

基于低维图表示学习的药物重定位预测方法

张 笛¹
ZHANG Di

摘 要

药物重定位作为提高药物研发效率的重要途径,已成为计算药物学研究的热点问题。传统方法在捕捉药物与靶点间复杂非线性关系方面存在一定局限性。为此,文章提出一种基于低维图表示学习的药物重定位预测方法。该方法利用图神经网络提取药物-靶点交互网络中的结构信息,并引入变分图自编码器和生成对抗网络联合学习低维潜在特征表示,从而增强特征表达能力与判别能力。随后,结合梯度提升决策树构建高效的药物-靶点相互作用预测模型。实验在多个公开数据集上进行,结果表明,所提方法在 ROC 和 PR 值等指标上均优于现有主流方法,具有良好的泛化能力与实际应用潜力,为药物重定位提供了一种有效的计算框架。

关键词

药物重定位;图表示学习;生成对抗网络;图神经网络;药物-靶点相互作用预测

doi: 10.3969/j.issn.1672-9528.2025.10.013

0 引言

随着新药研发成本的持续攀升与周期的显著延长,药物重定位(drug repositioning, DR)逐渐成为药物发现领域中的关键策略^[1]。通过挖掘已有药物在新适应症下的潜在治疗作用,药物重定位能够有效降低研发风险与成本,加快药物上市速度。与此同时,依托生物医学大数据与计算方法的发展,计算药物重定位技术在精准医疗和复杂疾病治疗中展现出广阔的应用前景。

在众多计算方法中,基于药物-靶点相互作用(drug-target interaction, DTI)预测的药物重定位手段^[2]受到广泛关注。传统方法如协同过滤^[3]、矩阵分解^[4]或基于结构相似性^[5]的建模方式,虽在一定程度上取得了进展,但通常依赖于稀疏的交互数据或忽视了药物与靶点间潜在的复杂非线性关系。近年来,图神经网络(graph neural networks, GNNs)因其强大的结构信息建模能力,在药物分子建模和 DTI 预测任务中取得了显著成效^[6-7],为提高药物重定位预测的准确性提供了新的方向。

尽管 GNN 能够提取高质量的结构特征,但在实际应用中仍面临特征维度高、冗余信息多以及模型训练不稳定等问题。为此,本文提出一种基于低维图表示学习的药物重定位预测方法。该方法融合变分图自编码器(variational

graph auto-encoder, VGAE)与生成对抗网络(generative adversarial network, GAN),有效压缩高维图结构信息,获取更具判别性的低维潜在表示;同时,引入基于梯度提升决策树(gradient boosted decision tree, GBDT)^[8]的 LightGBM(light gradient boosting machine)作为预测器,以增强模型的非线性拟合能力与泛化表现。

本文的主要贡献如下:

- (1) 提出一种融合图自编码与对抗生成机制的低维特征学习框架,有效提升图表示的紧凑性与表达能力;
- (2) 构建结合表征学习与集成学习的药物重定位预测模型,在多个基准数据集上实现性能优于现有主流方法;
- (3) 通过实证分析验证模型在不同场景下的稳健性和扩展性,展示其在实际药物研发中的应用潜力。

1 方法设计与低维图特征建模

1.1 问题建模与整体框架

药物重定位预测的本质是识别药物与目标蛋白之间可能存在的新的相互作用。本文将 DTI 预测问题建模为一个图表示学习任务,具体而言,在构建的异构生物网络中,药物和靶标被视为图中的节点,其间的已知相互作用构成边。在该图的基础上,通过引入图神经网络(graph neural networks, GNN)与多模态嵌入方法,提取药物与蛋白的低维语义特征表示,并进一步预测潜在的交互关系。

整体框架如图 1 所示,主要包括 4 个阶段:(1) 异构图构建;(2) 图特征提取;(3) 潜在交互预测;(4) 训练优化与策略融合。

1. 郑州财税金融职业学院大数据与人工智能学院
河南郑州 450048

[基金项目] 郑州财税金融职业学院校级课题支持
(ZCY2025004)

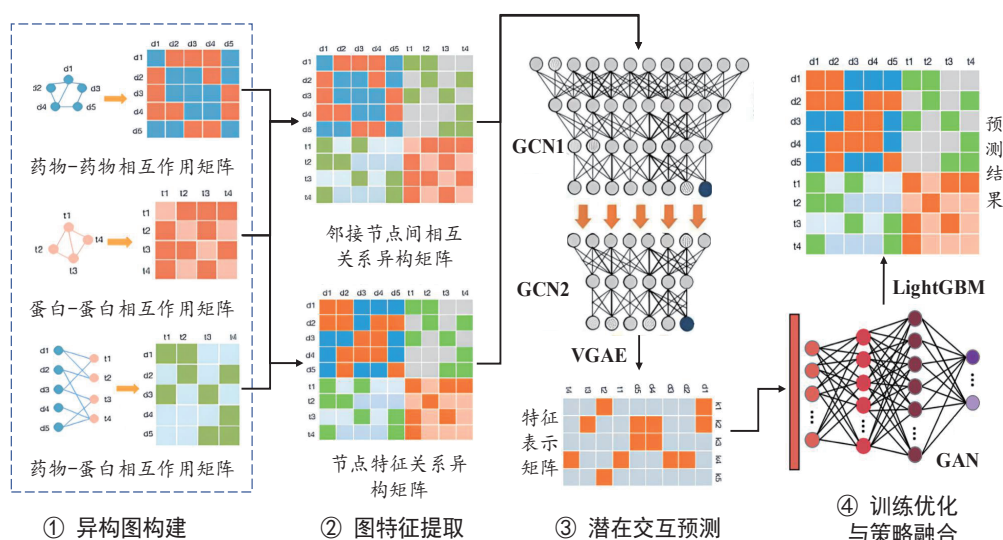


图1 整体技术框架

首先,利用公开数据库(如 DrugBank^[9]、KEGG^[10]等)构建包含药物结构相似性、靶点序列相似性、药物-靶标已知作用等的信息图谱;其次,借助图卷积神经网络(GCN)与变分图自编码器(VGAE)对节点进行低维特征嵌入;然后,基于生成对抗策略(GAN)优化特征分布,提升预测鲁棒性;最后,通过 LightGBM 集成学习方法进行交互评分与分类,得到最终的预测结果。

1.2 异构图构建与数据预处理

在建模过程中,首先需整合多源异构数据构建图结构,包括以下几类边信息:

药物-药物边:依据 Tanimoto 相似性计算 SMILES 结构的距离,连接结构相似度高的药物;

蛋白-蛋白边:通过序列比对算法(如 BLAST)构建蛋白序列相似网络;

药物-蛋白边:来源于公开数据库中已知的 DTI 交互记录;

额外生物边信息:如疾病相关性、路径关联、GO 功能注释等。

所有节点(药物或蛋白)均初始化为高维特征(药物为分子指纹,蛋白为序列特征),再通过归一化和 PCA 降维预处理,为后续图嵌入做准备。

1.3 图神经网络建模与低维嵌入

本研究采用改进的图卷积神经网络(GCN)提取节点的低维结构嵌入。在 GCN 中,节点的表示不仅由自身特征决定,也会受到其邻居节点影响。GCN 的更新规则为:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (1)$$

式中: $\tilde{A}$ 为邻接矩阵加单位矩阵; $\tilde{D}$ 为度矩阵; $H^{(l)}$ 为第 l 层

的节点特征; $W^{(l)}$ 为可训练权重; σ 为激活函数。本文采用两层 GCN, 获得稳定的结构表达。

考虑 GCN 无法充分刻画复杂非线性分布,进一步引入变分图自编码器(VGAE)对节点表示进行先验重建,并以 KL 散度约束潜在变量分布,从而提升嵌入空间的表达性与判别性。

1.4 生成对抗优化策略

为了避免图嵌入过拟合于训练集的已知交互分布,

本文引入生成对抗网络(GAN)作为正则机制,优化嵌入特征的分布。判别器 D 用于判断嵌入是否来自真实交互节点分布 P_{real} , 生成器 G 则引导生成模型生成更逼真的节点分布 P_{noise} 。其目标函数为:

$$\min_G \max_D \mathbb{E}_{x \sim P_{\text{real}}} [\log D(x)] + \mathbb{E}_{z \sim P_{\text{noise}}} [\log (1 - D(G(z)))] \quad (2)$$

生成器部分共享 GCN 的部分参数,使嵌入学习与 GAN 对抗过程协同进行。该策略提升了模型在冷启动条件下对未见药物或蛋白的泛化能力。

2 多源集成预测与分类优化

2.1 相似性融合与特征构建

基于图神经网络得到的节点低维表示后,下一步需对药物-靶点对进行预测。为充分利用不同视角的特征,本文采用以下三类信息构建预测样本:

(1) 图嵌入特征:来自 GCN/VGAE 的药物、蛋白向量,采用拼接、内积、差值等方式组合。

(2) 结构与序列相似性:引入结构/序列之间的余弦相似度作为补充特征。

(3) 统计交互模式特征:包括药物或蛋白浓度、关联疾病数量、交互多样性指标等。

最终每个药物-蛋白对被编码为一个高维特征向量,用于下游分类模型。

2.2 LightGBM 集成分类器

本文采用 LightGBM 作为最终分类器,它基于梯度提升决策树(GBDT)框架,具有训练速度快、支持稀疏特征和处理大规模数据能力强的优点。相比传统 XGBoost, LightGBM 采用基于直方图的决策树分裂方法,显著减少内存使用和训练时间。通过构建多棵弱分类器(即决策树)逐

步优化损失函数。在第 k 棵树构建时, 训练目标是拟合前 $k-1$ 棵树的预测误差, 也称为残差 r_k 。残差定义为:

$$r_k(i, j) = Y_{ij} - \sum_{m=0}^{k-1} h_m(i, j) \quad (3)$$

式中: $r_k(i, j)$ 表示第 k 棵树中药物 i 与靶标 j 的残差; Y_{ij} 表示标签值, 若存在相互作用则为 1, 反之为 0; $h_m(i, j)$ 表示第 m 棵树对该样本的预测结果。

每一棵新决策树 h_k 的训练目标即为最小化 r_k 所引起的损失。最终模型的预测输出由所有树的加权和决定。

$$\hat{Y}_{ij} = \sum_{k=1}^K h_k(i, j) \quad (4)$$

式中: K 为总树数。

通过这种逐步逼近方式, 模型能捕捉复杂的非线性关系与高阶交互特征。

2.3 融合预测策略与冷启动优化

在药物重定位的真实应用场景中, 冷启动问题是一项核心挑战, 尤其在面对新药物或新靶点时, 由于缺乏历史交互信息, 传统模型难以生成高质量的特征嵌入, 进而影响预测准确性。为缓解这一问题, 本文在模型设计中引入了两类融合策略, 旨在增强系统在冷启动条件下的泛化能力与稳定性。

首先, 基于置信加权的模型集成策略有效整合了多个异构模型的预测能力。具体而言, 本文联合使用 GCN、VGAE 和 LightGBM 三类模型, 对同一对药物-靶点样本分别进行预测, 并根据模型在验证集上的稳定性和预测置信度分配加权系数, 最终预测结果通过加权平均公式计算得出。

$$\hat{Y}_{ij} = \alpha \cdot \hat{Y}_{ij}^{\text{GCN}} + \beta \cdot \hat{Y}_{ij}^{\text{VGAE}} + \gamma \cdot \hat{Y}_{ij}^{\text{LightGBM}} \quad (5)$$

式中: $\alpha + \beta + \gamma = 1$ 依据其在训练过程中的 AUC 稳定性与方差进行自适应分配, 具有较高稳定性或低误差的模型会获得更高权重, 从而提升集成预测的整体鲁棒性。

其次, 针对节点特征稀缺问题, 本文设计了基于知识引导的特征补全机制。当某一药物或靶点节点因无交互记录而无法从图神经网络中获得有效表示时, 模型会首先依据结构相似性(如 Tanimoto 系数)或序列相似性(如 BLAST 比得分)检索其高相似邻居集合 N_i , 再通过加权融合邻居的嵌入表示替代当前节点特征。该机制可分为两种模式:

(1) 平均继承机制: 对所有相似邻居嵌入表示求平均, 适用于高相似度、低方差邻居场景;

(2) 注意力加权融合机制: a_{ij} 表示邻居 j 对目标节点 i 的影响程度, 计算公式为:

$$v'_i = \sum_{v_j \in N_i} a_{ij} v_j \quad (6)$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \quad (7)$$

$$e_{ij} = \text{LeakyReLU}(W[v_i || v_j]) \quad (8)$$

通过该方式, 即使在冷启动情形下, 也能从潜在关联邻居中提取特征, 生成具有判别性的节点表示, 进而维持模型的预测性能。

最后, 考虑到药物推荐在临床应用中对安全性和可解释性的高度要求, 本文还引入了预测置信区间估计机制。在模型完成预测后, 对每个药物-靶点对的预测分值得进行置信区间计算, 用以度量模型预测结果的不确定性。具体计算公式为:

$$\text{CI}_{ij} = \hat{Y}_{ij} \pm Z_{\alpha/2} \cdot \frac{\rho_{ij}}{\sqrt{n}} \quad (9)$$

式中: $\hat{Y}_{ij}$ 为平均预测值; ρ_{ij} 为多折预测结果的标准差; n 为交叉验证的折数; $Z_{\alpha/2}$ 为正态分布下的临界值(通常设为 1.96 对应 95% 置信度)。

置信区间的引入不仅提升了推荐系统的可解释性, 也为临床研究人员在候选关系筛选与实验验证中提供了明确的可信范围和风险判断依据。

3 实验评估

为全面验证本文所提出的基于低维图表示学习的药物重定位预测方法的有效性与鲁棒性, 本文设计了一系列实验, 从整体性能、模型模块贡献等多个维度进行评估, 并与主流的药物-靶点预测方法进行对比分析。

3.1 实验设置

数据集: 实验所用的药物-靶点交互数据来源于公开的 DrugBank 与 KEGG 数据库, 并构建了二元交互矩阵 $Y \in \{0, 1\}^{|D| \times |T|}$, 其中 $|D|$ 和 $|T|$ 分别表示药物与靶向蛋白的数量。还引入了药物分子结构特征(SMILES 序列经 RDKit 转为分子图)、靶点蛋白的一级结构序列、药物和蛋白的功能注释等信息用于图构建与特征初始化。最终构建的异构图中, 药物节点和蛋白节点分别具有 128 维初始特征。

评价指标: 本实验主要采用两种常用的分类性能评估指标: ROC 曲线(receiver operating characteristic curve)和 PR 曲线(precision-recall curve)。

其中, ROC 曲线以假阳性率(FPR)为横轴、真阳性率(TPR)为纵轴, 反映模型在不同分类阈值下的整体判别能力。相比之下, PR 曲线则更关注在正负样本极度不平衡的情境下模型的识别能力, 横轴为召回率(Recall), 纵轴为精确率(Precision), 适用于当前药物-靶点数据中正样本稀缺的情况。

为构造这两条曲线, 首先需将模型对每种药物与靶向蛋白的预测评分按照从高到低排序, 依次设置各评分为分类阈值, 计算不同阈值下的四类统计量: TP(真阳性)、FP(假阳性)、TN(真阴性)和 FN(假阴性)。

据此可计算 TPR、FPR、Recall 和 Precision 等核心指标,进一步绘制出 ROC 与 *PR* 曲线,并计算其下的面积,即 AUROC 和 AUPR 值,以量化整体预测效果。

实验环境与模型设置:所有实验在一台具备 Intel i7-9700K@3.6 GHz CPU、16 GB RAM 及 NVIDIA RTX 2080Ti 显卡的服务器上进行。编码模块使用 PyTorch 和 PyTorch Geometric 实现,LightGBM 使用其官方 Python 接口实现。模型训练参数如下设置:GCN/VGAE 隐藏层维度:128;VGAE 编码器为两层 GCN 结构,解码器采用内积解码;LightGBM 学习率 0.05,最大深度 6,树数量 200;模型训练迭代 100 轮,采用早停策略避免过拟合;所有模型采用五折交叉验证,实验重复 5 次取平均值。

3.2 实验结果与分析

(1) 模型整体性能

在本实验中,采用五折交叉验证的方式进行训练与测试,将每一轮预测结果的得分排序,构造多组 ROC 与 *PR* 曲线,并对评价指标进行平均,提升评估的稳健性。最终得到的曲线如图 2 所示。

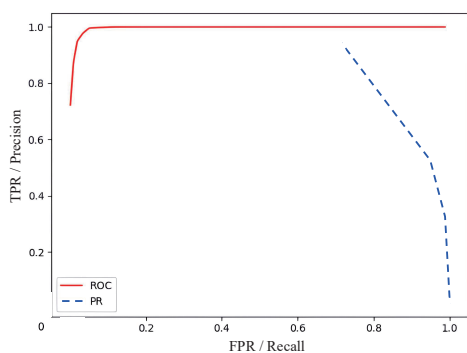


图 2 ROC 曲线与 *PR* 曲线

实验结果显示,本文模型在测试集上取得了 AUROC 为 0.983、AUPR 为 0.829 的优异成绩,表明该模型在判别药物是否与特定靶点相互作用方面具备极强的能力,尤其在正负样本比例极度失衡的情况下依然能保持较高的精确性与召回率。

(2) 先进方法对比

为了进一步验证所提出方法的有效性,本文选取了当前主流的药物-靶点相互作用预测模型进行对比实验,包括基于网络嵌入的 DTINET^[11]、融合多种特征并使用梯度提升树分类的 DTIGBDT^[12]、采用端到端深度交叉注意力机制建模的 MolTrans^[13]、集成图表示学习与协同矩阵因子的 NeoDTI^[14],以及经典深度学习架构构建的 DeepDTI^[15] 等方法。所有模型均在相同的数据集与实验设置下运行,以保证对比的公平性,并通过绘制 ROC 曲线与 *PR* 曲线(如图 3 与图 4 所示)进行性能评估。

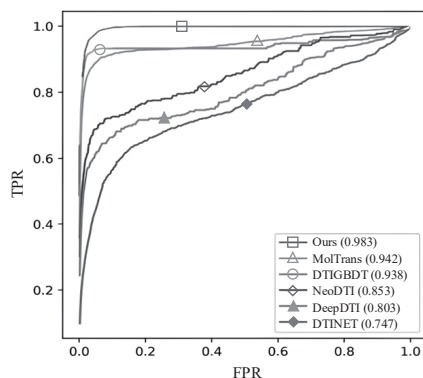


图 3 ROC 曲线对比结果

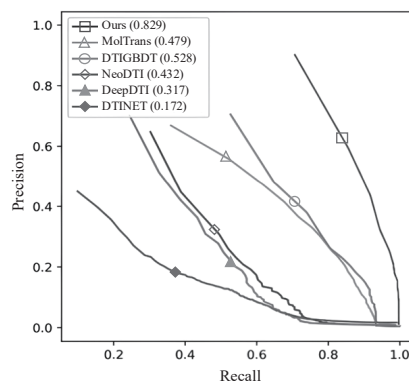


图 4 *PR* 曲线对比结果

从实验结果来看,本文提出的方法在两项关键指标上均显著优于对比方法。具体而言,在 ROC 曲线中,本文模型的 AUROC 值达到 0.983,显著领先于 MolTrans (0.942)、DTIGBDT (0.938)、NeoDTI (0.853)、DeepDTI (0.803) 和 DTINET (0.747)。而在更加关注正负样本不平衡问题的 *PR* 曲线上,本文模型同样表现优异,取得了 AUPR 值为 0.829,远高于 MolTrans (0.479)、DTIGBDT (0.528)、NeoDTI (0.432)、DeepDTI (0.317) 以及 DTINET (0.172)。

通过对比分析可以发现,本文模型性能优越的原因主要有以下两点。首先,引入了生成对抗网络(GAN)对药物与靶点的低维特征表示进行增强,有效缓解了节点特征稀疏和数据分布偏移的问题。其次,采用了 LightGBM 作为最终分类器,该方法具备动态调整样本权重的能力,使得在每轮迭代中能够更有效关注难分类样本,进一步提升了整体模型的鲁棒性与泛化能力。

(3) 消融实验对比

为进一步验证所提出方法中各个关键模块对模型性能的贡献,本文设计了一组消融实验,依次剔除模型中的核心组件,并在相同实验设置下评估性能变化,具体包括五种模型变体。

① Ours (完整模型): 包含 GCN 编码器、VGAE 结构、GAN 正则与 LightGBM 分类器;

② w/o GAN: 去除 GAN 模块,不进行生成对抗式特征

优化;

③ w/o VGAE: 去除变分图自编码器模块, 仅使用基础 GCN 编码器;

④ w/o Ensemble: 取消 GCN、VGAE 与 LightGBM 的集成机制, 仅使用单一预测器;

⑤ w/o LightGBM: 去除最终的 LightGBM 分类器, 仅使用深度模块进行预测。

各模型在同一测试集上分别绘制 ROC 曲线和 PR 曲线, 并计算对应的 AUROC 与 AUPR 值, 实验结果如表 1 所示。

表 1 消融实验对比结果

模型名称	AUROC	AUPR
Ours	0.983	0.829
w/o GAN	0.937	0.652
w/o VGAE	0.928	0.626
w/o Ensemble	0.963	0.721
w/o LightGBM	0.914	0.604

从上述实验结果可以看出, 移除任一模块都会导致模型性能的下降, 说明各组件在整体架构中均发挥着关键作用。具体而言: 去除 GAN 模块后, 模型性能略有下降, 说明 GAN 在优化药物与靶点嵌入的表示空间方面发挥了重要的对抗正则作用, 尤其是在处理特征稀疏或分布偏移问题时表现显著; 去除 VGAE 结构后, 模型的 AUPR 值下降更为明显, 表明 VGAE 对捕捉潜在图结构中的低维拓扑关系至关重要, 能有效增强图神经网络的表达能力; 取消集成机制后, 性能也有显著下降, 验证了融合多模型预测结果所带来的鲁棒性提升; 最终去除 LightGBM 分类器, 模型表现最差, 说明基于 GBDT 思想的分类器能够弥补深度模型对少数类样本识别能力的不足, 在正负样本不均衡的场景下具有有效优势。

4 总结

本文针对药物重定位任务中普遍存在的交互数据稀缺、样本不均衡以及冷启动问题, 提出了一种融合多种图表示学习与集成预测策略的低维特征预测方法。该方法结合了 GCN、VGAE 和 GAN 以增强节点的低维嵌入表示, 并引入 LightGBM 分类器以提升预测的稳定性和泛化能力。同时, 本文设计了基于邻居特征补全的冷启动机制与预测置信区间估计策略, 以增强模型在实际临床应用中的可解释性与可靠性。实验部分在两个公开数据集上进行评估, 结果显示本文方法在 ROC 和 PR 等多个评价指标上均优于现有主流方法, 且通过消融实验进一步验证了各模块的有效性。

参考文献:

[1] 李宗贤. 基于知识图谱的帕金森病药物重定位 [J]. 信息技术与信息化, 2022(7):28-32.

[2] 张然, 王学志, 汪嘉葭, 等. 药物 - 靶点相互作用预测的计算方法综述 [J]. 计算机工程与应用, 2023, 59(12):1-13.

[3] 陈浩, 秦玉芳, 陈明. 基于图神经网络协同过滤的药物疾病关联预测 [J]. 中国医学物理学杂志, 2023, 40(6): 780-787.

[4] 杨羽彤, 易琦, 孙正宝, 等. 矩阵分解方法应用研究进展 [J]. 云南大学学报 (自然科学版), 2025, 47(4):635-647.

[5] 丁琼, 于兰, 罗林, 等. 基于促黑色素合成分子的结构相似性筛选潜在白癜风的治疗药物 [J]. 石河子大学学报 (自然科学版), 2021, 39(6):762-770.

[6] 郭文博, 龙伟, 蒋林华, 等. 基于图神经网络的药物相互作用预测方法研究综述 [J]. 现代计算机, 2024, 30(20):36-40.

[7] 颜淑婷, 陈佳彤, 周小露, 等. 基于图神经网络的苦味分子预测方法研究 [J]. 电脑知识与技术, 2024, 20(25):5-7.

[8] 龚安, 耿航. 结合 GBDT 特征衍生与集成学习的客户忠诚度预测 [J]. 计算机与数字工程, 2024, 52(12):3610-3615.

[9] 赵仲麟, 吕东灿, 袁超, 等. 化学生物学课程中引入药物分析信息数据库 DrugBank 内容的探索 [J]. 农业网络信息, 2018(4):82-85.

[10] KANEHISA M, GOTO S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1):27-30.

[11] LUO Y N, ZHAO X B, ZHOU J T, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information[J]. Nature communications, 2017, 8(1):573.

[12] XUAN P, SUN C, ZHANG T G, et al. Gradient boosting decision tree-based method for predicting interactions between target genes and drugs[EB/OL](2019-05-31)[2025-06-19]. <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00459/full>.

[13] HUANG K X, XIAO C, SUN J M, et al. MolTrans: molecular interaction transformer for drug-target interaction prediction[J]. Bioinformatics, 2021, 37(6):830-836.

[14] WAN F P, HONG L X, XIAO A, et al. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions[J]. Bioinformatics, 2019, 35(1):104-111.

[15] TIAN Q Y, BILGIC B, FAN Q Y, et al. DeepDTI: high-fidelity six-direction diffusion tensor imaging using deep learning[J]. Neuroimage, 2020, 219:117017.

【作者简介】

张笛 (1995—), 女, 河南郑州人, 硕士, 助教, 研究方向: 人工智能、深度学习, email:se_diz@126.com。

(收稿日期: 2025-04-18 修回日期: 2025-10-09)