Statistics and Application 统计学与应用, 2025, 14(4), 1-9
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.144083

Hans汉斯

ERα靶向化合物筛选与优化：基于深度学习和多目标优化的ADMET模型

袁钰喜*，李林峰，李必嘉，廖礼航

重庆对外经贸学院数学与计算机学院，重庆

收稿日期：2025年2月27日；录用日期：2025年3月18日；发布日期：2025年3月31日

摘　要

目前，抗胰腺癌候选药物化合物在药物研发中面临时间和成本等诸多挑战。因此，本文提出一种融合Lasso回归与BP神经网络模型的方法，用于筛选和优化ERα靶向化合物。首先，使用Lasso回归筛选出与生物活性(pIC50)相关的重要分子描述符，并通过神经网络进行ADMET分类预测。实验结果表明，该方法能够有效提高药物活性和安全性的预测精度。从Lasso回归中筛选出的前20个重要特征对药物活性有显著影响，构建的随机森林回归模型在测试集上的准确率达到89%。并且筛选的特征在BP神经网络中ADMET分类任务中也表现良好，其中CYP3A4任务的准确率为91%。该方法为ERα拮抗剂的筛选和优化提供了可借鉴的思路。

关键词

Lasso回归，BP神经网络，ERα靶向化合物，ADMET分类

ERα Targeted Compound Screening and Optimization: ADMET Model Based on Deep Learning and Multi-Objective Optimization

Yuxi Yuan*, Linfeng Li, Bijia Li, Lihang Liao

School of Mathematics and Computer Science, Chongqing College of International Business and Economics, Chongqing

Received: Feb. 27th, 2025; accepted: Mar. 18th, 2025; published: Mar. 31st, 2025

*通讯作者。

文章引用: 袁钰喜, 李林峰, 李必嘉, 廖礼航. ERα 靶向化合物筛选与优化: 基于深度学习和多目标优化的 ADMET 模型[J]. 统计学与应用, 2025, 14(4): 1-9. DOI: 10.12677/sa.2025.144083

Abstract

Currently, anti-breast cancer drug candidate compounds are facing many heavy challenges in drug discovery such as time and cost. Therefore, in this paper, we propose an approach that integrates Lasso regression and BP neural network models for screening and optimizing ERα-targeting compounds. First, important molecular descriptors related to biological activity (pIC50) were screened using Lasso regression and predicted by neural network for ADMET classification. The experimental results showed that this method can effectively improve the prediction accuracy of drug activity and safety. The top 20 important features screened from Lasso regression had a significant effect on drug activity, and the accuracy of the constructed random forest regression model reached 89% on the test set. And the screened features also performed well in the ADMET classification task in BP neural network, with an accuracy of 91% in the CYP3A4 task. This method provides a referable idea for the screening and optimization of ERα antagonists.

Keywords

Lasso Regression, BP Neural Network, ERα Targeted Compound, ADMET Classification

1. 引言

胰腺癌[1]是全球最常见的恶性肿瘤之一，已成为女性癌症死亡的主要原因。随着胰腺癌发病率持续攀升，特别是在年轻女性群体中的增长，胰腺癌的早期检测与有效治疗愈发重要。雌激素受体(ER)特别是 Erα 在胰腺癌的发生和发展中扮演着关键角色，因此成为胰腺癌治疗的关键靶点。尽管现有的抗 ERα [2]药物(如他莫昔芬和雷诺昔芬)在治疗 ER 阳性胰腺癌方面取得了显著进展，然而耐药性及疗效降低依然是临床治疗面临的挑战。因此，开发新型的 ERα 拮抗剂依旧是胰腺癌治疗研究的热点之一。

随着药物研发的成本和周期日益增加，定量结构–活性关系(QSAR) [3]模型作为一种高效的药物筛选工具，逐渐在药物研发中获得广泛应用。此外，药物的药代动力学性质(ADMET) [4]同样是药物研发中的重要考虑因素，良好的 ADMET 性质对于药物的临床应用至关重要。国际上研究人员结合药物的分子结构描述符、酶活性、受体结合力等因素，通过多元回归分析、支持向量机(SVM)等方法建立了精准的 QSAR 模型，能够较为准确地预测化合物的生物活性[5]。ADMET 预测方面，国内的研究也取得了进展。一些学者通过建立药物的多维度描述符，结合机器学习算法，建立了可以预测 Caco-2 渗透性、CYP3A4 代谢稳定性等 ADMET 性质的模型[6]。这些研究为药物的早期筛选提供了重要的理论支持，并对药物的临床开发具有积极的推动作用。

尽管已有许多关于胰腺癌药物筛选(QSAR)模型和 ADMET 预测的研究取得了显著进展，但药物的非线性生物活性与 ADMET 性质之间的复杂关系仍然是研究中的难点。本文提出了一种基于 Lasso 回归的方法，通过在最小二乘回归的目标函数中加入 L1 正则化项(即系数的绝对值之和)，从而使某些回归系数变为零。该方法筛选出对模型预测最重要的特征，进而研究定量结构–活性关系。并在此基础上，构建了 BP 神经网络，用于对生物活性数据进行 ADMET 分类预测。

2. BP 神经网络[7]

BP 神经网络(Backpropagation Neural Network，即反向传播神经网络)是最常见的人工神经网络类型之一，广泛应用于机器学习和深度学习任务。它属于前馈神经网络(Feedforward Neural Network)范畴，包含输入层、多个隐藏层以及输出层，借助反向传播算法进行训练，能够自动学习输入数据与输出之间的复杂非线性关系。以下是对 BP 神经网络的详细介绍。

2.1. 前向传播

假设输入层有 n 个神经元(特征)，隐藏层有 m 个神经元。输入向量为 $x=\left[x_1,x_2,\cdots,x_n\right]^T$，隐藏层的输出为 $h=\left[h_1,h_2,\cdots,h_m\right]^T$。每个隐藏层神经元的输入是输入特征的加权和：

$$z_j=\sum_{i=1}^{n}w_{ij}x_i+b_j \tag{1}$$

其中，w_{ij} 为输入层第 i 个神经元到隐藏层第 j 个神经元的权重，b_j 是偏置项。通过激活函数计算隐藏层输出 h_j。

$$h_j=f\left(z_j\right)=f\left(\sum_{i=1}^{n}w_{ij}x_i+b_j\right) \tag{2}$$

其中，f 是激活函数 Sigmoid。同理，输出层神经元的输入 z_k，是隐藏层输出的加权和，再通过激活函数计算输出 y_k。

2.2. 计算损失函数

对于分类任务，损失函数使用交叉熵损失：

$$L=-\sum_{k=1}^{P}y_{true,k}log\left(y_k\right) \tag{3}$$

其中，$y_{true,k}$ 是真实标签的 One-hot 编码，y_k 是模型的预测概率。

2.3. 后向传播：计算梯度

后向传播的目的是计算损失函数对网络中每个参数(权重和偏置)的梯度，并使用这些梯度来更新参数。计算输出层的误差项(梯度)，反映了模型预测与真实标签之间的差异。对于每个输出层神经元 k，误差项为：

$$\delta_k=\frac{\partial L}{\partial z_k} \tag{4}$$

2.4. 参数更新

如果网络有多个隐藏层，可以继续通过链式法则将误差传播到输入层。对于多层网络，每一层的误差项由上一层的误差项计算而来。使用梯度下降法更新每层的权重和偏置。

$$w_{ij}\leftarrow w_{ij}-\eta\frac{\partial L}{\partial w_{ij}} \tag{5}$$

$$b_j\leftarrow b_j-\eta\frac{\partial L}{\partial b_j} \tag{6}$$

其中，η 是学习率。

3. 多目标优化的 ADMET 模型

当前，药物筛选面临生物相容性、代谢途径复杂性、毒性评估等多维度挑战，传统方法往往依赖大量实验以及漫长的测试周期。为提高筛选效率，本文提出了一种多目标优化的 ADMET 模型，利用 Lasso 回归筛选与化合物活性相关的重要分子特征，并结合 ADMET (吸收、分布、代谢、排泄和毒性)标准筛选符合条件的化合物。通过计算分子描述符并将其作为神经网络的输入，BP 神经网络依据这些特征预测药物的 ADMET 性质。经过训练后，模型能够准确预测药物的活性和安全性，从而帮助筛选出具有良好药效和安全性的候选化合物。进一步分析高活性化合物在各重要特征上的四分位范围，揭示在优化药物活性与安全性时哪些特征值至关重要，为药物设计提供数据支持，有助于识别兼具高活性与良好安全性的化合物。具体分析流程见图 1。

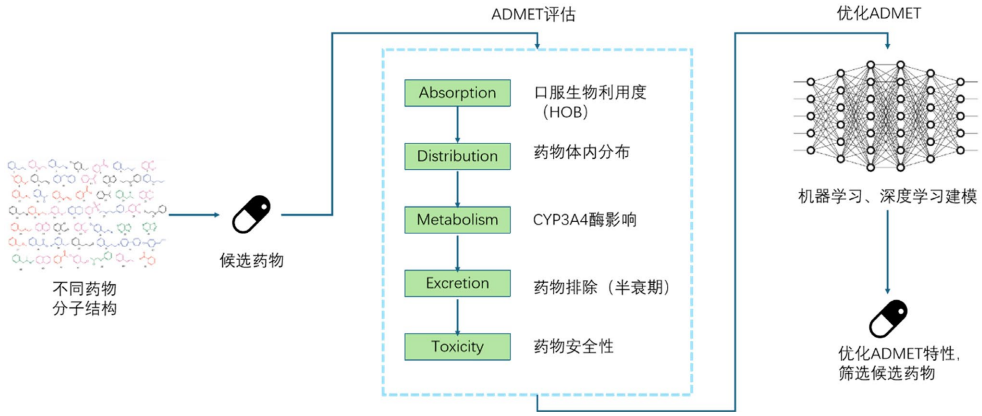

Figure 1. Workflow of QSAR and deep learning in drug discovery
图 1. QSAR 和深度学习药物研发

4. 实证分析

4.1. 分子描述符筛选与生物活性分析

本研究使用的数据集来源于 2021 年研究生数学竞赛，聚焦于抗胰腺癌候选药物的优化建模。数据集包含 Molecular_Descriptor、ERα_activity、ADMET 三个 EXCEL 表格。其中 Molecular_Descriptor：该数据集包括 1974 个化合物的 729 个分子描述符，每一行代表一个化合物，每一列对应一个不同的分子描述符。SMILES 列提供了每个化合物的化学结构信息，其他列则描述了化合物的结构和化学性质。*ERα _ activity*：该数据集记录了每个化合物的生物活性，每一行对应一个化合物，列包含了该化合物的 pIC50 值，表示其对 *ERα* 的生物活性。ADMET：该数据集包含了与 ADMET 相关的五个重要属性(吸收、分布、代谢、排泄和毒性)。每个化合物的属性 Caco-2 渗透性、CYP3A4 代谢、hERG 心脏毒性等通过 0 或 1 的标记表示是否符合良好的性质标准，其中 0 表示不符合，1 表示符合。因 Molecular_Descriptor 的变量多，这里只展示了 6 个变量以及 pIC50 的数据，详情见表 1。同时 ADMET 的重要属性分布见表 2。

Table 1. Statistical description of selected numerical data from Molecular_Descriptor
表 1. Molecular_Descriptor 部分数值型数据的统计描述

统计量	pIC50	ALogP	ALogp2	AMR	apol	naAromAtom	nAromBond

续表

count	1974	1974	1974	1974	1974	1974	1974
mean	6.59	1.11	3.29	116.56	60.63	15.45	16.19
std	1.42	1.43	12.83	31.57	19.45	5.16	5.64
min	2.46	−23.11	0.00	54.07	30.66	0.00	0.00
25%	5.38	0.38	0.41	88.30	44.43	12.00	12.00
50%	6.58	1.17	1.56	114.84	59.90	16.00	18.00
75%	7.57	1.95	4.02	141.42	74.42	18.00	18.00
max	10.37	5.18	533.84	517.43	359.66	30.00	34.00

Table 2. Distribution of ADMET dataset properties
表 2. ADMET 数据集分布描述

类别	Caco-2	CYP3A4	hERG	HOB	MN
0	1215	1461	1099	1465	1514
1	759	513	875	509	460
合计	1974	1974	1974	1974	1974

在数据处理过程中，首先提取分子描述符(去除 SMILES 列)与生物活性(pIC50)数据。随后，删除描述符数据中零值占比超 90%的特征列，以降低噪声干扰。接着，对描述符数据进行标准化处理，消除特征间的量纲差异，进而确保模型能够有效处理不同尺度的数据。可见图 2 标准化前后分子描述符数据箱线图。

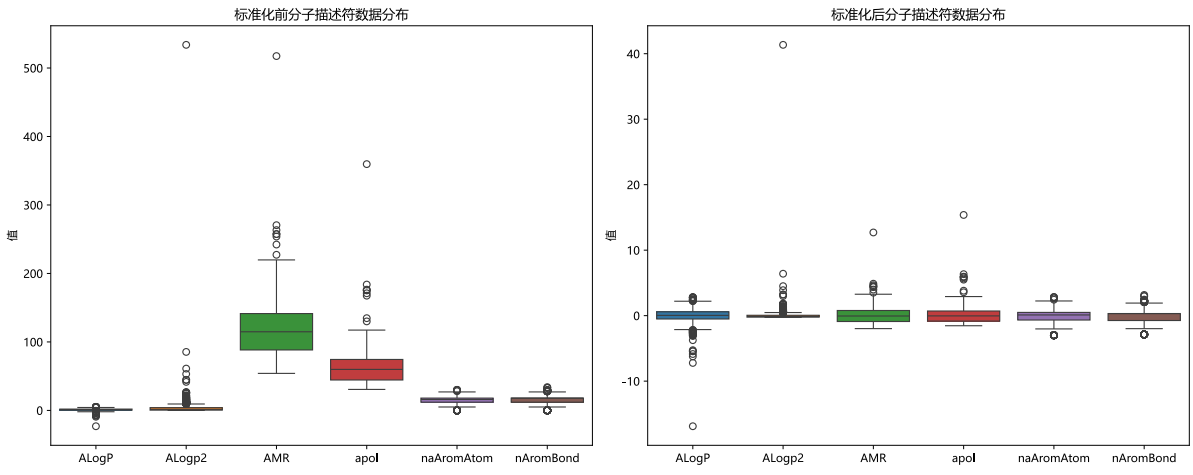

Figure 2. Boxplot of molecular descriptor data before and after standardization
图 2. 标准化前后分子描述符数据箱线图

接下来，采用 Lasso 回归(带 5 折交叉验证)进行变量选择，评估各分子描述符对生物活性(pIC50)的影响。通过 Lasso 回归的系数(其绝对值作为特征重要性)对特征进行了排序。为了便于展示，以下展示了前十个重要变量的回归系数路径图(见图 3)以及对应的分子描述符系数图(见图 4)。

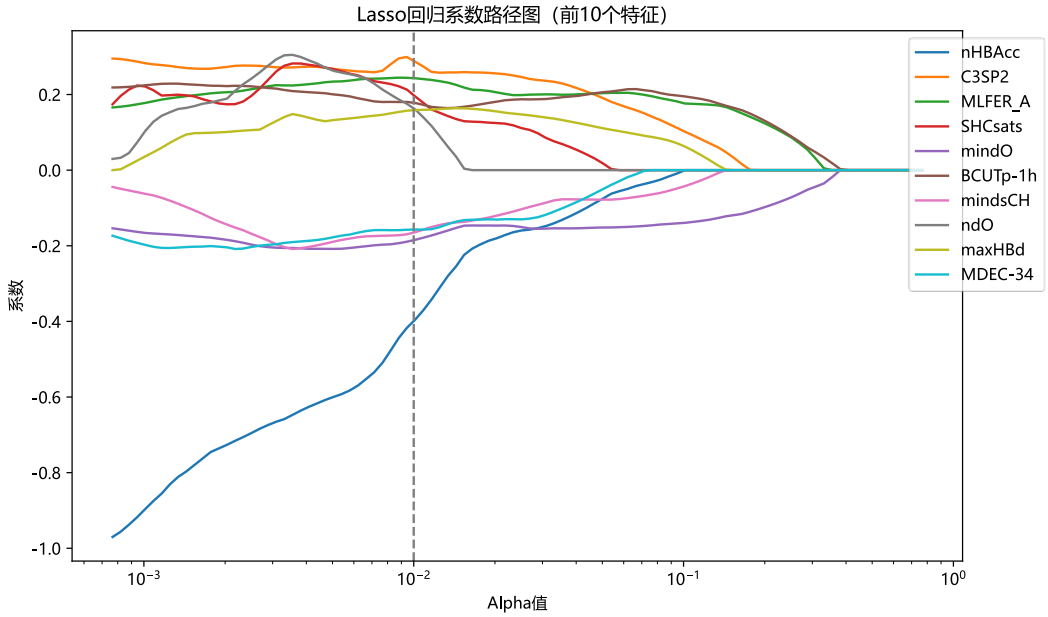

Figure 3. Lasso regression coefficient path plot
图 3. Lasso 回归系数路径图

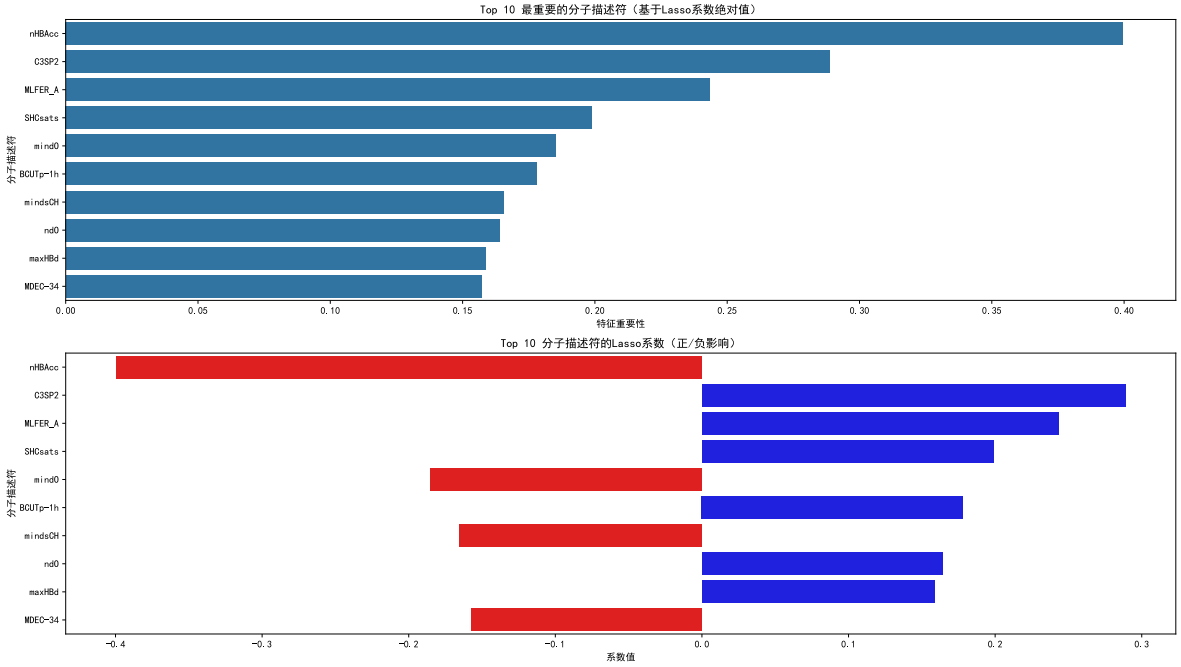

Figure 4. Coefficients of molecular descriptor features
图 4. 分子描述特征系数

通过 Lasso 回归分析，筛选出对生物活性(pIC50)影响最显著的前 20 个分子描述符，包括 nHBAcc (氢键受体数目)、C3SP2 (SP2 杂化的碳原子数目)、MLFER_A (分子线性自由能关系 A 项)、SHCsats (饱和度结构特征)等。这些特征的回归系数表明它们与 ERα 生物活性之间的关系，正系数 C3SP2 和 MLFER_A 表示这些特征与活性呈正相关，负系数 nHBAcc 和 mindO 则表示与活性呈负相关。

4.2. ERα 生物活性预测模型构建与 IC50 预测

结合上面的 4.1 Lasso 回归筛选的 20 个重要的分子描述符，将数据集划分为训练集和测试集，其中80%的数据用于训练，20%的数据用于测试，并构建了一个包含 100 棵树的随机森林回归模型，在训练集上进行预测，得到训练集的 pIC50 预测值，并绘制了测试集样本的真实值与预测值的对比图(见图 5)。

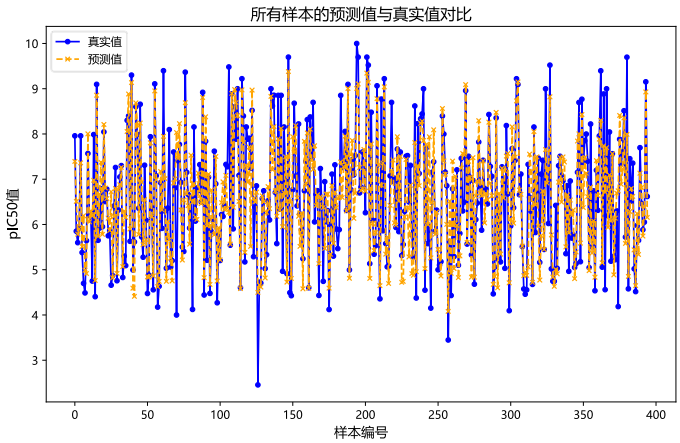

Figure 5. Comparison of predicted vs. actual pIC50 values
图 5. pIC50 预测值与真实值对比

在模型评估方面，真实值与预测值的对比图(如上图所示)提供了直观的预测结果展示。图中蓝色点代表真实值，橙色星号代表预测值，纵轴表示 pIC50 值，横轴为样本编号。通过对比可以发现，大部分样本的预测值与真实值较为接近，说明模型在大多数样本上具有较好的预测能力。

4.3. ADMET 分类预测模型构建与应用

这里基于 BP 神经网络进行化合物的 ADMET 分类预测，其中 BP 网络包含一个隐藏层，隐藏层具有 50 个神经元，输出层则有 2 个神经元，分别对应于分类任务的两个类别。在训练过程中，使用训练集数据作为输入，采用随机梯度下降(SGD)优化算法对模型进行训练，同时使用交叉熵损失函数计算损失。通过反向传播算法，网络更新参数，并进行 1000 次迭代训练(见图 6)。

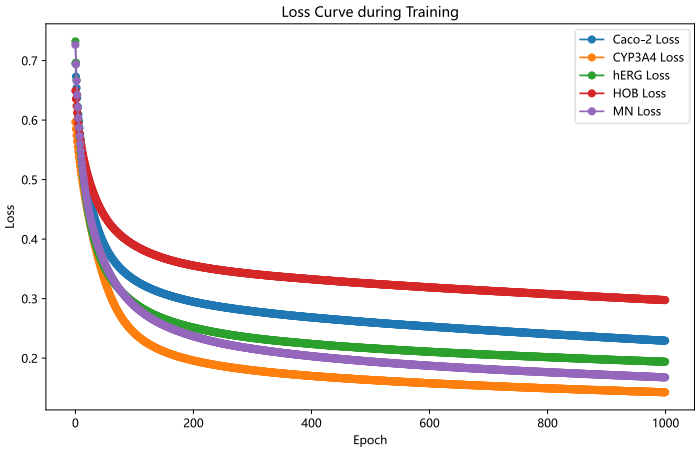

Figure 6. Training progress of BP neural network over 1000 iterations
图 6. 1000 次 BP 网络迭代训练

下表 3 是模型评估的结果：

Table 3. Performance evaluation of BP neural network prediction model
表 3. BP 神经网络预测模型评估

任务	准确率	召回率	精确度	F1 值
Caco-2	0.8887	0.8761	0.89	0.91
CYP3A4	0.914	0.9037	0.8	0.84
hERG	0.8567	0.8545	0.84	0.84
HOB	0.8432	0.7384	0.86	0.90
MN	0.8786	0.7673	0.91	0.69

基于 BP 神经网络进行 ADMET 分类预测，模型在不同任务中的表现有所不同。对于 CYP3A4 任务，模型表现最优，准确率为 91.4%，精确度和召回率均较高，F1 值为 0.84，表明该任务的分类效果较为理想。Caco-2 任务的准确率为 88.87%，F1 值为 0.91，显示了较强的综合表现。hERG 任务的准确率为 85.67%，精确度和召回率相对平衡，F1 值为 0.84，整体表现较好。MN 任务的准确率为 87.86%，精确度高达 91%，但召回率较低，F1 值为 0.69，表明该任务的分类效果稍逊，尤其是在识别正类样本方面。总体而言，模型在多数任务中表现优异，但在某些任务上仍有提升空间，特别是在提高召回率和 F1 值方面。

4.4. 分子描述符与 ERα 活性及 ADMET 性能分析

为了阐明哪些分子描述符能够提高化合物对抑制 ERα 的生物活性，同时具有更好的 ADMET 性质，需要计算出 pIC50 值的中位数，并标记出所有活性值高于中位数的化合物，作为“高活性”化合物。然后，将“高活性”化合物与符合良好 ADMET 条件的化合物进行筛选，只有同时满足这两个条件的化合物才会被标记为“优质化合物”。接着，对于这些符合条件的优质化合物，计算它们在当前特征值上的 25%和 75%分位数，以确定该特征值在优质化合物中的分布范围，并进一步计算这些化合物在该特征值上的中位数。最终，通过对比“优质化合物”与其他化合物的特征值分布，绘制对比图(见图 7)，从而揭示哪些分子描述符在优化药物活性和 ADMET 性质时起到了关键作用。

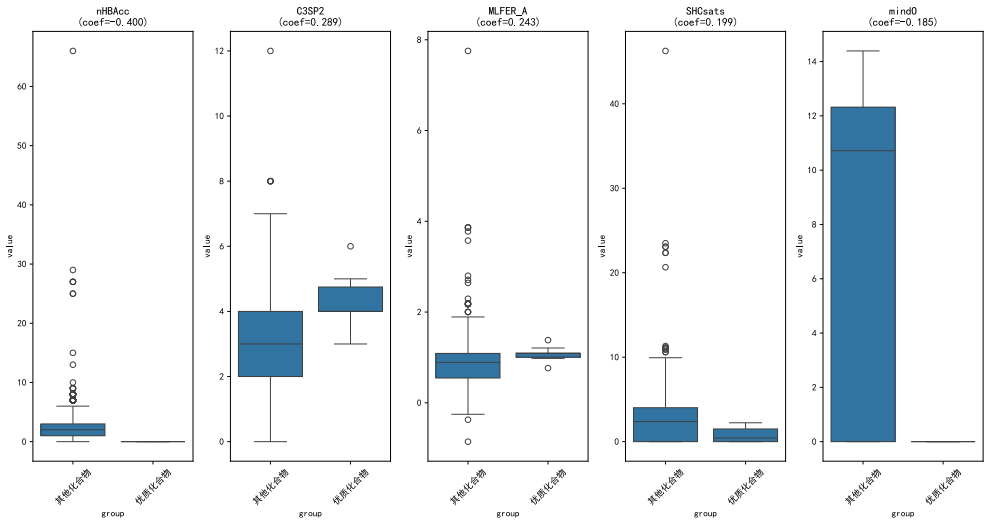

Figure 7. Optimal range of ADMET properties for selected compounds
图 7. 优选化合物最佳范围的 ADMET 性质

通过分析每个分子描述符的最佳范围和中位值，可以得出哪些特征对优质化合物的生物活性和ADMET性质至关重要。图中，C3SP2和MLFER_A的最优范围分别为4.00~4.75和1.01~1.10，表明这些特征在这一范围内能够显著提高化合物活性，并且它们的中位数值分别为4和1.089，显示出这一趋势在大多数优质化合物中的一致性。相反，nHBAcc和mindO的最优范围为0，且中位数值为0，表明较低的氢键受体数目和氧含量有助于化合物的优化。

5. 结论

本研究将Lasso回归与神经网络模型相结合，对与$ER\alpha$抑制活性及ADMET性能相关的分子描述符进行了分析。在Lasso回归中，筛选出了前20个关键分子描述符，发现C3SP2和MLFER_A等特征在特定范围内能够显著提高化合物活性，而nHBAcc和mindO则在取值较低时有助于优化活性。通过随机森林回归模型预测$ER\alpha$生物活性，准确率达到 89%，表明该模型能够较好地捕捉化合物的活性规律。在ADMET分类预测方面，BP神经网络模型在CYP3A4任务上表现最佳，准确率为91.4%，但在MN任务上的召回率较低，需进一步优化。总体而言，模型有效提高了药物活性和安全性的预测精度，为药物优化提供了有价值的参考。

基金项目

重庆对外经贸学院科学研究项目(KYZK2024052)。

参考文献

[1] 王懿辉, 王广兆. “最毒乳腺癌”治疗不再“千人一方” [N]. 健康报, 2024-11-26(002).

[2] Adams, B.D., Furneaux, H. and White, B.A. (2007) The Micro-Ribonucleic Acid (miRNA) Mir-206 Targets the Human Estrogen Receptor-A (ERα) and Represses ERα Messenger RNA and Protein Expression in Breast Cancer Cell Lines. *Molecular Endocrinology*, **21**, 1132-1147. https://doi.org/10.1210/me.2007-0022

[3] Verma, J., Khedkar, V. and Coutinho, E. (2010) 3D-QSAR in Drug Design—A Review. *Current Topics in Medicinal Chemistry*, **10**, 95-115. https://doi.org/10.2174/156802610790232260

[4] Ferreira, L.L.G. and Andricopulo, A.D. (2019) ADMET Modeling Approaches in Drug Discovery. *Drug Discovery Today*, **24**, 1157-1165. https://doi.org/10.1016/j.drudis.2019.03.015

[5] Feinberg, E.N., Joshi, E., Pande, V.S. and Cheng, A.C. (2020) Improvement in ADMET Prediction with Multitask Deep Featurization. *Journal of Medicinal Chemistry*, **63**, 8835-8848. https://doi.org/10.1021/acs.jmedchem.9b02187

[6] 杨紫媛, 李晨红, 唐晔翎, 等. 基于QSAR机器学习模型结合“久病致瘀”理论对丹参治疗慢性疼痛的分子机制研究[J]. 上海中医药大学学报, 2024, 38(2): 71-82.

[7] 郭海丁, 路志峰. 基于BP神经网络和遗传算法的结构优化设计[J]. 航空动力学报, 2003, 18(2): 216-220.