

融合门控机制与对比学习的化合物—蛋白质相互作用预测模型

贾晓田, 柳楠, 荆昊

(山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

摘要: 化合物与特定蛋白质的结合情况为确定药物疗效提供了有效参考,在药物发现中化合物—蛋白质相互作用(CPI)预测具有重要的应用价值。为了解决预测CPI在特征提取、信息融合方面的局限性,提出了融合门控机制与对比学习的化合物—蛋白质相互作用预测模型。首先,利用门控机制有效融合化合物与蛋白质特征,以准确捕捉二者之间的相互作用;其次,通过对比学习增强化合物与蛋白质的特征表示,提升模型区分相似与非相似样本的能力。实验表明,所提模型相较于GraphDTA、GNN-CPI及TransformerCPI等方法,在多个评价指标上提升显著。

关键词: 化合物—蛋白质相互作用;对比学习;门控机制;特征融合;预测模型

DOI:10.11907/rjdk.241926

扫描二维码阅读全文:

中图分类号:TP301

文献标识码:A

文章编号:1672-7800(2026)002-0046-06



Compound-Protein Interaction Prediction Model Integrating Gated Mechanisms and Contrastive Learning

JIA Xiaotian, LIU Nan, JING Hao

(School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China)

Abstract: Binding of compounds to specific proteins provides an effective reference for determining drug efficacy, and predicting compound protein interactions (CPI) has important application value in drug discovery. In order to address the limitations of predicting CPI in feature extraction and information fusion, a compound protein interaction prediction model combining gate control mechanism and contrastive learning was proposed. Firstly, utilizing gating mechanisms to effectively fuse compound and protein features, in order to accurately capture the interactions between the two; Secondly, enhancing the feature representation of compounds and proteins through contrastive learning improves the model's ability to distinguish between similar and dissimilar samples. Experiments have shown that the proposed model significantly improves multiple evaluation metrics compared to methods such as GraphDTA, GNN-CPI, and TransformerCPI.

Key Words: compound-protein interaction; contrastive learning; gating mechanism; feature fusion; prediction model

0 引言

化合物—蛋白质相互作用(Compound-Protein Interaction, CPI)的准确预测在药物发现过程中扮演着至关重要的角色。CPI预测可分析化合物的分子结构与蛋白质靶点

特性,识别出能与特定蛋白质有效结合的化合物,有助于筛选出潜在药物,减少实验筛选范围与成本,加速新药研发与应用进程^[1]。此外,CPI预测不仅有助于发现新的药物靶点,还可预测药物副作用、药物多靶点作用机制及进行药物重定位等^[2]。

传统生物实验方法耗时长、成本高、覆盖面有限,无法满足快速、全面筛选药物的需求。计算机辅助的预测技术

收稿日期:2024-12-16

基金项目:国家自然科学基金青年科学基金项目(61902221);山东省自然科学基金面上项目(ZR2018MF012);山东省研究生教育教学改革研究项目(SDYJG21173)

作者简介:贾晓田(2000-),女,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为深度学习、生物计算;柳楠(1980-),女,博士,CCF会员,山东建筑大学计算机科学与技术学院教授、硕士生导师,研究方向为生物计算、基因组片段填充问题、计算基因组学;荆昊(1997-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为蛋白质从头肽测序。本文通讯作者:柳楠。

可通过算法大规模筛选潜在的CPI候选者,提升效率、降低成本,但现有预测方法在特征提取、信息融合方面仍面临诸多挑战,在处理化合物与蛋白质的复杂特征时问题更为显著。尽管,深度学习技术在CPI预测中显示出了良好的应用潜力,能自动提取特征并进行高效学习,但性能仍需进一步提升。综上,CPI预测技术目前主要存在以下问题:①对化合物与蛋白质复杂特征的建模不足,难以有效捕获其潜在的关系;②现有方法在特征表达能力、鲁棒性方面不足,预测性能有待提升。

为此,本文提出融合门控机制与对比学习的化合物-蛋白质相互作用预测模型,不仅能解决当前方法的不足之处,还为CPI预测领域提供了新的研究视角。其中,门控机制通过动态筛选重要信息,有效融合了化合物与蛋白质特征,以准确捕捉二者间的相互作用;对比学习通过优化特征表征,以增强模型区分相似与非相似样本的能力,提升了预测性能和泛化能力。

1 相关工作

随着社会不断进步与发展,各类新型疾病不断涌现,对人类生命健康带来了巨大挑战和威胁。为了应对这些健康问题,药物发现成为开发新药治疗或治愈疾病的核心环节,而药物靶标的识别在药物发现中十分重要,其中CPI预测最为关键。CPI预测不仅在帮助研究人员发现潜在药物靶点方面发挥着重要作用,还对药物研发中的多个环节具有深远影响。通过精准的CPI预测,可有效推动药物靶向性研究、药物药效探索及其副作用和毒性预测,减少临床实验中的风险和成本。此外,CPI预测还为药物重定位提供了新的研究方向,使已有药物在新的治疗领域发挥作用,现已成为当前药物发现中的热点工作,推动了新药研发进程,为应对复杂疾病带来了新希望。

当前,CPI预测方法大致可分为3类:①生物化学实验方法,通过实验手段直接测量化合物与蛋白质之间的相互作用,尽管精确性高,但耗时长、成本昂贵^[2-6];②基于结构的计算方法,利用化合物与蛋白质的三维结构信息,结合分子对接与虚拟筛选技术计算、预测两者之间的相互作用^[7-10];③基于机器学习的预测方法,通过大规模数据训练模型来自动提取化合物与蛋白质特征,以高效进行CPI预测^[11-13]。

1.1 生物化学实验方法

宗可昕等^[2]基于动力学实验原理,深入研究了药物中化合物与靶标蛋白质之间的结合亲和力,揭示了两者的相互作用机制。孙国章等^[3]采用光谱学方法详细探讨化合物-蛋白质之间的相互作用问题,扩展了该领域的研究方法。Vilar等^[4]提出一种整合三维化学相似性、靶标信息和药物不良反应数据的创新方法,有效确定了药物与蛋白质之间的相互作用,深入分析了药物不良反应之间的关联

性。Molina等^[5]采用细胞热转变分析实验技术研究化合物与蛋白质的结合行为,为CPI领域提供了新的实验手段。张建锋^[6]通过电化学技术探索了化合物与蛋白质的相互作用机制,为CPI研究提供了电化学方面的新视角。上述方法各具特色,为CPI领域的深入研究奠定了基础。

1.2 基于结构的计算方法

李玲等^[7]通过分子对接技术筛选出能高效结合靶标蛋白的小分子化合物,为药物研发提供了精确的筛选工具。Ma等^[8]提出一种基于结构的虚拟筛选方法,通过计算手段实现药物重新定位,加速了药物发现进程。贾聪敏^[9]开发了一种利用分子振动特征识别药物靶点的方法,通过捕捉化合物与蛋白质之间的振动特性,提升了靶点识别精度。Chow等^[10]提出一种基于分子动力学的计算方法,定量评估化合物与蛋白质的结合能力,并分析二者间的相互作用。虽然,上述方法的预测准确性较高,但需要掌握较多专业知识,特别是算法、分子动力学领域的相关知识,因此间接增加了应用复杂性。

1.3 基于机器学习的方法

Jacob等^[11]提出一种改进的化学基因组学方法,结合化学亚结构与蛋白质家族间的张量积特征,捕捉了化合物与蛋白质之间的相互作用特性,并采用支持向量机(Support Vector Machine, SVM)进行精确预测。Bleakley等^[12]引入二分图局部模型法,通过度量化学结构与蛋白质序列之间的相似性,利用已知的交互数据训练SVM,以提升模型的CPI预测能力。Coelho等^[13]提出一种基于集成学习的创新方法,将随机森林与支持向量机相结合,通过多模型集成的方式提升CPI预测的准确性。虽然,上述方法在不同层面提升了CPI预测效果,但随着数据量、模型复杂度增加,仍存在许多问题亟待解决。

尽管,基于传统机器学习方法的CPI预测取得了一定的成果,但仍属于浅层学习,需依赖大量人工提取的特征。当面对复杂、大规模的数据时,传统方法的局限性尤为明显,难以充分挖掘深层次特征,导致预测性能受限。为此,Tian等^[14]利用深度神经网络从数据中自动学习化合物与蛋白质对的潜在特征,取得了较好的预测效果。Nguyen等^[15]引入图神经网络,通过建模药物与靶标之间的图结构,提升了药物-靶标亲和力的预测准确性。Tsubaki等^[16]结合图卷积网络(Graph Convolutional Network, GCN)与卷积神经网络(Convolutional Neural Network, CNN),使用GCN处理化合物的分子结构图,CNN则用于处理蛋白质的序列信息,较好地解决了CPI预测任务^[17,18]。然而,上述方法仍存在一定的局限性。例如,GCN在处理图数据时需全局访问图结构,随着网络层数增加,节点特征逐渐趋同,区分性减弱;CNN由于卷积核大小固定,局限于捕获滑动窗口内的局部信息,难以全面获取蛋白质序列中的全局依赖关系。

综上,尽管现有CPI预测方法在不同应用场景中具有

一定优势,但在处理高维复杂数据时仍面临诸多挑战,特别是在特征表示和信息融合方面存在改进空间。为此,本文融合门控机制(Gate Mechanism, GM)与对比学习(Contrastive Learning, CL),提出了一种化合物—蛋白质相互作用预测模型(GM-CL)^[19,20]。该模型引入更精细的特征学习和融合策略来提升预测性能,旨在弥补现有模型在处理复杂特征与数据融合方面的不足。首先,利用门控机制有效融合化合物与蛋白质特征,使模型更准确地捕捉二者之间的相互作用关系;其次,通过对比学习强化化合物与蛋白质的特征表示能力,特别是在处理相似与不相似样本时具有更强的区分能力。

2 GM-CL 模型

本文提出一种专用于预测化合物与蛋白质之间相互作用的计算模型 GM-CL,通过特征提取、融合步骤、对比学习技术来提升模型预测准确性与可靠性,如图 1 所示。

首先,对氨基酸序列与化合物分子进行特征提取,旨在更好地捕捉两者间的复杂特性。对于氨基酸序列,采用 n -gram 方法进行分割以保留局部序列特征,并通过长短期记忆网络(Long Short Term Memory Network, LSTM)捕捉序列中的有序性和长期依赖性,以获取丰富的序列表示^[21]。对于化合物分子,利用图神经网络(Graph Neural Network, GNN)提取原子特征与分子的结构信息,以充分捕捉分子图中的拓扑关系和局部结构^[22]。

其次,通过 GM 将提取到的特征进行融合,以有效过滤冗余信息,保留关键特征,从而增强特征表达精度。在此基础上,引入 CL 对相似与非相似样本进行对比学习,以提升模型的区分能力,使其更精确地识别化合物与蛋白质之间的相互作用关系。

最后,通过解码层将化合物与蛋白质进行特征拼接,以准确预测两者间的相互作用。

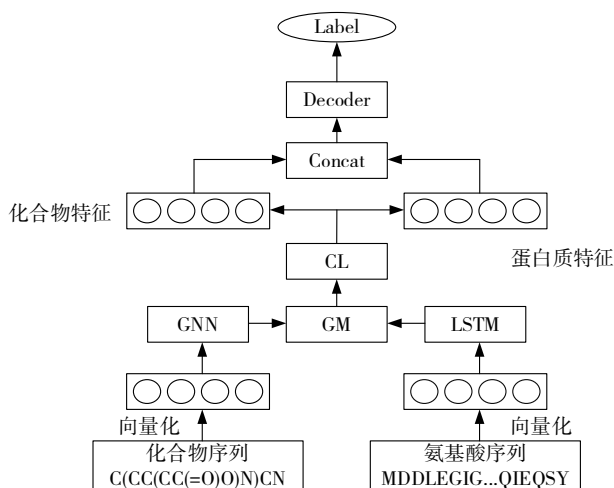


Fig. 1 GM-CL model architecture

图1 GM-CL 模型架构

2.1 编码层

2.1.1 氨基酸序列特征提取

本文采用 n -gram 方法提取、划分氨基酸序列特征,特征包括基本信息、物理化学性质和序列长度等。具体地, n -gram 既能有效捕获局部信息,保留氨基酸顺序关系,还能通过调整 n 值平衡模型提取全局与局部特征。实验证明,过小的 n 值会丢失序列上下文信息,过大的 n 值会引起模型发生冗余。为此,本文模型选择 $n=3$,原因为三联体能充分捕捉氨基酸之间的相互作用与局部结构信息,减少特征表示的稀疏性。

$$S' = [S_1 = [x_1, x_2, x_3], S_2 = [x_2, x_3, x_4], \dots,$$

$$S_{m-2} = [x_{m-2}, x_{m-1}, x_m]] \quad (1)$$

式中:将 $S' = [S_1, S_2, \dots, S_{m-2}]$ 中每个子序列映射为向量,即可得到 $Q = [q_1, q_2, \dots, q_{m-2}]$; q_i 表示第 i 个氨基酸子序列的特征表示。

本文模型基于 LSTM,可记忆、捕获序列中的长距离依赖关系,具有解决梯度消失问题的优点。同时,通过 LSTM 网络可得到特征矩阵,以捕捉序列中更深层次的上下文依赖关系。

$$H = [h_1, h_2, \dots, h_{m-2}] \quad (2)$$

式中: $H \in R^{N \times 3}$ 。

为了进一步保留氨基酸序列的时序特性,使用线性变换得到 D 实现特征的高维映射,使模型更丰富地表达序列特征,以提升后续网络的学习能力与表达效果。

$$D = H \times W \quad (3)$$

式中: $W \in R^{3 \times M}$; $H \in R^{N \times M}$ 。

2.1.2 化合物分子特征提取

化合物数据的格式为简化分子线性输入规范(Simplified Molecular Input Line Entry System, SMILES),包括原子类型、键类型、分子形状和分子量等。本文采用 Rdkit 工具将其转化为图结构,以更好地捕捉化合物的拓扑信息与原子之间的关系,增强模型对化合物特征的理解能力,提升化合物特征提取效果。为此,将数据转化后可生成原子的特征矩阵 $X \in R^{N \times C}$ 、结构矩阵 $A \in R^{N \times C}$, X 表示每个原子的特征信息, A 表示原子之间的连接关系。

具体而言, X 的第 i 行元素表示第 i 个原子 v_i 的特征, A 中的元素 a_{ij} 表示原子 v_i 与 v_j 之间的连接关系, $a_{ij}=1$ 表示存在连接,反之则为 0。

为了进一步增强特征表示,先对特征矩阵 X 进行线性变换得到初步的特征转换结果,实现原子特征的高维表示;再将特征矩阵 H 与结构矩阵 A 输入 GNN,逐层提取化合物的深层特征,每层的特征计算与更新表示如式(4)—式(6)所示。

$$H = X \times W_2 \quad (4)$$

$$e_{l+1} = \text{ReLU}(W_l \times H_l) \quad (5)$$

$$H_{l+1} = H_l + A \times e_{l+1} \quad (6)$$

式中: $W_2 \in R^{C \times M}$; $H \in R^{N \times M}$; H_l 表示第 l 层的原子特征

矩阵; $\mathbf{W}_l \in \mathbf{R}^{C \times N}$ 为第 l 层可学习的权重矩阵; ReLU 表示激活函数; $\mathbf{e}_{l+1} \in \mathbf{R}^{C \times M}$ 为第 $l+1$ 层的隐藏特征矩阵; $\mathbf{A}$ 为原子间的连接关系矩阵。

当所有层计算完成后,对每层的特征进行求和得到最终的化合物特征表示矩阵。通过这个过程,模型可充分捕获化合物分子结构中的局部、全局信息,为后续网络提供更丰富的特征表示。

$$\mathbf{V} = \sum_{i=1}^N \mathbf{H}_i \quad (7)$$

式中: $\mathbf{V} \in \mathbf{R}^{N \times M}$ 。

2.2 门控机制

将提取到的化合物特征矩阵 $\mathbf{V} \in \mathbf{R}^{N \times M}$ 与氨基酸特征矩阵 $\mathbf{H} \in \mathbf{R}^{N \times M}$ 输入到门控机制中过滤冗余信息、低相关性特征等噪声特征,为下层网络特征学习提供有效支持。

$$\mathbf{G}(\mathbf{V}, \mathbf{H}) = o(\mathbf{W}_g \times [\mathbf{V}; \mathbf{H}])b_g \quad (8)$$

式中: $\mathbf{V}, \mathbf{H}$ 分别表示化合物与氨基酸的特征矩阵,维度为 $N \times M$; $[\mathbf{V}; \mathbf{H}]$ 表示将 $\mathbf{V}$ 与 $\mathbf{H}$ 按列拼接成一个新的矩阵; $\mathbf{W}_g$ 为门控机制中的权重矩阵; b_g 为偏置项; 函数 o 为 Sigmoid 激活函数,用于将输出压缩到 $[0, 1]$ 区间,表示每个特征的重要性; $\mathbf{G}(\mathbf{V}, \mathbf{H})$ 为门控机制的输出矩阵,维度为 $N \times M$,每个元素都介于 0~1 之间,表征对应特征的重要性。

经过门控机制后,最终的过滤特征矩阵如式(9)所示。通过这种方式,门控机制能有效筛选出重要的特征,抑制不必要的噪声特征,提升模型特征表达效果。

$$\mathbf{F} = \mathbf{G}(\mathbf{V}, \mathbf{H}) \odot [\mathbf{V}; \mathbf{H}] \quad (9)$$

式中: $\odot$ 表示逐元素乘积 (Hadamard 乘积)。

2.3 对比学习

本文引入对比学习损失进一步提升特征表示的区分能力。在 CPI 预测中,锚点样本是某个特定化合物与蛋白质的特征向量对,正样本是与该锚点样本相互作用的真实化合物-蛋白质对,负样本则是无相互作用的随机化合物-蛋白质对。对比损失通过计算锚点样本与负样本的余弦相似度来衡量特征间的相似度,其目标是确保锚点样本与正样本的相似度比相较于负样本的相似度更高,且相差至少一个设定的边际值。如此,不断优化模型便能更准确地捕捉化合物与蛋白质的真实相互作用,提升模型对相似样本与非相似样本的区分能力。

对比损失基于锚点样本与负样本之间的余弦相似度进行度量,目标是确保锚点样本与正样本之间的相似度高于锚点样本与负样本之间的相似度,且至少高于一个边际值 (margin),具体计算公式如(10)所示。

$$L = \frac{1}{N} \sum_{i=1}^N \max(0, \text{margin} - \cos(\text{anchor}[i], \text{positive}[i]) + \cos(\text{anchor}[i], \text{negative}[i])) \quad (10)$$

式中: L 表示总的对比损失; N 为锚点样本数量; $\text{anchor}[i]$ 为第 i 个锚点样本的嵌入向量; $\text{positive}[i]$ 为第 i 个锚点样本对应的正样本的嵌入向量; $\text{negative}[i]$ 为第 i 个锚点样本对应的负样本的嵌入向量; $\cos(\cdot, \cdot)$ 表示两个向量之间的

余弦相似度; margin 为预先定义的边际值,用于确保正样本与负样本之间存在足够的区分度。

同时,通过优化损失函数使模型在学习过程中更好地区分不同特征,提升分类、回归任务的精度。在特征学习中,锚点样本与正样本的相似度通常来自于相同类别或相似结构的样本,负样本则来自不同类别或结构的样本。通过对比学习,模型可逐步调整特征空间,使相同类别的样本在嵌入空间中更接近,不同类别的样本则保持一定距离,从而增强了模型的鲁棒性与泛化能力。

2.4 解码层

在提取、处理蛋白质和化合物特征后,模型通过解码层进行最终的预测任务。解码层的主要作用是将特征映射到标签空间,以输出药物-蛋白质相互作用的预测结果。首先,采用全连接层对处理后的特征进行融合;其次,通过 Softmax 激活函数输出预测的概率分布,解码层的计算过程如式(11)所示。通过这种拼接操作,模型能融合蛋白质与化合物的特征,形成一个整体的特征表示。

$$\mathbf{F}' = \text{Softmax}(\mathbf{W}_1 \times \text{Concat}(\mathbf{P} \times \mathbf{C}) + b) \quad (11)$$

式中: $\mathbf{P}, \mathbf{C}$ 分别表示处理后的蛋白质特征矩阵与化合物特征矩阵; $\text{Concat}(\mathbf{P} \times \mathbf{C})$ 表示将这两者按列拼接,形成一个新的特征矩阵; $\mathbf{W}_1 \in \mathbf{R}^{2 \times M \times |\text{label}|}$ 是全连接层的权重矩阵; $2M$ 表示拼接后的特征维度; $|\text{label}|$ 表示药物-蛋白质相互作用标签的种类数; b 为偏置项,用于对线性变换的输出进行平移; Softmax 函数将线性变换的输出转化为概率分布,得出最终的预测结果 $\mathbf{F}'$ 。

3 实验结果与分析

3.1 数据集

实验依托 Human、Celegans 这两个常用的、具有平衡的正负样本的公开 CPI 数据集进行实验^[23,24]。为了便于模型训练、调优和评估,将两个数据集划分为训练集、验证集和测试集 3 个部分,训练集用于模型的初步训练,验证集用于在训练过程中对模型参数进行调整,测试集则用于评估模型的整体性能和泛化能力。

此外,为了更好地了解数据集的具体分布情况,为模型实验设计、调参提供参考,详细统计了 Human、Celegans 数据集的样本数量、蛋白质与化合物的分布情况,确保模型在不同阶段的表现具有充分的参考依据,可为后续实验结果的可靠性提供数据支撑。具体的数据分布情况如表 1 所示。

Table 1 Data distribution of Human and Celegans datasets

表 1 Human 与 Celegans 的数据分布情况

数据集	化合物数量	蛋白质数量	总样本数量	正样本数量	正样本比例/%
Human	1 709	2 043	6 212	3 364	54.2
Celegans	1 723	1 708	7 511	3 893	51.8

3.2 实验环境与参数设置

实验环境为 Ubuntu 18.04 系统,硬件配置为 Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz 处理器,64 GB 内存,NVIDIA GeForce GTX 2080Ti 显卡(12 GB 显存),为深度学习模型的训练、推理提供了充足的算力。特别是,在处理大规模数据和复杂网络结构时,显卡的并行计算能力可显著提升模型训练效率。

此外,实验代码为 Python 编写,基于 Pytorch 深度学习框架提供灵活的动态计算图和高效的 GPU 加速支持,可快速进行模型搭建、调试和训练。

3.3 实验结果

为了验证本文方法的有效性,研究比较了以下 5 种常见的基准模型,实验结果如表 2、表 3 所示。

Table 2 Performance of various models on the Human dataset

表 2 各模型在 Human 数据集上的性能表现

模型	AUC	Precision	Recall
GraphDTA	0.958	0.802	0.913
GNN-CPI	0.953	0.622	0.928
PWO-CPI	0.981	0.938	0.943
MolTrans	0.976	0.936	0.932
TransformerCPI	0.974	0.918	0.924
GM-CL	0.988	0.948	0.965

Table 3 Performance of various models on the Celegans dataset

表 3 各模型在 Celegans 数据集上的性能表现

模型	AUC	Precision	Recall
GraphDTA	0.973	0.927	0.913
GNN-CPI	0.974	0.923	0.927
PWO-CPI	0.982	0.960	0.953
MolTrans	0.980	0.948	0.960
TransformerCPI	0.981	0.952	0.951
GM-CL	0.986	0.968	0.964

(1) GraphDTA。该模型结合了图神经网络与传统 CNN,使用 SMILES 序列与蛋白质序列作为输入特征,通过多层神经网络进行亲和力预测。本文修改了最后的输出层,引入 sigmoid 激活函数将其转换为 CPI 任务的二分类模型。

(2) GNN-CPI。该模型选择分子的指纹信息和距离矩阵作为化合物的特征输入,利用 GNN 网络将两者信息融合后进行学习。本文在实验中使用与原始论文一致的超参数设置,并遵循了代码的默认配置。

(3) MolTrans。首先,构建一个大规模语料库,对化合物进行编码;其次,通过 Transformer 编码器对化合物与蛋白质序列进行语义学习。本文复现了模型的编码过程,保证各参数设定与原文一致。

(4) PWO-CPI。首先,该模型使用化合物的分子图像作为输入,通过卷积神经网络进行特征提取;其次,采用 word2vec 对蛋白质的 FASTA(一种用于记录核酸(DNA/RNA)或蛋白质序列的文本格式)序列进行编码。本文构建了具有相同卷积核大小与架构的 CNN 网络进行比较实验。

(5) TransformerCPI。该模型将分子序列与距离矩阵作为化合物特征输入,使用 Transformer 编码器学习化合物与蛋白质特征之间的作用。本文遵循了原文的模型结构与超参数配置,确保比较实验的结果公正、可信。

由表 2 可见,所提模型相较于 PWO-CPI,在 AUC、Precision、Recall 指标上分别提升 0.007、0.01、0.022,表明本文模型的性能更出色。由表 3 可见,所提模型相较于 PWO-CPI,在 AUC、Precision、Recall 指标上分别提升 0.004、0.08、0.004,在多种评价指标上优势明显,在不同数据集上的泛化能力更强。由表 2、表 3 可见,本文模型相较于现有 CPI 预测模型而言,在 AUC、Precision、Recall 指标上提升效果显著,更适用于药物预测领域。

3.4 消融实验

消融实验旨在评估每个组件对整体模型性能的贡献,通过移除模型某些模块来衡量各模块对最终结果的影响。为此,本文针对两个关键模块进行消融评估:①without GM 表示移除门控机制模块;②without CL 表示移除对比学习损失模块。由表 4 可知,移除 GM 模块后,模型的 AUC、Precision、Recall 分别降低 0.004、0.008、0.007;移除 CL 模块后,模型的 AUC、Precision、Recall 分别降低 0.010、0.015、0.013。由表 5 可知,移除 GM 模块后,模型的 AUC、Precision、Recall 分别降低 0.004、0.003、0.005;移除 CL 模块后,模型的 AUC、Precision、Recall 分别降低 0.012、0.014、0.016。

综上,GM 模块能帮助模型更好地捕捉化合物与蛋白质之间的相互作用,提升模型整体预测精度和稳定性;CL 模块增强了模型区分相似与非相似样本的能力,强化了化合物与蛋白质的特征表示,提升了模型预测性能,两个模块对模型的最终性能都具有不可忽视的贡献。

Table 4 Ablation experiment of GM-CL on the Human dataset

表 4 GM-CL 在 Human 数据集上的消融实验

Method	AUC	Precision	Recall
Without GM	0.984	0.940	0.958
Without CL	0.978	0.933	0.952
GM-CL	0.988	0.948	0.965

Table 5 Ablation experiment of GM-CL on the Celegans dataset

表 5 GM-CL 在 Celegans 数据集上的消融实验

Method	AUC	Precision	Recall
Without GM	0.982	0.965	0.959
Without CL	0.974	0.954	0.948
GM-CL	0.986	0.968	0.964

4 结语

本文提出融合门控机制与对比学习的化合物-蛋白质相互作用预测模型,旨在解决当前 CPI 预测模型中存在的特征提取、信息融合方面的局限性。首先,引入对比学习增强化合物与蛋白质的特征表示,提升模型区分相似与非相似样本的能力;其次,通过门控机制成功融合了化合物与蛋白质的特征,可捕捉两者间复杂的相互作用。

实验表明,所提方法相较于现有方法,在多项评价指标上的预测性能更优秀。虽然,所提模型解决了特征提取不充分的问题,但仍存在一些样本分布不均匀造成的长尾效应问题。未来,将深入探讨该问题,以更好地提升CPI模型性能,并探索对比学习在不同数据集和任务中的应用能力,以增强模型的泛化性和稳定性。

尽管门控机制在特征融合方面表现出色,但如何在更多样化的生物信息中提升自身性能仍然值得深入探讨,将有助于推动相关领域的研究发展,为实际应用提供更有效的解决方案。

参考文献:

- [1] XUE B, LI Y Z, LI M L. Research progress on predicting drug side effects based on cheminformatics methods[J]. Computer and Applied Chemistry, 2019, 36(5): 487-490.
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.
- [2] ZONG K X, ZHANG J, CHEN Y Y, et al. Screening and activity study of tumor inhibitors targeting PLK1 PBD[J]. Chinese Medical Biotechnology, 2016(1): 13-20.
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(1): 13-20.
- [3] SUN G Z, HE Y, YIN Z R, et al. Spectral study on the interaction between chrysanthemum flower extract and bovine serum albumin[J]. Journal of Shanxi University of Traditional Chinese Medicine, 2020, 21(3): 192-195.
孙国章,何莹,尹照瑞,等. 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.
- [4] 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: 1-12.
- [5] MOLINA D M, JAFARI R, IGNATUSHCHENKO M, et al. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay[J]. Science, 2013, 341(6141): 84-87.
- [6] ZHANG J F, ZHANG Z J, WANG Y, et al. Immobilized Ru(bpy)₃²⁺ electrochemiluminescence coupled with microdialysis sampling for on-line studying drug-protein interaction[J]. Analytical Chemistry, 2012, 40(5): 730-734.
张建锋,章竹君,王燕,等. 固定化联吡啶钌电化学发光结合微透析采样在线研究药物蛋白反应[J]. 分析化学, 2012, 40(5): 730-734.
- [7] LI L, LI J W, ZHANG Y M, et al. Research on the material basis and mechanism of action of Guiqi Baizhu Fang in treating COVID-19 based on molecular docking targeting ACE2 and IL-6R[J]. Journal of Gansu University of Traditional Chinese Medicine, 2020, 37(2): 1-9.
李玲,李佳蔚,张月梅,等. 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.
- [8] 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.
- [9] JIA C M. Study on drug target identification and activity prediction model based on molecular vibration features[D]. Beijing: Beijing University of Chinese Medicine, 2019.
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.
- [10] CHOW E, KLEPEIS J L, RENDLEMAN C A, et al. 9.6 new technologies for molecular dynamics simulations[EB/OL]. <https://moscow.sci-hub.se/2126/3653370521230458ce5754597b15e3a7/chow2012.pdf?download=true>.
- [11] JACOB L, VERT J P. Protein-ligand interaction prediction: an improved chemogenomics approach[J]. Bioinformatics, 2008, 24(19): 2149-2156.
- [12] BLEAKLEY K, YAMANISHI Y. Supervised prediction of drug-target interactions using bipartite local models[J]. Bioinformatics, 2009, 25(18): 2397-2403.
- [13] COELHO E D, ARRAIS J P, OLIVEIRA J L. Ensemble-based methodology for the prediction of drug-target interactions[C]// 29th International Symposium on Computer-Based Medical Systems, 2016: 36-41.
- [14] TIAN K, SHAO M, WANG Y, et al. Boosting compound-protein interaction prediction by deep learning[J]. Methods, 2016, 110: 64-72.
- [15] NGUYEN T, LE H, QUINN T P, et al. GraphDTA: predicting drug-target binding affinity with graph neural networks[J]. Bioinformatics, 2021, 37(8): 1140-1147.
- [16] 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.
- [17] KIPF T N, WELLMING M. Semi-supervised classification with graph convolutional networks[DB/OL]. <https://arxiv.org/abs/1609.02907>.
- [18] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [19] MENDELL L M. Constructing and deconstructing the gate theory of pain[J]. Pain, 2014, 155(2): 210-216.
- [20] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.
- [21] ZHAO Z, CHEN W, WU X, et al. LSTM network: a deep learning approach for short-term traffic forecast[J]. IET Intelligent Transport Systems, 2017, 11(2): 68-75.
- [22] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.
- [23] LI S H, JIA L. Prediction of compound-protein interactions based on graph attention networks and simple recurrent units[J]. Pattern Recognition and Artificial Intelligence, 2021, 34(6): 522-531.
李淑红,贾琳. 基于图注意力网络和简单循环单元的化合物-蛋白质交互预测[J]. 模式识别与人工智能, 2021, 34(6): 522-531.
- [24] LIU H, SUN J, GUAN J, et al. Improving compound-protein interaction prediction by building up highly credible negative samples[J]. Bioinformatics, 2015, 31(12): 221-229.

(责任编辑:刘嘉文)