and Kang. They provided extremely important assistance in the writing and competition process of this paper.


# REFERENCES


[1] LIU Zongchao; LI Zhexuan; ZHANG Yang; ZHOU Tong; ZHANG Jingying; YOU Weicheng; PAN Kaifeng; LI Wenqing.Interpretation of the 2020 Global Cancer Statistics Report [J]. Electronic Journal of Comprehensive Cancer Therapy, 2021, (02):1-14.

[2] Cai Tingting, Li Danyu, Huang Qingmei, Wu Fulei, Xia Haozhi, Yuan Changrong. Progress of a patient-reported quality of life measurement tool for breast cancer patients [J]. Shanghai Nursing, 2021, 21(10):47-51.

[3] Qiang Su. Research on QSAR of environmental toxicants based on data mining algorithm [C]. Shanghai University, 2013.

[4] Chen K. Research on categorical feature selection method for high-dimensional data [D]. Shandong University, 2021.

[5] ZHAO Tijun; SUN Lingling; NIU Yiguo; XIE Xiaoying; JIA Qingquan. Modeling method of photovoltaic output probability distribution based on improved nonparametric kernel density estimation [J]. Journal of Yanshan University, 2021, 45(05):430-437+448.

[6] DING Tao, JIANG Lesheng, SHI Zhengxiang, ZHAO Yang, MA Hui. Study on the relationship between environmental factors and milk production in dairy barns in cold areas based on random forest [J/OL]. Journal of Agricultural Machinery:1-11[2021-10-18].

[7] YANG Rishuang, NING Qian, LEI Yinjie, CHEN Bingcai. Rolling bearing fault diagnosis based on improved convolutional neural network and LightGBM [J]. Bearing, 2021(06):44-49.

[8] Geng Xiu-Lin, Huang Ting-Ting. Comparison and application of multi-objective influencing factor effects based on multiple multiple regression--an example of analyzing business activities [J]. Statistics and Information Forum, 2019, 34(10):100-107.

[9] Zhao Shubao; Jiang Chunmao. A fast KNN algorithm based on three-branch clustering [J]. Small Microcomputer Systems, 2021, (09):1845-1851.

[10] YANG Di; FANG Yang-Xin; ZHOU Yan. Research on new category classification based on MEB and SVM methods [J]. Journal of Guangxi Normal University (Natural Science Edition).

[11] LI Mengmeng, LIU Jingdang, LIANG Tianyi, TAN Liang, WANG Gang, ZHU Xi. A predictive model of sulfur element in magmatic sulfide deposits based on support vector machine algorithm [J/OL]. Journal of Jilin University (Earth Science Edition):1-15[2021-10-18].