DOI:10.12113/202307005

基于表示学习的图神经网络模型预测化合物-蛋白质相互作用

章广能[1],张育芳[2],张 宝[1*]
(1.南方医科大学 公共卫生学院,广州 511495;2.上海交通大学 数学科学学院,上海 200240)

摘 要:化合物-蛋白质互作的鉴定对药物发现、靶标鉴定,网络药理学和蛋白质功能的阐明等至关重要。本文开发了一种基于表示学习的图神经网络预测化合物-蛋白质互作模型。首先利用 Word2vec 表示学习方法自动提取化合物和蛋白质的特征;然后将特征输入构建图神经网络预测模型,并与传统机器学习方法和前人的先进方法对比。结果显示模型在曲线下面积,准确率等评价指标上表现出更好的结果。预测 Binding-DB 数据库中所有未知的化合物-蛋白质互作对的概率,其中预测得分排名前五的化合物-蛋白质互作对中有四个得到了外部证据的验证,进一步证明了模型的鲁棒性和有效性。本模型可以充分利用聚合邻居信息,节点特征和自适应地捕获化合物-蛋白质空间的拓扑结构,从而实现较高的模型精度。本研究成果为化合物和蛋白质互作鉴定的研究提供了新的思路和方法。

关键词:化合物-蛋白质相互作用;表示学习;图神经网络;药物发现
中图分类号:R96;TP18; **文献标志码**:A **文章编号**:1672-5565(2024)04-287-09

A graph convolutional network model based on representing learning for compound-protein interaction prediction

ZHANG Guangneng[1], ZHANG Yufang[2], ZHANG Bao[1*]
(1.*School of Public Health*, *Southern Medical University*, *Guangzhou* 511495, *China*;
2.*School of Mathematical Science*, *Shanghai Jiaotong University*, *Shanghai* 200240, *China*)

Abstract: The identification of compound-protein interactions is crucial for drug discovery, target identification, network pharmacology, and elucidation of protein function. In this paper, we develop a representation learning based graph neural network model for predicting compound-protein interactions. Firstly, Word2vec representation learning method is used to extract features of compounds and proteins automatically. Then the features are input to construct a graph neural network prediction model. Compared with traditional machine learning methods and previous advanced methods, this model shows better results in AUC, accuracy and other model evaluation indicators. Predict the probability of all unknown compound-protein interactions in the Binding-DB database, with four of the top five compound-protein interactions with the highest prediction score confirmed by external evidence. The robustness and effectiveness of the model are further proved. This model can fully utilize aggregated neighbor information, node features, and adaptively capture the topological structure of the compound protein space, thereby achieving high model accuracy. The results of this study provide a new idea and method for the study of compound-protein interaction identification.

Keywords: Compound-protein interactions; Representation learning; Graph neural network; Drug discovery

收稿日期:2023-07-21;**修回日期**:2023-09-19;**网络首发日期**:2023-10-11.
网络首发地址:https://link.cnki.net/urlid/23.1513.Q.20231009.1604.003
* **通信作者**:张宝,男,博士,教授,研究方向:病毒生物学,肿瘤发生机制.E-mail: zbggws2023@126.com.
引用格式:章广能,张育芳,张宝.基于表示学习的图神经网络模型预测化合物-蛋白质相互作用[J].生物信息学,2024,22(4):287-295.
ZHANG Guangneng,ZHANG Yufang,ZHANG Bao.A graph convolutional network model based on representing learning for compound-protein interaction prediction[J].Chinese Journal of Bioinformatics,2024,22(4):287-295.

化合物-蛋白质相互作用(Compound-protein interactions, CPIs)的鉴定对于新药研发过程中的靶点发现,网络药理学探索,蛋白质功能识别,药物重定位等至关重要[1-2]。现有的化合物-蛋白质相互作用预测方法主要分为三类:基于生物化学的湿实验(Wet tests)方法,基于结构的计算方法,基于机器学习(Machine learning, ML)和深度学习(Deep learning,DL)的计算方法。①湿实验如动力学实验,光谱分析,细胞热转变分析,电化学方法等[3-7]需要大量的人力,物力和经费的投入,是非常昂贵和耗时的过程;②基于结构的计算方法主要是通过分子对接[8-11]的方法,筛选能与靶蛋白有效结合的药物小分子化合物,但由于已知的配体数量不足、蛋白质的三维结构未知等原因,该方法的应用受到限制;③基于机器学习和深度学习的计算方法因其大规模测试能力和可靠的预测结果而最受关注[12-14]。

目前,研究人员已经开发了多种机器学习方法如朴素贝叶斯分类器(NBC)[15]、k-最邻近算法(KNN)[16]、随机森林(RF)[17]、支持向量机(SVM)[18]、梯度提升决策树(GBDT)[19]等,并将这些算法模型应用于 CPI 预测。Nanni 等[20]采用分子指纹和蛋白质描述符作为特征向量,输入到支持向量机模型中预测 CPIs。He 等[21]用 28 个常见的官能团表示化合物,用特别设计的伪氨基酸组成结合氨基酸组成表示蛋白质序列,再将通过使用最大相关最小冗余(mRMR)方法得到的最佳特征输入 K 最近邻模型(KNN)中预测药物-靶点相互作用关系。Farshid 等[22]利用靶点的结构信息和进化信息来生成包括二级结构图案,扭转角和结构概率,可访问表面积等在内的蛋白质特征,将这些特征和化合物分子指纹(811 位)组合后输入到其提出的iDTI-ESBoost模型中识别药物-靶点相互作用关系。

近几年来,作为机器学习中的一个新的研究范式,以深度神经网络[23]、卷积神经网络[24]、循环神经网络[25]等为代表的深度学习方法逐渐兴起,其因在语音识别、图像识别及自然语言处理等许多领域的出色表现而变得越来越受欢迎,将深度学习方法应用于药物发现的研究一直在不断增加,开发出了多种用于 CPI 预测的深度学习算法。Tian 等[26]提出利用深度神经网络(DNN)预测化合物-蛋白质互作对,蛋白质靶点特征使用氨基酸组成,伪位置特异性评分矩阵,联合三联组成,过渡和分布,自相关系数和结构特征等,药物化合物使用结构指纹进行编码。Xie 等[27]利用来自 LINCS 项目 L1000 数据库的转录组数据开发了一个基于深度学习算法的框架来预测潜在的药物靶标相互作用。上述方法[21-22,26-27]应用于 CPIs 的预测具有一定效果,但缺点在于特征提取需要大量人工的参与,因此当数据量很大时,这些方法就存在明显的局限性,不仅需要许多专家知识,而且涉及复杂的数据处理和算法优化过程,限制了 ML 和 DL 算法在 CPI 领域的进一步开发与应用。

近年来图神经网络(Graph neural network, GNN)尤其是图卷积网络(Graph convolutional network, GCN)的快速发展将深度学习的应用扩展到图领域,相关方法[28-30]也被应用到药物发现中。图(Graph)是一种数据格式,图中的节点表示图网络中的实体,边表示实体之间的连接关系。同质图(Homogeneous graphs)只有一种类型的节点和边,如蛋白互作[28]、药物间互作[29]、基因互作[30]等。相对应的异质图(Heterogeneous graphs, HetG)包括多种类型的图结构,如 CPI[31],RNA-疾病关联[32],生物知识图谱[33]等。Nguyen 等[34]提出使用图神经网络预测药物-靶标亲和力,预测性能较优。Tsubaki 等[35]将卷积神经网络(CNN)和图卷积网络(Graph convolutional network, GCN)结合起来,即使用 GCN 处理化合物分子图,CNN 处理蛋白质序列,从而进行端对端(End-to-End)的化合物-蛋白质相互作用预测,取得了较优结果,但预测性能和稳定性仍有进一步提升的空间。

GCN 应用节点之间的连接作为卷积滤波器来执行邻域聚合,在处理以 CPI 为代表的生物图(Graph)结构的数据时展现出优势。但 GCN 本身也存在一些缺陷,最为典型的是其在处理复杂大图时遇到的“邻居爆炸”现象(节点表示和随机梯度的计算复杂度会随着消息传递层数的增加而呈指数级上升),以及多层 GCN 堆叠所致的过平滑或过拟合问题(当神经网络更深入时,节点在聚合操作之后往往具有相似表示)。国内外学者提出了多种图采样技术以减少消息传递所牵涉的节点个数来降低训练成本,最常见的包括节点采样(如 GraphSAGE[36], PinSage[37], VRGCN[38])和层采样(如 FastGCN[39], ASGCN[40])等。

为了提升 CPI 预测的特征提取效率、泛化能力及稳定性,本文提出基于 Word2vec[41]表示学习和基于 GCN 的端对端化合物-蛋白质相互作用预测方法,利用节点特征和拓扑网络来构造图,并基于子图采样策略来实现,具体而言:Word2vec-GCN 基于图上的边(Edge)自适应地采样子图,并在每个子图上运行模型,这不仅可以避免邻居爆炸现象,还可以降低模型的复杂性和内存消耗。首先利用 Word2vec

表示学习方法学习到化合物与蛋白质的特征向量表示,再输入到图神经网络(GCN)中以预测化合物-蛋白质的相互作用,最后将预测结果分别进行交叉验证、独立测试及对比验证。在 Binding-DB 大规模数据集上运行的结果表明,本文模型相对于传统 ML、DL 方法和他人提出的先进方法,分类效果更优。

1　数据与处理

1.1　数据库

本论文主要选用结合亲和力数据库(The binding database,Binding-DB[42])来进行数据集的构建。Binding-DB数据库是加州大学实验室开发的一个免费的结合亲和力公共数据库,主要关注小分子类药化合物和蛋白质靶点的相互作用亲和力,同时,提供通路信息,化合物 ZINC 编号以及其他信息,研究者通过网络查询数据库可以获取化合物分子和蛋白质靶点之间的结合数据,可用于药物研发和结合预测模型的构建。Binding-DB数据库包含由 120 万种化合物和9 000种蛋白质靶标组成的 270 万个结合数据(Binding data),其中 58 万个化合物和4 500个蛋白质组成的 126 万个结合数据被人工标注。Binding-DB的数据来源包括 PDB,PubMed 相关文献报道数据,专利信息,PubChem BioAssay 数据库和 Chembl 记录数据,数据库定期进行更新并发布版本。

1.2　数据预处理

为了保证数据质量,本研究做如下处理:

1)IC50 值可以表明抑制剂与底物结合能力,考虑到成药性,去除 IC50 值大于 300 nM 的 CPI 对,同时去除 IC50 值缺失的 CPI 对。

2)考虑到小分子靶向药物主要是指分子量<1 000的有机化合物,因此去除数据库中的无机物和分子量>1 000 的类药分子。

3)为了降低蛋白质冗余度,解决“数据泄露”问题,将数据库中蛋白序列一致性>75%的蛋白质靶点进行过滤。

经过上述处理后的数据集包括了36 014个小分子化合物和 2 099 个蛋白质靶点,并可能产生超过 7 500万个 CPI 对。其中 83 676 条被确认的 CPI 对被称为正样本,其余尚未确认的作为未标记数据处理。按照 1∶1 的数量配比原则,从未标记数据集中随机挑选相同数量的 83 676 条样本,定义为负样本。随机挑选正负样本数据集中的 CPI 数据的 80%用作训练集,剩余 20%用作测试集。

2　方法

定义化合物分子集合 $=[M_1,M_2,\cdots,M_n]$、氨基酸序列集合 $=[P_1,P_2,\cdots,P_m]$,二者互作的标签集合 $=\{0,1\}$,其中,1 表示互作的类别标签,0 表示不互作的类别标签。化合物-蛋白质对的互作预测任务是一个二分类学习模型 $S:(M,P)\rightarrow L$,模型输入化合物分子 M 和氨基酸序列 P,输出类别标签 L。

本文方法的整体流程如图 1 所示,主要过程如下:

1)基于 Word2vec 表示学习方法实现对化合物分子和蛋白质序列的特征向量化的表示。

2)将特征输入到图神经网络中,反复调优测试,得到分类预测结果。

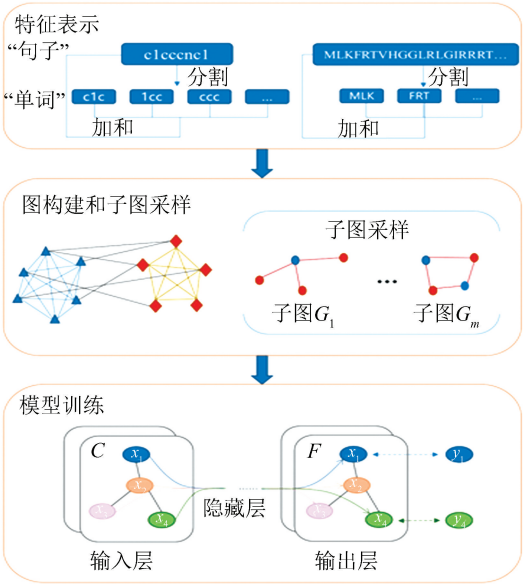

图 1　本文方法流程图

Fig.1　Workflow chart of the proposed method

2.1　Word2vec 表示学习

Word2vec 是一款基于统计语言模型构建的利用深度学习来训练生成词向量的工具。Word2vec 的主要原理是通过预测上下文或目标单词的方法学习单词的向量表示,具体为通过一个简单的神经网络来学习单词的向量表示,该神经网络包括一个输入层(One-hot 热编码向量),一个隐藏层(单词嵌入向量)和一个输出层(Softmax 激活函数)。Word2vec 有两种实现形式——连续词袋模型(Continuous bag-of-Words, CBOW)和跳字模型(Skip-gram)。本论文采用 Skip-gram 模型进行化合

物药物和蛋白质靶标的向量计算。

神经网络的训练过程为:输入层(1×V 维)→输入矩阵 * 权重矩阵 W(V×N)= 隐藏层 H(1×N)→隐藏层 * W′(N×V)→输出层(C * 1×V)→softmax 输出概率→反向传播训练→最优化权重参数矩阵 W→得到词的词向量,完成词嵌入。

为了表示药物分子,将 SMILES 视为“句子”,将每三个不重叠的字符(代表原子、化合键、分支等)视为“单词”。对于蛋白质序列,我们将它们视为“句子”,每三个不重叠的氨基酸视为一个“单词”。

本研究 Word2vec 的执行主要是利用 Python 工具包 Gensim 完成。选用的方法为 Skip-Gram,设置词向量的维度 = 100,窗口大小 = 3,负采样(Negative sampling)的数量 = 12,其余使用默认参数。

2.2 图卷积神经网络(GCN)

定义图结构 $G = (V,E)$:V 表示节点的集合,$V =$ {化合物分子集合,氨基酸序列集合} $= [M_1, M_2, \cdots, M_n, P_1, P_2, \cdots, P_n]$,$E$ 表示边的集合,$e_{ij} = (e_i, e_j) \in E$ 表明化合物与蛋白质存在相互作用。A 代表图 G 的邻接矩阵,$A_{ij} = 1$ 代表化合物节点 M_i 和蛋白质节点 P_j 之间存在边。节点特征由 Word2vec 方法得到。

图卷积的核心思想是利用边的信息对节点信息进行聚合,从而生成新的节点表示。计算公式如下:

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

上述公式中:$\widetilde{A} = A + I_N$ 为添加了自连接的邻接矩阵,I_N 是单位矩阵;$\widetilde{D}$ 是 $\widetilde{A}$ 的度矩阵,$\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$;$W^{(l)}$ 是 l^{th}层特定的可训练权重矩阵;$\sigma(\cdot)$ 表示激活函数,如 $ReLU(\cdot) = max(0, \cdot)$,实际为一个关系调用函数,用来判断输入参数对应下一层的哪一个节点;$H^{(l)} \in \mathbb{R}^{N \times D}$ 是第 l^{th}层的输入激活矩阵,初始输入 $H^{(0)} = X$。

GCN 训练流程为:

1)输入图结构信息,经过 GCN 前向传播最终输出每个节点的特征,输出结果与期望值之间产生误差。

2)进行反向传播,计算有标签的节点的训练损失,对误差运行梯度下降算法来动态调整权重参数直至最优,实现监督分类。

3)将反向传播后得到的全局权重矩阵代入 GCN 前向传播公式,计算得到与期望值最接近(即最精确)的预测值输出结果,整个模型达到最优。

上述 GCN 训练中使用了子图采样方法来代替传统 GCN 算法中通常使用的微批量(Mini-batch)构造。该方法主要思想是从训练图中抽取几个子图,并在每个子图上建立一个完整的 GCN。这种方法使样本中的节点能够提供信息并相互帮助,而不会在 GCN 层中传播期间涉及批外的节点,使用该算法可以避免传统 GCN 算法中经常出现的“邻居爆炸”问题。为了确保训练过程的准确性,该采样方法使用了一种基于拓扑结构的边缘采样器,采样器以很大的概率对边缘进行采样,通过考虑同一子图中的节点相互影响的程度来计算边缘概率。为了避免偏差,我们在聚合节点信息和计算微批量损失的过程中对边缘采样概率进行了归一化,来减少图神经网络中的偏差;通过使全图卷积网络中节点聚合的方差最小化定义了最佳边缘采样概率。本研究 GCN 模型基于 Pytorch 框架构建完成,相关的算法和数据已在 github 上公开(网址 https://github.com/zgn002/Word2vec-GCN)。

对于均衡数据集,选取曲线下面积(Area under curve,AUC)、准确率、精密度,召回率和 F1 分数作为模型的评价指标。本研究进行了 10 次 5 折交叉验证,其模型表现结果均值作为最终算法准确性的估计。

3 结果分析

3.1 Word2vec-GCN 在 Binding DB 数据集上的分类性能

图 2 展示 Word2vec-GCN 模型在 Binding-DB 数据集上的分类性能总结。结果表明,本模型在 CPI 预测任务中训练集 AUC、准确性,精确度和召回率以及 F1 分数分别达到0.998 7,0.995,0.996 2,0.997 和 0.995 2。较高的预测准确度表明,本模型不仅实现了 CPI 异质性特征的有效提取与融合,还可以更好地聚合邻居信息、节点特征和自适应地捕获化合物-蛋白质空间的拓扑结构。

同时,模型在测试集上 AUC、准确性、精确度和召回率以及 F1 分数分别达到0.995 7,0.991 3,0.990 8,0.994 和0.993 2。表明了模型具有较高的鲁棒性和可扩展性。模型在测试集上准确性降低可能有以下两点原因:①由于负样本是随机选择未标记的数据而构建,出现假阴性的化合物-蛋白质互作对的概率在较小的数据集中会更高;②测试集中化合物-蛋白质互作对的分布是非常不均衡的,与训练集的分布有所偏离,这也会影响模型的准确性。

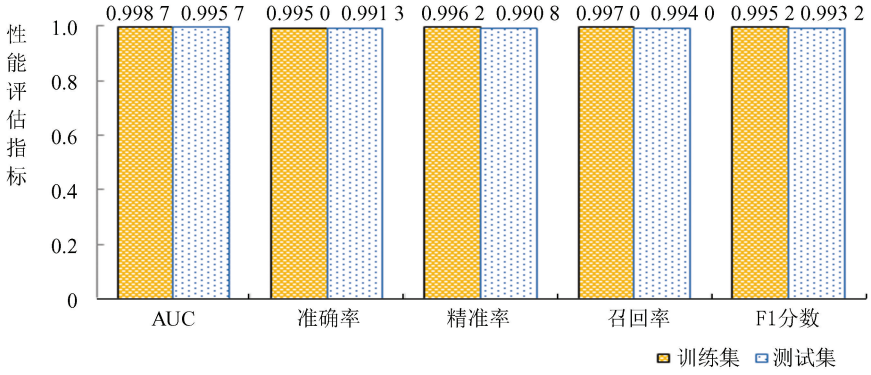

图 2　Word2vec-GCN 在 Binding-DB 数据库训练集和测试集上的分类性能

Fig. 2　Results of classification performance of our methods on Binding-DB training sets and test sets

3.2　Word2vec-GCN 和传统 ML 和 DL 方法模型表现对比

为了进一步说明 Word2vec-GCN 在化合物-蛋白质互作预测中的优越性,还将其与一些经典的机器学习和深度学习算法进行了比较,包括随机森林(RF)、支持向量机(SVM)、深度神经网络(DNN)、梯度提升决策树(GBDT)。上述模型的预测结果如表 1 所示。根据结果,本文模型优于其他经典的 ML 和 DL 算法,本文方法的 AUC 分别比 SVM,GBDT,RF 和 DNN 高 46.75%,2.59%,2.30%和 5.28%。

基于 BindingDB 测试集,树模型方法(即 GBDT 和 RF 模型)对 CPI 的分类任务没有产生重要差异,并且所有方法都获得了同样高的 AUC 分数、准确度、精确性,召回率和 F1 分数。作为深度学习方法,GCN 和 DNN 模型在网络结构和参数数量上较为复杂。与 DNN 模型相比,GCN 更善于处理以 CPI 为例的图结构数据。虽然在上述评估方法中,基于树的方法如 GBDT 和 RF 都优于 DNN 架构,但由于 DNN 模型参数复杂,进一步微调(Fine-tuning)可能会提高 DNN 模型的预测性能。而 SVM 方法在 CPIs 预测任务中取得了五个模型中最差的模型性能,原因可能在于 SVM 通常在线性数据集上表现良好。尽管许多核函数,如高斯径向基(Gaussian RBF)核函数,也可以处理非线性情况,但它们往往需要大量的相似性特征和大量的计算资源,而在最坏的情况下,一个完全不可分割的数据,如环状数据,使用 SVM 往往表现不佳。

表 1　Word2vec-GCN 在与四个经典 ML/DL 分类器的模型表现结果比较

Table 1　Results of classification performance of our method compared with four classic ML and DL methods on BindingDB testsets

模型	AUC	准确率	精确率	召回率	F1 分数
Word2vec-GCN	0.995 7	0.991 3	0.990 8	0.994 0	0.993 2
SVM	0.678 5	0.671 4	0.660 6	0.667 5	0.670 1
GBDT	0.972 3	0.973 0	0.965 9	0.976 7	0.970 1
RF	0.972 7	0.976 5	0.970 8	0.974 0	0.969 2
DNN	0.945 8	0.946 2	0.947 5	0.938 3	0.939 7

3.3　Word2vec-GCN 与他人先进方法的模型表现对比

为了进一步证明本文所提出方法的优越性,Word2vec-GCN 方法与以下五种他人先进方法进行了比较。

1)Tsubaki 等[35]开发了一种新的 CPI 预测方法 GNN-CNN,该方法通过结合用于化合物结构提取的 GCN 和用于蛋白质序列表示的 CNN,进行端到端表示学习。

2)Tian 等[26]提出了一种 DL-CPI 方法,该方法使用增强深度神经网络来有效地学习化合物-蛋白质对的表示。DL-CPI 可以通过分层提取来学习化合物-蛋白对的有用特征,从而实现良好的预测性能。

3)Ye 等[43]提出了一种基于多元分类策略的药物-靶标相互作用预测方法—MCSDTI,首先根据相互作用的数量将靶点分为两组,然后设计并采用不同的分类策略来准确预测每组药物-靶点相互作用。

4) Rayhan 等[44]提出的 CFSBoost 是一种简单且计算成本低廉的集成增压分类模型,使用进化和结构特征识别和预测药物-靶标相互作用。

5) Zhang 等[45]同样基于 Word2vec 方法提出了 SPVec 模型学习药物-靶点相互作用对的特征表示,并将特征向量输入 GBDT 模型中在 DrugBank[46]独立测试集中预测二者相互作用。

本方法与上述 5 种先进方法的模型表现结果比较如表 2 所示。本方法在所有模型性能评估指标上都优于上述五种方法,AUC 评分、准确性、精确度,召回率和 F1 评分分别比次优方法高 2.63%,2.06%,2.44%,1.93%和 2.51%。在使用相同 Word2vec 特征提取方法前提下,GCN 模型结果优于 SPVec-GBDT 模型,说明了 GCN 在表示化合物、蛋白质等生物分子结构以及分子之间的功能关系和集成异质数据方面的优秀性能。Word2vec-GCN 方法不仅可以自适应地聚合邻居信息与节点特征,还可以利用化合物-蛋白质空间的拓扑结构。与同为端对端表示学习的 GNN-CNN 方法相比,Word2vec 比 CNN 更加适用于蛋白质序列的特征提取。这可能是由于 CNN 固定大小的卷积核只能捕获蛋白质序列滑动窗口内的子序列,而所有子序列的依赖关系难以获取。与 CFSBoosting 和 MCSDTI 方法相比,本方法不仅取得了更优的模型表现,而且省去了手工构建特征的繁琐过程。总之,表 2 体现了本方法的有效性和高效性。

表 2 Word2vec-GCN 方法与 5 种先进方法的模型表现结果比较

Table 2 Results of classification performance of our method compared with five state-of-art models

模型	AUC	准确率	精确率	召回率	F1 分数
SubGCN-CPI	0.995 7	0.991 3	0.990 8	0.994 0	0.993 2
SPVec-GBDT	0.970 2	0.971 3	0.967 2	0.974 7	0.968 9
DL-CPI	0.909 9	0.902 1	0.893 7	0.916 4	0.912 3
GNN-CNN	0.933 2	0.930 1	0.929 6	0.923 7	0.925 1
CFSBoosting	0.803 4	0.826 7	0.806 5	0.819 0	0.824 0
MCSDTI	0.861 7	0.833 2	0.828 7	0.824 8	0.819 7

3.4 Word2vec-GCN 预测潜在的化合物-蛋白质相互作用

从前几节讨论可以得出,本文提出的基于表示学习的图神经网络 Word2vec-GCN 方法在化合物-蛋白质互作预测任务中展现出了优秀的模型性能和良好的泛化能力。在真实世界中,准确的预测出潜在化合物-蛋白质互作对于促进药物重定位,提高靶点识别效率,加速药物研发过程意义重大。本研究使用 Word2vec-GCN 方法对 Binding-DB 数据集中的所有未标记 CPIs 存在相互作用的概率进行了预测,预测得分越高,说明互作可能性越大。其得分前五名的化合物-蛋白质对如表 3 所示。通过检索多个生物信息学数据库(如 PubChem, ChEMBL, KEGG 等)和相关文献,最终有 4 个化合物-蛋白质对得到外部证据的证实。为了直观表现出化合物-蛋白质对相互作用,基于结构的分子对接图如图 3 所示。由于 P33402 蛋白数据库中未收录三级结构,故未在图 3 中进行展示。排名第三的化合物-蛋白质作用对虽然未在外部数据库中找到实验数据,分子对接结果表明了二者的相互作用。进一步证明 Word2vec-GCN 方法预测出的五个药物-靶标对的有效性。

表 3 Word2vec-GCN 预测排名前五的新化合物-蛋白质对

Table 3 Top five novel CPIs predicted by Word2vec-GCN

序号	化合物 CID	UniProt ID	靶标名	验证来源
1	17600651	Q13115	双特异性磷酸酶	PubChem
2	5353788	P33402	鸟苷酸环化酶可溶性亚基 α-2	PubChem
3	59033885	Q9BY41	组蛋白脱乙酰酶	分子对接
4	293961	P03366	HIV-2 蛋白酶	KEGG
5	5329721	P14635	G2/有丝分裂特异性细胞周期蛋白 B1	UniProt

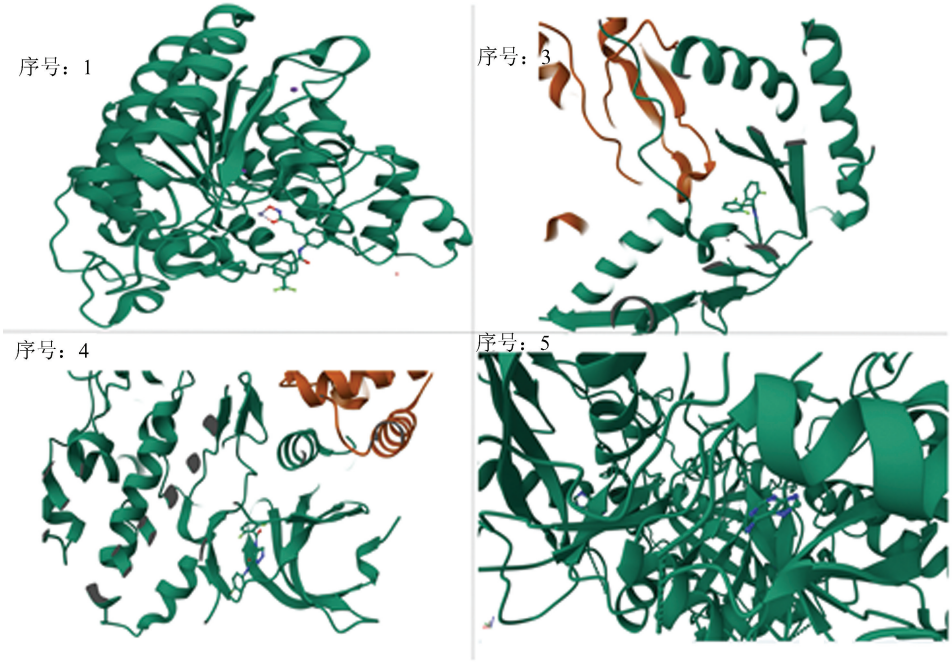

图 3　Word2Vec-GCN 预测排名 1,3,4,5 的新化合物-蛋白质对分子对接结果

Fig. 3　Molecular docking results of rank 1,3,4,5 novel DTIs predicted by Word2Vec-GCN

4　讨　论

Word2vec-GCN 方法优势可分为三个方面:①无需专家知识,表示学习自动表征化合物和蛋白质连续、稠密、低维特征,有效避免维数灾难、信息有限等传统特征提取方法的局限性;②GCN 模型可以充分利用聚合邻居信息、节点特征和自适应地捕获化合物-蛋白质空间的拓扑结构,从而实现较高的模型精度,同时兼具鲁棒性和可扩展性;③该方法是端对端(End-to-End)学习,即直接从原始数据得到预测结果,极大节省人力、时间和成本。

本文研究内容不足之处在于:①Word2vec 表示学习方法本质上属于深度学习方法,深度学习方法的"黑箱模型"使得计算得出的特征向量每一维度代表的含义难以解释,目前暂时缺少适用于像 CPI 预测等生物问题的可解释性模型方法;②相对于 ML 方法,GCN 方法的参数众多,调参工作量大,后续研究可以尝试引入自动化调参工具,进一步提高 Word2vec-GCN 方法的预测效率;③仅采用计算方法预测,对于化合物-蛋白质互作发生机制的研究不够深入全面,缺少体内或体外实验支撑。在后续研究中,需要与生物实验结合,提高模型的实用性与准确性。

5　结　论

1)模型在训练集和测试集上均取得较高的模型表现,表明了模型具有优秀的预测性能、鲁棒性和可扩展性。

2)与经典的 ML 和 DL 方法(即 SVM,GBDT,RF 和 DNN)和最先进的 MDA 预测方法(即 DL-CPI,GNN-CNN,CFSBoosting 和 MCSDTI)相比,本研究所提出的 Word2vec-GCN 模型可以实现更好的预测性能。

3)本模型重新预测了 Binding DB 中所有未标记的 CPI,并且来自其他数据库的外部证据证实了前五个预测的新 CPI 中的四个。这表明,Word2vec-GCN 可以在新发现的化合物和蛋白质靶点中发现可靠的 CPI,可以促进药物重定位、提高靶点识别效率、加速药物研发过程。

参考文献(References)

[1]薛斌,李益洲,李梦龙. 基于化学信息学方法预测药物副作用的研究进展[J]. 计算机与应用化学, 2019,36(5):487-490.DOI: 10.16866/j.com.app.chem201905009.
XUE Bin, LI Yizhou, LI Menglong. Research progress in predicting drug side effects based on chemical informatics methods[J]. Computers and Applied Chemistry, 2019,36(5):487-490. DOI: 10.16866/j.com.app.chem201905009.

[2]YABUUCHI H, NIIJIMA S, TAKEMATSU H. Analysis of multiple compound-protein interactions reveals novel bioactive molecules[J]. Molecular Systems Biology, 2011,7(1):472.DOI: 10.1038/msb.2011.5.

[3]宗可昕, 张晶, 陈云雨,等. PLK1 PBD 靶向抗肿瘤抑制

剂的筛选及活性研究[J]. 中国医药生物技术,2016,11(1):13-20. DOI:10.3969/j.issn.1673-713X.2016.01.004.
ZONG Kexin, ZHANG Jing, CHEN Yunyu, et al. Screening and anti-tumor activity of PLK1 PBD targeted antitumor inhibitors [J]. China Medical Biotechnology, 2016,11(1):13-20. DOI:10.3969/j.issn.1673-713X.2016.01.004.

[4]孙国章,何莹,尹照瑞. 灯盏花素与牛血清白蛋白相互作用的光谱学研究[J]. 山西中医药大学学报, 2020,21(3):192-195. DOI:10.19763/j.cnki.2096-7403.2020.03.10.
SUN Guozhang, HE Ying, YIN Zhaorui. Spectroscopic study of the interaction between breviscapine and bovine serum albumin[J]. Journal of Shanxi University of Chinese Medicine, 2020, 21 (3): 192-195. DOI: 10.19763/j.cnki.2096-7403.2020.03.10.

[5]JAFARI R, ALMQVIST H, AXELSSON H, et al. The cellular thermal shift assay for evaluating drug target interactions in cells[J]. Nature Protocols, 2014,9: 2100-2122. DOI: 10.1038/NPROT.2014.138.

[6]VILAR S, HRIPCSAK G. Leveraging 3D chemical similarity, target and phenotypic data in the identification of drug-protein and drug-adverse effect associations[J]. Journal of Cheminformatics, 2016, 8:35. DOI: 10.1186/s13321-016-0147-1.

[7]MOLINA D M, JAFARI R, IGNATUSHENKO M. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay[J]. Science,2013,341(6141):84-87. DOI:10.1126/science.1233606.

[8]李玲,李佳蔚,张月梅.基于分子对接预测靶点 ACE2 和 IL-6R 研究归芪白术方治疗新型冠状病毒肺炎的物质基础及其作用机制[J]. 甘肃中医药大学学报, 2020, 37(2): 1-9. DOI:10.16841/j.issn1003-8450.2020.02.01.
LI Ling, LI Jiawei, ZHANG Yuemei, et al. Material basis and action mechanism of Guiqi Baizhu Fang in the treatment of COVID-19 based on molecular docking prediction for target ACE2 and I-6R[J]. Journal of Gansu University of Chinese Medicine, 2020, 37 (2): 1-9. DOI: 10.16841/j.issn1003-8450.2020.02.01.

[9]MA D L, CHAN D S H, LEUNG C H. Drug repositioning by structure based virtual screening[J]. Chemical Society Reviews,2013,42(5):2130-2141. DOI:10.1039/c2cs35357a.

[10]贾聪敏.基于分子振动特征的药物靶点识别及活性预测模型研究[D].北京:北京中医药大学,2019.
JIA Congmin. Research on drug target recognition and activity prediction model based on molecular vibration characteristics[D]. Beijing: Beijing University of Chinese Medicine,2019.

[11]CHOW E, KLEPEIS L. New technologies for molecular dynamics simulations[J]. Comprehensive Biophysics,2012,9:86-104. DOI: 10.1016/B978-0-12-374920-8.00908-5.

[12]JACOB L, VERT J P. Protein-Ligand Interaction Prediction: An improved chemogenomics approach [J]. Bioinformatics, 2008, 24(19):2149-2156. DOI:10.1093/bioinformatics/btn409.

[13]BLEAKLEY K, YAMANISHI Y. Supervised prediction of drug-target interactions using bipartite local models[J]. Bioinformatics, 2009, 25(18): 2397-2403. DOI: 10.1093/bioinformatics/btp433.

[14]COELHO E D, ARRAIS J P, OLIVEIRA J L. Ensemble-based methodology for the prediction of drug-target interactions[C]. Process of the 29th IEEE International Symposium on Computer-Based Medical Systems. Washington, USA: IEEE, 2016: 36-41. DOI: 10.1109/CBMS.2016.67.

[15]MADHUKAR N S, KHADE P K, HUANG L, et al. A bayesian machine learning approach for drug target identification using diverse data types[J]. Nature Communication, 2019,10(1):5221. DOI:10.1038/s41467-019-12928-6.

[16]郭真俊.基于异构网络的药物靶标交互作用预测方法研究[D].长春:长春工业大学,2022.
GUO Zhenjun. Research on prediction method of drug target interaction based on heterogeneous networks[D]. Changchun: Changchun University of Technology,2022.

[17]HO T K. The random subspace method for constructing decision forests[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998,20(8):832-844. DOI: 10.1109/34.709601.

[18]CORTES C, VAPNIK V. Support-Vector Networks[J]. Machine Learning,1995,20(3):273-297. DOI: 10.1007/BF00994018.

[19]FRIEDMAN J H. Greedy function approximation: A gradient boosting machine [J]. Annals of Statistics, 2001, 29(5):1189-1232. DOI:10.1214/aos/101320345.

[20]NANNI L, LUMIN A, BRAHNAM S. A set of descriptors for identifying the protein-drug interaction in cellular networking [J]. Journal of Theoretical Biology, 2014, 359(24):120-128. DOI:10.1016/j.jtbi.2014.06.008.

[21]HE Zhisong, ZHANG Jian, SHI Xiaohe, et al. Predicting drug-target interaction networks based on functional groups and biological features[J]. PloS one, 2010,5(3):e9603. DOI:10.1371/journal.pone.0009603.

[22]RAYHAN F, AHMED S, SHATABDA S, et al. iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting[J]. Scientific Reports,2017,7(1):17731. DOI:10.1038/s41598-017-18025-2.

[23]LIU Weibo, WANG Zidong, LIU Xiaohui, et al. A survey of deep neural network architectures and their applications [J]. Neuro Computing, 2017,234(234):11-26. DOI:10.1016/j.neucom.2016.12.038.

[24]KIM, P. Convolutional neural network[M]. Berkeley, CA:

MATLAB Deep Learning, 2017.

[25] MARHON S, CAMEROM C, KREMER S. Recurrent neural networks[M]. Berlin, Heidelberg: Handbook on Neural Information Processing,2013: 49.

[26] TIAN Kai, SHAO Mingyu, WANG Yang. Boosting compound-protein interaction prediction by deep learning[J]. Methods, 2016,110: 64-72. DOI: 10.1016/j.ymeth.2016.06.024

[27] XIE Lingwei, HE Song, SONG Xinyu. et al.Deep learning-based transcriptome data classification for drug-target interaction prediction[J]. BMC Genomics, 2018, 19(7): 667. DOI:10.1186/s12864-018-5031-0.

[28] YUAN Qianmu, CHEN Jianwen, ZHAO Huiying Structure-aware protein-protein interaction site prediction using deep graph convolutional network [J]. Bioinformatics, 2022, 38(1):125-132. DOI:10.1093/bioinformatics/btab643.

[29] CHANHEE P, JINUK P, SANGHYUN P. AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction[J]. Expert Systems with Applications, 2020,159:113538. DOI:10.1016/j.eswa.2020.113538.

[30] YUAN Ye, BAR-JOSEPH Z. GCNG: Graph convolutional networks for inferring gene interaction from spatial transcriptomics data[J]. Genome Biology,2020, 21:300. DOI:10.1186/s13059-020-02214-w.

[31] WAN Fangping, HONG Lixiang, XIAO An, 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. DOI: 10.1093/bioinformatics/bty543.

[32] RAO A, VG S, JOSEPH T, et al. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks[J]. BMC Med Genomics, 2108, 11(1):57.DOI:10.1186/s12920-018-0372-8.

[33] MONA A. Neuro-symbolic representation learning on biological knowledge graphs[J]. Bioinformatics, 2017, 33(17): 2723-2730. DOI:10.1093/bioinformatics/btx275.

[34] NGUYEN T, LE H, QUINN T P. GraphDTA: Predicting drug-target binding affinity with graph neural networks[J]. Bioinformatics, 2020,37(8):1140-1147. DOI: 10. 1093/bioinformatics/btaa921.

[35] TSUBAKI M, TOMII K, SESE J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. Bioinformatics, 2019, 35(2): 309-318. DOI: 10. 1093/bioinformatics/bty535.

[36] HAMILTON W, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]. Neural Information Processing Systems. Long Beach, California, USA, 2017: 1025-1035.DOI:10.48550/arXiv.1706.02216.

[37] REX Y, RUINING H, KAIFENG C, et al. Graph convolutional neural networks for web-scale recommender systems [C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining London, United Kingdom:ACM,2017: 974-983. DOI:10.1145/3219819.3219890.

[38] CHEN Jianfei, ZHU Jun, SONG Le. Stochastic training of graph convolutional networks with variance reduction[J]. arXiv:1710.10568v3, 2018. DOI: 10.48550/arXiv.1710.10568.

[39] CHEN Jie, MA Tengfei, XIAO Cao. FastGCN: Fast learning with graph convolutional networks via importance sampling[J].arXiv: 1801.10247, 2018. DOI: 10.48550/arXiv.1801.10247.

[40] YU Z. AS-GCN: Adaptive semantic architecture of graph convolutional networks for Text-Rich networks[C]. 2021 IEEE International Conference on Data Mining (ICDM). Auckland, New Zealand:IEEE, 2021:837-846. DOI: 10.1109/ICDM51629.2021.00095.

[41] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv: 1301.3781, 2013. DOI: 10.48550/arXiv.1301.3781.

[42] GILSON M K, LIU Tiqing, BAITALUK M, et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology[J]. Nucleic Acids Research, 2016, 44(D1): D1045-D1053. DOI:10.1093/nar/gkv1072.

[43] YE Qing, ZHANG Xiaolong, LIN Xiaoli. Drug-target interaction prediction via multiple classification strategies[J]. BMC Bioinformatics, 2022, 22(12):461. DOI: 10.1186/s12859-021-04366-3.

[44] RAYHAN F, AHMED S, FARID D, et al. CFSBoost: Cumulative feature subspace boosting for drug-target interaction prediction[J]. Journal of Theoretical Biology, 2019, 464:1-8. DOI:10.1016/j.jtbi.2018.12.024.

[45] ZHANG Yufang, WANG Xiangeng, KAUSHIK A C, et al. SPVec: A word2vec-inspired feature representation method for drug-target interaction prediction[J]. Frontiers in Chemistry, 2020,7:895. DOI:10.3389/fchem.2019.00895.

[46] DAVID S. DrugBank: A knowledgebase for drugs, drug actions and drug targets[J]. Nucleic Acids Research, 2008, 36:901-906. DOI: 10.1093/nar/gkm958.