基于深度学习和多种机器学习算法预测人体细胞色素P450抑制剂活性

林明德,韩伟杰,徐小贺,戴晓雯,陈亚东*

(中国药科大学理学院,医药大数据与人工智能研究院,南京 211198)

摘 要 人体细胞色素P450(CYP)受到抑制会导致药物-药物相互作用,从而产生严重的不良反应。因此,准确预测给定化合物对特定CYP亚型的抑制能力至关重要。本研究基于不同的分子表征,比较了11种机器学习方法和2种深度学习模型,实验结果表明,基于RDKit_2d + Morgan的CatBoost机器学习模型在准确率和马修斯系数方面优于其他模型,甚至优于先前发表的模型。此外,实验结果还显示,CatBoost模型不仅性能佳,而且计算资源消耗较低。最后,本文将表现较好的前3名模型结合为co_model,其在性能方面稍微优于单独使用CatBoost模型。

关键词 细胞色素P450;机器学习;深度学习;CatBoost

中图分类号 TP18;R914 **文献标志码** A **文章编号** 1000 -5048(2023)03 -0333 -11

doi:10. 11665/j. issn. 1000 -5048. 2023033103

引用本文 林明德,韩伟杰,徐小贺,等.基于深度学习和多种机器学习算法预测人体细胞色素P450抑制剂活性[J].中国药科大学学报,2023,54(3):333 - 343.

Cite this article as: LIN Mingde, HAN Weijie, XU Xiaohe, *et al*. Activity prediction of human cytochrome P450 inhibitors based on multiple deep learning and machine learning methods[J].*J China Pharm Univ*, 2023, 54(3): 333 - 343.

Activity prediction of human cytochrome P450 inhibitors based on multiple deep learning and machine learning methods

LIN Mingde, HAN Weijie, XU Xiaohe, DAI Xiaowen, CHEN Yadong*

Institute of Medical Big Data and Artificial Intelligence, School of Science, China Pharmaceutical University, Nanjing 211198, China

Abstract Inhibition of human cytochrome P450 (CYP) can lead to drug-drug interactions, resulting in serious adverse reactions. It is therefore crucial to accurately predict the inhibitory power of a given compound against a particular CYP isoform. This study compared 11 machine learning methods and 2 deep learning models based on different molecular representations. The experimental results showed that the CatBoost machine learning model based on RDKit_2d+Morgan outperformed other models in terms of accuracy and Mathews coefficient, and even outperformed previously published models. Moreover, the experimental results also showed that the CatBoost model not only had superior performance, but also consumed less computational resources. Finally, this study combined the top 3 performing models as co_model, which slightly outperformed the CatBoost model alone in terms of performance.

Key words cytochrome P450; machine learning; deep learning; CatBoost

人体细胞色素P450(CYP)广泛存在于细菌、真菌、植物和动物中。CYP超家族有57种血红蛋白亚型,主要存在于肝细胞[1]中。在人体中,5种主要的CYP亚型(CYP1A2、CYP2C9、CYP2C19、CYP2D6和CYP3A4)负责内源性和外源性化合物(如脂肪酸、类固醇和毒素等)以及90%常用药物的氧化还原,是影响药物功效和毒性的重要因素之一[2]。因此,抑制CYP将可能引发药物与药物之间的相互作用,导致严重的不良反应[3-4]。基于上述原因,了解小分子的代谢途径对于设计安全有

收稿日期 2023-03-31 * **通信作者** Tel:13951918023 E-mail:ydchen@cpu. edu. cn

效的药物至关重要。

各种基于结构和配体的药物设计方法已经被用于预测CYP底物的模型[5-6]。基于结构的药物设计方法有以下3种:分子对接、分子动力学模拟和药效团图谱,可以用于预测酶的底物选择性[7]。然而,这些技术面临CYP空腔灵活性和疏水性的挑战。因此,基于结构的方法主要用于研究特殊蛋白质和配体的相互作用[8-9]。与基于结构的方法相反,基于配体的方法不考虑CYP结合位点的情况。尽管CYP与配体相互作用的结构仍不确定,但基于配体的方法对于预测CYP小分子具有相当高的预测精度[10-11]。基于配体的药物设计方法包括以下两种:定量构效关系(QSAR)和3D-QSAR。最常用的是QSAR模型,通过建立分子结构和生物活性之间的定量关系,预测新化合物的生物活性。此外,一系列的机器学习方法,可视为QSAR的升级,已被广泛用于预测CYP底物[12-13]。

目前,由于现有的CYP数据不足以产生良好的且广泛适用的回归模型,因此大多数已开发的模型为分类模型[14]。例如,Cheng等[15]基于Pubchem AID1851数据集,通过反向传播人工神经网络融合多个机器学习分类器,开发了5种分类模型,包括支持向量机、C4.5决策树、k-近邻和朴素贝叶斯。对于CYP1A2、CYP2C9、CYP2C19、CYP2D6和CYP3A4,测试集的预测准确率(ACC)分别为73.1%、86.7%、81.0%、87.8%和76.0%。Pan等[16]同样基于Pubchem AID1851数据集,探索了分子全息图和MACCS描述符,并建立了预测CYP1A2抑制剂活性的支持向量机分类模型,该模型在测试集上的准确率为71%。深度学习技术也已应用于CYP活性预测中,例如,Wu等[17]比较了深度神经网络、卷积神经网络、随机森林和不同梯度增强决策树区分抑制剂和非抑制剂的能力。他们通过Pubchem AID1851数据集进行模型训练,并使用来自AID410、AID883、AID899、AID891和AID884的数据进行模型测试。实验结果表明,极端梯度增强方法的表现略优于神经网络(预测准确率约为90%)。Li等[18]使用与Wu等相同的数据集训练了一个多任务自动编码器深度神经网络分类模型。在外部测试集上,他们的模型获得的马修斯相关系数(MCC)在0.37~0.89之间。实验结论是,深度神经网络不仅在训练集上表现良好,而且在外部测试集上也优于其他机器学习方法。

然而,目前CYP数据集存在不平衡的问题,这使得建立的模型无法很好地应对这一关键问题,导致训练出的模型准确率无法达到最佳,并且评价指标过于依赖ACC和受试者工作特征曲线下面积(AUC)等,对MCC的衡量没有充分重视,特别是在存在不平衡的二分类问题中,这往往无法客观反映模型的性能。

本研究通过Pubchem上的大数据集,并基于不同的分子表征方法,建立了深度学习和机器学习模型,比较了针对CYP活性预测的最佳模型和其最佳描述符,并主要使用MCC作为重要指标来评估模型的优劣。

1 方 法

1.1 数据集准备

本实验使用的所有数据集均从Pubchem生物分析数据库中(https://Pubchem.ncbi.nlm.nih.gov/bioassay)下载而来。每个数据集都包含了化合物的活性评分、效价、曲线描述、拟合对数IC_{50}和拟合r^2。第一个数据集(AID:1851)包含了17 143种化合物。该数据集使用多种人体CYP同工酶(包括CYP1A2、CYP2C9、CYP2C19、CYP2D6和CYP3A4)来测量荧光素底物到荧光素脱烷基的过程。在进行实验时,通过加入荧光素酶检测试剂并测量荧光素的发光程度,来观察不同浓度的化合物对发光率的影响,从而确定这些化合物对这5种亚型的效力[18],这些化合物的数据将作为本实验的训练集。对于其他5项数据集,采用相同的实验方案,但每项只检测一个同工酶(AID:410用于CYP1A2,AID:883用于CYP2C9,AID:899用于CYP2C19,AID:891用于CYP2D6,以及AID:884用于CYP3A4),这些化合物的数据将作为本实验的测试集。

1.2 抑制剂和非抑制剂的标记

在每个数据集中,样本标记分为3种不同的类别,包括“active”(活性)、“inactive”(非活性)和“inconclusive”(不确定)。首先,本研究排除了标记为“inconclusive”的数据,只有标记为“active”或“inactive”的化合物进行进一步处理。然后,根据IC_{50}[15]、Pubchem活性评分[15]和浓度-反应曲线[19](参见表1),将所有化合物分为两类:抑制剂和非

抑制剂。由于 IC_{50} 是用浓度-反应曲线计算得出来的，所以可能会受到异常值的影响。因此，本研究结合另外两个标准（Score 和 Curve class）来进行鉴别。对于通过 IC_{50} 区分出的抑制剂与非抑制剂，如果某个化合物无法同时符合这两个鉴定标准，那么该化合物将被直接删除。最后，本研究还删除了数据集中重复出现的化合物 SMILES。最终的数据集数量如表 2 所示。测试集和训练集的数据是相互独立的，即测试集中的化合物不会同时出现在训练集中。

Table 1　Classification criteria for cytochrome P450 (CYP) isoform inhibitors and non-inhibitors

Criteria	Inhibitor	Non-inhibitor
IC_{50}	≤10 μmol/L	>57 μmol/L
Score	≥40	0
Curve class	−1.1, −1.2, −2.1	4

Table 2　Details of the 5 CYP isoform datasets

Isoform	Dataset			
	Class	Train	Test	Total
CYP1A2	Noninhibitor	6 575	325	6 904
	Inhibitor	4 364	91	4 455
	Total	10 939	416	11 355
CYP2C9	Noninhibitor	7 852	1 039	8 891
	Inhibitor	2 958	96	3 054
	Total	10 810	1 135	11 945
CYP2C19	Noninhibitor	6 722	940	7 662
	Inhibitor	4 925	230	5 155
	Total	11 647	1 170	12 817
CYP2D6	Noninhibitor	10 459	1 346	11 805
	Inhibitor	1 473	130	1 603
	Total	11 932	1 476	13 407
CYP3A4	Toninhibitor	6 909	2 524	9 433
	Inhibitor	3 460	558	4 018
	Total	10 366	3 082	13 448

1.3　模　型

1.3.1　支持向量机（support vector machine，SVM）　SVM 是一种广义线性分类器，通过监督学习的方式对数据进行二元分类。它在特征空间中寻找一个超平面作为决策边界，使得不同类别样本点之间的间隔最大化，从而提高分类的准确性和泛化能力[20]。

1.3.2　k-近邻（k-nearest neighbor，KNN）　KNN 是一种简单直观的机器学习算法，可用于分类和回归问题。它基于邻居之间的距离来确定样本的类别或输出值。KNN 通过计算待分类样本与训练集样本之间的距离，并选择最近的 k 个邻居，根据它们的类别进行投票或计算平均值来预测样本的类别或输出值[21]。

1.3.3　决策树（decision tree，DT）　DT 是一种分层决策结构，可用于分类和回归。它通过建立一棵树状结构来对数据进行划分和预测。决策树的每个节点代表一个特征，根据该特征对数据进行划分，直到达到叶子节点，叶子节点代表数据的类别或输出值。决策树具有推理速度快且可解释性强的特点，是一种应用非常广泛的算法[22]。

1.3.4　随机森林（random forest，RF）　RF 是一种非线性的基于决策树的集成方法，是决策树的 Bagging 扩展变体，它通过在决策树的训练过程中引入随机特征选择，提高最终集成模型的泛化能力。RF 具有高预测精度，对异常值和噪声的容忍度高，不易过拟合的特点。此外，它能够处理具有高维特征的输入样本，无需降维，因此成为 QSAR 建模中最流行的算法之一[23]。

1.3.5　轻量级梯度提升机（light gradient boosting machine，LightGBM）　LightGBM 是一种实现 GBDT 算法的框架，GBDT 是机器学习中长期存在的模型，其主要思想是使用弱分类器（决策树）迭代训练获得最优模型。该模型具有训练效果好、过拟合可能性较小的优点。LightGBM 支持高效的并行训练，具有更快的训练速度、更低的内存消耗、更好的准确性以及分布式支持快速处理大量数据等优点[24]。

1.3.6　梯度提升决策树（gradient boosting decision tree，GBDT）　GBDT 也称为多元加性回归树，是一种迭代决策树算法。与传统的 Boosting 方法不同，GBDT 的每次计算都是为了减少先前构建的树学习器的残差，而不是专注于重新加权错误分类的样本。为了最小化残差，GBDT 构建了一个决策树学习器以及残差梯度的方向。GBDT 通过累加所有树的预测结果来得出预测结果，而累加过程是通过回归而不是分类来完成。因此，与 RF 不同，GBDT 的树是 CART 回归树，而不是分类树，并且这些树只能串行生成[25]。

1.3.7　极端梯度提升（extreme gradient boosting，XGBoost）　XGBoost 是一种强大的机器学习算法，

属于集成学习中的梯度提升方法。它通过迭代构建多个弱学习器,并将它们组合成一个强大的预测模型。XGBoost通过优化目标函数,使用梯度提升算法逐步提高模型的性能,并提供正则化选项,包括L1和L2正则化,用于控制模型的复杂度,降低模型的方差,防止过拟合[26]。

1.3.8 AdaBoost (Ada) AdaBoost是由Freund和Schapire于1995年首次提出的迭代算法,是Boosting家族的成员之一,也是最常用的机器学习方法之一。Ada的核心思想是为同一个训练集训练多个不同的弱分类器,然后将这些弱分类器集合到一个强分类器中。该分类器具有简单、检测速度快、分类准确率高和不易过拟合等优点[27]。

1.3.9 CatBoost CatBoost算法是俄罗斯搜索巨头Yandex于2017年开源的机器学习库,是一种Boosting系列算法。CatBoost、XGBoost和LightGBM模型也被称为GBDT的三大主流工具,都是基于GBDT算法框架进行改进实现[28]。在本研究中,CatBoost模型表现最为出色,在所有训练模型中具有最快的训练速度、最高准确度和最少的内存消耗。此外,该模型具有调节参数少,操作简单的特点,其中的参数scale_pos_weight能有效处理CYP数据中的不平衡问题。本文通过交叉验证对CatBoost设定了关键参数(depth:8,interaction:2 000,l2_leaf_reg:3,learing rate:1),其余参数保持原始值设定。

1.3.10 逻辑回归(logistic regression,LR) LR是一种用于解决二元分类问题的机器学习方法。LR算法的核心思想是通过对输入特征的线性组合应用逻辑函数,将输入映射到一个介于0和1之间的概率。这个概率可以被解释为样本属于某个类别的概率。在训练阶段,LR通过最大似然估计或梯度下降等优化算法来学习模型参数,使得模型能够最大程度地拟合训练数据[29]。

1.3.11 极端随机树(extreme random trees,ET) ET算法采用了类似决策树的结构,通过组合多个弱分类器来构建一个强分类器。它通过迭代训练,每一轮都基于前一轮的结果调整样本的权重,使得前一轮分类错误的样本在下一轮中得到更多关注。最终的预测结果是基于所有弱分类器的加权组合。[30]。

1.3.12 深度神经网络(deep learning neural network, DNN) DNN是深度学习的一种框架,由多层计算节点组成,按照不同层的位置,内部神经网络可以分为输入层、隐藏层和输出层。神经元和神经元层的数量取决于数据集中描述符的数量、化合物的数量和输出的类型[31]。本实验基于DNN建立了单任务学习方法和多任务学习方法。两种模型都包含了一个输入层,3个隐藏层和一个输出层,唯一的区别是多任务学习模型在输入层后面增加了一个共享层,用于共享参数。这些模型的超参数如表3所示。通过逐个实验,根据在测试集上的ACC和MCC确定了模型的最佳参数,其余超参数根据原始值设定。

Table 3 Hyperparameter setting based on single-task and multi-task learning

Hyperparameter	Setting	Hyperparameter	Setting
Optimizer	Adam (lr = 1e − 3)	Share layer	520 (1 250, 256)
Dropout rate	0.3 (0.2, 0.4, 0.5)	Tower_1	168 (520, 256, 168)
Sampling	RUS (ROS, SMOT)	Tower_2	168 (520, 256, 168)
Epoch	250 (150, 250, 500)	Tower_3	64 (256, 168, 64)
Loss	Crossentropy	Output layer	2

1.4 模型评价指标

本研究采用马修斯相关系数(Matthews correlation coefficient, MCC)作为模型性能的度量。MCC是一种综合考虑真阳性(true positive,TP)、真阴性(true negative,TN)、假阳性(false positive,FP)和假阴性(false negative,FN)预测数量的测量方法,因此被认为是一种稳健的度量,适用于评价不平衡数据的模型[32]。根据公式(1)可得:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

MCC的范围从-1到+1,其中-1为负相关,0为无相关,+1为完美相关。根据经验,获得MCC大于0.5的模型通常被认为是性能良好的模型。

本研究还使用了准确率（accuracy，ACC）作为第二个性能指标，它能直观地反映模型预测正确样本的比例。其他在本研究中使用的性能指标包括敏感性（sensitivity，Se），特异性（specificity，Sp），精确度（precision，Pr），受试者工作特征曲线下面积（area under curve，AUC）。

1.5　描述符的计算和修剪

本研究计算了3种类型的特征，包括所有化合物的二维分子描述符（2D）、Morgan指纹和MACCS指纹，并将它们作为建模的分子表征。此外，合并2D描述符和Morgan指纹合并为第4个分子表征。所有类型的分子表征都是使用RDKit库（http://www. RDKit. org/）在Python中计算生成的。尽管本实验计算了大量的描述符，但并非所有的描述符都对模型都有帮助。存在不相关的和冗余的描述符变量会影响模型的泛化性能，并可能导致过度拟合。为了建立一个可靠的模型，在进行特征选择之前，本研究对已生成的分子描述符和分子指纹做了3次预处理。预处理步骤如下：（1）填充空值：使用相应描述符的平均值来填充缺失值；（2）低方差过滤：删除方差为0或接近于0的描述符，这些描述符的变量对于不同分子具有相同值，因此可以清除；（3）高相关过滤：利用SelectKBest方法过滤具有相关性的特征。选择与标签最相关的特征的最优超参数k，生成与统计量相匹配的新特征矩阵。这种方法可以过滤掉模型的噪声值，提高模型特征的相关性和有效性。表4详细描述了经过预处理后不同分子描述符集的信息。数据集处理的整体工作流程如图1所示。

特征选择方面，本研究采用互信息来筛选描述符并评估两个随机变量[33-35]之间的相关性。在概率论和信息论中，互信息或反式信息量化了两个随机变量X和Y[36]之间的依赖性，其公式（2）如下所示：

$$I(X;\ Y) = \sum_{x \in X}\sum_{y \in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (2)$$

其中$p(x,y)$是X和Y的联合概率分布函数，$p(x)$和$p(y)$分别是X和Y的边际概率分布函数。只有在两个随机变量是独立的情况下，互信息为零，数值越大表示依赖性越强。

归一化是指以均值μ为中心，然后以标准差σ为尺度对数据进行处理，最终使数据呈现出均值μ = 0、标准差σ = 1的正态分布。其公式（3）如下：

$$x' = \frac{x - \mu}{\sigma} \quad (3)$$

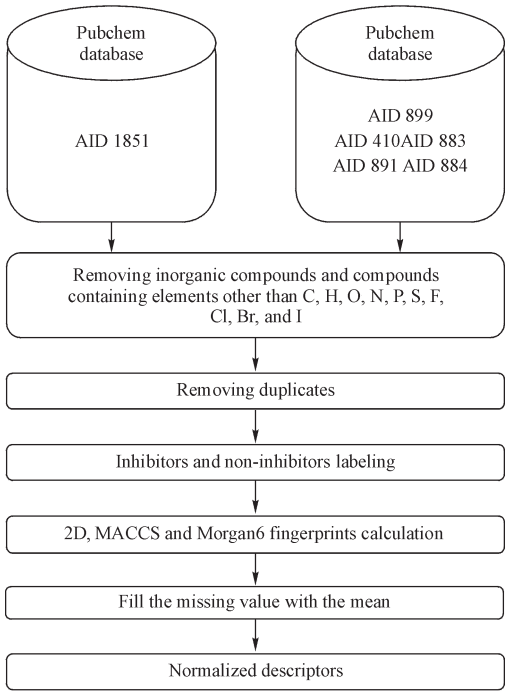

Figure 1　Overall workflow of CYP dataset processing

MACCS: Molecular access system

Table 4　Number of molecular descriptors for the 5 CYP isoforms after preprocessing

Isoform	MACCS	Morgan	RDKit_2d	RDKit_2d+ Morgan
CYP1A2	118	604	170	813
CYP2C9	107	594	171	740
CYP2C19	110	582	172	750
CYP2D6	109	549	158	716
CYP3A4	107	597	164	714

2　结　果

2.1　基于不同方法的CYP抑制分类模型的比较

使用了Morgan指纹作为特征，并采用了11种不同的机器学习方法来建立CYP1A2、CYP2C9、CYP2C19、CYP2D6和CYP3A4的二分类模型。这些方法包括7种代表性的集成学习方法（CatBoost、LightGBM、AdaBoost、XGBoost、ET、RF、GBDT）和4种单一学习器（KNN、DT、LR和SVM）。通过对5个训练集进行5折交叉验证，评估了不同模型在

5个测试集上的预测精度,具体结果如图2和图3所示。结果显示,CatBoost模型在CYP1A2、CYP2D6和CYP3A4亚型的预测中表现最好,其准确率分别为0.94、0.93和0.89。MCC分别为0.82、0.46和0.59。另外,LightGBM模型对CYP2C19亚型的预测效果较好,准确率为0.93,与CatBoost模型相当,且其MCC为0.46,稍高于CatBoost模型的0.43;而XGBBoost模型在CYP2C9亚型的预测中取得了良好的结果,准确率为0.93,其MCC为0.50也高于CatBoost模型的0.44。综合以上实验结果来看,CatBoost、LightGBM和XGBoost模型在CYP抑制分类中表现较佳。尽管这3种模型在不同亚型上的性能有所差异,但是差距并不明显,无法明确突出某个模型的优势。

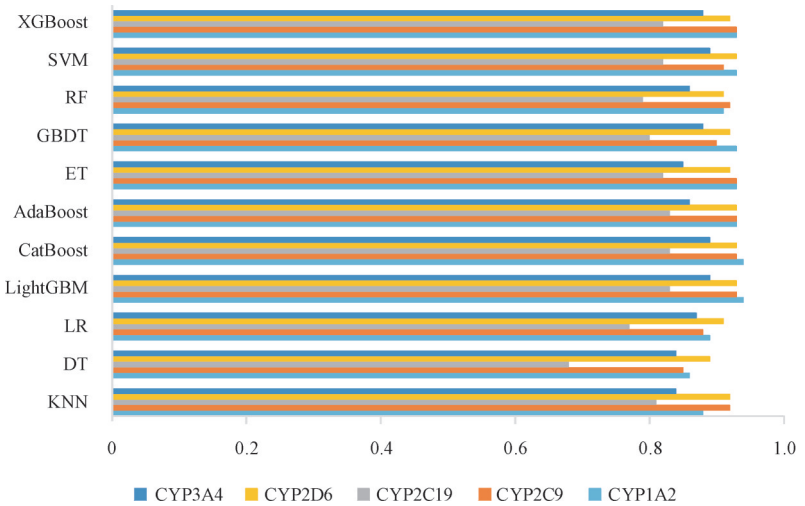

Figure 2 Cross-validated the accuracy (ACC) of different models for the training set of 5 CYP isoforms

XGBoost: Extreme gradient boosting; SVM: Support vector machine; RF: Ranadom forest; GBDT: Gradient Boosting decision tree; ET: Extreme random trees; LightGBM: Light gradient boosting machine; LR: Logistic regression; DT: Decision tree; KNN: k-nearest neighbor

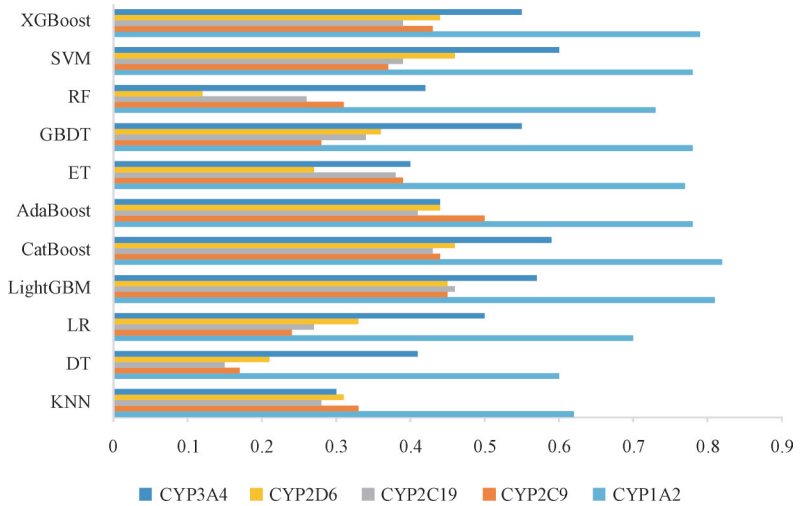

Figure 3 Cross-validated the Matthews correlation coefficient (MCC) of different models for the training set of 5 CYP isoforms

2.2 基于相同核数下不同模型的运行时间比较

根据前面的实验结果,得出了有3种性能较佳的机器学习模型,但无法具体突出某个模型的优势。因此,本研究进一步比较这3种模型在资源利用方面的差异。本研究关注的是在有限的资源下如何获得较好的结果。为了扩大比较范围,还加入了在上述实验中指标仅次于最佳模型的SVM和AdaBoost模型。将使用CYP1A2亚型的数据集作为代表,分别运行这5种表现较优的模型,并计算在相同核数(本次实验为10核)情况下的运行时间,通过模型的运行时长来评估其资源占用情况(即如果一个模型的运行时间是另一个模型的2

倍，则可以推导出在相同时间内，该模型运行所占用的核数是另一个模型的 2 倍）。在计算各个模型的运行时间之前，本实验都会在封装好的模型前后加上开始计时和结束计时的代码，以准确计算各模型的运行时长。

实验结果显示（图 4），在保持各模型运行条件保持相同的环境下，CatBoost 模型以最快的速度完成计算，仅用时 55 s。而与之性能相当的 XGBoost 模型却需要 2 956 s 的训练时间，差距巨大，近乎 60 倍。换句话说，在相同时间内，训练这两个模型所需的计算资源，CatBoost 模型仅需要 XGBoost 模型的 1.7% 就足够。这可以归因于以下几个原因：

（1）特定优化技术：CatBoost 模型使用了一些特定的优化技术，如对称树布局和特征直方图近似算法等，以加速模型的训练和预测过程。这些技术可以减少内存使用并提高计算效率。

（2）类别特征处理：CatBoost 模型在处理类别特征时采用了一种基于特征哈希技术的编码方式。这种编码方式能够在不引入过多的内存开销的情况下，有效地将类别特征转换为数值表示，提高了训练和预测的速度。

（3）多线程支持：CatBoost 模型支持多线程训练，可以同时利用多个 CPU 核心进行并行计算。这样可以加快模型训练的速度，尤其是在处理大规模数据集时更为明显。

综上所述，尽管 CatBoost 模型与 XGBoost 模型表现性能相当，各项指标相差无几，但是从资源利用的角度来看，CatBoost 模型远远优于 XGBoost 模型。同样，它也优于与之性能相当的 LightGBM 模型，以及附加的 AdaBoost 模型和 SVM 模型。

2.3　基于不同描述符集的 CatBoost 模型比较

在 QSAR 建模中，分子的结构通过分子描述符进行编码，因此选择合适的描述符对于开发可靠的 QSAR 模型至关重要。本实验的重点是确定最适合 CatBoost 模型的分子表示方式。首先，比较了仅使用单一分子描述符（RDKit_2d）或任何一套分子指纹（MACCS 和 Morgan）的 CatBoost 模型的预测能力。然后，又结合了 Morgan 指纹和 RDKit_2d 描述符来开发 CatBoost 模型，以比较基于单一的描述符或指纹与结合二者开发的 CatBoost 模型的优劣。实验结果如图 5 所示，基于 RDKit_2d+Morgan 的模型性能最优。这表明该描述符与分子特征之间存在较强的相关性，能够提供丰富的信息用于描述分子特征，同时也说明该描述符与 CatBoost 模型的匹配度较高（具体模型指标值见表 5 和表 6）。表 7 展示了基于 RDKit_2d+Morgan 描述符的 CatBoost 模型的各项指标值。从表中可以看到，该模型不仅在 ACC 和 MCC 指标上表现良好，而且在 AUC（0.88~0.98）以及其他指标上也显示出优越性，这表明 CatBoost 模型不仅具有良好的分类性能，而且具备较高的鉴别能力，能够有效区分正例和负例。

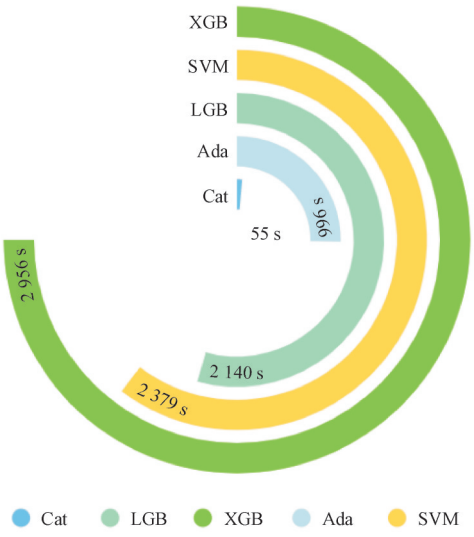

Figure 4　Time required to train each model based on the CYP1A2 dataset

XGB: Extreme gradient boosting; LGB: Light gradient boosting machine; Ada: AdaBoost; Cat: CatBoost

2.4　CatBoost 模型与深度学习算法的比较

基于前面的实验结果，在机器学习领域中，基于 Morgan+RDKit_2d 的 CatBoost 模型表现最佳。因此，本研究进一步将机器学习中的 CatBoost 模型与深度学习中的单任务 DNN 模型与多任务 DNN 模型进行比较。实验结果如表 8 所示，无论基于单任务还是多任务 DNN 模型，对于任一亚型，它们的 ACC 和 MCC 均低于 CatBoost 模型，而其他指标如 AUC、Pr 和 Sp 等也没有超过 CatBoost 模型。这表明，在全面考虑真假阳性与真假阴性的情况下，CatBoost 模型最能准确地对 CYP 化合物的抑制剂和非抑制剂进行分类，尽管在敏感性上可能存在微小差距，说明 CatBoost 模型在个别亚型数据集上对真正例的识别略有误差，但这种误差并不影响

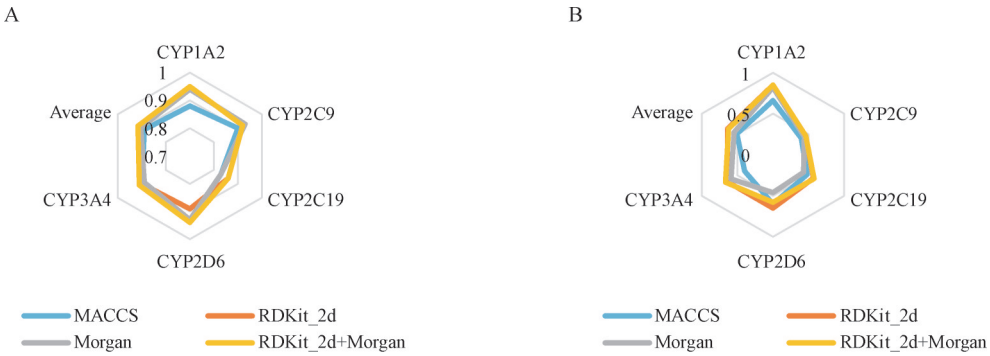

Figure 5 Accuracy (ACC) values (A) and MCC values (B) of the CatBoost models based on different sets of descriptors for the 5 CYP isoform test sets

Table 5 ACC values of the CatBoost models based on different sets of descriptors for the 5 CYP isoform test sets

Isoform	MACCS	RDKit_2d	Morgan	RDKit_2d+Morgan
CYP1A2	0.88	0.95	0.94	0.95
CYP2C9	0.90	0.92	0.93	0.92
CYP2C19	0.83	0.86	0.83	0.86
CYP2D6	0.94	0.89	0.93	0.94
CYP3A4	0.90	0.90	0.89	0.91
Average	0.89	0.904	0.904	0.916

Table 6 MCC values of the CatBoost models based on different sets of descriptors for the 5 CYP isoform test sets

Isoform	MACCS	RDKit_2d	Morgan	RDKit_2d+Morgan
CYP1A2	0.66	0.85	0.82	0.85
CYP2C9	0.4	0.46	0.44	0.46
CYP2C19	0.48	0.57	0.43	0.58
CYP2D6	0.6	0.65	0.46	0.58
CYP3A4	0.4	0.65	0.59	0.67
Average	0.508	0.636	0.548	0.628

Table 7 Various indicators of the CatBoost classifier for the 5 CYP isoform test sets

Isoform	Sp	Se	Pr	ACC	MCC	AUC
CYP1A2	0.97	0.86	0.89	0.94	0.84	0.98
CYP2C9	0.96	0.48	0.53	0.92	0.46	0.88
CYP2C19	0.9	0.68	0.65	0.86	0.58	0.89
CYP2D6	0.98	0.52	0.72	0.94	0.58	0.92
CYP3A4	0.97	0.62	0.84	0.91	0.67	0.95

Sp: Specificity; Se: Sensitivity; Pr: Precision

CatBoost模型成为预测CYP活性的最佳模型。

2.4 集成模型

考虑到不同算法的原理可能会对同一分子产生完全相反的预测[37],因此,本文针对每个亚型建立了一个集成模型(co_model)。co_model结合之前3个表现较好的分类器的结果,以确定化合物是否具有活性。在这个实验中,规定当其中的3个单分类器中至少有两个预测某个分子对特定靶点具有活性时,该分子才被判定为活性分子。通过对比实验发现,co_model的预测结果略优于单个分类器的预测结果。co_model在AUC指标上显示出了显著差异,表明与单个模型相比,co_model的性能有所提高,详细结果见表9。

Table 8 Performance comparison of 5 CYP isoforms in DNN, Mul_DNN and CatBoost models

Isoforms	Method	ACC	MCC	AUC	Sp	Se	Pr
CYP1A2	DNN	0.91	0.74	0.96	0.94	0.8	0.79
	Mul_DNN	0.91	0.75	0.96	0.94	0.82	0.78
	CatBoost	**0.94**	**0.84**	**0.98**	**0.97**	**0.86**	**0.89**
CYP2C9	DNN	0.9	0.38	0.8	0.94	0.45	0.42
	Mul_DNN	0.86	0.36	0.86	0.89	**0.58**	0.32
	CatBoost	**0.92**	**0.46**	**0.88**	**0.96**	0.48	**0.53**
CYP2C19	DNN	0.83	0.46	0.84	0.90	0.55	0.57
	Mul_DNN	0.84	0.51	0.87	0.90	0.61	0.60
	CatBoost	0.86	0.58	0.89	0.90	0.68	0.65
CYP2D6	DNN	0.83	0.46	0.84	0.90	**0.55**	**0.57**
	Mul_DNN	0.92	0.47	0.86	0.96	0.5	0.53
	CatBoost	**0.94**	**0.58**	**0.92**	**0.98**	0.52	0.72
CYP3A4	DNN	0.91	0.66	0.93	0.96	**0.64**	0.80
	Mul_DNN	0.89	0.60	0.92	0.95	0.61	0.73
	CatBoost	**0.91**	**0.67**	**0.95**	**0.97**	0.62	**0.84**

DNN: Deep neural networks; Mul_DNN: Multitask deep neural networks

2.5 CatBoost与之前报道的模型比较

本实验对比了CatBoost模型与之前报道的Li等[18]、Su等[38]、Sun等[39]和Wu等[17]所开发的CYP抑制剂分类器。所有这些分类器都是在AID 1851

Table 9　Performance comparison between co_model and CatBoost for 5 CYP isoforms

Item	CYP1A2		CYP2C9		CYP2C19		CYP2D6		CYP3A4		Difference
	Cat	Co	Cat	Co	Cat	Co	Cat	Co	Cat	Co	
ACC	0.94	0.96	0.96	0.97	0.9	0.91	0.98	0.99	0.97	0.98	0.01
MCC	0.84	0.87	0.48	0.45	0.68	0.7	0.52	0.41	0.62	0.61	−0.02
AUC	0.98	0.96	0.53	0.62	0.65	0.67	0.72	0.83	0.84	0.85	0.04
Sp	0.97	0.98	0.92	0.93	0.86	0.87	0.94	0.94	0.91	0.91	0.01
Se	0.86	0.88	0.46	0.49	0.58	0.6	0.58	0.56	0.67	0.67	0.01
Pr	0.89	0.92	0.88	0.88	0.89	0.9	0.92	0.92	0.95	0.9	0.00

Cat: CatBoost; Co: co_model

数据集上进行训练，并在 AID410、AID883、AID899、AID891、AID884 数据集上进行测试。实验结果（图 6）显示，除了 Wu 等[17]和 Li 等[18]开发的分类器对 CYP1A2 亚型的预测准确率超过 CatBoost 模型，其他 4 个亚型的所有分类模型均低于 CatBoost 模型的预测性能。

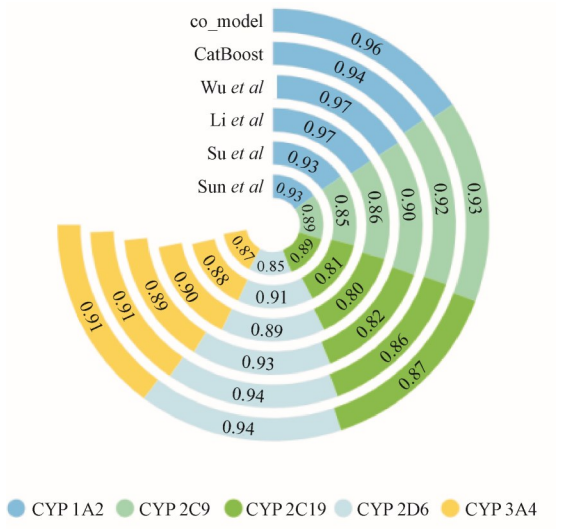

Figure 6　Performance of the 5 CYP isoforms in the CatBoost model compared to previously published models

3　总结与展望

本研究旨在解决 CYP 抑制剂的二分类问题，通过建立了 11 种机器学习模型和 2 种深度学习模型。研究结果表明，在使用不同机器学习算法创建的模型中，CatBoost、LightGBM 和 XGBoost 模型的性能相当。进一步比较这 3 种模型所需的计算资源，实验结果显示 CatBoost 模型消耗的资源远少于 LightGBM 和 XGBoost 模型，仅占用了 XGBoost 模型资源的 1.7%，却能获得相当甚至更好的预测结果。接下来比较了基于不同分子表征的 CatBoost 模型，结果显示基于 RDKit_2d+Morgan 的 CatBoost 模型性能最优。随后，将该模型与基于 DNN 的单任务和多任务学习算法进行比较，CatBoost 模型仍然表现优于深度学习模型。因此，基于 RDKit_2d+Morgan 的 CatBoost 模型在预测 CYP 小分子活性方面表现更佳。此外，本研究将表现相当的 CatBoost、LightGBM 和 XGBoost 模型集成成一个 co_model，其预测准确率略优于单个 CatBoost 模型。将 CatBoost 模型和 co_model 与已发表的模型进行比较，除了 CYP1A2 亚型的数据集外，在其他亚型的数据集中，CatBoost 模型和 co_model 的准确率均优于已发表的模型。

综上所述，本研究认为基于 RDKit_2d+Morgan 的 CatBoost 模型在预测 CYP 小分子活性方面表现出色，并且在计算资源成本上迈出了重要的一步。利用训练好的 CatBoost 模型，可以预测任何化合物，判断其是否为人体内 CYP 的底物，是抑制剂还是诱导剂。这对于早期药物活性预测提供了巨大的帮助。未来将进一步优化模型性能，提高预测的准确性和可靠性，并将其应用于更广泛的化合物库和药物研发中。

References

[1] Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics[J]. *Science*, 1999, **286** (5439): 487-491.

[2] Feiters MC, Rowan AE, Nolte R. ChemInform abstract: from simple to supramolecular cytochrome P450 mimics[J]. *Chem Soc Rev*, 2000, **29**(6): 375-384.

[3] du Souich P. In human therapy, is the drug-drug interaction or the adverse drug reaction the issue[J]? *J Can De Pharmacol Clin*, 2001, **8**(3): 153-161.

[4] Williams JA, Hyland R, Jones BC, *et al*. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic

explanation for typically observed low exposure (AUCi/AUC) ratios[J]. *Drug MeTable Dispos*, 2004, **32**(11): 1201-1208.

[5] Khakar PS. Two-dimensional (2D) *in silico* models for absorption, distribution, metabolism, excretion and toxicity (ADME/T) in drug discovery[J]. *Curr Top Med Chem*, 2010, **10**(1): 116-126.

[6] Dai H, Xu Q, Xiong Y, *et al*. Improved prediction of *Michaelis* constants in CYP450-mediated reactions by resilient back propagation algorithm[J]. *Curr Drug Metab*, 2016, **17**(7): 673-680.

[7] Kato H. Computational prediction of cytochrome P450 inhibition and induction[J]. *Drug MeTable Pharmacokinet*, 2020, **35**(1): 30-44.

[8] Leach AG, Kidley NJ. Cytochrome P450 substrate recognition and binding[M]// *Drug Metabolism Prediction*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA,2014: 103-132.

[9] Oostenbrink C. Structure-based methods for predicting the sites and products of metabolism[M]// *Drug Metabolism Prediction*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2014: 243-264.

[10] Kirchmair J, Williamson MJ, Tyzack JD, *et al*. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms[J]. *J Chem Inf Model*, 2012, **52**(3): 617-648.

[11] Shan XQ, Wang XG, Li CD, *et al*. Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method[J]. *J Chem Inf Model*, 2019, **59**(11): 4577-4586.

[12] Xiong Y, Qiao YH, Kihara D, *et al*. Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates[J]. *Curr Drug Metab*, 2019, **20**(3): 229-235.

[13] Tyzack JD, Hunt PA, Segall MD. Predicting regioselectivity and lability of cytochrome P450 metabolism using quantum mechanical simulations[J]. *J Chem Inf Model*, 2016, **56**(11): 2180-2193.

[14] Gleeson MP, Davis AM, Chohan KK, *et al*. Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models[J]. *J Comput Aided Mol Des*, 2007, **21**(10/11): 559-573.

[15] Cheng FX, Yu Y, Shen J, *et al*. Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers [J]. *J Chem Inf Model*, 2011, **51**(5): 996-1011.

[16] Pan XC, Chao L, Qu SJ, *et al*. An improved large-scale prediction model of CYP1A2 inhibitors by using combined fragment descriptors[J]. *RSC Adv*, 2015, **5**(102): 84232-84237.

[17] Wu ZX, Lei TL, Shen C, *et al*. ADMET evaluation in drug discovery. 19. reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches[J]. *J Chem Inf Model*, 2019, **59**(11): 4587-4601.

[18] Li X, Xu YJ, Lai LH, *et al*. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network[J]. *Mol Pharm*, 2018, **15**(10): 4336-4345.

[19] Inglese J, Auld DS, Jadhav A, *et al*. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries[J]. *Proc Natl Acad Sci U S A*, 2006, **103**(31): 11473-11478.

[20] Zhao XW, Ma ZQ, Yin MH. Using support vector machine and evolutionary profiles to predict antifreeze protein sequences[J]. *Int J Mol Sci*, 2012, **13**(2): 2196-2207.

[21] Hu LY, Huang MW, Ke SW, *et al*. The distance function effect on k-nearest neighbor classification for medical datasets[J]. *Springerplus*, 2016, **5**(1): 1304.

[22] Tong WD, Hong HX, Fang H, *et al*. Decision forest: combining the predictions of multiple independent decision tree models[J]. *J Chem Inf Comput Sci*, 2003, **43**(2): 525-531.

[23] Breiman L. Random Forests[J]. *Mach Learn* , 2001, **45**: 5-32.

[24] Ke G, Meng Q, Finley T, *et al*. LightGBM: a highly efficient gradient Boosting decision tree[C]// Advances in Neural Information Processing Systems 30. Long Beach:Curran Associates Inc., 2017: 3149-3157.

[25] Friedman JH. Greedy function approximation: a gradient Boosting machine[J]. *Ann Statist*, 2001, **29**(5): 1189-1232.

[26] Chen TQ, Guestrin C. XGBoost: a scalable tree Boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016: 785-794.

[27] Xing HJ, Liu WT. Robust AdaBoost based ensemble of one-class support vector machines[J]. *Inf Fusion*, 2020, **55**: 45-58.

[28] Prokhorenkova L, Gusev G, Vorobev A, *et al*. CatBoost: unbiased Boosting with categorical features[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2018: 6639-6649.

[29] Connelly L. Logistic regression[J]. *Med Surg Nurs*, 2020, **29**(5): 353-354.

[30] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. *Mach Learn*, 2006, **63**(1): 3-42.

[31] Moon T, Chi MH, Kim DH, *et al*. Quantitative structure-activity relationships (QSAR) study of flavonoid derivatives for inhibition of cytochrome P450 1A2[J]. *Quant Struct Act Relatio*, 2000, **19**(3): 257-263.

[32] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation[J]. *arXiv*, 2020:010.16061.

[33] Vergara JR, Estévez PA. A review of feature selection methods based on mutual information[J]. *Neural Comput Applic*, 2014, **24**(1): 175-186.

[34] Bachman P, Hjelm RD, Buchwalter W. Learning representations by maximizing mutual information across views[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, New York: Curran Associates Inc, 2019:15535-15545.

[35] Kwak N, Choi CH. Input feature selection by mutual information based on Parzen window[J]. *IEEE Trans Pattern Anal Mach Intell*, 2002, **24**(12): 1667-1671.

[36] Cai CP, Guo PF, Zhou YD, *et al*. Deep learning-based prediction of drug-induced cardiotoxicity[J]. *J Chem Inf Model*, 2019, **59**(3): 1073-1084.

[37] Xing GM, Liang L, Deng CL, *et al*. Activity prediction of small molecule inhibitors for antirheumatoid arthritis targets based on artificial intelligence[J]. *ACS Comb Sci*, 2020, **22**(12): 873-886.

[38] Su BH, Tu YS, Lin C, *et al*. Rule-based prediction models of cytochrome P450 inhibition[J]. *J Chem Inf Model*, 2015, **55**(7): 1426-1434.

[39] Sun HM, Veith H, Xia MH, *et al*. Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data[J]. *J Chem Inf Model*, 2011, **51**(10): 2474-2481.

[**专家介绍**] 陈亚东,教授,医药大数据与人工智能、药物化学专业博士生导师。江苏省“青蓝工程”优秀青年骨干教师,江苏省“青蓝工程”中青年学术带头人,美国密歇根大学医学院综合癌症中心访问学者。主要研究方向:基于人工智能的药物分子设计及其应用研究,基于重大疾病的原创小分子药物发现研究。主持和参与了多项国家自然科学基金、国家重大科技专项“重大新药创制”等科研项目;课题组与国内多家药企合作进行新药研发。获国内专利 14 项,PCT 专利 3 项;发表 SCI 论文 100 多篇。2015 年研究团队发现的 1.1 类抗肿瘤新药以 1.5 亿元人民币里程金转让给上海复星医药,目前在美国、澳大利亚和中国大陆地区进行 I 期临床试验,2019 年底获美国 FDA 授予孤儿药资格认定。