

# 基于神经网络抗乳腺癌候选药物优化建模

钱古椿

上海理工大学机械工程学院, 上海

收稿日期: 2022年9月19日; 录用日期: 2022年11月4日; 发布日期: 2022年11月14日

---

## 摘要

ER $\alpha$ 在乳腺发育过程中具有十分重要的作用, 被认为是治疗乳腺癌的重要靶标。在化合物候选过程中, 其不仅需要具备良好的生物活性, 还需要具备良好的药代动力学性质和安全性。本文采用数据挖掘技术, 根据提供的ER $\alpha$ 拮抗剂信息, 构建化合物生物活性的定量预测模型和ADMET性质的分类预测模型, 从而为同时优化ER $\alpha$ 拮抗剂的生物活性和ADMET性质提供预测服务, 具有一定的现实意义。

---

## 关键词

ER $\alpha$ 拮抗剂, 随机森林, Spearman相关系数, PSO-BP神经网络, Logistic回归

---

# Optimization Modeling of Anti Breast Cancer Drug Candidates Based on Neural Network

Guchun Qian

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Sep. 19<sup>th</sup>, 2022; accepted: Nov. 4<sup>th</sup>, 2022; published: Nov. 14<sup>th</sup>, 2022

---

## Abstract

ER $\alpha$  plays a very important role in breast development and is considered as an important target for the treatment of breast cancer. In the candidate process of compounds, they need not only good biological activity, but also good pharmacokinetic properties and safety. In this paper, data mining technology is adopted according to the provided ER $\alpha$  Antagonist information, to build a quantitative prediction model of compound biological activity and a classification prediction model of ADMET properties, so as to simultaneously optimize ER $\alpha$ . The biological activity and ADMET properties of antagonists provide prediction services, which has certain practical significance.

## Keywords

**ER $\alpha$  Antagonist, Random Forest, Spearman Correlation Coefficient, PSO-BP Neural Network, Logistic Regression**

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

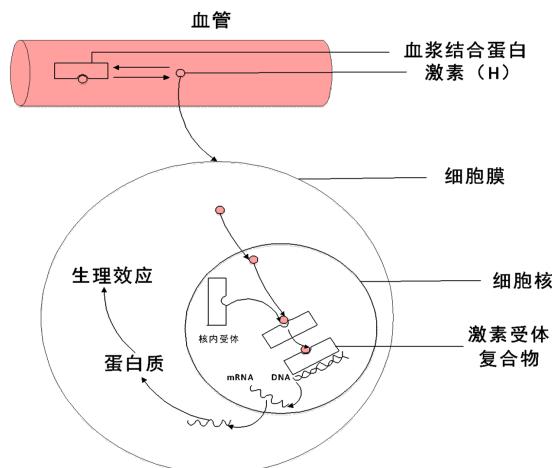
<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

据肿瘤数据库 GLOBOCAN 近期发布的全球癌症统计报告数据显示，乳腺癌已超肺癌，成为全球癌症新发病例数首位，同时也是中国人数癌症发病的第四大原因。乳腺癌的发展与雌激素受体密切相关，经典的雌激素受体，包括 ER $\alpha$ 、ER $\beta$ 这两种亚型，并且两者的结构相似，它们位于细胞核内，通过调节特异性靶基因的转录发挥“基因型”调节效应，雌激素受体如图 1 所示。



**Figure 1. Estrogen receptor**  
**图 1. 雌激素受体**

有研究发现，雌激素受体  $\alpha$  亚型(Estrogen receptors alpha, ER $\alpha$ )在不超过 10% 的正常乳腺上皮细胞中表达，但大约在 50%~80% 的乳腺肿瘤细胞中表达；而对 ER $\alpha$  基因缺失小鼠的实验结果表明，ER $\alpha$  [1]确实在乳腺发育过程中扮演了十分重要的角色。目前，抗激素治疗常用于 ER $\alpha$  表达的乳腺癌患者，其通过调节雌激素受体活性来控制体内雌激素水平。因此，ER $\alpha$  被认为是治疗乳腺癌的重要靶标，能够拮抗 ER $\alpha$  活性的化合物可能是治疗乳腺癌的候选药物，具体做法如图 2 所示。

一个化合物想要成为候选药物，除了需要具备良好的生物活性(此处指抗乳腺癌活性)外，还需要在人体内具备良好的药代动力学性质和安全性，合称为 ADMET (Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性)性质。其中，ADME 主要指化合物的药代动力学性质，描述了化合物在生物体内的浓度随时间变化的规律，T 主要指化合物可能在人体内产生的毒副作用。一个化合物的活性再好，如果其 ADMET 性质不佳，比如很难被人体吸收，或者体内代谢速度太快，或者具有某种毒性，那么其仍然难以成为药物，因而还需要进行 ADMET 性质优化。一般需着重考虑化合物的 5 种 ADMET 性质，即小肠上皮细胞渗透性、细胞色素 P450 酶、化合物心脏安全性评价、人体口服生物利用度、微核试验。

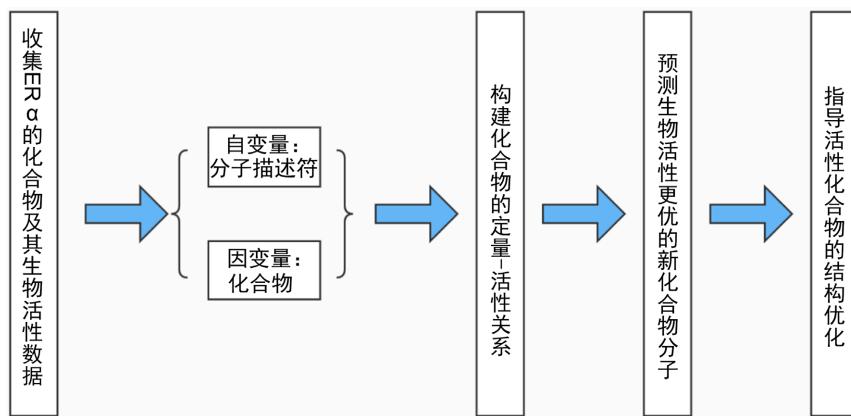


Figure 2. Selection process of candidate drugs

图 2. 候选药物的选取过程

## 2. 寻找影响生物活性物的主要因素

### 2.1. 数据预处理

为保证变量降维的可靠性,需预先对变量数据进行处理,检查是否存在空值的位点,如存在,则需使用当前变量数据均值替代缺失值;检查是否存在变量异常值,经查阅文献,若当前变量数据 65% 以上为恒定值,则可将该变量判定为无意义变量,故将该变量当作异常变量进行剔除。

#### 1) 缺失值处理

本文给出 1974 个化合物对应 729 个分子描述变量的原始数据,将其数据表导入 SPSS 软件中,进行遍历搜索,若该位点为空,则判定为缺失值,最终搜索的缺失值数目为 0。

#### 2) 异常值处理

缺失值处理结束之后,为了数据的可靠性,再进行异常值处理,由于各分子描述符变量所对应不同的化合物,故各分子描述符变量下的数据并没有离群这一说法,因为并不是同一样本下测得的数据。经查阅相关文献,当某个分子描述符变量下存在 65% 以上的数据都相同,则表明,该分子描述符变量对各个化合物的生物活性没有规律性影响,故可将该分字符变量当作异常变量,并进行剔除操作。

#### 3) 数据标准化

在变量降维之前,需要对各变量数据进行标准化操作,主要是因为各个变量的性质、量纲、可用性等特性可能会存在差异,导致降维无法得出具有代表性的变量。标准化能使各变量具有均等权重,同时,下文需要通过随机森林回归算法对预处理后的 247 个变量进行降维,故数据标准化具有重要作用。

目前,数据标准化的方法有很多,一般有极值法、标准差法以及三折线法等等。然而不同的标准化方法,可能对系统变量的评价结果也会有所不同。本文采用标准化方法——标准差法(Z-Score 标准化)。通过标准化法[2]对这 247 个变量进行标准化处理,对应的计算公式为

$$\bar{x}_i = \frac{x_i - \mu}{\delta} \quad (1)$$

式中,  $\bar{x}_i$  为标准化后的第  $i$  个化合物对应的数据;  $x_i$  为第  $i$  个化合物;  $\mu$  为该变量的所有化合物下的数据均值;  $\delta$  为该变量的所有化合物下的数据标准差。

### 2.2. 变量降维

随机森林[3]进行回归分析时,可以在样本分类的基础上进行计算,由此获得各个重要性特征变量对

目标值的影响程度即贡献度。贡献度越大，说明该特征变量对目标值的影响程度越大；反之则越小。关于某节点  $k$  的贡献度计算公式如下式所示。

$$n_k = \omega_k * G_k - \omega_{\text{left}} * G_{\text{left}} - \omega_{\text{right}} * G_{\text{right}} \quad (2)$$

式中， $\omega_k$ ：节点  $k$  中训练样本与总训练样本数目的比例。

$\omega_{\text{left}}$ ：节点  $k$  左节点中训练样本与总训练样本数目的比例。

$\omega_{\text{right}}$ ：节点  $k$  右节点中训练样本与总训练样本数目的比例。

$G_k$ ：节点  $k$  处的不纯度。

$G_{\text{left}}$ ：节点  $k$  左节点的不纯度。

$G_{\text{right}}$ ：节点  $k$  右节点的不纯度。

由上述公式，可得出任一特征变量对目标值的贡献度公式为：

$$f_i = \frac{\sum_{j \in \text{nodes split on feature } i} n_j}{\sum_{k \in \text{all nodes}} n_k} \quad (3)$$

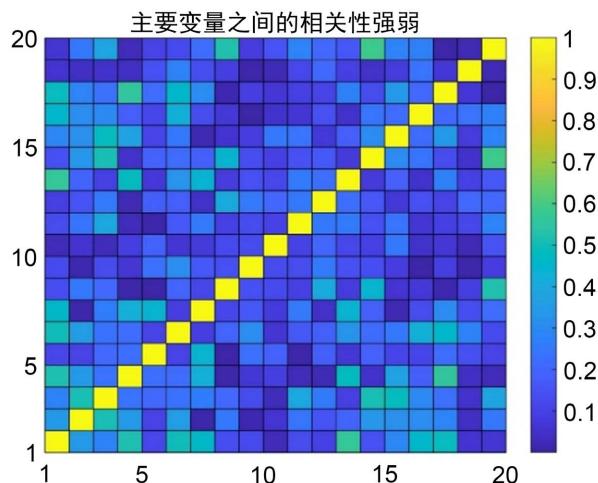
借助 Matlab 工具对随机森林回归算法进行应用，对预处理过后的 247 个变量进行贡献度计算，贡献度分布情况如图 5 所示。并将这 247 个变量针对贡献度进行由高到低排列，并设定贡献度阈值为 0.2%，由此初步筛选得到贡献度排名前 60 个变量。

在不确定变量数据是否满足正态分布的前提下，不可直接使用 Person 相关性分析，本文采用 Spearman 相关系数[4]表示两变量间的关联程度，由此通过 Matlab 程序进行高耦合变量过滤。在统计学中，Spearman 相关系数通常用来衡量两个变量间依赖性的非参数指标，用希腊字母  $\rho$  表示。对与样本容量为  $n$  的样本，相关系数  $\rho$  为：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (4)$$

式中， $x_i, y_i$  代表秩次， $\rho$  表示相关系数， $|\rho|$  越接近 1，表明两变量相关程度越大。

通过 Matlab 得出上文 60 个变量间的相关系数，然后进行两两筛选，对具有高耦合性的变量组剔除低贡献度的变量，最后综合得出 20 个主要变量，以确保其独立性，主要变量间的相关度分布图如图 3 所示。



**Figure 3.** Distribution of correlation between main variables  
**图 3.** 主要变量间的相关度分布图

### 3. 建立化合物对 $ER\alpha$ 生物活性的定量预测模型

#### 3.1. BPER $\alpha$ 生物活性预测模型建立与验证

##### 1) 数据整理

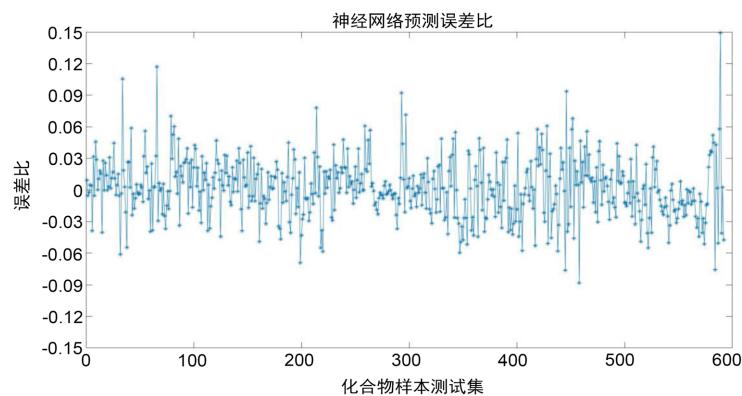
将上文中选取的 20 个主要变量对应的 1974 个化合物样本数据进行整理, 然后将数据导入 SPSS 软件中进行标准化处理, 处理以后的数据再导入 Matlab 中, 将其中 70% 化合物样本数据作为训练集, 将 30% 化合物样本数据作为测试集。

##### 2) 网络结构建立

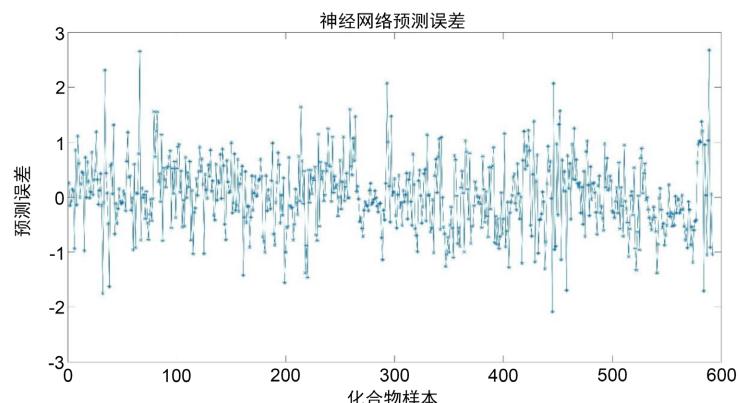
① 设置输入输出层: 将问题 1 中选取的 20 个主要变量所对应的 1974 个化合物样本数据, 经标准化之后导入 Matlab 中作为输入, 输入层神经元数目设置为 20 个; 将  $pIC_{50}$  值导入 Matlab 作为输出, 输出层神经元数目设置为 1 个。

② 设置隐含层: 隐含层的确定对整个网络训练过程起十分重要的作用。隐含层设置过少会导致网络训练性能达不到要求; 而隐含层设置过多又会导致过拟合现象。经反复训练, 最终确定隐含层神经元数目为 20 个。

根据 BP (Back Propagation) 网络[5]建立  $ER\alpha$  生物活性预测模型后, 得出 20 个主要变量对应的化合物样本测试集平均误差百分比在 4% 左右, 个别样本所到最大误差百分比在 12% 左右, 由此说明建立的预测模型比较良好。化合物样本总体误差比如图 4, 预测误差如图 5, 预测值与真实值如图 6。



**Figure 4.** Overall error ratio of compound sample  
**图 4.** 化合物样本总体误差比



**Figure 5.** Compound sample error  
**图 5.** 化合物样本误差

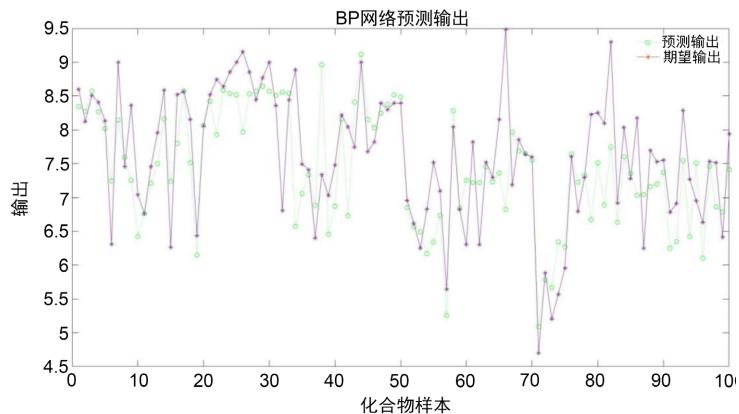


Figure 6. Comparison between predicted value and real value

图 6. 预测值与真实值的比较情况

### 3.2. ER $\alpha$ 生物活性结果预测

根据上文 Spearman 相关系数对  $\text{pIC}_{50}$  值进行预测, 得出  $\text{pIC}_{50}$  预测结果, 可知  $\text{pIC}_{50}$  值是  $\text{IC}_{50}$  值的负对数, 因此这两目标变量之间存在很强的相关性。再通过非线性拟合[6]得到  $\text{pIC}_{50}$  与  $\text{IC}_{50}$  之间的数学表达式, 进而求得  $\text{IC}_{50}$  列中的值。

此外, 均方误差(Mean Squared Error, MSE)也是评价预测模型好坏的重要指标, 是预测值与真实值差值平方的平均值, 本文模型的均方误差为 0.254, 误差较小, 进一步说明本文建立的预测模型良好。

通过非线性拟合[6]得到  $\text{pIC}_{50}$  与  $\text{IC}_{50}$  之间的数学表达式, 假设  $\text{pIC}_{50}$  与  $\text{IC}_{50}$  之间的逻辑关系式为:

$$Y(\text{pIC}_{50}) = a * \ln(X(\text{IC}_{50})) + b \quad (5)$$

将文中  $\text{pIC}_{50}$  值与  $\text{IC}_{50}$  值导入 Matlab 中, 进行拟合, 拟合函数曲线如图 7 所示。由曲线拟合情况和基本信息可以看出, 函数拟合优秀。

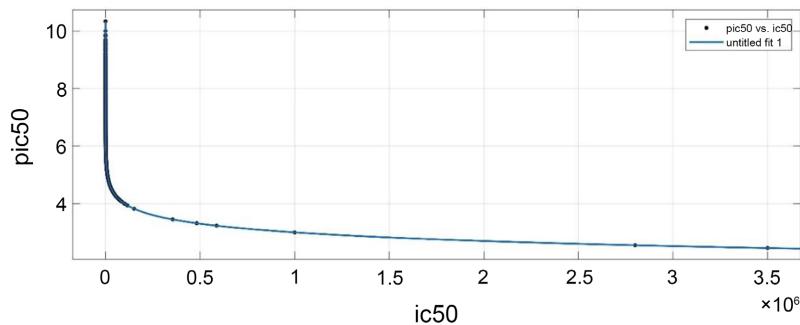


Figure 7. Fitting function curve

图 7. 拟合函数曲线

## 4. 建立各 ADMET 性质的预测模型

Logistic 回归为概率型非线性回归模型[7], 是研究二分类观察结果与一些影响因素之间关系的一种多变量分析方法。

考虑具有  $n$  个独立变量的向量  $x = (x_1, x_2, \dots, x_n)$ , 设条件概率  $P(y=1|x) = p$  为根据观测量相对于某事件  $x$  发生的概率。Logistic 回归模型可以表示为:

$$P(y=1|x) = \pi(x) = \frac{1}{1+e^{-g(x)}} \quad (6)$$

这里的  $f(x) = \frac{1}{1+e^{-x}}$  称为 Logistic 函数。其中  $g(x) = w_0 + w_1x_1 + \dots + w_nx_n$

在  $x$  条件下  $y$  不发生的概率为：

$$P(y=0|x) = 1 - P(y=1|x) = 1 - \frac{1}{1+e^{-g(x)}} = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (7)$$

故事件发生的概率与不发生的概率之比如下式，也称作发生比。

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{p}{1-p} = e^{g(x)} \quad (8)$$

对发生比取对数可得到：

$$\ln\left(\frac{p}{1-p}\right) = g(x) = w_0 + w_1x_1 + \dots + w_nx_n \quad (9)$$

根据 Logistic 回归对得出的各组 20 个主要变量进行计算得出各组回归模型的  $w_0, w_1, \dots, w_{20}$ ，如下表 1 所示。并对 1974 个化合物的 ADMET 数据，分组进行训练，并用测试集数据进行预测。各组 ADMET 性质数据均分布在 0~1 之间，各组数据分布情况如下图 8 所示。

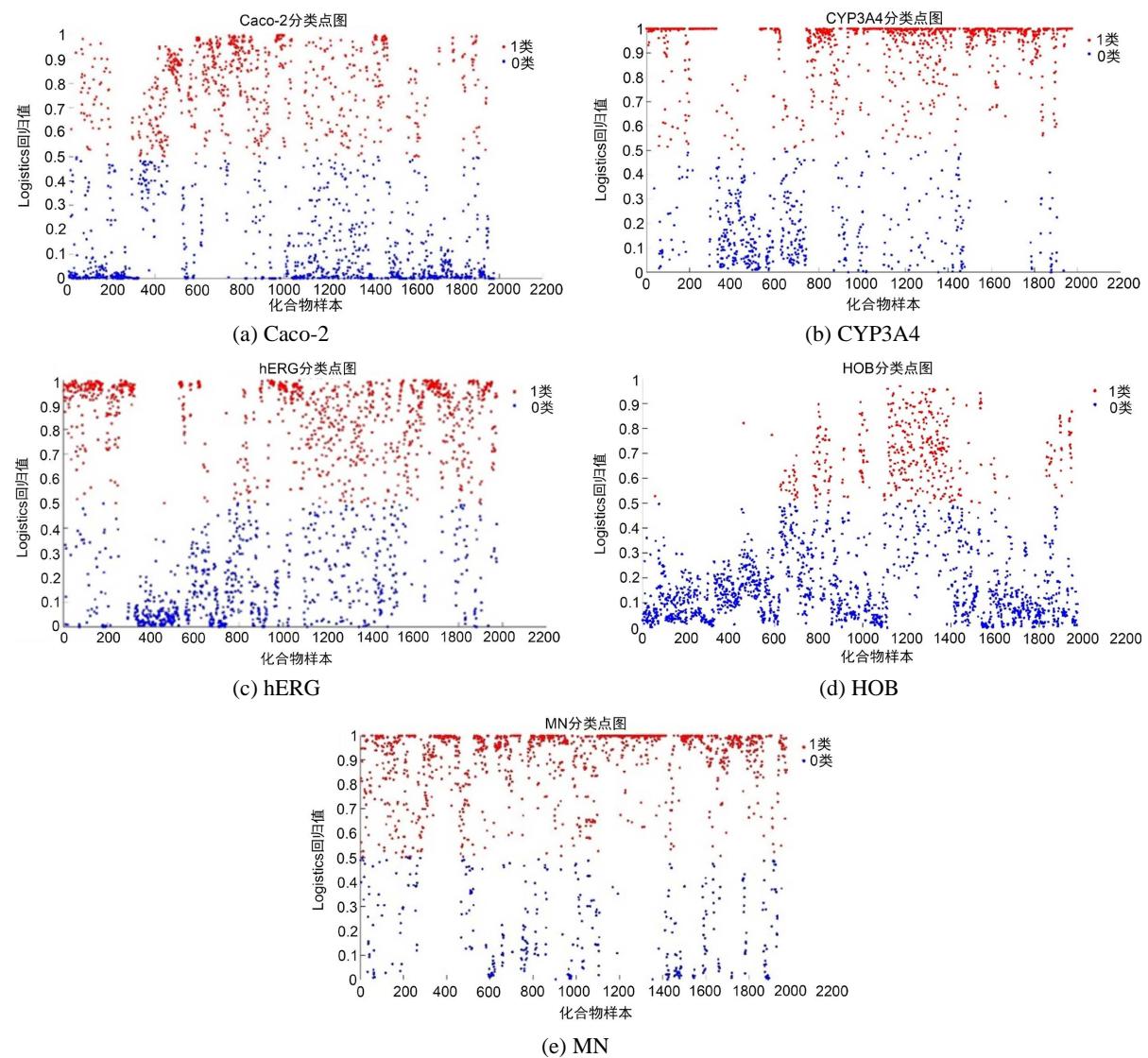
**Table 1.**  $w_i$  Values of regression models in each group

**表 1.** 各组回归模型的  $w_i$  值

$w_i$ 下标值	第 1 组 $w_i$ 值	第 2 组 $w_i$ 值	第 3 组 $w_i$ 值	第 4 组 $w_i$ 值	第 5 组 $w_i$ 值
0	-162.82	52.215	-1.582	4131.48	9.085
1	1.086	11.584	-0.003	0.834	-0.004
2	2.134	-2.932	-2.61	-2.162	-34.812
3	344.954	0.027	-3.406	0.469	-3.461
4	-1.625	-11.851	0.157	0.884	0.099
5	-5.664	0.047	-11.113	0.23	-0.019
6	-0.334	-1.311	-0.887	0.183	-76.952
7	0.516	-20.689	0.007	-0.356	1.212
8	0.019	0.006	-0.194	7.552	-0.712
9	-5.742	-2.718	0.172	0.114	0.283
10	-0.079	-0.644	-0.062	-0.429	0.158
11	2.459	-0.346	0.632	-0.094	0.131
12	0.121	-0.007	3.908	-5791.158	17.499
13	0.513	-0.248	-0.219	-5796.788	33.444
14	-0.045	0.04	-0.672	7.136	-0.121
15	-0.369	0	0.25	0.348	3.103
16	-0.167	-0.15	-0.141	1.353	0.006
17	-0.716	0.423	0.174	-0.05	0.054

Continued

18	-330.865	-1.779	-1.061	7.434	0.051
19	5.642	0.489	0.144	-5.413	1.421
20	-0.049	0.342	0.104	0.312	-14.151



**Figure 8.** Distribution of ADMET property data in each group  
**图 8.** 各组 ADMET 性质数据分布情况

由图 8 可知, 1974 个化合物样本的 Logistic 回归值分布在 0~1 之间, 以 0.5 为分界线, 大于 0.5 的数据归 1, 小于 0.5 的数据归 0。通过数据处理得出: Caco-2 组测试数据的正确性为 89.9%; CYP3A4 测试数据的正确性为 87.6%; hERG 测试数据的正确性为 88.2%; HOB 测试数据的正确性为 91.7%; MN 测试数据的正确性为 86.7%, 说明本文回归模型较好。

## 5. ER $\alpha$ 和 ADMET 性质同性影响的较优变量

首先对 ADMET 中的 Caco-2、CYP3A4、hERG、HOB、MN 进行正向化以保证各个指标值为 1 时性

质为优, 值为 0 时性质为差。进行多指标综合评价[8]后, 对数据预处理后的 247 个操作符变量根据各个指标(共 6 个)求得相应的 spearman 系数, 找到对  $ER\alpha$  和 ADMET 中三个以上具有同相关性的变量, 思路流程图如图所示。为确定各个变量的取值范围, 本文通过题目给出的数据样本, 选取 90% 样本分布区间作为变量的取值范围。然后将上述找到的变量与 ADMET 各性质间建立回归预测模型, 进而得出 ADMET 性质的综合评价方程, 再通过 PSO 粒子群算法寻 ADMET 性质优值及最优解, 最后通过此时的变量最优值去验证  $ER\alpha$  生物活性的优劣性。

### 1) 数据整理与 ADMET 评价模型建立

将 8.2 中选取的 10 个主要变量分别与 Caco-2、CYP3A4、hERG、HOB、MN 建立 Logistic 回归预测模型, 为建立 ADMET 的评价模型, 令:

$$F = F_1 + F_2 + F_3 + F_4 + F_5 \quad (10)$$

$F$ ——ADMET 的评价指标;

$F_1$ ——Caco-2 的评价指标;

$F_2$ ——CYP3A4 的评价指标;

$F_3$ ——hERG 的评价指标;

$F_4$ ——HOB 的评价指标;

$F_5$ ——MN 的评价指标;

各指标的 Logistic 回归方程形式为

$$F(x) = 1 - \frac{1}{1 + e^{-g(x)}} = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (11)$$

其中:

$$g(x) = w_0 + w_1 x_1 + \cdots + w_n x_n \quad (12)$$

由此得到各评价指标的回归系数, 如下表 2 所示:

**Table 2.** Regression coefficient of each evaluation index

**表 2.** 各评价指标的回归系数

回归系数	$F_1$	$F_2$	$F_3$	$F_4$	$F_5$
$w_0$	39.3	-11.08	-12.46	-0.2	27.58
$w_1$	-0.58	0.7	-0.67	1.81	-3.54
$w_2$	3.79	-0.33	-1.85	2.8	-3.06
$w_3$	1.43	0.23	0.8	-0.35	0.58
$w_4$	0.037	-0.48	0.11	0.35	-0.1
$w_5$	-0.88	-0.037	1.67	-1.6	2.52
$w_6$	2.26	-0.12	5.37	3.95	-1.45
$w_7$	-92.3	-42	63.32	-56.3	-35
$w_8$	48.4	47.79	-46.4	53.9	4.4
$w_9$	38.7	-1.02	-8.52	11.18	0.26
$w_{10}$	17.36	-7.23	3.85	10.26	-2.71

各回归预测正确百分比为:  $F_1$  模型为 86.75%;  $F_2$  模型为 89.58%;  $F_3$  模型为 85.32%;  $F_4$  模型为 87.45%;  $F_5$  模型为 91.53%, 说明各性质回归模型较好。

## 2) PSO 算法寻 ADMET 性质优值及最优解

PSO 又叫粒子群算法, 源于对鸟类捕食行为的研究, 其核心思想是利用群体中的个体对信息的共享使整个群体的运动在问题求解空间中产生从无序到有序的演化过程, 从而获得问题的可行解。粒子群算法流程图如下图 9 所示:

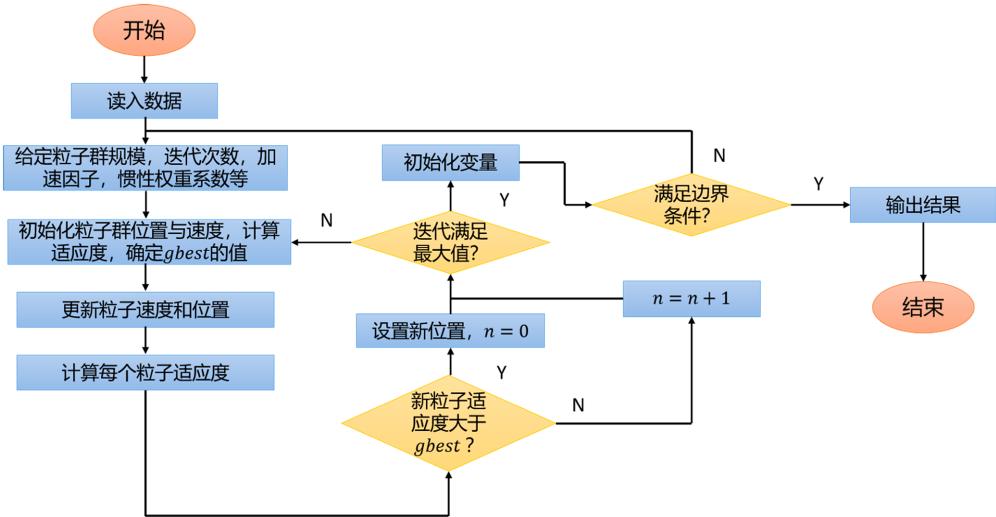


Figure 9. Flow chart of particle swarm optimization

图 9. 粒子群算法流程图

PSO 算法核心公式如下:

$$v_i^d = wv_i^{d-1} + c_1r_1(pbest_i^d - x_i^d) + c_2r_2(gbest^d - x_i^d) \quad (13)$$

其中:

$c_1$ : 粒子的个体学习因子;

$c_2$ : 粒子的社会学习因子;

$w$ : 速度的惯性权重;

$v_i^d$ : 第  $d$  次迭代时, 第  $i$  个粒子的速度;

$x_i^d$ : 第  $d$  次迭代时, 第  $i$  个粒子所在的位置;

$pbest_i^d$ : 到第  $d$  次迭代为止, 第  $i$  个粒子经过的最好的位置;

$gbest^d$ : 到第  $d$  次迭代为止, 所有粒子经过的最好位置;

$r_1, r_2$ :  $[0, 1]$  上的随机数。

最终通过粒子群寻优, 得到 ADMET 性质[9]的模型值为 5, 即说明 Caco-2、CYP3A4、hERG、HOB、MN 各性质都为优, 此时得出 10 个变量的最优数值如下表 3 所示:

Table 3. Optimal values of 10 variables

表 3. 10 个变量的最优数值

序号	变量名称	最优数值
1	nBonds	22.78
2	SP-1	9.56
3	SP-6	6.07

Continued

4	naaCH	4.06
5	WTPT-1	44.59
6	hmin	0.24
7	ETA_Epsilon_2	0.77
8	ETA_Epsilon_5	0.83
9	ETA_dPsi_A	0.16
10	ETA_Shape_P	0.17

### 3) 检验 ER $\alpha$ 生物活性

将上述变量所得得最优值代入问题 2 建立的 BP 神经网络中, 计算得出此时的  $pIC_{50}$  值为 6.8, 该值说明 ER $\alpha$  生物活性较好, 故本题筛选出的这 10 个变量及求出的最优值符合题目要求。

## 6. 模型评价与改进

### 6.1. 模型评价

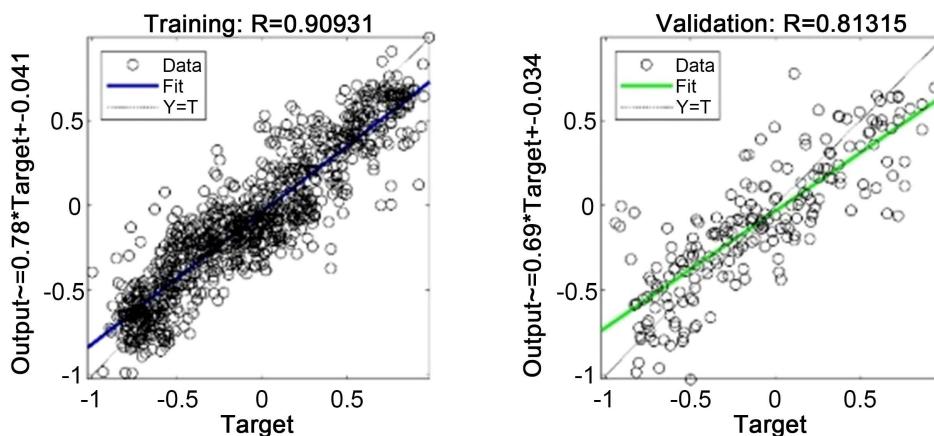
1) 模型充分考虑到 729 个分子描述符变量的异常情况, 各分子描述符变量与生物活性之间存在高度非线性, 以及各变量之间存在高度耦合性, 故先对 729 个变量进行异常值筛选, 得出 247 个对生物活性存在较大影响的变量, 之后使用随机森林回归、Spearman 相关系数进行二次降维, 确保了变量与生物活性间的相关性和变量间的独立性。

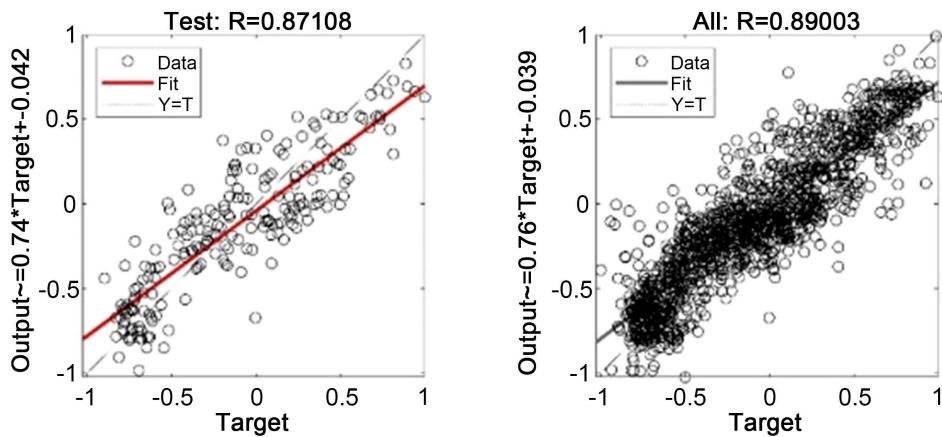
2) 由于分子描述符变量与生物活性间存在较强的非线性关系, 因此本文利用三层 BP 神经网络, 再根据自学习的网络对  $pIC_{50}$  进行预测, 最后得出预测误差比相对较小, 预测模型精度表现良好。

3) 由于 ADMET 的 5 种性质的数值符合“0~1”分布, 故采用 logistic 回归对其建立分类预测模型, 得出 5 类模型的正确预测百分比均大于 85%, 说明模型精度表现良好。

### 6.2. 模型改进

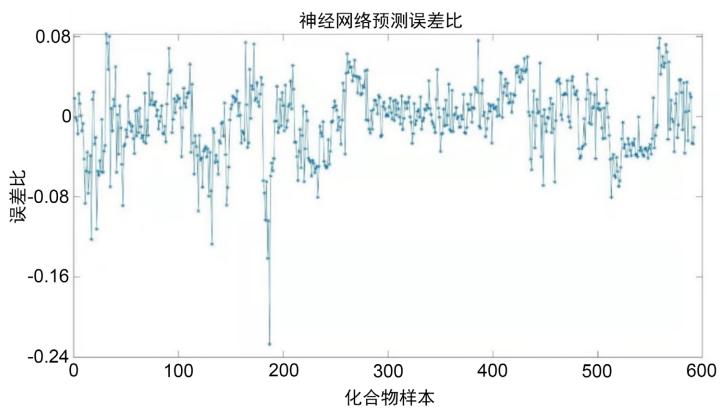
由于 BP 神经网络计算的误差是通过不断地逐层传递来修正各层间地权值与阈值, 因此本文基于 PSO 算法全局搜索 BP 神经网络的权值与阈值, 并将得到的优化初值作为 BP 网络的初始权值与阈值, 由此对文中的生物活性进行预测, 最终得出训练样本的拟合优度为 90.93%, 测试集的拟合优度为 87.11%, 说明网络训练效果比较理想, 拟合优度如下图 10 所示。





**Figure 10.** Goodness of fit  
**图 10.** 拟合优度

通过 PSO 优化后的 BP 网络, 对 ER $\alpha$  生物活性进行预测, 得出化合物样本测试集平均误差比为 3.74%, 比优化前的平均误差比降低了 19.05%, 说明优化结果比较理想。化合物样本预测误差比如图 11 所示。



**Figure 11.** Optimized compound sample prediction error ratio  
**图 11.** 优化后的化合物样本预测误差比

## 7. 结语

- 1) 本文充分利用随机森林回归、Spearman 相关性分析、BP 神经网络、Logistic 回归、PSO 算法研究了如何建立乳腺癌候选药物的预测模型。
- 2) 根据本文所用的降维方法, 得到 ER $\alpha$  生物活性的 20 个主要变量, 严格满足题目要求, 既保证变量数据与目标数据的高相关性, 且保证变量数据间的独立性, 适用于高维度且大样本的变量筛选, 对医疗药物成分数据建模具有较高参考意义。
- 3) 本文采用 BP 神经网络模型表达非线性系统, 采用粒子群算法优化 BP 神经网络初始权值, 改善预测误差, 适用于医疗药物的影响预测。
- 4) 本文采用 logistic 回归算法用于 0~1 分类系统的预测, 并得到预测准确率为 85% 以上的 5 类预测模型。
- 5) 通过同时考虑分子符描述变量对生物活性 ER $\alpha$  与 ADMET 性质产生的正相关或负相关影响, 筛选出满足条件的变量, 并通过样本数据集, 给出变量的取值范围, 并给出获得最佳药物性能时变量的取值范围, 若想得出更精确的取值范围, 则需对这些变量一一进行实验探究。

## 参考文献

- [1] 刘蒙蒙, 袁丹, 周正平, 孙小杰, 刘俊江. ER 拮抗剂对人子宫内膜样癌 JEC 细胞中 ER $\alpha$ 、ER $\beta$ 、p57~(kip2)表达的影响[J]. 临床与实验病理学杂志, 2018, 34(3): 288-294.
- [2] 白晓媛, 李世奇. 数据安全管理框架国际标准化研究[J]. 信息技术与标准化, 2021(8): 33-35+47.
- [3] 曾双. 随机森林模型在房地产评估中的适用性分析[J]. 中国管理信息化, 2021, 24(19): 29-30.
- [4] 兰文宝, 车畅, 陶成云. 基于斯皮尔曼等级相关的单演谱成分选择及其在 SAR 目标识别中的应用[J]. 电波科学学报, 2020, 35(3): 414-421.
- [5] 王景刚, 于雷, 李洪辉, 吴昌哲, 吴鹏. BP 神经网络深度对辐射计温湿度反演精度的影响[J]. 科学技术创新, 2021(29): 1-3.
- [6] 薛立娟, 邱旭阳, 李晓骅, 张博阳. 基于非线性拟合方法的雷达探测俯仰角估计[J]. 科学技术创新, 2021(4): 4-6.
- [7] 田圆圆. 危重症产妇行急诊剖宫产终止妊娠危险因素的 Logistic 回归分析及母婴结局观察[J]. 中国妇幼保健, 2021, 36(20): 4800-4803.
- [8] 张少华, 郑明强. 海绵型道路多指标综合评价方法研究[J]. 企业科技与发展, 2020(4): 110-111.
- [9] 顾耀文, 张博文, 郑思, 杨丰春, 李姣. 基于图注意力网络的药物 ADMET 分类预测模型构建方法[J]. 数据分析与知识发现, 2021, 5(8): 76-85.