Computer Science and Application 计算机科学与应用, 2024, 14(2), 298-307
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142030

Hans汉斯

基于机器学习的抗乳腺癌药物活性预测模型

贺　冰[1]，甘俊毅[2]

[1]同济大学电子与信息工程学院，上海
[2]同济大学软件学院，上海

收稿日期：2024年1月22日；录用日期：2024年2月21日；发布日期：2024年2月27日

摘　要

乳腺癌是一种由乳腺上皮细胞不受控制地增殖，并最终导致恶性变化的疾病。作为女性最常见的恶性肿瘤之一，乳腺癌的发病率与雌激素受体密切相关。雌激素受体α亚型($ER\alpha$)被视为治疗乳腺癌的关键靶标，因此能够拮抗$ER\alpha$活性的化合物被认为是潜在的乳腺癌治疗药物。在药物研发阶段，建立化合物分子结构描述符与生物活性值得的定量关系模型对指导新药物的设计和优化具有重要意义。这不仅可以节约研发资源，还有望加速新药物的上市进程，为乳腺癌治疗领域的研究和开发提供重要的支持。然而，化合物分子描述符种类繁多，直接进行活性预测效果不佳。因此，本文首先提出了一种分子描述符筛选方法，使用基于最大期望算法的高斯混合模型进行分子描述符分布检测，接着根据少量化合物分子描述符和生物活性的对应关系，使用随机森林降维算法选取候选分子描述符组，最后使用斯皮尔曼相关性系数计算剔除掉相关性较高的分子描述符，最终得到对生物活性影响大且相关性低的分子描述符组。接着，本文使用随机森林回归算法预测药物活性，并创新地利用遗传算法求解映射函数，对原始回归目标进行均衡化处理。在化合物$ER\alpha$生物活性数据集上的实验表明，我们的生物活性预测模型取得了很好的效果。

关键词

随机森林，遗传算法，药物活性预测

Prediction Model of Anti-Breast Cancer Drug Activity Based on Machine Learning

Bing He[1], Junyi Gan[2]

[1]School of Electronic and Information Engineering, Tongji University, Shanghai
[2]School of Software, Tongji University, Shanghai

Received: Jan. 22nd, 2024; accepted: Feb. 21st, 2024; published: Feb. 27th, 2024

Abstract

Breast cancer is a disease characterized by uncontrolled proliferation of epithelial cells in the

文章引用: 贺冰, 甘俊毅. 基于机器学习的抗乳腺癌药物活性预测模型[J]. 计算机科学与应用, 2024, 14(2): 298-307.
DOI: 10.12677/csa.2024.142030

breast, leading to malignant transformation. As one of the most common malignancies in women, the incidence of breast cancer is closely associated with estrogen receptors. The estrogen receptor alpha subtype (*ERα*) is considered a crucial target for treating breast cancer, and compounds capable of antagonizing *ERα* activity are regarded as potential therapeutic agents. During the drug development phase, establishing a quantitative relationship model between compound molecular structure descriptors and bioactivity values is of paramount importance in guiding the design and optimization of new drugs. This not only conserves research and development resources but also holds the promise of expediting the market entry of new drugs, providing essential support for research and development in the field of breast cancer treatment. However, due to the diverse types of compound molecular descriptors, direct prediction of activity yields suboptimal results. Therefore, we introduce a molecular descriptor screening method. It employs a Gaussian mixture model based on the Expectation-Maximization algorithm for molecular descriptor distribution detection. Subsequently, based on the correspondence between a small number of compound molecular descriptors and bioactivity, a random forest dimensionality reduction algorithm is used to select a candidate set of molecular descriptors. Finally, the Spearman correlation coefficient is employed to eliminate highly correlated descriptors, resulting in a set of molecular descriptors with significant impact on bioactivity and low correlation. Next, we utilize a random forest regression algorithm for predicting drug activity and innovatively employ a genetic algorithm to solve the mapping function, balancing the original regression target. Experimental results on a dataset of *ERα* bioactivity of compounds demonstrate the effectiveness of our bioactivity prediction model.

Keywords

Random Forest, Genetic Algorithm, Drug Activity Prediction

1. 引言

乳腺癌是一种异常乳腺细胞生长失控并形成肿瘤的疾病。如果不加以控制，肿瘤会扩散到全身并致命。乳腺癌是女性最常见的恶性肿瘤之一[1]，其发病率与雌激素受体密切相关。研究表明，雌激素受体 α 亚型(*ERα*)在正常乳腺上皮细胞中的表达不超过 10%，但在乳腺肿瘤细胞中的表达约为 50%至 80%。对于*ERα*基因缺失小鼠的实验结果显示，*ERα* 在乳腺发育过程中扮演着至关重要的角色。目前，对于基因 *ERα* 表达的乳腺癌患者，会采用抗激素治疗，通过调节雌激素受体活性来控制体内雌激素水平。因此，*ERα* 被认为是治疗乳腺癌的重要靶标，能够拮抗 *ERα* 活性的化合物可能成为治疗乳腺癌的候选药物。

在药物研发领域，为了提高效率并降低成本，通常采用建立化合物活性预测模型的方法，以筛选潜在活性化合物。具体而言，针对与某一疾病相关的靶标(如 *ERα*)，研究者会收集一系列作用于该靶标的化合物及其生物活性数据。然后，以一系列分子结构描述符作为自变量，将化合物的生物活性值作为因变量，构建化合物的定量结构－活性关系(Quantitative Structure-Activity Relationship, QSAR)模型[2]。A 该模型可以用于预测具有更好生物活性的新化合物分子，或者指导已有活性化合物的结构优化。

然而，面对复杂多样的化合物分子描述符，直接进行活性预测的效果并不理想。因此，本文首先提出了一个活性敏感的分子描述符筛选算法。传统的主成分分析方法虽然能够方便快捷地实现降维操作，

但却忽略了所选分子描述符与生物活性之间的关系，可能导致所选描述符虽能代表化合物的部分特征，但与生物活性的特征关联性较弱。单一方法实现降维操作也容易忽略所选分子描述符之间的相关性，使得选出的分子描述符之间可能存在高度耦合关系。为了解决这两个问题，本文首先采用基于 EM 算法(Expectation-Maximization Algorithm, EM 算法)的高斯混合模型(Gaussian Mixture Model, GMM)对分子描述符进行聚类，以便将具有相似特征的分子描述符尽可能聚为一类。随后，利用信息增益理论计算每种分子描述符对生物活性值的信息增益，从聚类结果中的每类分子描述符中挑选信息增益最大的一项，形成一组能够代表不同特征信息且对生物活性影响显著的分子描述符。若聚类的类别数远大于所需选择的分子描述符变量数，则说明大多数分子描述符变量都可以代表不同特征信息，此时可直接选取信息增益最高的分子描述符变量组成一组，而不受聚类类别的影响。最后，进一步通过斯皮尔曼相关性系数计算所挑选的分子描述符之间的相关性，剔除候选分子描述符组中信息增益相对较低且相关性系数较高的分子描述符变量，最终筛选出对生物活性影响显著且相互之间具有低耦合关系的分子描述符。

在化合物活性预测领域，回归随机森林(Regression Random Forest, RFR)方法[3]得到广泛应用，其具有抗干扰能力和泛化能力强、可以平衡不均衡数据集带来的误差等优点[4]。然而，回归随机森林使用均方误差(Mean Squared Error, MSE)作为优化指标，在训练过程中，RFR 通过最小化 MSE 来调整树模型的参数，以提高对训练数据的拟合效果。当样本分布较为集中时，即目标变量的取值范围较小或样本在某个区域内密集分布时，均方误差可能过于敏感。同时均方误差对离群值非常敏感，因为它是每个样本与预测值之差的平方的平均值。如果存在一些样本在分布中相对较远，其残差较大，那么均方误差就会受到这些离群值的影响而增大。模型更倾向于在训练数据中呈现较大残差的区域进行过度拟合，而这可能并不代表整体样本分布的真实情况。因此，本文建立了基于遗传算法的改进随机森林模型，以提高预测结果的准确性。具体来说，本文在使用回归随机森林方法构建$ER\alpha$生物活性定量预测模型的基础上，考虑建立一个映射函数，使得回归目标变量均衡化，从而提高回归随机森林对映射后预测变量的学习性能。然后，再对模型的预测输出使用映射函数的反函数映射至原始目标分布空间。映射函数的求解可以简化为一个目标优化问题，本文采用遗传算法进行求解，以获得一个性能良好的映射函数，从而提高预测模型的准确率。

2. 相关工作

2.1. 基于 EM 算法的高斯混合模型

高斯混合模型是指多个高斯分布函数的线性组合，理论上可以拟合出任意形式的分布，通常用于解决同一集合下的数据包含多个不同的分布的情况。该方法假设输入样本服从 K 个参数未知的高斯分布，服从同一分布的则被聚为一类，并利用 EM 最大算法进行迭代拟合数据分布。

2.2. 回归随机森林

回归随机森林模型[3]由多个决策树组成，而决策树中每一个非叶结点都对应着一个切分变量，也即，一个分子描述符。某一个分子描述符i的重要性f_i定义为所有用这个分子描述符作为切分变量的结点的重要性之和与所有结点的重要性之和的比值：

$$f_i = \frac{\sum_{j \in nodessplitonfeaturei} n_j}{\sum_{k \in alinodes} n_k} \tag{1}$$

其中，n_j 和 n_k 分别为结点 j 和结点 k 的重要性。某一个结点k所对应的分子描述符的重要性定义为：

$$n_k = w_k \times G_k - w_{left} \times G_{left} - w_{right} \times G_{right} \tag{2}$$

其中，w_k，w_{left}，w_{right} 分别为结点k及其左右子结点中训练样本个数与总训练样本个数的比例，G_k，G_{left}，G_{right} 为结点k及其左右子结点的不纯度。不纯度是在训练决策树的过程中衡量一个切分变量以及切分点的优劣的指标。某一结点的不纯度为各子结点的不纯度的加权和$G\left(x_i,v_{ij}\right)$：

$$G\left(x_i,v_{ij}\right)=\frac{n_{left}}{N_S}H\left(X_{left}\right)+\frac{n_{right}}{N_S}H\left(X_{right}\right) \tag{3}$$

x_i 为某一个切分变量，v_{ij} 为切分变量的一个切分值，n_{left} 为切分后左子树的训练样本个数，n_{right} 为切分后右子树的训练样本个数，N_S 为当前树的所有训练样本个数。$H\left(X\right)$为衡量节点不纯度的函数，在回归任务中，不纯度函数为均方误差(MSE)：

$$H\left(X_m\right)=\frac{1}{N_m}\sum_{i\in N_m}\left(y-\bar{y}_m\right)^2 \tag{4}$$

或平均绝对误差(MAE)：

$$H\left(X_m\right)=\frac{1}{N_m}\sum_{i\in N_m}\left|y-\bar{y}_m\right| \tag{5}$$

为了使所有特征的重要性之和为 1，对每一个特征的重要性使用进行归一化：

$$f_{n_i}=\frac{f_i}{\sum_{j\in allfeatures}f_j} \tag{6}$$

2.3. 斯皮尔曼相关系数

斯皮尔曼相关系数也称斯皮尔曼秩相关系数，是衡量两个变量的依赖性的非参数指标。秩可以理解为一种顺序或者排序，根据原始数据的排序位置进行求解。它利用单调方程评价两个统计变量的相关性。

3. 算法原理及模型结构

3.1. 分子描述符筛选算法

本文提出的分子描述符筛选算法由三个关键模块组成，分别为聚类分析、信息增益筛选和相关性过滤。这三个模块共同构成了一个综合而有效的筛选框架，用于挑选出对生物活性影响显著的、具有代表性的分子描述符。

首先，聚类分析模块采用基于 EM 算法的 GMM 对分子描述符进行聚类操作。通过将具有相似特征的分子描述符划分到同一类别，该模块实现了对化合物特征的有效整理，使得相似性较高的描述符被聚集在一起。这有助于减少描述符的冗余性，提高后续分析的效率。

化合物的分子描述符变量可以表述为：

$$x_i=\left(x^1,x^2,\cdots,x^d\right) \tag{7}$$

包含 K 个高斯分布函数的高斯混合模型可以表述为：

$$p\left(x\right)=\sum_{k=1}^{K}\omega_k N\left(x\middle|\mu_k,\sum_k\right) \tag{8}$$

其中，ω_k、μ_k、$\sum_k$ 分别代表第 k 个高斯密度函数的权重、均值和协方差矩阵。$\sum_{k=1}^{K}\omega_k=1,\ 0\le\omega_k\le1$，$N\left(x\middle|\mu_k,\sum_k\right)$表示第 k 个高斯密度函数：

$$N\left(x\left|\mu_k,\sum_k\right.\right)=\frac{1}{\sqrt{\left|\sum_k\right|2\pi^{\frac{d}{2}}}}\exp\left\{-\frac{1}{2}\left(x-\mu_k\right)^T\sum_k^{-1}\left(x-\mu_k\right)\right\} \tag{9}$$

高斯混合分布的参数集合可以表示为：

$$\theta=\left\{\omega_1,\cdots,\omega_K,\mu_1,\cdots,\mu_K,\Sigma_1,\cdots,\Sigma_K\right\} \tag{10}$$

其估计值 $\overline{\omega}_k,\overline{\mu}_k,\overline{\Sigma}_k$ 可由下式计算：

$$L\left(X\right)=\ln^n\left[\prod_{i=1}p_M\left(x_i\right)\right]=\sum_{i=1}^{n}\ln\left[\sum_{k=1}^{K}\omega_k\cdot p_M\left(x_i\left|\mu_k,\Sigma_k\right.\right)\right] \tag{11}$$

则第 i 个化合物 x_i 属于第 k 个高斯分布，即第 k 类的可能性为：

$$\gamma_{ik}=p_M\left(z_i=k\left|x_i\right.\right)=\frac{P\left(z_i=k\right)\cdot p_M\left(x_i\left|z_i=k\right.\right)}{p_M\left(x_i\right)}=\frac{\overline{\phi}_k\varphi\left(x_i\left|\overline{\mu}_k,\overline{\sum_k}\right.\right)}{\sum_{k=1}^{K}\overline{\phi}_k\varphi\left(x_i\left|\overline{\mu}_k,\overline{\sum_k}\right.\right)} \tag{12}$$

最终第 i 个样本的类别为：

$$\lambda_i=\underset{k\in\{1,2,\cdots,K\}}{\arg\max}\,\gamma_{ik} \tag{13}$$

其次，信息增益筛选模块会对每个聚类中的分子描述符计算其对生物活性值的信息增益。回归随机森林方法可以在训练过程中衡量每个分子描述符对于预测精度的贡献程度[5]。因此，本文利用随机森林训练过程中的每个分子描述符对预测精度的信息增益，选择具有最大信息增益的分子描述符，该模块从每个聚类中精心挑选出最富信息、最能够代表该类别特征的描述符。这种筛选机制有助于确保所选描述符不仅能够代表整体数据特征，同时对生物活性的影响更为显著。

最后，相关性过滤模块使用斯皮尔曼相关性系数计算所挑选的分子描述符之间的相关性。通过剔除信息增益较低且相关性较高的分子描述符变量，该模块筛选出最终对生物活性影响显著且相互之间具有低耦合关系的分子描述符。这一步骤有助于消除潜在的冗余信息，保留关键的特征描述符，提高建模的准确性和可解释性。斯皮尔曼相关系数是衡量两个变量的依赖性的非参数指标。它利用单调方程评价两个统计变量的相关性。如果数据中没有重复值，并且当两个变量完全单调相关时，斯皮尔曼相关系数则为+1 或−1。其计算公式为：

$$\rho=\frac{\sum_i\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sqrt{\sum_i\left(x_i-\overline{x}\right)}\sqrt{\sum_i\left(x_i-\overline{x}\right)}} \tag{14}$$

其中 x_i,y_i 表示分别变量 X，Y 的秩序。

综合而言，这三个模块相互协作，构建了一个全面的分子描述符筛选框架，为药物研发提供了有力的工具，可以更精确地挑选出具有生物活性相关性的关键分子特征。

3.2. 基于遗传算法的改进随机森林模型

随机森林属于 Bagging 类算法，是一种集成学习方法。集成学习方法的主要思想是训练多个弱模型并将多个弱模型组成一个强模型。强模型的性能要明显优于比单个弱模型。当弱模型选择为决策树时，就是随机森林算法。随机森林可用于分类和回归任务。RFR 是由多个二叉决策树组成的，训练 RFR 也就是训练多个二叉决策树。

首先，将训练数据集通过 bootstrap 以及 subsample 采样划分为 n 个子数据集。bootstrap 是统计学习

中的一种重采样技术。在对训练数据集 D 进行采样时，D 中的每个样本会以 63.3%的概率被抽中，这样可以划分出多个不同的数据集，从而训练出多个不同的子模型，增加最终模型预测结果的鲁棒性和稳定性。假设 D_b 是对输入的训练样本集合 D 进行 bootstrap 抽样后得到的样本集合，那么 subsample 会根据输入参数 sample_sz 的大小从 D_b 中采样 sample_sz 个样本组成集合 D_{bs}。

其次，单独地对每个子数据集训练决策树模型。在训练二叉决策树的时候以切分后节点的不纯度来衡量一个切分变量以及切分点的优劣。在活性预测中，使用均方误差作为不纯度函数。

在决策树中某一结点的训练过程等价与以下优化问题：

$$\left(x^*, v^*\right) = \arg\min_{x,v} G\left(x_i, v_{ij}\right) \tag{15}$$

分子活性的预测结果由随机森林内部所有决策树输出的预测结果取平均值得到。

由于回归随机森林使用均方误差作为优化指标，依据均方误差的大小对特征进行空间划分，因此随机森林回归结果对样本分布较为敏感，对于预测目标值之间的差距较大且分布较为均匀的样本有较好效果。当分子活性的分布较为集中时，不利于随机森林的学习。因此，本文将回归目标变量进行均衡化操作，即寻找一个映射函数 $f_m : y \to y'$，将原始的化合物活性预测值 y 映射到 y'，使得回归随机森林对映射后的预测变量的学习表现较好。使用映射后的目标变量值对模型进行训练，再对训练好的模型的预测输出使用映射函数的反函数 $f_m^{-1} : y' \to y$ 映射至原始目标分布空间。

我们使用多项式函数作为映射函数 f_m：

$$f_m\left(y\right) = a_n y^n + a_{n-1} y^{n-1} + \cdots + a_1 y + a_0 \tag{16}$$

为保证映射关系始终为单射，即 $y \to y'$ 和 $y' \to y$ 的过程都是一一对应的，我们假设映射函数 f_m 的导数恒大于 0。我们使用遗传算法寻找最优的映射关系，并设置导数恒为正的约束，优化目标为最小化使用 f_m 映射后的活性值训练的随机森林模型的均方误差，通过不断迭代，最终找到一个性能较好的映射函数。

求解得到多项式系数后，使用该多项式函数对训练集中的活性值进行变换，重新训练回归随机森林。在进行预测时，将随机森林模型的输出通过反函数 f_m^{-1} 映射回原始化合物活性指标值。

4. 实验

4.1. 数据集介绍

本文使用的数据集来自 2021 年华为杯研究生数学建模大赛提供的“Molecular_Descriptor.xlsx”和“*ERα*_activity.xlsx”。该数据提供了 1974 个化合物及其 *ERα* 活性，每个化合物有 729 个分子描述符。

4.2. 实验设置

在实验验证阶段，本文从 729 个分子描述符中利用信息增益筛选出的候选分子描述符为 30，去除相关性较高的分子描述符之后的数量设置为 20。其余实验设置如 4.3 节中所述。

4.3. 结果与讨论

4.3.1. 聚类分析结果

通过对 1974 个化合物的 504 个分子描述符利用高斯混合模型和 EM 算法进行分类，并利用 AIC 及 BIC 准则(值越大分类效果越好)尝试获得最优的分类类别数，可以发现如图 1 所示，AIC 及 BIC 值基本上随着类别数的增加而不断增加。因此本文认为绝大多数分子描述符均属于各自不同的类别，其在数据上的表现各不相同，可以根据增益信息的高低随意选择分子描述符。

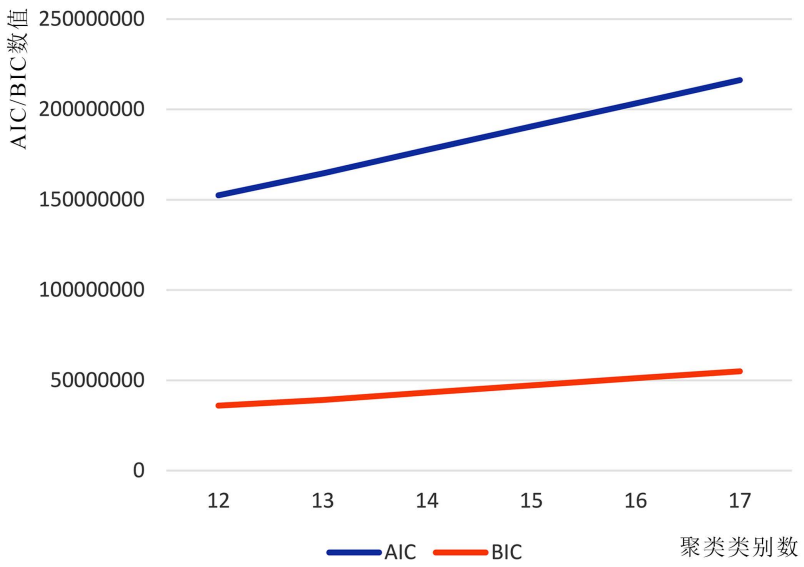

Figure 1. Variation of AIC/BIC values with changes in cluster numbers
图 1. AIC/BIC 值随聚类数的变化情况

4.3.2. 信息增益筛选结果

使用 python 语言 Scikit-learn 库[6]中的 RandomForestRegressor 实现回归随机森林模型。训练完成后模型输出的最重要的 30 个分子描述符及其重要性如表 1 所示：

Table 1. Top 30 Molecular Descriptors with the greatest influence on the activity of compounds to inhibit *ERα*
表 1. 对化合物抑制 *ERα* 的活性影响最大的 30 个分子描述符及其重要性

分子描述符名称	重要性	分子描述符名称	重要性
MDEC-23	0.28976	VCH-5	0.01022
minsOH	0.09127	SP-6	0.00969
C1SP2	0.07222	WTPT-2	0.00923
maxssO	0.05466	ATSc5	0.00858
TopoPSA	0.02286	ATSc2	0.00737
MDEO-12	0.02030	BCUTp-1h	0.00698
minHBa	0.02006	WTPT-5	0.00694
SHBint10	0.01885	SHsOH	0.00686
BCUTc-1h	0.01873	minsssN	0.00682
MLFER_S	0.01811	nHBint6	0.00660
maxHBd	0.01405	XLogP	0.00630
ndssC	0.01250	maxHdsCH	0.00628
SP-3	0.01175	SPC-6	0.00616
MDEC-22	0.01101	BCUTp-1l	0.00612
SsOH	0.01027	ETA_Eta_F_L	0.00585

4.3.3. 相关性过滤结果

通过随机森林降维，本文筛选出了三十种对生物活性影响较大的分子描述符，但由于这些分子描述

符之间可能具有较大的相关性，造成一些特征的重复表示和另一些特征的缺失。因此本文利用斯皮尔曼相关系数计算分子描述符的相关性，筛去某些相关性系数较大的分子描述符。

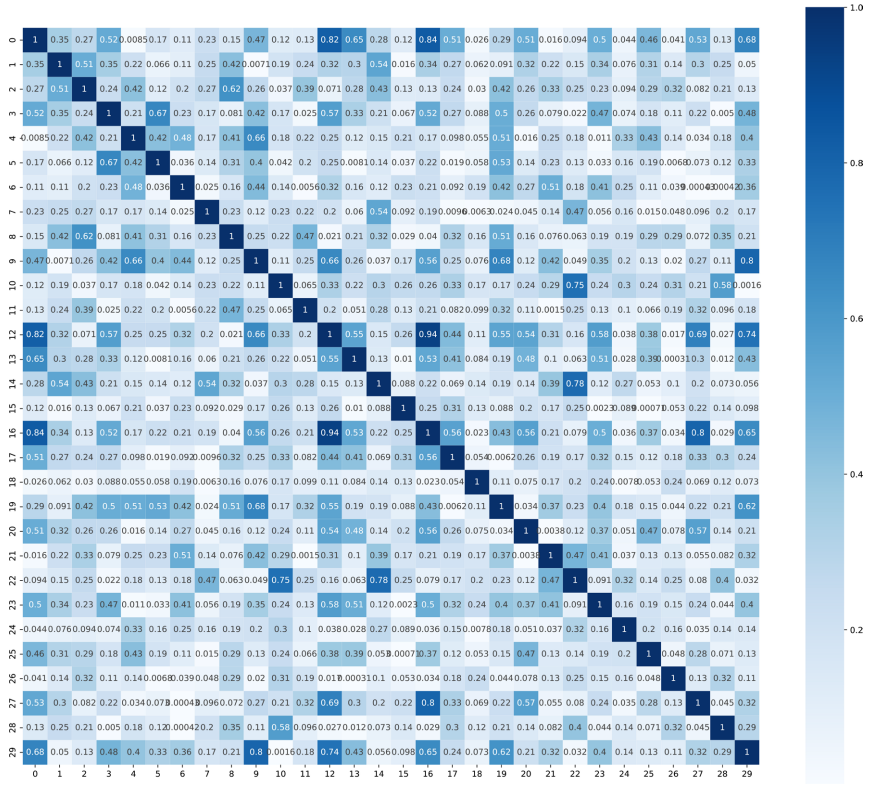

Figure 2. Correlation coefficient matrix for 30 molecular descriptor variables
图 2. 30 个分子描述符变量的相关系数矩阵

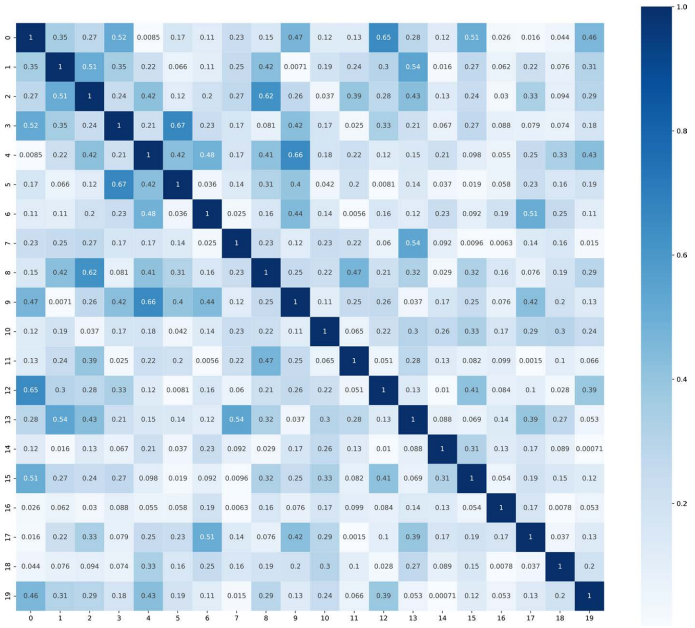

Figure 3. Correlation coefficient matrix between the final molecular descriptors
图 3. 最终筛选的分子描述符间的相关系数矩阵

30 种分子描述符间的变量相关系数矩阵，如图 2 所示。横纵坐标均为 30 个分子描述符变量的序号，其中序号越小的代表该分子描述符的信息增益越高，而颜色越深的方格代表所对应的两个分子描述符间的相关性越大。我们从信息增益较高的分子描述符开始筛选，保留其中相关性较为正常的描述符，在排名前二十的描述符中仅删去了斯皮尔曼相关系数较大的 12、16、19 列,然后删去了信息增益后十名中排名较为靠后且相关系数较大的 20、22、23、26、27、28 和 29 列，最终得到了经信息增益和斯皮尔曼系数共同筛选的 20 个分子描述符，在对生物活性有重大影响的同时保留了分子描述符间的独立性和特征代表性。最终结果图 3 所示，可以观察到除对角线外整体图形中方格的颜色均较浅，说明分子描述符间的相关性并不强。

本文首先利用基于 EM 算法的高斯混合聚类模型分析发现，729 种分子描述符间的数据表现特征和分布并不相似，减少了使用同类分子描述符造成的特征权重过高或过低问题，同时使用随机森林降维和斯皮尔曼相关系数对分子描述符进行选择，得到了在对生物活性有重大影响的同时保留了独立性和特征代表性的二十个分子描述符。这些分子描述符的名称及信息增益信息如表 2 所示。

Table 2. 20 Molecular Descriptors with the greatest influence on the activity of compounds to inhibit *ERα*
表 2. 对化合物抑制 *ERα* 的活性影响最大的 20 个分子描述符及其重要性

分子描述符名称	重要性	分子描述符名称	重要性
MDEC-23	0.28976	maxHBd	0.01405
minsOH	0.09127	ndssC	0.01250
C1SP2	0.07222	MDEC-22	0.01101
maxssO	0.05466	SsOH	0.01027
TopoPSA	0.02286	VCH-5	0.01022
MDEO-12	0.02030	WTPT-2	0.00923
minHBa	0.02006	ATSc5	0.00858
SHBint10	0.01885	WTPT-5	0.00694
BCUTc-1h	0.01873	nHBint6	0.00660
MLFER_S	0.01811	XLogP	0.00630

4.3.4. 随机森林模型活性预测结果

考虑到选取出的 20 个分子描述符均为信息增益较高且不相关的分子描述符。本文将 20 个分子描述符变量作为样本的 20 维特征，训练集和测试集的比例为 8:2，使用 sklearn 库中的 RandomForestRegressor 实现回归随机森林训练。以均方误差作为模型评价指标，回归随机森林模型在测试集上的均方误差为 0.493。同时，为了验证选出的 20 个分子描述符的有效性，本文分别使用决策树和支持向量机模型，选取 20 个和 729 个分子描述符进行了训练，得到的测试集均方误差如表 3 所示。由表中数据对比可以发现，随机森林模型的预测性能要明显优于决策树和支持向量机。由于随机森林具有擅长处理高维数据的特点，因此，使用全部(729 个)分子描述符的预测均方误差要小于只使用 20 个分子描述符的均方误差，但总体来说差距不大，这也侧面证明了选取的 20 个分子描述符是对化合物活性预测有较大影响的分子描述符。而支持向量机模型使用 20 个分子描述符训练得到的预测均方误差小于 729 个分子描述符训练得到的预测均方误差，验证了我们选取的分子描述符的有效性。

Table 3. The MSE of each model when different numbers of molecular descriptors are selected
表 3. 选取不同数量的分子描述符时各模型的预测均方误差

分子描述符数量	随机森林	决策树	支持向量机
20	0.493	0.905	1.189
729	0.466	0.966	1.439

4.3.5. 基于遗传算法的改进随机森林模型活性预测结果

考虑到计算的复杂性，我们设置多项式函数 f_m 的最高阶数为四阶，共包括五个系数，使用遗传算法设置变异率为 0.001，每一代保留一百个样本，迭代五百次，以最小化使用 f_m 映射后的 pIC50 值训练的随机森林模型的测试均方误差为优化目标。最终得到的最优多项式系数 $a_4 = -0.01094136$，$a_3 = 0.00196174$，$a_2 = 0.06468567$，$a_1 = 1.08033715$，$a_0 = -0.05813918$。经过该组系数决定的多项式变换后训练出的随机森林模型的回归均方误差为 0.382，相较于原均方误差 0.493，降低了 22%。

5. 总结

在药物发现领域，由于化合物的化学空间巨大，无法通过物理实验逐一测试每个可能的化合物。因此，计算模型的出现成为降低药物发现成本的重要途径。特别是通过预测化合物的检测结果，计算模型可以以极低的成本实现这一目标。本文使用基于遗传算法的改进随机森林模型进行化合物拮抗 *ERα* 活性预测模型。随机森林通过将具有相同特征的样本分到同一个叶子结点并对结点上的所有样本求均值来实现，使得结构相似的化合物能够在随机森林中聚集，从而提高基于分子描述符的化学特征相似性样本的预测准确率。在构建定量预测模型时，本文又使用遗传算法建立了映射函数，以均衡化回归目标变量，从而提高回归随机森林对映射后预测变量的学习性能。这一方法有望进一步提升模型的预测能力，为药物发现提供更准确的指导。

参考文献

[1] Siegel, R.L., Miller, K.D., Fuchs, H.E., *et al*. (2022) Cancer Statistics, 2022. *CA*: *A Cancer Journal for Clinicians*, **72**, 7-33. https://doi.org/10.3322/caac.21708

[2] Gramatica, P. (2020) Principles of QSAR Modeling. *International Journal of Quantitative Structure-Property Relationships*, **5**, 61-97. https://doi.org/10.4018/IJQSPR.20200701.oa1

[3] Svetnik, V., Liaw, A., Tong, C., *et al*. (2003) A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, **43**, 1947-1958. https://doi.org/10.1021/ci034160g

[4] Menze, B.H., Kelm, B.M., Masuch, R., *et al*. (2009) A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. *BMC bioinformatics*, **10**, Article No. 213. https://doi.org/10.1186/1471-2105-10-213

[5] 何冰, 罗勇, 李秉轲, 薛英, 余洛汀, 邱小龙, 杨登贵. 基于分子描述符和机器学习方法预测和虚拟筛选乳腺癌靶向蛋白 HEC1 抑制剂[J]. 物理化学学报, 2015, 31(9): 1795-1802.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al*. (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.