Statistics and Application 统计学与应用, 2022, 11(6), 1338-1347
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.116139

Hans汉斯

抗癌药物活性的预测和优化分析

——以乳腺癌为例

侯　爽

江西财经大学，江西　南昌

收稿日期：2022年11月7日；录用日期：2022年11月27日；发布日期：2022年12月9日

摘　要

本文利用机器学习方法，探究了统计模型在制药领域的应用价值。以治疗乳腺癌的候选药物为例，在研发过程中，主要考虑如下两种指标：1) 找出能够抑制ERα活性的化合物，尽量降低ERα活性；2) 药物应当具备ADMET性质。本文采用数据挖掘技术和基本统计方法，首先基于随机森林算法从504个变量中筛选出20个主要变量，利用XGBoost算法建立化合物结构(即分子描述符)与ERα生物活性之间的非线性关系，预测ERα生物活性，结果显示用该方法预测的准确率为92.74%。再用粒子群算法建立优化模型，在保证化合物具有较好的生物活性和ADMET性质的前提下，计算出分子描述符的具体数值，该研究结果为药物设计提供了一定的参考依据。

关键词

随机森林，XGBoost，粒子群算法，优化问题

Prediction and Optimization of Anticancer Drug Activity

—A Case Study of Breast Cancer

Shuang Hou

Jiangxi University of Finance and Economics, Nanchang Jiangxi

Received: Nov. 7th, 2022; accepted: Nov. 27th, 2022; published: Dec. 9th, 2022

Abstract

This paper uses machine learning methods to explore the application value of statistical models in

文章引用: 候爽. 抗癌药物活性的预测和优化分析[J]. 统计学与应用, 2022, 11(6): 1338-1347.
DOI: 10.12677/sa.2022.116139

the pharmaceutical field. Taking drug candidates for the treatment of breast cancer as an example, in the research and development process, the following two indicators are mainly considered: 1) find out the compounds that can inhibit ER*α* activity and reduce ER*α* activity as much as possible; 2) The drug should have ADMET properties. In this paper, data mining technology and basic statistical methods are used, firstly, 20 main variables are selected from 504 variables based on random forest algorithm, and the nonlinear relationship between compound structure (*i.e.* molecular descriptor) and ER*α* biological activity is established by XGBoost algorithm to predict ER*α* biological activity, and the results show that the accuracy of prediction by this method is 92.74%. Then, the particle swarm algorithm is used to establish an optimization model, and the specific value of the molecular descriptor is calculated under the premise of ensuring that the compound has good biological activity and ADMET properties, and the results of this study provide a certain reference for drug design.

Keywords

Random Forest, XGBoost, Particle Swarm Optimization Algorithm, Optimization Problem

1. 引言

乳腺癌作为全球最常见的肿瘤疾病，其治疗方式和药物研发进程一直备受关注。国际癌症研究机构(IARC)发布的数据显示，截至 2020 年中国女性新增癌症人群中，乳腺癌患者占五分之一，乳腺癌已经成为了女性中发病率最高的癌症。相关研究表明，乳腺癌的疾病越早被发现，患者存活的机会就越大，因此提高抗乳腺癌药物的生物活性是药物研发的关键问题。

经研究发现，雌激素受体 α (ERα)作为转录因子，通过调节各种基因的转录来控制恶性肿瘤的发展[1]。因此 ERα 被认为是乳腺癌增殖的重要受体，也是乳腺癌中一个重要的预后因素，能够拮抗 ERα 的药物对于乳腺癌的治疗是必要的[2]。其次，ADMET 性质也是药物研发过程中最关键的问题之一[3]，即吸收(Absorption)、分布(Distribution)、代谢(Metabolism)、排泄(Excretion)、毒性(Toxicity)。然而在药物研发期间，分析化合物分子的生物活性和 ADMET 性质费时且费力，因此通常采用建立化合物生物活性预测模型(QSAR)的方法来筛选潜在活性化合物，国际上已有很多学者依靠数据挖掘和机器学习方法解决这一问题。Zang Q. [4]等(2013)使用随机森林特征选择方法提取与 Erα 拮抗剂活性最相关的结构特征，并使用 SVM 结合 RF 算法从大量描述符中筛选识别，构建 QSAR 模型。Zekri A. [5]等(2020)基于多元线性回归 MLR 方法，设计了一个稳健可靠的 QSAR 模型来预测抗癌药物的活性。Chang K. [6]等(2021)筛选出 12 个分子描述符，分别构建了基于 BP 神经网络、决策树和随机森林的复合 ERα 生物活性定量预测模型。认为随机森林模型可用于预测具有更好生物活性的新化合物分子。Babiker S. [7] (2020)等使用逻辑回归模型分析导致沙特妇女患乳腺癌的危险因素。

目前，针对抗乳腺癌候选药物的研究大多数倾向于预测活性，对药物活性与 ADMET 性质综合分析的文献较少；虽然近年有学者利用机器学习方法进行预测建模，但预测精度还存在提升空间。因此，本文构建集成学习 XGBoost 模型，建立化合物生物活性的定量预测模型，并采用粒子群算法，以药物活性达到最佳且至少满足三种 ADMET 性质为条件，求解化合物分子描述符的最优解。

2. 数据来源与处理

2.1. 数据来源

本文数据来源于 DrugBank 药物分子数据库，数据集包括 1974 个化合物的 729 个分子描述符和生物活性 pIC50 值。使用 train_test_split 函数来将 1974 个化合物以 4:1 划分为训练集和测试集，训练集样本数为 1579，测试集样本数为 395。在训练集上训练模型，再用测试集的数据来考察模型的预测效果。

2.2. 数据处理

由于数据表中分子描述符信息变量过多，在筛选并提取主要变量之前，需要将数据降维，弱化潜在的相关性。具体处理方法如下：

1) 缺失值处理：首先对数据进行预处理，将全为 0 的变量进行剔除，这相当于第一次的变量选择。最终剔除分子描述符变量 225 个，同时进一步针对化合物进行数据检测，发现没有其他缺失值。此时的数据为 1974 个化合物的 504 个分子描述信息。

2) 变量筛选：经过综合考虑各种降维方法的特点，本文采用随机森林实现对变量的筛选。随机森林在处理小样本量、高维特征空间和复杂数据结构方面具有独特的优势。对于每一个变量，随机森林的决策树都可以度量由该变量所导致的分裂准则函数(残差平方和或基尼指数)的下降幅度，然后针对此下降幅度，对每棵决策树进行平均，即为对该变量重要性的度量。最终依据随机森林方法进行变量重要性排序，选取了前 20 个变量作为建立 ERα 生物活性的定量预测模型的变量，所筛选的 20 个分子描述符见表 1。

Table 1. Results of variable screening
表 1. 变量筛选结果

排名	分子名称	分子描述
1	MDEC.23	所有二级和三级之间的分子距离边
2	Lipoaffinity Index	脂亲和指数
3	maxHsOH	最大原子类型 HE 态：-OH
4	minsssN	最小原子型电子态：>NH-+
5	minHsOH	最小原子类型电子态：>N-
6	SsOH	原子型电子态之和：-OH
7	BCUTc.1l	nhigh 最低部分电荷加权 BCUT
8	MLFER_A	总溶质氢键酸度或总溶质氢键酸度
9	BCUTp.1h	nlow 最高极化率加权 BCUT
10	ATSp2	ATS 自相关描述符，极化率加权
11	maxsOH	最大原子型电子态：-OH
12	minsOH	最小原子型电子态：-OH
13	McGowan_Volume	麦高文特征体积
14	VP.1	价路径，1 阶
15	nHBAcc	氢键受体数量
16	ECCEN	结合距离和邻接信息的拓扑描述符
17	nBonds2	键的总数(包括与氢的键)
18	BCUTc.1h	nlow 最高部分电荷加权 BCUT
19	C2SP2	双键碳与另外两个碳结合
20	SaaCH	原子类型 E 态之和：:CH:

3) 变量合理性评价：从变量降维流程来看，先删除数值为空的分子描述符变量，共删除 225 个变量，再采用随机森林法得到在对化合物的生物活性影响性排名前 20 的分子描述符。可视化结果见图 1，颜色越深表明相关程度越大，容易导致模型估计失真和预测结果无效等问题。由图可知，20 个变量之间的自相关性程度较低，表明化合物分描述变量具有较好的独立性，有利于下一步分析。

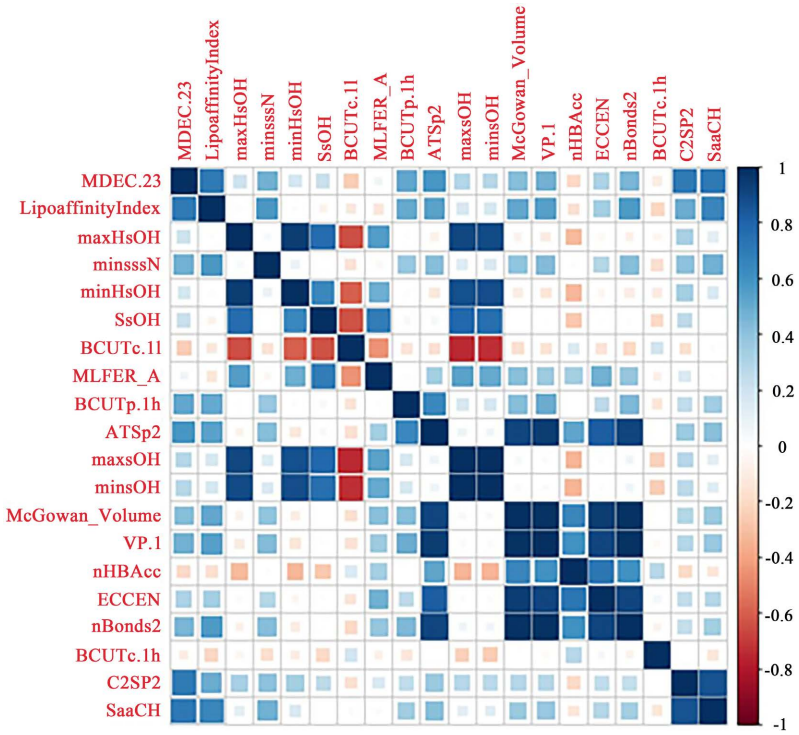

Figure 1. Variable autocorrelation plot
图 1. 变量自相关图

3. 预测模型构建

3.1. XGBoost 活性预测

XGBoost 是一种基于 GBDT 并进行了许多优化的极致梯度提升决策树算法[8]。XGBoost 能够通过对每个附加的新树进行优化，将弱学习器变成强学习器(提升)。将树修剪到定义的最大深度，然后再向后修剪，直至损失函数低于设定的值，从而使分类模型生成更少的误报、轻松标记数据和准确分类数据。其目标函数为：

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right)w_j^2\right] + \gamma T$$

其中：$G_j = \sum_{i\in I_j} g_i$，$H_j = \sum_{i\in I_j} h_i$，且 g_i 为样本i的一阶偏导数，h_i 为样本 i 的二阶偏导数。

本文选取上文所筛选出的 20 个分子描述符变量的数据与变量 pIC50 的数据进行合并，形成一个 1974 × 21 的矩阵，再从中选取 90%数据作为二次划分的训练集，而其余 197 组数据用来测试。用 XGBoost 算法进行拟合，并针对模型中的部分参数进行了改进：max_depth 表示“树的最大深度”，设置为 8；eta 学习效率默认 0.3，在每次提升后起到收缩步长的作用，得到新特征的权重，使提升过程更加具有鲁棒性；其他参数按照模型初始值设置。利用 R 软件编程实现的预测值和误差曲线如图 2 所示。

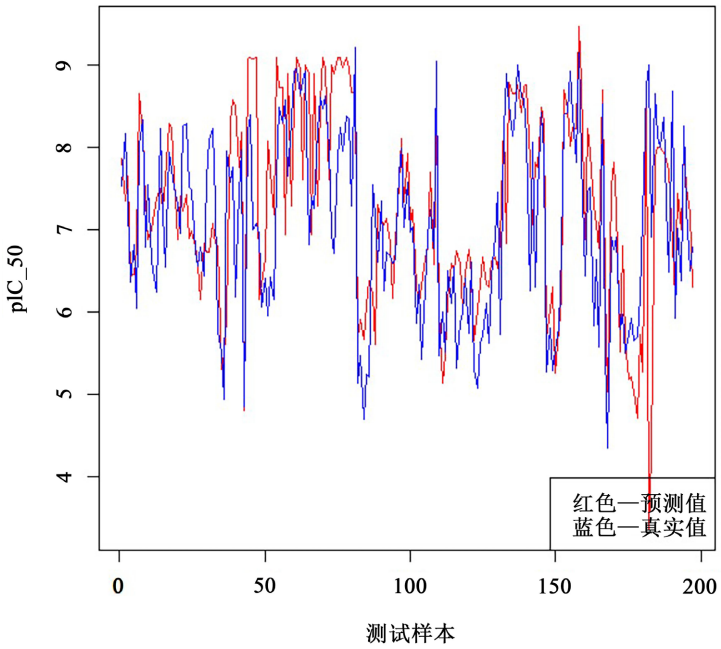

Figure 2. XGBoost fitting effect
图 2. XGBoost 拟合效果

将得到的 197 组 pIC50 预测值与同组样本实际值做比较。图 2 表明，拟合值大体上与真实情况相近，虽然存在着个别偏离真实值较严重的样本点，但整体趋势一致，出现这种情况的原因可能是因为存在试验误差或受到了参数设定的影响。图 3 描绘了残差波动的情况，拟合误差在 0 值上下波动，模型误差分布比较集中，大部分样本的误差仅在[−2, 2]之间，说明 XGBoost 模型比较稳定。

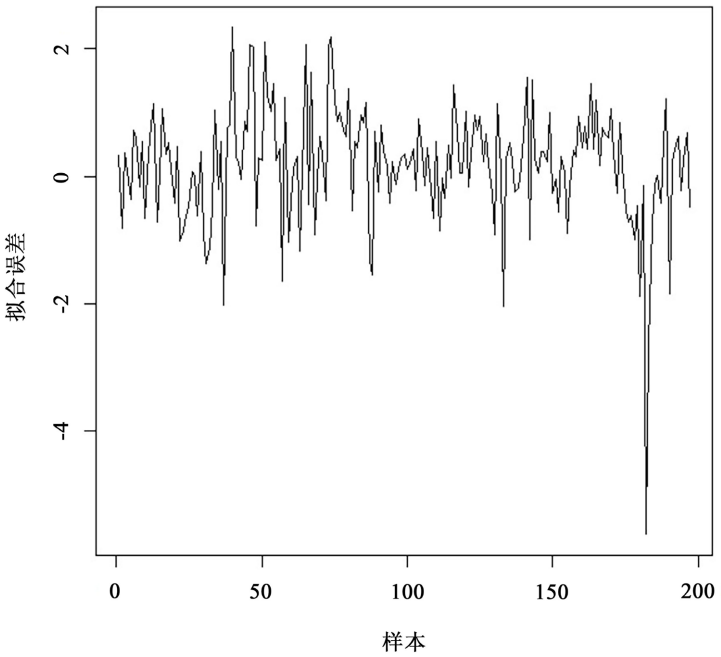

Figure 3. Residual plot of XGBoost fitting
图 3. XGBoost 拟合残差图

3.2. 模型评价

为了探究 XGBoost 算法的拟合效果，本文选取测试集中的 30 组样本，用上述方法计算了模型的拟合误差和精度，计算公式如下：

$$\text{error} = \frac{\sum\left(\frac{\hat{y}_i - y_i}{y_i}\right)}{n}$$

$$\text{精度} = 1 - \frac{|\text{误差值}|}{\text{真实值}}$$

其中，$\hat{y}_i$ 表示第 i 个样本的估计值，y_i 表示第 i 个样本的实际值。通过计算每个测试样本的误差值，计算了模型的预测精度，具体数值如表 2 所示。通过验证测试集数据，预测趋势与真实趋势一致，误差在[0, 0.8]之间，XGBoost 模型的预测效果较为良好，且预测精度较高。

Table 2. Evaluation table of algorithm prediction effect
表 2. 算法预测效果评价表

测试序号	XGBoost 预测结果	真实值	误差值	预测精度
1	7.86	7.53	0.33	95.62%
2	7.35	8.17	−0.83	89.84%
3	7.65	7.28	0.37	94.92%
4	6.44	6.36	0.08	98.74%
5	6.45	6.82	−0.36	94.72%
6	6.78	6.05	0.73	87.93%
7	8.65	8.04	0.61	92.41%
8	8.17	8.33	−0.16	98.08%
9	7.33	6.78	0.55	91.89%
10	6.89	7.54	−0.66	91.25%
11	7.00	6.63	0.37	94.42%
12	7.22	6.42	0.80	87.54%
13	7.38	6.25	1.13	81.92%
14	7.50	8.23	−0.73	91.13%
15	7.26	7.20	0.06	99.17%
16	7.61	6.55	1.07	83.66%
17	8.29	7.94	0.35	95.59%
18	8.25	7.72	0.53	93.13%
19	7.63	7.51	0.11	98.54%
20	6.88	7.30	−0.42	94.25%
21	7.42	6.95	0.47	93.24%
22	7.23	8.25	−1.02	87.64%

Continued

23	7.42	8.28	−0.86	89.61%
24	6.89	7.52	−0.63	91.62%
25	6.99	7.48	−0.49	93.45%
26	6.81	6.74	0.07	98.96%
27	6.57	6.56	0.01	99.85%
28	6.16	6.79	−0.63	90.72%
29	6.83	6.44	0.39	93.94%
30	6.73	7.62	−0.89	88.32%

4. 优化建模

4.1. ADMET 性质样本选择

对 1974 个化合物依据 Caco-2 (化合物的小肠上皮细胞渗透性)、CYP3A4 (化合物的代谢稳定性)、hERG (化合物的心脏毒性)、HOB (人体口服生物利用度)、MN (化合物的遗传毒性)这五个药代动力学性质进行筛选。

数据中 CYP3A4 为“1”代表该化合物能够被 CYP3A4 代谢，为“0”表示该化合物不能被 CYP3A4 代谢，而化合物能够被 CYP3A4 代谢表示药代动力学性质差；同时数据 MN 为“1”代表该化合物具有遗传毒性、药代动力学性质差，为“0”则代表该化合物不具有遗传毒性、药代动力学性质好。因此对数据进行处理，表 3 为满足不同 ADMET 性质个数的化合物样本数。其中，在给定的五种 ADMET 性质中，有 543 个化合物样本至少满足了三种性质。

Table 3. Properties of compound ADMET
表 3. 化合物 ADMET 性质满足情况

化合物 ADMET 性质满足个数	化合物样本数
0	571
1	429
2	431
3	364
4	162
5	17

4.2. 优化目标及约束设定

1) 决策变量

本文所建立的基于 XGBoost 算法的化合物对 ERα 生物活性的定量预测模型的主要变量主要有 20 个，因此该粒子群算法模型中共有 20 个决策变量，记为：

$$X=\{x_1,x_2,\cdots,x_{20}\}$$

2) 目标函数及约束条件

本文的优化目标在于得出化合物的哪些分子描述符，以及这些分子描述符在什么取值或者处于什么

取值范围时，能够使化合物对抑制 ERα 具有更好的生物活性，同时具有更好的 ADMET 性质。化合物的筛选满足了好的 ADMET 性质，因此本文以生物活性 pIC50 为主要的优化目标，即要求化合物的生物活性尽可能大。同时，考虑到提出的 20 个分子描述符在操作时需要给定范围，不能无限大或无限小，因此设定约束为：

$$0 \le x_i \le \max x_{ib},\ i = 1,2,\cdots,20;\ b = 1,2,\cdots,543$$

另外，需要保证经过 XGBoost 预测的生物活性大于原来的生物活性值，于是提出约束：

$$\min \text{pIC50} < \text{pIC50}_{\text{predict}} \le \max \text{pIC50}$$

所以目标函数及约束条件为：

$$\max : \text{pIC50}_{\text{predict}}$$

$$s.t.\begin{cases} 0 \le x_i \le \max x_{ib},\ i = 1,2,\cdots,20;\ b = 1,2,\cdots,542,543 \\ \min \text{pIC50} \le \text{pIC50}_{\text{predict}} \le \max \text{pIC50} \\ \text{pIC50}_{\text{predict}} = f_{\text{XGBoost}}\left(x_i\right),\ i = 1,2,\cdots,20 \end{cases}$$

其中，$f_{\text{XGBoost}}\left(x_i\right)$ 表示化合物分子描述符与生物活性的 XGBboost 算法预测模型。

4.3. 粒子群算法参数设定

粒子群算法是一种受到群居动物集体行为的启发而产生的优化方法。假设一群粒子进入了函数的搜索空间，每个粒子在搜索时，需要考虑个人搜索的历史最佳位置和其他粒子搜索的全剧最佳位置。每一次搜索中，各粒子通过迭代来更新自己的位置和速度生成一个新的序列，再用新序列中的最优的粒子替换某个粒子的位置，从而搜寻全局最优解。主要变量设定及相关参数如表 4 所示。

Table 4. Parameter setting
表 4. 参数设定

参数名称	符号	参数值
种群大小	m	543
维数	D	20
权重因子	ω	0.9
学习因子	C_1:C_2	2
粒子的最大速度	V_{MAX}	0.5
最大迭代步数	t	500
初始化粒子位置	x_0	随机数
初始化粒子速度	v_0	随机数

4.4. 模型求解

针对上述建立的复杂多约束优化模型，本文使用 R 语言编写粒子群算法程序进行求解。在演化过程中，样本的平均迭代次数为 16 次，并未超过预设的参数，表明模型参数设置合理，模型求解性能较好，比较准确快速。本文对 543 个数据样本求解生物活性最大时的主要变量对应的取值。可求得在满足约束条件下，每个化合物分子描述符最优取值统计结果如表 5 所示。

Table 5. Optimal values of molecular descriptors
表 5. 分子描述符最优取值表

分子描述符	分子描述	最优解
MDEC.23	所有二级和三级之间的分子距离边	38.603
Lipoaffinity Index	脂亲和指数	16.0788
maxHsOH	最大原子类型 HE 态：-OH	11.362
minsssN	最小原子型电子态：>NH-+	2.5055
minHsOH	最小原子类型电子态：>N-	0
SsOH	原子型电子态之和：-OH	0
BCUTc.1l	nhigh 最低部分电荷加权 BCUT	−0.416
MLFER_A	总溶质氢键酸度或总溶质氢键酸度	2.197
BCUTp.1h	nlow 最高极化率加权 BCUT	16.748
ATSp2	ATS 自相关描述符，极化率加权	3535.7
maxsOH	最大原子型电子态：-OH	0
minsOH	最小原子型电子态：-OH	11.166
McGowan_Volume	麦高文特征体积	3.837
VP.1	价路径，1 阶	4.699
nHBAcc	氢键受体数量	0
ECCEN	结合距离和邻接信息的拓扑描述符	206
nBonds2	键的总数(包括与氢的键)	22
BCUTc.1h	nlow 最高部分电荷加权 BCUT	0.0721
C2SP2	双键碳与另外两个碳结合	18
SaaCH	原子类型 E 态之和：:CH:	31.95

在满足至少三种 ADMET 性质较好的前提下，pIC50 活性值达到最大时，20 个变量的最优解如表 5 所示，其中 ATSp2 的数值应为 3535.7，ECCEN 的数值为 206，minHsOH、SsOH、maxsOH、nHBAcc 四个变量的最优解为 0，其余 14 个分子描述符的最优解分布在[0, 50]范围内，数值较接近。

通过观察每个化合物分子描述符最优取值，一方面可以为抗乳腺癌药物的设计提供理论支撑，另一方面也可为 ERα 拮抗剂的优化提供参考依据。该优化结果从理论上具备了可靠性和实用价值，能够减少研发成本和资源浪费，未来可用于设计具有更好生物活性的新化合物分子，对乳腺癌药物的研制具有一定的现实意义。

5. 结论

针对抗乳腺癌药物研发过程中的生物活性预测和求解化合物最优值问题，本文首先用随机森林方法降解数据维度，并对所有化合物的重要性排序，筛选出前 20 个对生物活性有显著影响的分子描述符，建立了分子描述符和生物活性之间的数学关系。基于选定模型，在保证较高生物活性和良好药物性质的前提下，确定各分子描述符的最佳取值。全文主要研究结果如下：

1) 本文自变量为候选药物的生物活性，因变量为构成化合物的 20 个分子描述符，用 XGBoost 模型拟合二者关系，通过观察预测精度的高低来检验模型是否合适。根据表 2，选取的预测方法对数据拟合

能力较强，估计误差在 0 值上下平稳波动，因此变量选择的结果较为合适且模型合理。

2) 以化合物生物活性最佳为目标，通过设定约束条件和粒子群参数值，计算了化合物分子描述符的具体值。取值结果如表 5 所示，其中 ATSp2 取值最大，为 3535.7。大部分描述符的范围在[0, 50]区间内取值，此时能够使化合物在保证良好药代动力学性质下，对抑制 ERα 具有更好的效果。

参考文献

[1] Saba, A., Muhammad, S., Khera, R.A., *et al*. (2022) Identification of Halogen-Based Derivatives as Potent Inhibitors of Estrogen Receptor Alpha of Breast Cancer: An *In-Silico* Investigation. *Journal of Computational Biophysics and Chemistry*, **21**, 181-205. https://doi.org/10.1142/S2737416522500090

[2] Bentzon, N., Düring, M., During, R.B., *et al*. (2008) Prognostic Effect of Estrogen Receptor Status across Age in Primary Breast Cancer. *International Journal of Cancer*, **122**, 1089-1094. https://doi.org/10.1002/ijc.22892

[3] 沈杰. 药物 ADMET 理论预测方法开发和靶向雌激素受体的药物设计研究[D]: [博士学位论文]. 上海: 华东理工大学, 2011.

[4] Zang, Q., Rotroff, D.M. and Judson, R.S. (2013) Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure-Activity Relationship and Machine Learning Methods. *Journal of Chemical Information and Modeling*, **53**, 3244-3261. https://doi.org/10.1021/ci400527b

[5] Zekri, A., Harkati, D., Kenouche, S., *et al*. (2020) QSAR Modeling, Docking, ADME and Reactivity of Indazole Derivatives as Antagonizes of Estrogen Receptor Alpha (ER-α) Positive in Breast Cancer. *Journal of Molecular Structure*, **1217**, Article ID: 128442. https://doi.org/10.1016/j.molstruc.2020.128442

[6] Chang, K., Liu, S., Yan, H., *et al*. (2021) Quantitative Structure-Activity Relationship Modeling of Estrogen Receptor Alpha Bioactivity Based on Multiple Algorithms. 2021 4*th International Conference on Algorithms, Computing and Artificial Intelligence*, 1-6. https://doi.org/10.1145/3508546.3508572

[7] Babiker, S., Nasir, O., Alotaibi, S.H., *et al*. (2020) Prospective Breast Cancer Risk Factors Prediction in Saudi Women. *Saudi Journal of Biological Sciences*, **27**, 1624-1631. https://doi.org/10.1016/j.sjbs.2020.02.012

[8] Chen, T.Q. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. 22*nd SIGKDD Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785