Advances in Applied Mathematics 应用数学进展, 2022, 11(4), 1578-1587
Published Online April 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.114172

Hans汉斯

基于图神经网络和XGBoost的抗乳腺癌候选药物预测模型研究

牛　淇

上海理工大学理学院，上海

收稿日期：2022年3月12日；录用日期：2022年4月5日；发布日期：2022年4月14日

摘　要

药物研发过程中，通过筛选影响显著的化合物继而合成抗癌药物无疑能够保证抗癌药物研发的高效性和针对性。本文针对收集到的与乳腺癌相关的ERα活性的1974种化合物，首先利用基于遗传算法的随机森林模型筛选出前20个对生物活性最具有显著影响的分子描述符，其次以此选择分子描述符变量构建定量预测模型得到预测结果，随后构建化合物的分类预测模型。结果表明该模型预测具有很强的实践意义，采用的预测策略是有效的，可为抗乳腺癌药物的研发提供借鉴。

关键词

抗乳腺癌药物，遗传算法，随机森林，图神经网络，XGBoost

Study on Anti-Breast Cancer Drug Candidate Forecast Model Based on Graph Neural Network and XGBoost

Qi Niu

College of Science, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 12th, 2022; accepted: Apr. 5th, 2022; published: Apr. 14th, 2022

Abstract

In the process of drug development, screening compounds with significant influence and then synthesizing anticancer drugs can undoubtedly ensure the high efficiency and pertinence of anti-cancer drug development. Aiming at the collected 1974 compounds with ER*α* activity related to

文章引用: 牛淇. 基于图神经网络和 XGBoost 的抗乳腺癌候选药物预测模型研究[J]. 应用数学进展, 2022, 11(4): 1578-1587. DOI: 10.12677/aam.2022.114172

breast cancer, the random forest model based on genetic algorithm was used to screen out the top 20 molecular descriptors with the most significant impact on biological activity, and then the molecular descriptor variables were selected based on this. A quantitative prediction model is constructed to obtain the prediction result, and then a classification prediction model of the compound is constructed. The results show that the model prediction has strong practical significance, and the prediction strategy adopted is effective, which can provide reference for the research and development of anti-breast cancer drugs.

Keywords

Anti-Breast Cancer Drugs, Genetic Algorithm, Random Forest, Graph Neural Network, XGBoost

1. 引言

乳腺癌是当今临床治疗中一种常见的恶性肿瘤，其发病率逐年上升。世界卫生组织 GLOBOCAN 2018 年全球癌症统计报告显示，乳腺癌位居全球女性癌症发病率和死亡率的第 1 位，它严重威胁着女性的身心健康[1]。女性乳腺癌已经成为一个全球性的重点公共医疗问题，为解决这个问题，药物研发领域对大量的候选药物进行了研究分析。研究发现，乳腺癌的发展与雌激素受体 α 亚型(Estrogen receptors alpha, Erα)息息相关，生物实验表明 ERα 在治疗乳腺癌中有极其重要的地位。截止到现在，医疗界的专家学者在 ERα 表达的病患中，通过控制雌激素受体活性进而控制患者体内的雌激素水平，起到治疗乳腺癌的作用。因此可以利用 ERα 的特性，研究针对性强的抗癌靶细胞药，与此同时，乳腺癌研发的备选药物也可以是拮抗 ERα 的活性化合物[2]。

一个化合物想要成为治疗乳腺癌的候选药物，需要满足三点，良好的生物活性、良好的药代动力学性质以及安全性，记为 ADMET (Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性)性质，前四个名词代表着化合物在生物体内浓度与时间的变化关系，保证了药代动力学性质，T 是观察化合物是否产生毒性物质，浓度是否可以接受，保证了药物的安全性[3]。抗乳腺癌候选药物从研发到投入使用的过程中，如果仅仅采用实验的方式去评估化合物的生物活性、药代动力学性质和安全性，需要花费大量时间成本和经济成本。因此，在药物研发中，为了避免不必要的资源浪费，通常采用建立化合物活性预测模型的方法来筛选潜在活性化合物。研究机构通过把体外研究技术和计算机运算模型结合起来，针对与疾病相关的某个靶标(此处为 ERα)，收集一系列作用于该靶标的化合物及其生物活性数据，然后选取一系列分子结构描述符作为自变量，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)模型，然后使用该模型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[4]。此外，如果一个化合物的 ADMET 性质不佳，比如不容易被人体吸收，或者在体内代谢速度太快，或者具有某种毒性，那么其仍然难以成为药物，因而对其进行药物 ADMET 性质优化也是非常有必要的[5]。

二十一世纪以来，随着信息技术的飞速发展，数据挖掘技术已经广泛应用于医药学研究，目前最常用的研究方式是利用计算机辅助的人工智能算法对药物生物活性和 ADMET 性质进行预测分析，研究学者通过数据拟合数学模型，进而通过优化的方法得到最优解，增加了研究的价值性，大大降低了研究成

本，提高研发成功几率，且更有利于对候选药物在生物体内发挥的作用进行探索，有效避免因药物产生的副作用和毒性导致的人体疾病，可指导临床治疗时的合理用药[6]。由此可见，使用计算机辅助的人工智能算法进行理论预测抗乳腺癌候选药物的生物活性和 ADMET 性质是极具现实意义的。

本文基于重构损失的图神经网络进行建模分析，凭借强大的端到端的学习能力，这类模型可以非常友好地支持有监督的学习方式。与较为常见的随机森林算法相比，使用图神经网络，用预先计算的分子描述符(或特征)来构建分子特性预测模型，其性能有明显提高，计算时间大幅度缩短。在运用 XGBoost 算法求解时，通过 CART 回归树模型进行拟合运算，激活函数得出目标函数值和对应的最优值。

2. 数据收集

针对乳腺癌治疗靶标 ERα，从阿尔伯塔大学的 DrugBank 药物分子数据库中获取了 1974 种化合物对乳腺癌治疗靶标 ERα 的生物活性和 ADMET 性质数据，DrugBank 数据库是一个整合了生物信息学和化学信息学资源，并提供详细的药物数据与药物靶标信息及其机制的全面分子信息，包括药物化学、药理学、药代动力学、ADME 及其相互作用信息。相比于其他数据库来说，该数据库的信息更加准确，所有的药物靶点信息均是实验验证过的。

3. 筛选主要的分子描述符

3.1. 数据处理

针对收集到的 1974 个化合物对 ERα 的生物活性值的描述，从提供的 729 个分子描述符信息中选出 20 个对生物活性影响最大的分子描述符。为方便建模计算，我们首先将所有变量数据进行编码，将变量按照从 1~729 的顺序编号，即表 1 所示：

Table 1. Sequential numbering feature parameters
表 1. 顺序编号特征参数

符号	特征参数
1	nAcid
2	ALogP
3	ALogp2
4	AMR
5	apol
⋮	⋮
728	XLogP
729	Zagreb

在此编号的基础上，我们即有了一个 1974 × 729 的样本矩阵，并根据此矩阵建立模型并进行求解。由于传统的分类算法在低维度的数据集上面能够获得比较理想的分类效果，但是在高维度的数据集上它的分类性能则会出现较大的下降。高维数据结构复杂，包含更多的非信息和噪声，随机森林(Random Forests, RF)算法可以在对样本分类的同时给出变量的重要性评分(Variable Importance Measures, VIMs)作为特征筛选的依据[7]。RF 采用了特征子空间来构建模型，所以构建的模型难免会混杂很多的噪音，而利

用这些包含噪音的模型进行预测分类将会降低随机森林算法的分类效果。使得 VIMs 不稳定，而那些真正对分类有意义的变量可能会因在筛选后得到的变量子集中排序较后而无法被选入[8]。遗传算法(Genetic Algorithm, GA)是按照随机搜索策略进行特征筛选的，可以由不同的染色体提供多样化的特征筛选结果[9]，采用适当的遗传算法与随机森林模型相结合将有可能降低“噪声”对筛选结果的影响，同时保证筛选得到的特征变量集有较小的假发现率(False Discovery Rate, FDR)。所以本文考虑使用基于遗传算法的随机森林模型(GARF)，用于 1974 个化合物的 729 个分子描述符的筛选，而最终挑选出前 20 个对生物活性最具有显著影响的分子描述符。

3.2. 基于遗传算法的随机森林模型筛选分子描述符

GARF 采用 RF 模型对变量在样本分类中的作用进行评价，以 Permutation 方法确定特征筛选界值作为最终确定特征变量的依据。为减少噪声变量对 RF 变量评价结果的干扰，每个 RF 模型仅包含由 GA 算法选取的部分变量，并且在遗传过程中加入了变量筛选步骤以进一步降低噪声变量影响，尽量减少 RF 模型过拟合的可能。遗传过程内的变量筛选中，采用 Permutation 方法获得组间无差异变量重要性评分的经验分布，根据该经验分布自适应确定变量筛选界值[10]。使用 Python 应用遗传算法－随机森林模型，运行程序得到 729 个变量的贡献度排名，由大到小将贡献度排序。考虑到下一步的高相关性滤波操作会对进一步变量进行降维，故先筛选出排名处于前 26 的变量，如表 2 所示。

Table 2. Filtered 26 feature parameters
表 2. 筛选后的 26 个特征参数

编号	参数名	影响因子	编号	参数名	影响因子
26	ATSc5	0.058302811	25	ATSc4	0.010925728
637	ETA_EtaP_B_RC	0.053953952	115	nHBint6	0.010017568
635	ETA_EtaP_B	0.050063064	68	SCH-7	0.009997138
615	ETA_BetaP	0.045284284	80	VC-5	0.009802765
106	ECCEN	0.03896089	269	SdsCH	0.009565066
614	ETA_Beta	0.028631826	632	ETA_Eta_F_L	0.009105144
532	maxssO	0.015266195	277	SssssC	0.008573706
533	maxaaO	0.012867972	546	C2SP2	0.00846456
42	BCUTp-1l	0.012714979	325	VABC	0.00824946
234	SHBint6	0.01259159	289	ETA_dPsi_B	0.00761864
414	minaaO	0.012397447	594	minHBd	0.00716846
658	MDEC-14	0.01141645	338	SPC-4	0.006835646
507	maxdsCH	0.010948259	198	SP-1	0.006364866

为了提高遗传算法的收敛性，采用最优保留策略，将最大适应度的个体直接保留到下一代。每次更新种群时将群体中最差的个体进行替换成上一代的最优个体，以防止当前种群中适应度较好的个体被淘汰。将每一代中获得的变量评价结果取中位数，作为 GARF 特征筛选方法对变量的最终评价，记为VI_{gene}。

计算最后一代种群中每条染色体中包含的基因个数；y 取其平均值记为 M_{gene}，作为 Permutation 抽样参数；z 从数据集中随机抽取 M_{gene} 个变量，将分类标签随机打乱，建立随机森林模型，记录变量重要性评分，重复进行 $2000/M_{gene}$ 次，共获得 2000 个变量重要性评分；以上述 2000 个变量重要性评的百分位数 P_{95} 或 P_{99} 作为 GARF 算法特征筛选界值，如 VI_{gene} 大于该界值则将该变量识别为特征变量。最后得出最终的 20 个参数指标，如表 3 所示；根据所提取的特征变量之间的距离相关系数计算结果对变量之间的相关程度进行可视化，如图 1 所示。可见选取的 20 个变量之间相关性低，独立性较好，这 20 个分子描述符能够尽可能地描述化合物的生物活性。

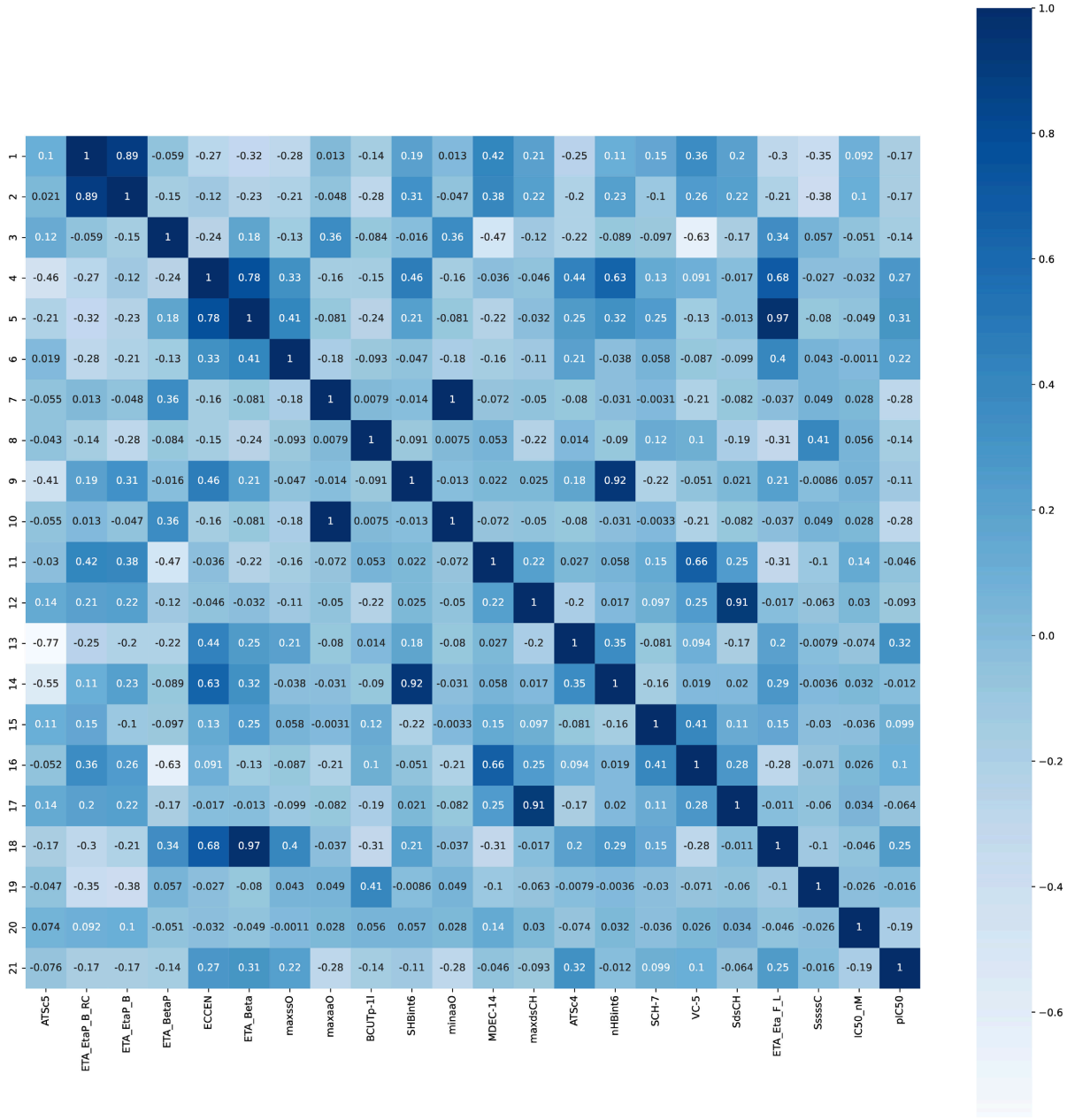

Figure 1. The strength of the correlation between the main variables
图 1. 主要变量之间的相关性强弱

Table 3. Filtered 20 feature parameters
表 3. 筛选后的 20 个特征参数

编号	参数名	影响因子	编号	参数名	影响因子
26	ATSc5	0.058302811	414	minaaO	0.012397447
637	ETA_EtaP_B_RC	0.053953952	658	MDEC-14	0.01141645
635	ETA_EtaP_B	0.050063064	507	maxdsCH	0.010948259
615	ETA_BetaP	0.045284284	25	ATSc4	0.010925728
106	ECCEN	0.03896089	115	nHBint6	0.010017568
614	ETA_Beta	0.028631826	68	SCH-7	0.009997138
532	maxssO	0.015266195	80	VC-5	0.009802765
533	maxaaO	0.012867972	269	SdsCH	0.009565066
42	BCUTp-1l	0.012714979	632	ETA_Eta_F_L	0.009105144
234	SHBint6	0.01259159	277	SssssC	0.008573706

4. 基于图神经网络模型预测生物活性结果

在对分子描述符数据进行降维处理后，大大减少了数据量，鉴于化合物是由原子和化学键构成的，它们天然就是一种图数据的形式，图神经网络(GNN)能够自动化地同时学到图的特征信息与结构信息[11]，这一特点与分子描述符数据性质对候选药物影响方式的特点相吻合，因此可以选择图神经网络进行训练学习，并对 50 个化合物进行 IC50_nM 值和对应的 pIC50 值预测。本节分析基于图神经网络的生物活性值预测方法，并通过与相应的基于重构损失设计的损失函数结合起来实现无监督图表示学习，避免传统图神经网络有监督的学习方式。

基于重构损失构建的 GNN 模型进行生物活性值预测，其基本框架包括推断模型(编码器)、生成模型(解码器)、损失函数。

1) 推断模型

$$q\left(Z\left|X,A\right.\right)=\prod_{i=1}^{N}q\left(z_i\left|X,A\right.\right)$$

$$q\left(z_i\left|X,A\right.\right)=\mathcal{N}\left(z_i\left|\mu_i,diag\left(\sigma_i^2\right)\right.\right)$$

这里使用两个 GNN 对 μ 、 σ 进行拟合：

$$\mu=GNN_{\mu}\left(X,A\right),log\sigma=GNN_{\sigma}\left(X,A\right)$$

2) 生成模型

$$p\left(A\left|Z\right.\right)=\prod_{i=1}^{N}\prod_{j=1}^{N}p\left(A_{ij}\left|z_i,z_j\right.\right)$$

$$p\left(A_{ij}=1\left|z_i,z_j\right.\right)=\sigma\left(z_i^{T}z_j\right)$$

这里也简单使用了两个节点表示向量的内积来拟合邻接关系。

3) 损失函数

$$L=L_{recon}+L_{kl}=-E_{q(Z|X,A)}\left[logp\left(A\left|Z\right.\right)\right]+KL\left[q\left(Z\left|X,A\right.\right)\left\|p\left(Z\right)\right.\right]$$

同样地，隐变量 z 的先验分布选用标准正态分布：

$$p(Z)=\prod_{i=1} p(z_i)=\prod_{i=1} \mathcal{N}(z_i,\mathbf{0},\mathbf{I})$$

通过建立的化合物对 ERα 生物活性的定量预测模型，对 50 个化合物的生物活性值进行预测。得到的 IC50 值和对应的 pIC50 值，详见表 4。

Table 4. IC50_nM value and corresponding pIC50 value prediction results
表 4. IC50_nM 值及对应 pIC50 值的最优预测结果

序号	IC50_nM 值	pIC50 值	序号	IC50_nM 值	pIC50 值
1	2.5	8.146578	26	0.7	7.464932
2	7.5	8.265166	27	1.4	7.413951
3	3.1	8.259882	28	3.6	7.390989
4	3.9	7.7889285	29	1.7	7.4282365
5	7.4	7.887341	30	1	7.4666038
6	490	7.861537	31	4.4	7.409476
7	1	8.144045	32	155	7.401839
8	35	7.973138	33	3.6	7.3934574
9	4.3	7.8698697	34	1.3	7.396768
10	91	7.7398953	35	32	7.4053774
11	172	7.692528	36	39	7.429451
12	35	7.8928227	37	397	7.594025
13	11	7.802808	38	46	7.3883576
14	2.6	7.7979064	39	94	7.376009
15	542	7.7384086	40	33	7.2936335
16	3	7.901499	41	6	8.092378
17	2.7	7.7395024	42	9	8.224709
18	7	8.017715	43	18	8.384666
19	366	8.125411	44	1	8.31978
20	8.8	7.818233	45	21	8.429541
21	3	7.590331	46	15	8.0946045
22	1.8	7.525207	47	4	8.40542
23	2.3	7.4395084	48	5	7.807101
24	1.4	7.5160265	49	4	7.474482
25	1	7.4091053	50	4	7.4112115

5. 基于 XGBoost 的分类预测模型

由于 XGBoost 的计算效率高、分类准确性强，且随着迭代次数的增加准确性大幅提升，对构建化合物的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型具有较好的预测效果，能有效降低预测误差，取得较高的预测精度[12]。因此，本文将使用 XGBoost 算法构建化合物的 ADMET 预测模型，分别对 5 个指标进行预测分析。

通过 CART 回归树模型进行拟合计算，具体模型参考如下：

$$\widehat{y_i} = \sum_{i=1}^{n} f_t(x_i), f \in F$$

其中，n 表示树数量；f_t 表示函数空间 F 独立函数；$\widehat{y_i}$ 表示预测值；x_i 表示输入 i 个数据；F 表示 CART 总和。

迭代过程如下

$$\widehat{y_i}^{(0)} = 0$$

$$\widehat{y_i}^{(1)} = f_1(x_i) = \widehat{y_i}^{(0)} + f_1(x_i)$$

$$\widehat{y_i}^{(t)} = f_1(x_i) = \widehat{y_i}^{(t-1)} + f_1(x_i)$$

XGBoost 的目标函数如下所示：

$$X_{obj} = \sum_{i=1}^{n} l(y, \hat{y}) + \sum_{k=1}^{k} \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j-1}^{T} \omega_j^2$$

$$T^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(X_i)\right) + \Omega(f_t)$$

目标函数近似为

$$T^{(t)} = \sum_{i=1}^{n}\left[\left(y_i, \widehat{y_i}^{(t-1)} + f_t(X_i)\right) + \frac{1}{2} f_t^2(X_i)\right] + \Omega(f_t)$$

激活函数：

$$\begin{aligned} X_{obj} &= \sum_{i-1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_t f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{i-1}^{n}\left[g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i)\right] + \Omega(f_t) + \lambda T + \lambda \frac{1}{2}\sum_{j-1}^{T} w_j^2 \\ &= \sum_{j-1}^{T}\left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \lambda T \end{aligned}$$

其中，X_{obj} 为目标函数；$g_i = \partial \hat{y}^{i-1} l\left(y_i, \hat{y}^{i-1}\right)$ 为一阶导数；$h_i = \partial^2 \hat{y}^{i-1} l\left(y_i, \hat{y}^{i-1}\right)$ 为二阶导数。

所得的最优 ω 和目标函数值分别如下所示：

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}$$

$$X_{obj} = -\frac{1}{2}\sum_{j-1}^{T}\frac{G_j}{H_j+\lambda} + \lambda T$$

其中，$G_j = \sum_{i\in I_j} g_i; H_j = \sum_{i\in I_j} h_i$ 。

从图 2 的分子描述符变量间的相关性图可以看到，除了四个角是色块密集且颜色对应的相关系数绝对值接近 1，相关度高，其余大部分区域色块较为稀疏，且颜色对应的相关系数值近于 0，相关度低。所以综合看来，这些分子描述符之间的相关性较弱，以此为变量建模更具代表性与说服力。

从图 3 标准误差表中可得，黄线代表的是由常用的长记忆神经网络算法模型预测的结果，除了在横坐标为 50 和 120 处标准差偏低，大部分都是要比 XGBoost 模型预测的结果标准差大，由此可以看出本文所采用的 XGBoost 模型的预测效果要优于长记忆神经网络(LSTM)，更准确更有说服力。

根据所构建的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型，可以由化合物分子的结构式对新化合物的 ADMET 性质进行相应预测，从而判断新化合物的性质好坏，对药物性质判断提供一定的参考价值。

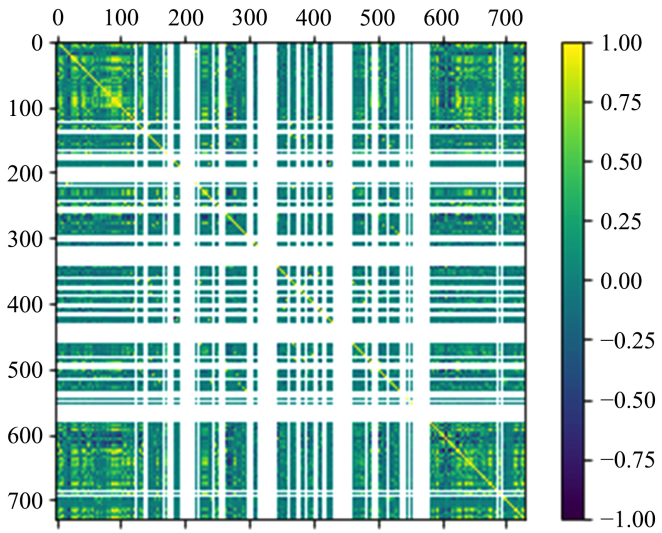

Figure 2. The strength of the correlation between the main variables
图 2. 主要变量之间的相关性强弱

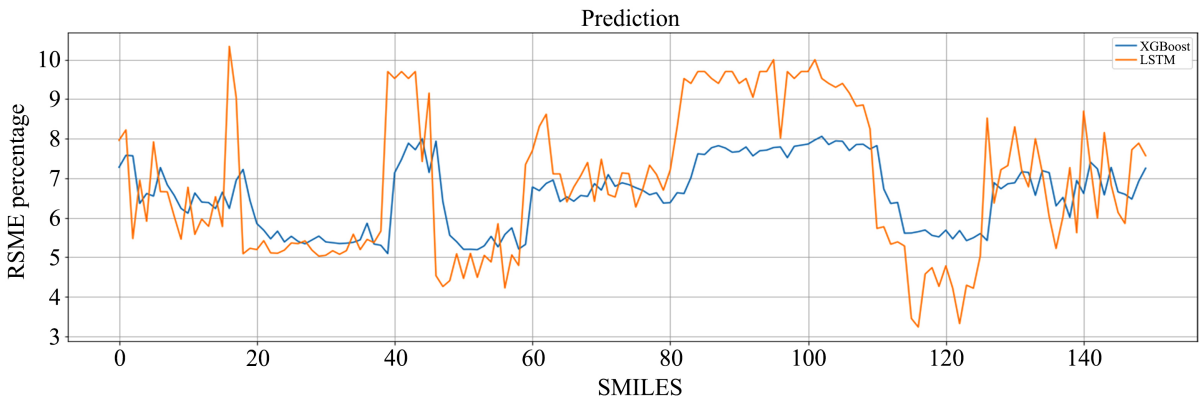

Figure 3. RSME percentage for LSTM and XGBoost
图 3. LSTM 和 XGBoost 的 RSME 百分比

6. 结论

针对抗乳腺癌候选药物研发过程中的生物活性和 ADMET 性质预测问题，本文提出的两个预测模型精度、鲁棒性表现优秀，采用的方法结合了遗传算法和随机森林，与一般的方法相比，预测的结果更加均衡，降低了单一方法预测结果产生的偏差。对于研制乳腺癌的治疗候选药物，减少乳腺癌的致死率以及防治其他肿瘤疾病等人体生命健康的研究具有一定的指导性。

参考文献

[1] Freddie, B., Jacques, F., Isabelle, S.A., *et al.* (2018) Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **68**, 394-424. https://doi.org/10.3322/caac.21492

[2] 孙博, 蒙艺灵, 温涛, 等. 多肽修饰的金纳米颗粒对小鼠三阴性乳腺癌细胞的靶向光热效应研究[J]. 北京生物医学工程, 2021, 40(5): 441-446.

[3] 牛东梅. 乳腺癌影像诊断中多层螺旋 CT 与钼靶摄影的应用对比研究[J]. 系统医学, 2021, 6(1): 122-124.

[4] 顾耀文, 张博文, 郑思, 等. 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.

[5] Lei, T., Sun, H., Kang, Y., *et al.* (2017) ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning-Approaches. *Molecular Pharmaceutics*, **14**, 3935-3953. https://doi.org/10.1021/acs.molpharmaceut.7b00631

[6] Wenzel, J., Matter, H. and Schmidt, F. (2019) Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, **59**, 1253-1268. https://doi.org/10.1021/acs.jcim.8b00785

[7] 陈禹, 毛莺池. 基于随机森林和遗传算法的 Ceph 参数自动调优[J]. 计算机应用, 2020, 40(2): 347-351.

[8] Wu, Z., Ramsundar, B. and Feinberg, E.N. (2018) MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science*, **9**, 513-530. https://doi.org/10.1039/C7SC02664A

[9] 白茹, 滕奇志, 杨晓敏, 等. 基于 SVM 和 GA 的药物与人血清白蛋白结合的预测[J]. 计算机工程与应用, 2009, 45(12): 226-228, 248. https://doi.org/10.3778/j.issn.1002-8331.2009.12.072

[10] Mirza, Q.A.K., Awan, I. and Younas, M. (2018) CloudIntell: An Intelligent Malware Detection System. *Future Generations Computer Systems*: *FGCS*, **86**, 1042-1053. https://doi.org/10.1016/j.future.2017.07.016

[11] Qiao, K., Zeng, L., Chen, J., *et al.* (2018) Wire Segmentation for Printed Circuit Board Using Deep Convolutional Neural Network and Graph Cut Model. *IET Image Processing*, **12**, 793-800. https://doi.org/10.1049/iet-ipr.2017.1208

[12] Wang, Y., Li, Z., Gao, J., *et al.* (2020) Deep Neural Network-Based Wi-Fi/Pedestrian Dead Reckoning Indoor Positioning System Using Adaptive Robust Factor Graph Model. *IET Radar*, *Sonar & Navigation*, **14**, 36-47. https://doi.org/10.1049/iet-rsn.2019.0260