

中国科学技术大学

硕士学位论文



基于深度学习的 药物相互作用预测方法研究

作者姓名： 孙晋辰

学科专业： 计算机科学与技术

导 师： 郑浩然副教授

完成时间： 二〇二五年五月二十九日

University of Science and Technology of China

A dissertation for master's degree



Research on Prediction Methods for Drug Interactions Based on Deep Learning

Author: Sun Jincheng

Speciality: Computer Science and Technology

Supervisor: Associate Prof. Zheng Haoran

Completion date: May 29, 2025

摘要

药物相互作用预测旨在预测药物-药物相互作用 (Drug-Drug Interaction, DDI)、药物-靶点相互作用 (Drug-Target Interaction, DTI) 等与药物相关的相互作用, 其对于潜在靶点识别、新药研发与药物再利用等任务具有重要意义。近年来, 基于深度学习的预测方法已取得显著进展, 但现有研究仍存在药物子结构可解释性不足、靶点编码模态单一以及相互作用信息挖掘不充分等问题, 相关方法的预测性能、泛化能力以及预测结果的可解释性有待进一步改善。针对上述挑战, 本文结合深度学习技术对 DDI 预测与 DTI 预测开展研究, 主要工作如下:

(1) 针对现有 DDI 预测方法存在的感知子结构可解释性不佳、化学子结构相互孤立的问题, 提出了一种基于图表示学习的 DDI 预测方法 HDN-DDI。该方法采用“原子-子结构-分子”层次化分子图表征药物分子, 并通过双视图表示学习提取药物的分子内层次化特征和分子间相互作用特征, 从而提高 DDI 预测的准确性。实验结果表明, HDN-DDI 在热启动场景中达到了 97.93% 的预测准确率, 在冷启动和实际应用场景中实现了多项指标 2% 至 6% 的有效提升。

(2) 针对现有 DTI 预测方法存在的二维图结构利用不足、相互作用信息挖掘不充分的问题, 给出了一种基于二维图结构的 DTI 预测方法 GDN-DTI。该方法通过二维图结构生成模块将药物与靶点的一维序列转化为二维图, 并通过双视图表示学习捕获分子内结构信息与分子间相互作用信息, 最终利用相互作用得分矩阵实现信息整合与 DTI 预测。实验结果表明, GDN-DTI 在两个数据集上的受试者工作特征曲线下面积 (Area Under the Receiver Operating Characteristic Curve, AUROC) 和精确率-召回率曲线下面积 (Area Under the Precision-Recall Curve, AUPR) 均高于 90%, 并且有多项指标优于现有方法。

(3) 针对现有 DTI 预测方法存在的编码模态单一、多模态融合不足的问题, 给出了一种基于多模态信息融合的 DTI 预测方法 MIF-DTI。该方法通过序列编码模块和图编码模块分别对药物与靶点的一维序列和二维图结构进行双视图编码, 并通过协注意力机制融合不同模态的全局特征, 从而提升 DTI 预测的准确性。实验结果表明, MIF-DTI 在三个数据集上均达到领域最优性能, 并且在准确率、AUROC 和 AUPR 等指标上实现了 1% 至 3% 的有效提升。

关键词: 药物-药物相互作用预测; 药物-靶点相互作用预测; 层次化分子图; 接触图; 双视图表示学习

ABSTRACT

Drug interactions prediction aims to forecast drug-drug interactions (DDI), drug-target interactions (DTI) and other drug-related interactions, which play a crucial role in potential target identification, novel drug development, and drug repurposing. In recent years, deep learning-based approaches have achieved notable progress in this domain. However, existing methods still face several challenges, including insufficient interpretability of drug substructures, unimodal encoding of target proteins, and inadequate exploitation of interaction information, which hinder further improvements in predictive performance, generalization capability, and interpretability of the results. In response to the above challenges, this dissertation conducts research on DDI prediction and DTI prediction based on deep learning technology. The main works are as follows:

(1) To tackle the challenges in current DDI prediction methods such as the poor interpretability of perceived substructures and the isolation of chemical substructures, a DDI prediction method based on graph representation learning called HDN-DDI is proposed. HDN-DDI employs hierarchical molecular graphs that incorporate atomic-level, substructure-level and molecular-level nodes to represent drug molecules, and leverages dual-view representation learning to extract both intramolecular hierarchical features and intermolecular interaction features, thereby enhancing the accuracy of DDI prediction. Experimental results indicate that HDN-DDI achieves a prediction accuracy of 97.93% in the warm-start scenario and exhibits a significant improvement of 2% to 6% in the cold-start and real-world application scenarios.

(2) To overcome the limitations associated with the inadequate utilization of two-dimensional graph structures and the insufficient mining of interaction information in existing DTI prediction methods, a DTI prediction method based on two-dimensional graph structure called GDN-DTI is proposed. This method converts the one-dimensional sequences of drugs and targets into two-dimensional graphs via a graph structure generation module and applies dual-view representation learning to capture the intramolecular structure features and intermolecular interaction features. An interaction score matrix is subsequently computed to integrate information and facilitate DTI prediction. Experimental results demonstrate that GDN-DTI attains both the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) exceeding 90% on two datasets and outperforms existing methods across multiple evaluation metrics.

(3) To address the challenges of single-modality encoding and insufficient multimodal fusion in existing DTI prediction methods, a DTI prediction method based on multimodal information fusion called MIF-DTI is proposed. MIF-DTI employs both a sequence encoding module and a graph encoding module to perform dual-view encoding of the one-dimensional sequences and two-dimensional graphs of drugs and targets. A co-attention mechanism is then employed to fuse global features across different modalities, thereby enhancing the accuracy of DTI prediction. Experimental results show that MIF-DTI achieves state-of-the-art performance across three datasets, with improvements of 1% to 3% in accuracy, AUROC, AUPR and so on.

KEY WORDS: Drug-Drug Interaction Prediction, Drug-Target Interaction Prediction, Hierarchical Molecular Graph, Contact Map, Dual-view Representation Learning

目 录

第 1 章 绪论.....	1
1.1 研究背景.....	1
1.2 研究现状.....	2
1.2.1 药物-药物相互作用预测	3
1.2.2 药物-靶点相互作用预测	5
1.3 本文主要工作.....	7
1.4 内容安排.....	9
第 2 章 相关知识介绍	11
2.1 药物的表示形式.....	11
2.2 靶点的表示形式.....	13
2.3 基于深度学习的分子编码模型	15
2.3.1 基于卷积神经网络的编码模型.....	15
2.3.2 基于注意力机制的编码模型.....	16
2.3.3 基于图注意力网络的编码模型.....	17
2.4 本章小结.....	18
第 3 章 基于图表示学习的药物-药物相互作用预测方法.....	19
3.1 引言.....	19
3.2 方法模型.....	20
3.2.1 HDN-DDI 整体框架.....	20
3.2.2 模型输入.....	20
3.2.3 子结构提取模块.....	21
3.2.4 HDN 编码器	23
3.2.5 HDN 解码器	24
3.2.6 损失函数.....	25
3.3 实验及结果分析.....	25
3.3.1 实验设置.....	25
3.3.2 热启动场景下的实验结果.....	27
3.3.3 冷启动场景下的实验结果.....	28
3.3.4 消融实验.....	29
3.3.5 实际应用场景评估.....	30
3.3.6 可视化分析.....	32
3.4 讨论.....	33
3.5 本章小结.....	34
第 4 章 基于二维图结构的药物-靶点相互作用预测方法.....	35

4.1 引言.....	35
4.2 方法模型.....	36
4.2.1 GDN-DTI 整体框架.....	36
4.2.2 二维图结构生成模块.....	36
4.2.3 编码模块输入.....	37
4.2.4 GDN 编码器.....	37
4.2.5 GDN 解码器.....	40
4.2.6 损失函数.....	41
4.3 实验及结果分析.....	41
4.3.1 实验设置.....	41
4.3.2 实验结果.....	43
4.3.3 消融实验.....	44
4.4 讨论.....	46
4.5 本章小结.....	47
第 5 章 基于多模态信息融合的药物-靶点相互作用预测方法.....	49
5.1 引言.....	49
5.2 方法模型.....	50
5.2.1 MIF-DTI 整体框架.....	50
5.2.2 序列编码模块.....	51
5.2.3 图编码模块.....	54
5.2.4 MIF 解码模块.....	54
5.2.5 损失函数.....	55
5.3 实验及结果分析.....	56
5.3.1 实验设置.....	56
5.3.2 实验结果.....	57
5.3.3 消融实验.....	58
5.4 本章小结.....	60
第 6 章 总结与展望.....	61
6.1 工作总结.....	61
6.2 未来展望.....	63
参考文献.....	64
附录 A 补充材料.....	73

插图清单

图 1.1	研究内容之间的关系	8
图 2.1	水杨酸乙酯的分子图	11
图 2.2	水杨酸乙酯的不同表示形式	12
图 2.3	靶点的不同表示形式	14
图 2.4	靶点 P35228 的二维距离图构建流程	15
图 3.1	HDN-DDI 总体框架图	20
图 3.2	子结构提取模块中药物分子的分解过程	21
图 3.3	子结构提取模块中层次化分子图的构建过程	22
图 3.4	HDN 编码块内部结构示意图	22
图 3.5	HDN-DDI 及基线模型的预测曲线	31
图 3.6	HDN-DDI 针对 3 组 DDI 的子结构识别情况	32
图 4.1	GDN-DTI 总体框架图	36
图 4.2	GDN 编码块内部结构示意图	38
图 4.3	GDN-DTI 及其变体在训练过程中 AUROC 与 AUPR 的变化曲线	45
图 4.4	靶点 P45059 的预测接触图与实际距离图的邻接矩阵	46
图 5.1	靶点 P35228 的距离图邻接矩阵及其局部放大图	49
图 5.2	MIF-DTI 总体框架图	50
图 5.3	MIF-DTI 中序列编码模块的示意图	51
图 5.4	MIF-1D 编码块的结构示意图	52
图 5.5	MIF-1D 编码块内多层卷积神经网络示意图	52
图 5.6	MIF-DTI 及其变体在训练过程中 AUROC 与 AUPR 的变化曲线	60
图 A.1	靶点 P00488 的预测接触图与实际距离图的邻接矩阵	74

附表清单

表 3.1	热启动场景下 HDN-DDI 等模型在不同数据集上的表现 (%)	27
表 3.2	冷启动场景下 HDN-DDI 等模型在 S1 和 S2 划分上的表现 (%)	29
表 3.3	冷启动场景下 HDN-DDI 及其变体在 S1 和 S2 划分上的表现 (%)	30
表 3.4	HDN-DDI 等模型在 FDA 新批准药物上的表现 (%)	31
表 4.1	GDN-DTI 等模型在 DrugBank 数据集上的表现 (%)	43
表 4.2	GDN-DTI 等模型在 BioSNAP 数据集上的表现 (%)	44
表 4.3	GDN-DTI 及其变体在 DrugBank 数据集上的表现 (%)	45
表 5.1	MIF-DTI 等模型在 DrugBank 数据集上的表现 (%)	57
表 5.2	MIF-DTI 等模型在 BioSNAP 数据集上的表现 (%)	58
表 5.3	MIF-DTI 等模型在 Davis 数据集上的表现 (%)	58
表 5.4	MIF-DTI 及其变体在 DrugBank 数据集上的表现 (%)	59
表 A.1	层次化节点的特征含义	73
表 A.2	HDN-DDI 超参数搜索空间	73
表 A.3	氨基酸节点的理化特征	74

第1章 绪论

1.1 研究背景

在疾病的预防、治疗以及恢复等各个环节，药物都发挥着不可替代的作用。药物通常是指能够调控生物体内生理过程或病理过程的化学物质^[1]，其在临床应用中具有广泛的适应症。例如，阿司匹林等抗血小板聚集药物可有效预防血栓形成^[2]，而亚硝酸戊酯等有机硝酸盐则能有效缓解心绞痛^[3]。尽管现阶段药物极大改善了人类的医疗条件，但药物治疗仍然存在以下问题：

首先，现阶段仍存在大量未知的药物-靶点相互作用。药物通常需要与特定的靶点（如蛋白质与核酸等）相结合并产生相互作用才能发挥预期疗效，即药物-靶点相互作用（Drug-Target Interaction, DTI）是药理活性的前提^[4]。截至2016年，美国食品药品监督管理局已批准了大约2110种小分子药物和4964种实验药物^[5]，而在约20 000种人类蛋白质中只有3150种蛋白质与这些药物存在相互作用^[6]，这意味着人类现有的化合物和生物活性数据库中仍存在着大量未被发现的DTI。

针对任一药物，如果不明确该药物的靶点范围，一方面会使得药物无法得到充分利用，即药物虽然具有治疗多类疾病的潜力，但实际上仅被用于治疗单种疾病^[7]，另一方面则容易使人们对药物的实际治疗效果以及副作用失去控制，即药物与预期外的靶点产生相互作用，从而引发药物不良反应（Adverse Drug Reactions, ADR）^[8]。例如，抗风湿药托法替布（Tofacitinib）在近20年内ADR报告数为146 014例，其常见的ADR如疼痛、关节痛以及疲劳等均未在药物说明书中进行标注，且该药物的ADR发生率在不同性别与不同年龄的人群中存在显著差异^[9]。与之类似，静脉麻醉药丙泊酚（Propofol）在不同个体中的药效差异较大，且静脉注射疼痛、血流动力学不稳定及呼吸抑制等ADR发生率高，其中注射疼痛ADR发生率超过60%，常见于女性患者与外周静脉注射部位^[10]。

其次，联合用药治疗中出现的药物-药物相互作用易导致药物疗效降低或引发严重副反应。在各类疾病的治疗过程中，由于疾病表达网络复杂以及靶点易突变，针对单一靶点的单药治疗效果往往十分有限，而联合用药治疗则是一种更常见的治疗模式^[11]。通过联合用药，两种或两种以上的药物可针对疾病发病过程中涉及到的多个靶点、代谢途径或细胞过程发挥作用，从而同时阻断疾病的多个表达通路，达到单一用药所不具备的治疗效果^[12]，例如减少用药剂量或减缓耐药性的发展^[13-15]。然而，由于不同药物之间存在协同或拮抗等药物-药物相互作用（Drug-Drug Interaction, DDI），联合用药有可能引发呼吸抑制或代谢减缓等副作用^[16-17]。例如，昂丹司琼（Zofran）是一种用于预防恶心和呕吐的药物，多非利特（Tikosyn）常用于稳定心律，当两种药物被同时服用时，患者可能出现心

跳停歇时间变长、头晕或昏厥等状况^[18-19]。据统计，由 DDI 引起的药物不良事件在所有药物不良事件中占比高达 30%^[20]。

综上所述，现阶段的药物治疗仍面临以下挑战：（1）大量未知的药物-靶点相互作用限制了药物的应用范围并且可能引发药物不良反应；（2）在联合用药方案中，部分药物-药物相互作用可能降低治疗效果或产生严重副作用。

针对药物安全性问题，传统研究^[21]主要依赖于药代动力学和药效学等理论框架，同时结合药物理化实验和临床试验进行评估，然而此类方法通常针对某一种药物进行研究，成本高、耗时长且不适用于大规模的药物组合^[22-24]。随着生物学数据的快速积累以及计算资源的不断优化，机器学习与深度学习方法在分子信息学领域得到了广泛应用^[25-26]，并且在药物设计^[27]以及蛋白质空间结构预测^[28]等方向上取得了突破性进展。与此同时，药物相互作用预测^[29-30]也逐渐受到越来越多学者的关注。

作为生物信息学与药物学交叉领域的重要研究方向，药物相互作用预测旨在预测药物与药物、药物与靶点以及药物与基因之间的相互作用^[4]。准确预测这些相互作用有助于探索药物的潜在作用途径，从而为新药研发和现有药物再利用等下游任务提供理论支撑^[18]。2022 年，Cheng 等人^[31]提出了一种端到端的深度学习方法 IIFDTI，该方法结合药物-靶点对的独立特征以及子结构之间的交互特征实现了较为准确的 DTI 预测。2023 年，Lin 等人^[32]将关系特征整合到药物表示中并构建出深度学习模型 R2-DDI，从而显著提升了 DDI 预测的性能。大量研究表明，利用机器学习与深度学习预测 DDI 以及 DTI 等药物相互作用具备可行性，并且具有重要的应用价值^[31-34]。

在此背景下，本文拟利用卷积神经网络（Convolutional neural network, CNN）与图注意力网络（Graph Attention Network, GAT）等深度学习方法，从 DDI 预测以及 DTI 预测两个方面对药物相互作用预测方法进行深入研究。通过构建高效的计算模型，本文期望为新药物研发、药物潜在副作用分析以及药物再利用等下游任务提供计算依据，从而提升临床用药的安全性与有效性。

1.2 研究现状

在药物相互作用预测领域，DDI 预测和 DTI 预测是两个具有代表性的任务。尽管两者的核心目标相同，即挖掘药物与其它分子之间的相互作用信息，但其编码对象和任务复杂性存在显著差异^[4]。具体而言，DDI 预测任务仅涉及药物小分子的编码学习，而 DTI 预测任务需要同时对药物小分子以及蛋白质大分子进行编码，即 DTI 预测任务更为复杂。基于此，现有工作通常针对其中一项任务展开研究，因此本节分别阐述 DDI 预测领域与 DTI 预测领域的研究现状。

1.2.1 药物-药物相互作用预测

DDI 预测在识别联合用药中潜在的药物不良反应、完善药物数据库以及提升联合治疗的安全性等方面具有重要意义。近年来,基于机器学习和深度学习的方法在 DDI 预测领域取得了显著的进展^[35-37]。现有的 DDI 预测方法主要可以归纳为以下四类:基于网络拓扑关系的预测方法、基于自然语言处理(Natural Language Processing, NLP)的文本挖掘方法、基于分子化学结构的预测方法以及融合多种策略的预测方法^[18-19,38]。

基于网络的方法通常以标签传播、连接预测及多源知识融合等方式进行 DDI 预测^[18]。2019 年, Karim 等人^[39]提出 KG-DDI, 该模型采用多种嵌入方法将节点表示嵌入到预先构建的知识图谱中。2021 年, Yu 等人^[40]提出 SumGNN, 该模型利用子图提取模块对知识图谱的相关子图进行锚定, 并通过自注意力机制进行子图摘要及路径推理, 最后利用多通道集成模块实现外部知识集成, 从而预测不同类型的 DDI。多项研究^[39-41]表明, 基于网络的 DDI 预测方法能够有效利用药物的外部生物学知识, 从而显著改善 DDI 预测模型的性能。然而, 由于新型药物往往缺乏相应的拓扑关系及外部知识, 该类方法在预测涉及新型药物的 DDI 时存在较大局限性。

基于 NLP 的方法通常依托生物学文本挖掘技术进行 DDI 预测^[18]。2020 年, Lee 等人^[42]利用生物学领域的文本语料对 BERT 模型进行训练, 首次将预训练模型引入生物学领域。2022 年, He 等人^[43]提出 3DGT-DDI, 该模型通过三维图神经网络挖掘分子的三维结构信息, 通过预训练文本注意力机制捕获分子的位置信息, 从而提升了 DDI 预测的准确性。实验结果表明, 3DGT-DDI 的 F1 值可达到 84.48%, 优于此前的最优模型。基于 NLP 的方法主要应用于生物学文献挖掘以及 DDI 数据库的构建, 即从已有的非结构化文本中自动提取结构化的 DDI 信息, 该类方法同样难以有效预测涉及新型药物的 DDI。

基于分子结构的方法主要通过相似性分析、一维序列表示或分子图结构进行 DDI 预测。基于相似性的方法假设具有相似结构的药物可能表现出相似的 DDI, 并通过计算分子在结构、表型、功能及副作用等方面的相似性来构建预测模型^[19]。例如, 2018 年, Ryu 等人^[44]提出 DeepDDI, 该模型利用结构相似性计算模块与深度神经网络对 DDI 进行预测。2020 年, Deng 等人^[35]提出 DDIDML, 该模型首先使用四种类型的药物特征来计算药物之间的相似性, 随后采用神经网络提取药物的跨模态特征, 最终对 DDI 进行分类。

基于序列的方法通常针对药物的一维表示例如简化分子输入行系统(Simplified Molecular Input Line Entry System, SMILES)序列与扩展连通性指纹(Extended Connectivity Fingerprints, ECFP)等进行编码, 从而实现 DDI 预测。2018 年,

Preuer 等人^[45]利用多层感知机对药物的 ECFP 表示以及癌细胞系的表达特征进行编码,从而预测药物之间的协同情况。2023 年, Yang 等人^[46]提出 CNN-Siam, 该方法利用一维卷积神经网络及残差连接对药物的 SMILES 序列进行编码, 其 DDI 预测性能优于基线模型。该类方法对模型的学习能力要求较低, 因此能够利用简单的机器学习模型如随机森林与多层感知机等进行预测^[19]。然而, 由于 SMILES 序列和 ECFP 表示等一维表示方式所包含的信息有限, 基于序列的方法在模型性能方面存在瓶颈。

由于药物分子可表示为由原子和化学键组成的分子图, 其中原子为节点, 化学键为边, 基于分子图的 DDI 预测方法^[32,47-50]逐渐成为主流方法。例如, 2019 年, Xu 等人^[49]提出 MR-GNN, 该方法利用图神经网络对分子图进行编码, 并结合双向长短期记忆网络对不同深度的分子图表示进行序列建模, 最终通过神经网络融合分子图编码与序列编码并实现 DDI 预测。2023 年, Lin 等人^[32]提出 R2-DDI, 该方法利用 DeeperGCN^[51]对分子图进行编码并将关系特征融入分子表示, 同时采用多项特征损失函数以确保模型训练的一致性。实验结果表明, 该方法的表现优于 MR-GNN^[49]和 SSI-DDI^[52]等已有模型。

药物化学领域的相关研究表明, 药物分子由不同官能团或化学子结构组成, 这些官能团及子结构决定了药物的药代动力学特性与药效学特性, 并最终决定药物之间的相互作用^[53]。受此启发, 基于药物子结构的 DDI 预测方法^[52,54-55]被陆续提出。2021 年, Nyamabo 等人^[52]提出 SSI-DDI, 该模型利用多层图注意力网络对分子图进行编码并提取分子中的子结构, 同时利用协注意力机制预测子结构之间的相互作用, 从而实现子结构级别的 DDI 预测。2022 年, Nyamabo 等人^[54]利用门控信息传递网络 GMPNN 从药物的分子图表示中学习不同大小和形状的化学子结构并基于这些子结构进行 DDI 预测。2023 年, Li 等人^[55]提出一种用于 DDI 预测的双视图表示学习框架 DSN-DDI, 该框架通过迭代使用局部和全局表示学习模块从单个药物和药物对中学习子结构, 其识别准确率相较于已有模型提升约 4%。在双视图表示学习框架的基础上, Ning 等人^[56]通过双线性网络捕获原子级相互作用信息, Shen 等人^[57]针对热启动与冷启动场景设计不同的双视图策略, 进一步提升了模型预测表现。

基于子结构的 DDI 预测方法能够有效利用药物分子中的子结构信息改善模型的预测性能。由于该类方法直接对分子图进行编码, 无需依赖药物的其它特征, 因此其能够更好地预测涉及新药物的 DDI。近年来, DSN-DDI^[55]与 BDN-DDI^[56]等多个模型在热启动和冷启动场景中均取得了更好的预测表现, 这进一步验证了该类方法的有效性。然而, 现有方法仍然存在一些不足, 例如 SSI-DDI 和 SA-DDI^[58]模型以图神经网络输出的节点表示作为子结构表示, 而 GMPNN-CS^[54]与 DSN-DDI 模型则以池化后的全局图表示作为子结构表示。尽管这些方

法在一定程度上提升了模型的可解释性，但其学习到的子结构往往缺乏明确的化学意义，仍需要额外的人工解释。

针对这一问题，2023年，Fan等人^[12]提出PMSE-DDI，该模型利用化学子结构分解识别（Breakdown Identification of Chemical Substructures, BRICS）算法将药物分解为多个化学子结构，并分别对药物分子以及分解所得的子结构进行编码，从而实现DDI预测。不同于已有方法，PMSE-DDI首次显式利用药物的化学子结构进行DDI预测，在提升模型可解释性的同时保持了子结构的完整性。然而，该方法在编码过程中未能充分学习子结构之间的邻接信息与相互作用信息，预测性能有待进一步提升。因此，如何利用具备明确化学意义的子结构并充分挖掘子结构之间的相互作用信息，从而实现更精确的DDI预测，仍然是值得深入研究和改进的方向。

1.2.2 药物-靶点相互作用预测

药物-靶点相互作用预测（即DTI预测）是药物设计以及药物重定位中的核心任务，其旨在识别药物与靶点（如蛋白质与基因）之间的相互作用。近年来，药物与蛋白质靶点相互作用预测以及药物与基因靶点相互作用预测逐渐被视为两个独立的研究方向^[4]。目前，DTI预测通常特指药物与蛋白质靶点之间的相互作用预测^[59-60]，而药物与基因靶点之间的相互作用预测则被归类为药物-基因相互作用（Drug-Gene Interaction, DGI）预测^[61-62]。本文所研究的DTI预测问题主要聚焦于预测药物与蛋白质靶点之间的相互作用。

随着基因组学与蛋白质组学等生物学数据的不断积累以及计算能力的持续提升，计算方法已成为药物研发中识别DTI的关键技术之一^[63]。计算机模拟方法能够显著缩小候选化合物的搜索范围，从而降低研究成本，加速药物研发，并为深入理解药物可能引发的潜在副作用提供参考依据^[64]。

目前，DTI预测方法可基本归纳为三类：基于配体的方法、基于结构的方法和基于机器学习与深度学习的方法^[23,59]。基于配体的方法通常基于相似性假设，即如果两个配体的化学性质相似，则它们的作用靶点也较为相似^[65]。具体而言，如果已知药物A与靶点B之间存在相互作用，则与A相似的药物可能与靶点B之间也存在相互作用；同理，与B相似的靶点可能与药物A之间存在相互作用；进一步而言，与A相似的药物可能和与B相似的靶点之间存在相互作用。该方法依赖于药物与靶点的先验知识以及合理的相似性度量，当某一靶点的已知配体数量不足时，则无法准确预测出对应的配体^[66]。

基于结构的方法利用化合物和靶点的三维结构来识别DTI，例如分子对接、分子动力学模拟以及结合自由能预测等^[67]。1982年，Kuntz等人^[68]首次提出DOCK，其利用几何形状匹配算法评估药物与靶点的结合情况。该软件能够对结

合情况进行可视化,自推出后受到较多使用,至今已更新至 6.11 版本^[69]。2024 年, Milon 等人^[70]利用三角空间关系键表征药物与靶点的三维结构并计算键的交叉情况,从而对 DTI 以及药靶结合位点进行预测。尽管该类方法能够实现较为准确的预测,但其仅适用于三维结构已知的靶点,无法适用于 G 蛋白偶联受体等缺乏结构结晶的靶点^[71],并且计算耗时较长,因此其应用范围依然有限。

基于机器学习与深度学习的方法通常包括两项步骤:特征提取和相互作用预测,即首先将药物和靶点的结构表示分别转化为特征向量,随后再进行模型训练及预测。部分研究利用深度学习对相似性矩阵^[64]与三维空间结构^[72]进行编码,然而大多数研究仍以药物与靶点的序列表示和图表示作为信息源^[71]。

早期基于深度学习的方法通常从药物的 SMILES 序列、ECFP 表示以及靶点的氨基酸序列中提取信息,并且通过浅层神经网络实现 DTI 预测。2018 年, Ozturk 等人^[73]提出 DeepDTA,该模型使用 CNN 对药物的 SMILES 序列以及靶点的氨基酸序列进行编码并通过全连接层实现 DTI 预测。2019 年, Lee 等人^[74]提出 DeepConv-DTI,该模型利用 CNN 对 ECFP 表示以及氨基酸序列进行编码,实现了更准确的 DTI 预测。

随着注意力机制与 Transformer^[75]等序列编码模型的提出,基于深度学习的方法逐渐侧重于药物与靶点的分解以及相互作用信息的挖掘,从而改善方法表现。2021 年, Huang 等人^[76]提出 MolTrans,该方法通过无标签数据构建药物与靶点的片段库,随后通过 Transformer 对片段序列进行编码并计算片段之间的相互作用矩阵,最终利用 CNN 进一步池化并通过全连接层完成预测。2022 年, Zhao 等人^[77]提出 HyperAttentionDTI,该模型通过两个 CNN 块分别提取药物和蛋白质特征,利用超注意力机制捕捉相互作用信息并利用全连接层实现预测。2023 年, Dou 等人^[78]提出 BCM-DTI,该模型利用支链挖掘算法将药物分解为若干子结构,利用类别片段挖掘算法将靶点分解为多个氨基酸片段,最终利用两个 CNN 模块分别学习子结构与片段的特征。2024 年, Zhang 等人^[59]进一步提出 FMCA-DTI,该模型利用碎片挖掘算法与多头交叉注意力机制学习药物与靶点之间的相互信息。上述多项研究^[59,73-74,76-78]表明, CNN 与 Transformer 等深度学习方法能够有效应用于药物序列与靶点序列的编码,从而提升 DTI 预测表现。

近年来,一些研究^[60,72,79]开始对药物与靶点的二维图或三维结构进行编码,从而充分利用其拓扑结构和空间构象。在分子信息学领域,二维图是一类介于一维序列与三维结构之间的表示形式^[80]。与 SMILES 序列等一维序列(由原子或氨基酸构成的序列)相比,二维图可以反映更多的拓扑结构信息,例如分子中的苯环;与三维结构相比,二维图对计算与存储资源的需求更低,能够以更高的效率表征药物与靶点的空间构象^[79]。常见的二维图表示形式包括药物的分子图^[18]、靶点的氨基酸距离图^[80]以及氨基酸接触图^[81]等。

2020年, Zheng等人^[80]提出 DrugVQA, 该模型利用长短期记忆网络从药物的 SMILES 序列中提取特征, 利用二维动态 CNN 从靶点的氨基酸距离图中提取特征, 从而实现 DTI 预测。2022年, Li 等人^[82]提出 MINN-DTI, 该模型利用消息传递网络对药物的分子图进行编码, 利用靶点预处理网络对靶点的氨基酸距离图进行编码, 随后通过 Transformer 捕获两者的相互作用信息。2024年, Koh 等人^[79]提出 PSICHIC, 该模型采用图神经网络对药物的分子图与靶点的预测接触图进行编码, 同时采用注意力机制学习相互作用特征。相关实验^[79-80,82]表明, 相比于基于序列的方法, 基于二维图结构的 DTI 预测方法可以更有效地地捕获拓扑结构信息, 因此能够在 DUD-E^[83]等大规模数据集上取得更优表现。

尽管基于深度序列的方法能够有效从一维序列和二维图中提取信息, 但现有方法通常只利用单种表示形式实现 DTI 预测。一方面, 单一信息源所表征的信息有限。例如, 一维序列无法体现药物与靶点的拓扑结构与三维构象, 而三维结构无法反映两者的功能属性及作用机制, 这在一定程度上增大了下游模型的预测难度。另一方面, 当关键信息缺失时, 相关方法的预测准确度和适用性将受到显著影响。例如, 对于新出现的靶点, 基于配体的方法缺乏足够的已知配体, 基于三维结构的方法缺乏三维构象, 两类方法均难以实现准确预测。因此, 如何有效融合序列信息、图结构信息、三维结构信息以及外部知识表示, 从而提升 DTI 预测方法的准确性和泛化能力, 已然成为 DTI 预测领域一项新的挑战^[64,84]。

1.3 本文主要工作

如上文所述, DDI 预测和 DTI 预测是药物相互作用预测领域中两个具有代表性的任务。尽管两者在具体目标与任务难度上存在差异, 但均涉及分子表示学习与相互作用信息挖掘, 因此可采用统一的框架进行研究。

针对药物相互作用预测领域存在的问题与挑战, 本研究利用图注意力网络与注意力机制等深度学习方法构建了一个适用于 DDI 预测和 DTI 预测的计算框架。首先, 本文提出了一种基于图表示学习的 DDI 预测方法 HDN-DDI。随后, 在 HDN-DDI 的基础上, 本文将双视图表示学习方法迁移至 DTI 预测领域, 并采用靶点二维图这一更有效的表示形式, 从而给出了一种基于二维图结构的预测方法 GDN-DTI。最终, 在 GDN-DTI 的基础上, 本文进一步引入一维序列编码, 通过协注意力机制实现不同模态药物与靶点表示的融合, 从而给出了一种基于多模态信息融合的预测方法 MIF-DTI。整体研究思路以及各部分研究内容之间的关系如图1.1所示。各部分的主要研究内容如下:

(1) 基于图表示学习的药物-药物相互作用预测。现有的大多数基于子结构的 DDI 预测方法^[52,54-55,58]以数据驱动的方式从药物分子中提取虚拟的子结构,

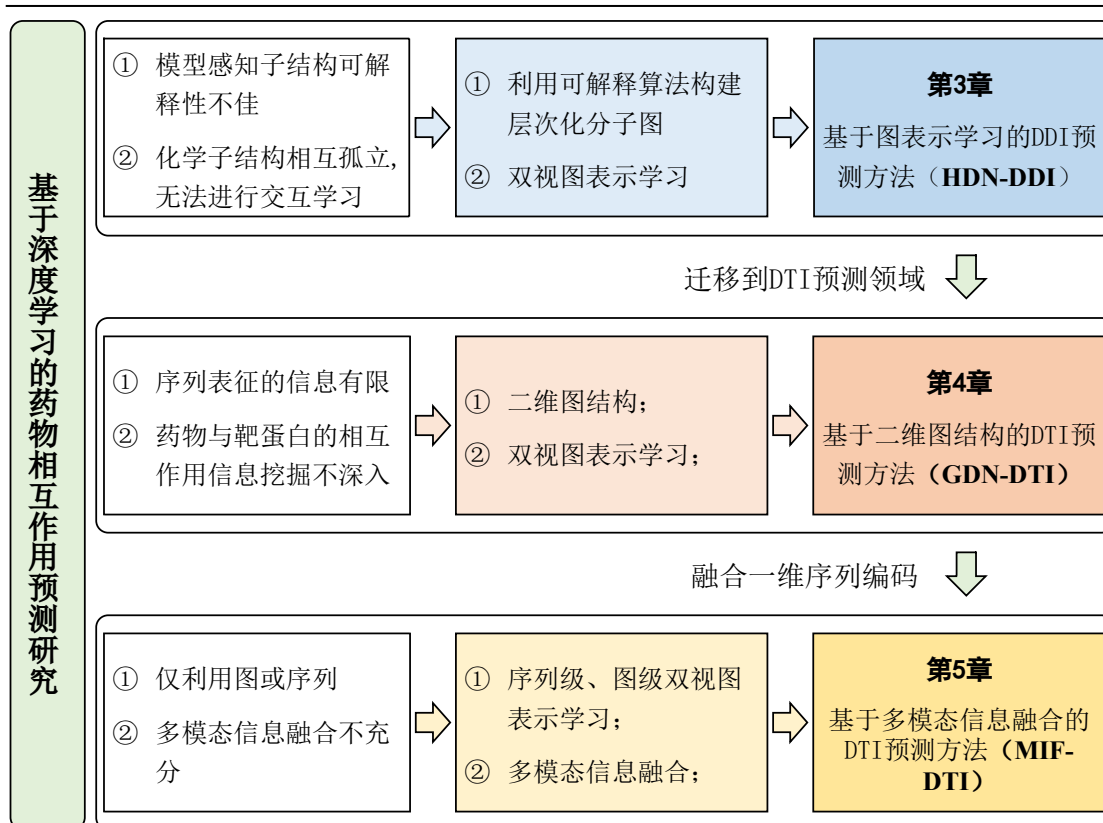


图 1.1 研究内容之间的关系

但该类子结构往往缺乏化学可解释性，仍需要额外的人工解析。此外，部分方法^[12]利用可解释的算法将药物分解为多个真实的化学子结构，然而此类方法未考虑子结构之间的关联性，模型仅能学习子结构内部的信息而无法捕捉子结构之间的交互特征。针对现有方法存在的感知子结构可解释性不佳、化学子结构相互孤立的问题，本文提出了一种基于层次化分子图与双视图表示学习的 DDI 预测方法，即 HDN-DDI。该方法命名含义如下：H 表示层次化分子图（Hierarchical Molecular Graphs），D 表示双视图表示学习（Dual-view Representation Learning），N 表示网络（Network），DDI 则表示药物-药物相互作用（Drug-Drug Interaction）。HDN-DDI 以两个给定的药物分子为输入，首先通过可解释的子结构提取模块将药物分解为多个实际存在的化学子结构，从而构建“原子-子结构-分子”层次化分子图，随后通过层次化视图层获取药物分子的层次化表示，通过相互作用视图层获取药物分子的相互作用表示，最终通过协注意力机制融合不同深度的分子表示，从而实现 DDI 预测。实验结果表明，HDN-DDI 在不同场景下均能达到最优表现，并且在热启动场景中的预测准确率达到了 97.93% 与 99.38%。

（2）基于二维图结构的药物-靶点相互作用预测。当前主流的 DTI 预测方法^[63,76,85]通常从药物与靶点的一维序列中提取特征。由于序列包含的信息有限，此类方法难以从序列中捕获药物与靶点的拓扑结构信息。为克服该问题，部分研究^[80,82]引入二维图结构以增强表征能力，但其侧重于药物与靶点的内部结构

信息，而未充分挖掘分子间的相互作用信息。针对现有方法存在的序列表征信息不足、相互作用信息挖掘不充分的问题，本文在 HDN-DDI 的基础上给出了一种基于二维图结构与双视图表示学习的 DTI 预测模型 GDN-DTI，其中 G 代表图 (Graphs)，D 代表双视图表示学习 (Dual-view Representation Learning)，N 表示网络 (Network)，DTI 表示药物-靶点相互作用 (Drug-Target Interaction)。在具体实现层面，GDN-DTI 以药物的 SMILES 序列与靶点的氨基酸序列为输入，首先通过二维图结构生成模块构建药物的层次化分子图和靶点的二维接触图，随后通过双视图表示学习方法捕获分子内结构信息与分子间相互作用信息，并通过药物-靶点相互作用得分矩阵实现多尺度全局表征融合，最终利用全连接层计算 DTI 概率。实验结果表明，GDN-DTI 在两个数据集上均达到了 83% 的预测准确率，并且在多项评估指标上优于现有最佳模型。

(3) 基于多模态信息融合的药物-靶点相互作用预测。目前大部分 DTI 预测方法^[59,63,78]聚焦于一维序列编码，同时也有部分方法^[79-80,82]尝试从药物分子图与靶点接触图等二维图以及药物与靶点的三维结构中挖掘关键信息。一维序列和二维图等不同的表示形式在表示效率与信息量等方面各有优势^[71]，然而大多数工作仅关注某一模态的表示形式，涉及不同模态信息融合的研究仍然较少。针对现有方法存在的编码信息源单一、多模态信息融合不足的问题，本文在 GDN-DTI 的基础上进一步给出了一种基于多模态信息融合 (Multimodal Information Fusion, MIF) 的 DTI 预测模型，即 MIF-DTI。该模型通过序列编码模块对药物的 SMILES 序列与靶点的氨基酸序列进行编码，从而学习序列信息与序列级相互作用信息，同时通过图编码模块构建药物的层次化分子图与靶点的二维接触图并对其进行编码，从而学习图结构信息与图级相互作用信息，最终通过协注意力机制以及相互作用得分矩阵对不同模态与不同深度的全局表示进行融合，从而实现 DTI 预测。实验结果表明，MIF-DTI 在 3 个数据集上均达到了最优表现，并且在准确率、AUROC 以及 AUPR 等多项指标上实现了 1% 至 3% 的有效提升。消融实验进一步验证了序列编码模块、图编码模块以及 MIF 解码模块的有效性。

1.4 内容安排

第 1 章：绪论。本章首先概述了当前药物治疗面临的主要挑战，介绍了药物相互作用预测的研究背景，随后总结了药物-药物相互作用预测和药物-靶点相互作用预测领域的研究进展，归纳了现有方法存在的问题，最终阐述了全文的研究思路、主要研究内容以及论文的内容安排。

第 2 章：相关知识介绍。本研究首先介绍了药物的 SMILES 序列、分子图以及三维结构等常见的药物表示形式，随后介绍了靶点的氨基酸序列、距离图与三

维结构等常见的靶点表示形式，最后介绍了本研究所采用的卷积神经网络与图注意力网络等深度学习方法，从而为读者提供必要的知识铺垫。

第3章：基于图表示学习的药物-药物相互作用预测。本章首先回顾了基于子结构的DDI预测方法，指出了现有方法存在的不足，在此基础上提出了一种基于图表示学习的DDI预测方法HDN-DDI，并详细阐述了HDN-DDI的整体架构以及各模块原理。在实验环节，本章在不同场景下将HDN-DDI与多个基线模型进行了详细对比，随后通过消融实验进一步验证了HDN-DDI中层次化分子图以及双视图表示学习的有效性。此外，本章结合实际DDI案例对HDN-DDI的预测过程进行了可视化分析。最后，本章对HDN-DDI的优势与局限性进行了讨论，并为今后的DDI预测研究指出了一些改进方向。

第4章：基于二维图结构的药物-靶点相互作用预测。本章首先分析了基于一维序列与二维图的DTI预测方法，分析了现有方法的局限性，在此基础上给出了一种基于二维图结构与双视图表示学习的DTI预测方法GDN-DTI，并详细介绍了GDN-DTI的整体架构以及其中的二维图结构生成模块、GDN编码器和GDN解码器的设计原理。在实验环节，本章在两个数据集上对GDN-DTI与多个基线模型进行对比，并通过消融实验进一步分析了GDN-DTI中各模块的作用。此外，本章还深入分析了GDN-DTI与基线模型的性能差异及其原因，并初步探讨了二维接触图与一维序列的关系，为后续研究提供了一定的参考依据。

第5章：基于多模态信息融合的药物-靶点相互作用预测。本章首先分析了现有的基于序列与图的DTI预测方法的不足，并探讨了二维图表示与一维序列表示的各自优势及其互补性，在此基础上给出了一种基于双视图表示学习与多模态信息融合的DTI预测方法MIF-DTI，并详细介绍了MIF-DTI的整体架构以及序列编码模块、图编码模块和MIF解码模块的方法原理。最后，本章在三个数据集上对MIF-DTI与多个基线模型进行了详细对比，并通过消融研究进一步验证了MIF-DTI中各模块的有效性。

第6章：总结与展望。本章对全文的研究内容、方法及实验结论进行了系统性总结，并结合本研究对药物相互作用领域未来的工作方向进行了展望。

第2章 相关知识介绍

药物和靶点可以通过多种形式进行表示，包括一维序列、二维图结构以及三维空间构型等。不同的表示方式各有优势，并且在一定程度上影响着模型的预测性能与泛化能力^[19]。本研究利用了药物的简化分子输入行系统（Simplified Molecular Input Line Entry System, SMILES）表示、层次化分子图以及靶点的氨基酸序列、氨基酸接触图等多种表示形式，并且采用了卷积神经网络与图注意力网络等多种深度学习模型实现药物与靶点的分子编码。为增强研究的逻辑性，本章将分别从常见的药物表示形式、靶点表示形式以及相关的深度学习模型三个方面系统介绍相关基础知识。

2.1 药物的表示形式

在药物相互作用预测与药物设计等与药物相关的任务中，药物的表示形式对模型编码与预测具有重要影响^[19,38]。常见的药物表示形式包括一维序列、分子图、三维结构、“药物-靶点-疾病”同源网络以及“药物-药物”相似性嵌入等，其中前三种形式侧重于反映药物内部的结构信息，其获取方式更便捷，适用范围更广，而同源网络与相似性嵌入则着重反映药物的作用机制（例如药物的靶点范围、适用疾病以及与其它药物的相似度），其具备更丰富的语义信息，但高度依赖于现有的医学知识库，应用场景存在限制^[19]。本研究聚焦于利用结构信息进行药物相互作用预测，以确保所提出的方法能够应用于新型药物，因此本节着重叙述前三类表示形式。

受分子特性的影响，二维图（即分子图）成为最常见且直观的药物表示形式。在分子图中，药物分子中的原子被视为节点，原子之间的化学键被视为边，如图2.1所示。进一步而言，根据原子的理化特性为图中的每个节点构建特征嵌入，根据化学键的类型（例如单键、双键或三键等）为图中的边设定类型，根据原子的邻接关系建立邻接矩阵，即可将分子图转化为可编码的图，并通过图神经网络与图注意力网络等深度学习模型进行编码。

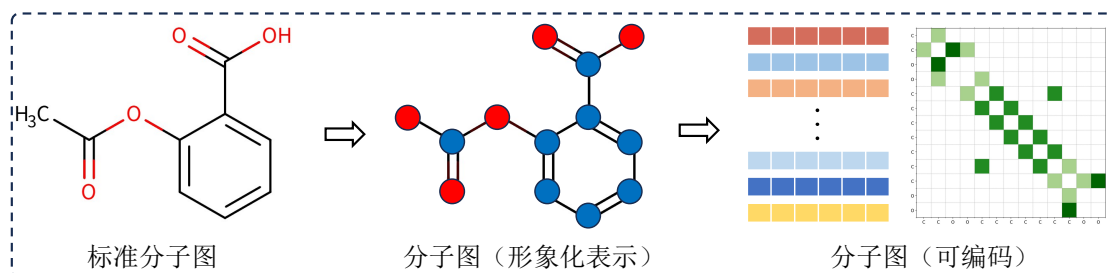


图 2.1 水杨酸乙酯的分子图

在常规分子图的基础上，Zhang 等人^[86]利用 BRICS 算法分解药物分子，进一步提出了子结构树；Zang 等人^[87]则提出了层次化分子表示。相较于常规分子图，子结构树能够表征药物分子中的子结构组成以及子结构之间的邻接关系，层次化分子表示同时包含了节点级、子结构级与图级共三个层次的表示。实验结果表明，在分子图的基础上进一步整合子结构的信息，可以降低模型训练的难度，同时改善模型的编码效率，提升模型性能^[86-87]。

分子图能够直观反映药物分子的结构，然而其不便于存储与传输。相比于分子图，SMILES 序列等一维表示形式的信息表征效率更高，因此同样被广泛使用。就定义而言，SMILES 序列即简化分子线性输入规范，其初步形成于 20 世纪 80 年代，最终在 2007 年形成开放标准^[88]。在 SMILES 序列中，所有字符均为美国信息交换标准码（American Standard Code for Information Interchange, ASCII）。具体而言，氢原子与单键不予显示，其余原子通过 ASCII 码进行表示，双键采用等号表示，分支采用括号表示，环的起点与终点则采用数字表示^[19]。遵循该标准，通过对分子图进行深度优先遍历，即可生成对应的 SMILES 序列^[88]。

以水杨酸乙酯为例，以该分子最左侧的甲基“C”为起点，依次遍历羟基“C(=O)”、氧原子“O”、苯环“C1=CC=CC=C1”以及羧基“C(O)=O”，即可形成完整的 SMILES 序列“CC(=O)OC1=CC=CC=C1C(O)=O”，如图 2.2 所示。在获取 SMILES 序列后，可采取与文本序列相同的方式进行编码，即获取字符嵌入，随后利用 CNN 与 Transformer 编码器^[75]等方法进行编码。

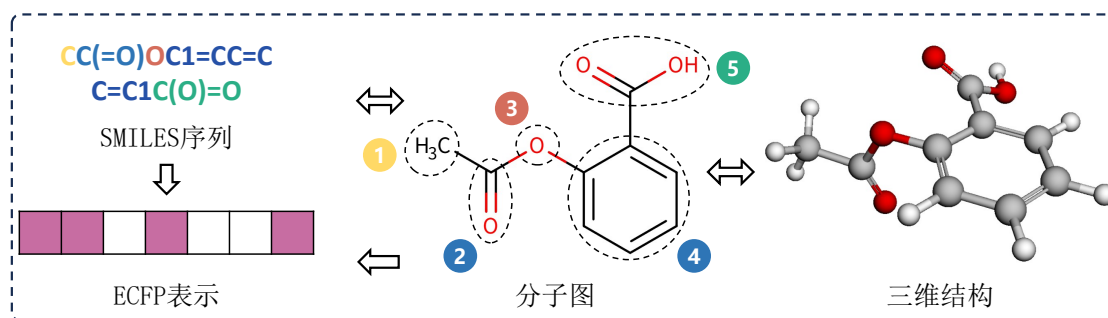


图 2.2 水杨酸乙酯的不同表示形式

除了 SMILES 序列，常见的药物分子一维序列还包括自引用嵌入式字符串（SELF-referencing Embedded Strings, SELFIES）与 SMILES 任意目标规范（SMILES Arbitrary Target Specification, SMARTS）等，其中 SELFIES 序列的表示规则更复杂，因此能够解决 SMILES 易生成非法字符串的问题，并且对药物分子的区分度更高^[89]；SMARTS 序列为用于子结构搜索的扩展 SMILES 表示，其常用于模糊查询以及化学数据库搜索^[90]。此外，扩展性连接指纹（Extended-Connectivity Fingerprint, ECFP）与 RDKit 指纹^[91]也是常见的分子线性表示。这类线性表示通常根据分子结构提取得出，信息表征能力更强，例如 ECFP 是由原

子邻域信息生成的分子指纹，常用于机器学习任务^[45]，但该类表示在生成过程中存在信息损失，无法根据分子指纹还原完整的分子结构^[88]。

一维序列与分子图可以相对高效地表征分子的结构信息，而药物的三维结构可以更全面地反映药物的空间构象信息（例如化学键的扭角与分子的手性）。如图2.2所示，药物的三维结构与分子图较为相似，但前者包含了原子的三维坐标，能够进一步体现空间信息，因此在药靶结合位点预测与靶向药物设计等任务中的应用较多。虽然药物的三维结构所表征的信息总量最多，但其需要更多的存储空间，获取难度相对较大，并且对计算资源以及编码模型的要求更高，因此其在药物相互作用预测领域尚未得到充分应用^[18]。

总体而言，SMILES 序列、分子图与三维结构可从不同层面表征药物的结构信息，并且三者各具优势。其中，分子图是最直观的表达形式。相比于 SMILES 序列，分子图能够反映药物分子的拓扑结构信息（即原子的邻接关系）；相比于三维结构，分子图的表示与编码难度更低，因此在药物相互作用预测任务中得到了广泛应用^[54-56,79]。SMILES 序列可通过 RDKit 等工具^[91]对分子图自动生成，相比于 SELFIES 序列，其规则更为简洁，信息表征效率更高；与 ECFP 等分子指纹相比，其能够完整还原分子结构，因此在数据集构建、存储与传输中得到较多应用。三维结构尽管能够全面反映分子的空间构型，但其获取过程复杂，编码难度较大，且表征效率相对较低，因此应用场景受限。

在此背景下，为保证方法的准确性与泛化能力，本研究拟采用 SMILES 序列与分子图对药物分子进行表征，分别通过卷积神经网络与图注意力网络进行编码，并在 DTI 预测任务中进一步探索不同分子表示方式的融合。

2.2 靶点的表示形式

在 DTI 预测与蛋白质空间结构预测^[81]等与蛋白质相关的任务中，蛋白质的表示始终是一项重要的环节^[71,92]。与药物分子相似，常见的蛋白质靶点表示形式同样包括一维序列、二维图、三维结构以及“药物-靶点”、“基因-靶点”与“靶点-疾病”同源网络。其中，前三类表示只能表征靶点的结构信息，但获取方式相对简单，对应方法的迁移能力较好。同源网络可以表征靶点的配体信息与关联疾病等生物学知识，能够为编码模型提供更多信息，但高度依赖于同源知识库，对应方法在面对缺乏同源知识的新型靶点时无法达到预期作用，例如基于配体的 DTI 预测方法迁移难度高，并且无法预测涉及新型靶点的 DTI^[71]。本节将着重叙述与靶点结构相关的前三类表示形式，其关系如图2.3所示。

蛋白质由大量氨基酸组成，它们决定了蛋白质的空间结构，进而影响蛋白质的功能^[93]。此外，氨基酸序列作为一种线性表示，易于存储与传输，因此是目

前最常使用的蛋白质表示形式。例如，BioSNAP 等数据集^[6,94-95]以氨基酸序列的形式记录蛋白质靶点，而 DeepDTA 等模型^[63,73-74]从氨基酸序列中提取信息。氨基酸序列反映了氨基酸的线性排列顺序，其中每个字符均表示一种氨基酸，例如 A 表示丙氨酸（Alanine），E 表示谷氨酸（Glutamic acid）。该序列的优势在于信息表征效率高，获取难度低，并且编码方法成熟，可通过 CNN 与 Transformer 等模型进行编码，因此广泛应用于 DTI 预测等与蛋白质相关的任务^[60,77]。

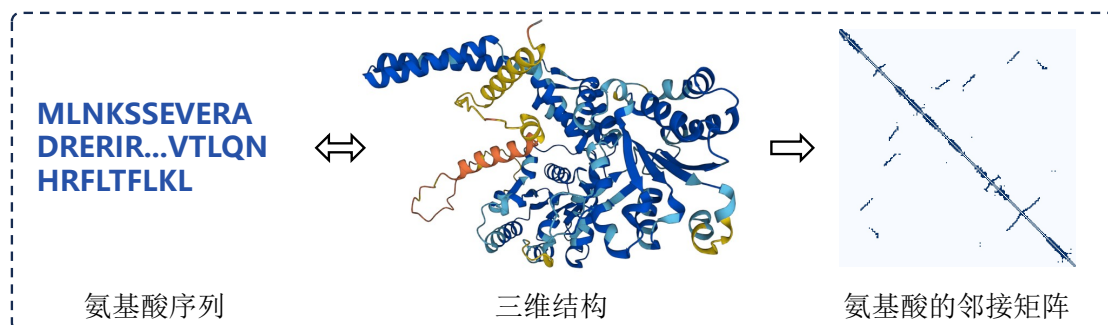


图 2.3 靶点的不同表示形式

除了一级结构（即氨基酸序列），蛋白质的功能还受到二级结构（如 α -螺旋与 β -折叠等）、三级结构（即亚基）以及四级结构（即多个亚基组成的复合体）等空间构象的显著影响，这些空间构象可统一视为蛋白质的三维结构^[93]。近年来，随着计算方法与硬件设施不断改进，AlphaFold^[28]与 ESM-2^[81]等蛋白质结构预测模型陆续出现，蛋白质的三维结构逐渐受到更多关注。

蛋白质的三维结构通常以点云的形式呈现，此外也有三维嵌入与 α 形状纵横比等形式^[92]。在点云中，每个氨基酸被视为一个节点，并且具有特定的标识与三维坐标。相比于其它形式，三维结构表征的结构信息最为全面，但其对存储空间需求更大，主要用于药物与靶点的模拟对接^[72,92]。由于蛋白质靶点通常包含数百个甚至数千个氨基酸，使用深度学习模型对三维结构进行编码的难度较大，同时现阶段仍存在 G 蛋白偶联受体等缺乏结构结晶的靶点，因此蛋白质的三维结构在基于深度学习的 DTI 预测领域尚未得到充分利用^[71,92]。

根据蛋白质的三维结构计算 α -碳原子或相邻残基之间的空间距离，即可得到氨基酸的距离矩阵^[92]，通过指定的阈值来判别两个氨基酸是否相邻，可以进一步得到邻接矩阵。随后，将氨基酸视为节点，并根据邻接情况构建边，即可获取蛋白质的二维表示形式（蛋白质距离图），如图2.4所示。相较于氨基酸序列，蛋白质距离图能够表征更多的空间结构信息；相较于三维结构，其具有更高的信息表征效率，因此目前已有部分工作尝试基于距离图进行蛋白质表示学习^[80,82]。

蛋白质距离图虽然具备一定的优势，但仍然需要根据三维结构计算得出。如果蛋白质的三维结构未知，则其二维距离图同样无法确定。针对该问题，Lin 等人^[81]提出了 ESM-2 模型，该模型在由氨基酸序列预测蛋白质三维结构的过程中

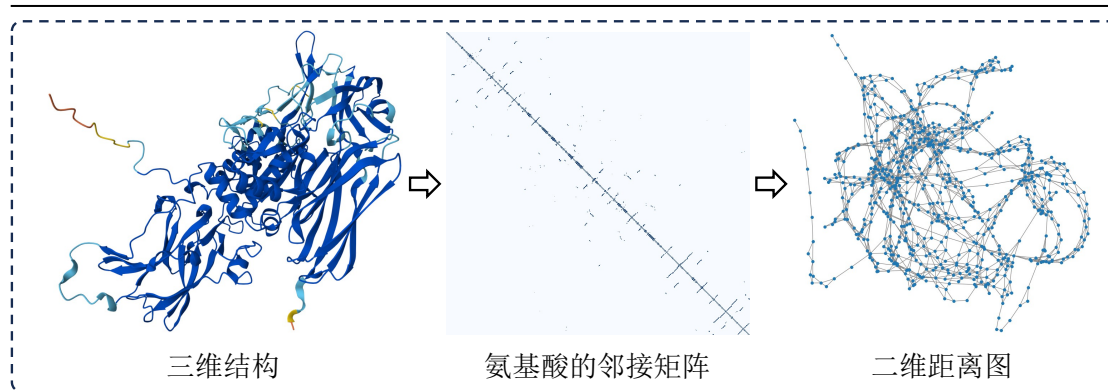


图 2.4 靶点 P35228 的二维距离图构建流程

能够同时产生对应的预测接触图。接触图与实际距离图本质相同，两者均反映了氨基酸之间的邻接情况。相关实验结果表明，ESM-2 模型预测生成的接触图同样能反映蛋白质的三维结构^[79,81]。相比于距离图，预测接触图可根据一维序列直接生成，获取方式更为简便。

总体而言，氨基酸序列、二维图与三维结构均可表征靶点的结构信息，三者各具优势。氨基酸序列是最常见的表示形式，其易于存储与编码，并且获取难度低，对应方法的迁移能力更好，但序列表征信息相对有限^[96]。三维结构能够完整表征靶点的空间构象，然而其获取难度与编码难度较高，现有模型难以对其有效编码。二维图能够兼顾表征信息量与表征信息效率，并且可通过计算氨基酸之间的空间距离或通过模型预测而获得，获取方式相对便捷。此外，二维图可采用图神经网络等模型进行编码，因此其编码难度低于三维结构。

在此背景下，本研究拟采用氨基酸序列与二维接触图对靶点进行表征，分别通过卷积神经网络与图注意力网络进行编码，并在 DTI 预测任务中进一步探索不同靶点表示方式的融合，以确保当样本数量较少时，方法能够从氨基酸序列中提取有效信息，当样本数量充足时，方法能够从二维图中提取更多结构信息，从而实现更准确的预测。

2.3 基于深度学习的分子编码模型

针对药物与靶点的分子编码，本研究利用了多种深度学习模型，本节将依次介绍卷积神经网络、注意力机制与图注意力网络等深度学习模型的计算原理。

2.3.1 基于卷积神经网络的编码模型

卷积神经网络 (Convolutional Neural Network, CNN) 是一种受生物视觉感知机制启发的深度学习模型。1998 年, LeCun 等人提出代表性方法 LeNet-5^[97]。2015 年以后, 大量的 CNN 方法被提出并在计算机视觉任务中展现出卓越性能^[98-99]。在生物信息学领域, CNN 也被广泛用于药物与靶点的序列编码^[45,73-74]。

CNN 的基本运算包括卷积、激活与池化^[92]。卷积运算是指以固定大小和步长的卷积核在输入特征图（如图像矩阵或序列嵌入）上滑动，计算局部的加权和；激活运算利用 ReLU 等激活函数对卷积结果施加非线性变换；池化运算则在激活的特征图上，以固定大小和步长的滑动窗口计算区域内的最大值或平均值，通过下采样操作减小空间维度并保留重要特征。在图像处理任务中，CNN 通常在每次卷积运算后随即进行激活与局部池化；而在序列建模过程中，CNN 在每次卷积后仅进行激活，在所有卷积完成后才执行全局池化运算^[92]。

以序列编码为例，记序列长度为 L ，嵌入维度为 d ，则输入序列可表示为 $\mathbf{S} \in \mathbb{R}^{L \times d}$ 。给定卷积核大小为 k ，步长为 1，输出通道数（即输出维度）为 d ，则卷积操作的输出特征图为 $\mathbf{C} \in \mathbb{R}^{(L-k+1) \times d}$ ，其在第 i 行的嵌入表示 $\mathbf{C}_i$ 的计算过程如式（2.1）所示。

$$\mathbf{C}_i = \sum_{j=0}^{k-1} \mathbf{S}_{i+j} \mathbf{W}_j + \mathbf{b}, \quad i = 0, 1, \dots, L-k \quad (2.1)$$

其中 $\mathbf{S}_{i+j} \in \mathbb{R}^d$ 为序列表示 $\mathbf{S}$ 在第 $i+j$ 行的嵌入向量， $\mathbf{W} \in \mathbb{R}^{k \times d \times d}$ 为可训练的权重， $\mathbf{W}_j \in \mathbb{R}^{d \times d}$ 为第 j 个权重矩阵， $\mathbf{b} \in \mathbb{R}^d$ 为可训练的偏置项。

在完成卷积操作后，对卷积输出的特征图 $\mathbf{C}$ 施加激活函数（例如 ReLU），得到激活后的特征图 $\mathbf{C}' = \text{ReLU}(\mathbf{C})$ ，随后进入下一次“卷积-激活”循环。在所有卷积与激活操作完成后，对最终的激活特征图进行全局池化。假设“卷积-激活”过程仅执行一次，且采用最大池化策略，最终得到全局表示 $\mathbf{P} \in \mathbb{R}^d$ ，其中第 i 维的值即 $\mathbf{P}_i$ 的运算过程如式（2.2）所示。

$$\mathbf{P}_i = \max \left(\mathbf{C}'_{0,i}, \mathbf{C}'_{1,i}, \mathbf{C}'_{2,i}, \dots, \mathbf{C}'_{L-k,i} \right), \quad i = 0, 1, \dots, d \quad (2.2)$$

其中 $\mathbf{C}'_{x,i}$ 表示 $\mathbf{C}'$ 在第 x 行第 i 列的值，并且有 $x \in \{0, 1, 2, \dots, L-k\}$ 。

通过上述运算，CNN 实现了局部片段的卷积提取与全局序列的特征聚合，从而为后续任务提供有效的特征表示。在本研究的第 5 章中，CNN 的三项主要运算步骤被拆分，即第 5 章中的 $\text{CNN}(\cdot)$ 仅表示卷积操作，计算原理同式（2.1）。

2.3.2 基于注意力机制的编码模型

注意力机制（Attention Mechanism）源于对人类感知系统的模拟，其在处理大量信息时聚焦于与当前任务最相关的部分，从而提高建模效率。此外，注意力机制在一定程度上具备可解释性，因此已成为深度学习领域的重要组成部分^[100]。2017 年，Vaswani 等人^[75]提出基于自注意力机制的 Transformer 模型，该模型显著提升了模型在序列建模任务中的表现，并推动了注意力机制在自然语言处理等领域的广泛应用^[76,81,85]。

注意力机制以查询（Query）、键（Key）与值（Value）三组向量为输入，通过计算查询与键之间的相似性得到注意力权重，常见的计算方式包括加性注意

力^[101]、缩放点积注意力^[75]以及连接注意力等，随后对注意力权重进行归一化，常见的归一化函数包括 softmax 与 sigmoid^[100]，最终根据归一化的权重计算值的加权和，从而捕获与当前查询最相关的信息。

以缩放点积注意力与 softmax 函数为例，记查询矩阵为 $\mathbf{Q} \in \mathbb{R}^{L_q \times d}$ ，键与值矩阵分别为 $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times d}$ ，则输出矩阵 $\mathbf{S} \in \mathbb{R}^{L_q \times d}$ 的计算过程如式 (2.3) 所示。

$$\mathbf{S} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right) \mathbf{V} = \mathbf{A}\mathbf{V} \quad (2.3)$$

其中 $\text{softmax}(\cdot)$ 表示行归一化， $\mathbf{A} \in \mathbb{R}^{L_q \times L}$ 为归一化后的注意力权重矩阵，其任一行的和均为 1，即对于任意 $i \in \{0, 1, \dots, L_q - 1\}$ ，均有 $\sum_{j=0}^L \mathbf{A}_{ij} = 1$ 。

在实际应用中，多数方法会在计算之前对查询、键与值执行线性转换，以将其投影到相同的特征空间，从而改善模型捕获关键信息的能力。在常规机制的基础上，Vaswani 等人^[75]进一步提出了自注意力（Self Attention）机制与多头注意力（Multi-Head Attention）机制，其中前者是指查询、键与值三者相同的注意力机制，而后者则通过多个注意力头捕获不同子空间的信息，从而提升模型对复杂依赖关系的建模能力。

除了多头注意力机制，本研究也利用了协注意力（Co-Attention）机制。协注意力机制旨在计算两类表示的关联强度^[100]，该机制由 Lu 等人^[102]于 2016 年首次提出并被用于生成图像与问题的特征图。2019 年，Deac 等人^[50]将其引入 DDI 预测领域，以挖掘药物之间的相互作用信息。2021 年以后，协注意力机制在 DDI 预测领域得到较多应用^[52,54-55]，主要用于计算子结构之间的相互作用强度。

记输入的两类表示分别为 $\mathbf{X} \in \mathbb{R}^{L_x \times d}$ 与 $\mathbf{Y} \in \mathbb{R}^{L_y \times d}$ ，输出的得分矩阵为 $\mathbf{C} \in \mathbb{R}^{L_x \times L_y}$ ，则协注意力的计算过程如式 (2.4) 所示。

$$\mathbf{C}_{ij} = \mathbf{a}^\top \tanh(\mathbf{W}_x \mathbf{X}_i + \mathbf{W}_y \mathbf{Y}_j) \quad (2.4)$$

其中 $\mathbf{C}_{ij}$ 为得分矩阵 $\mathbf{C}$ 在第 i 行第 j 列的值， $\mathbf{a} \in \mathbb{R}^d$ 为可学习的权重向量， $\mathbf{W}_x, \mathbf{W}_y \in \mathbb{R}^{d \times d}$ 为可学习的权重矩阵， $\mathbf{X}_i$ 为矩阵 $\mathbf{X}$ 第 i 行的表示， $\mathbf{Y}_j$ 为矩阵 $\mathbf{Y}$ 第 j 行的表示，且有 $i \in \{0, 1, \dots, L_x - 1\}$ 以及 $j \in \{0, 1, \dots, L_y - 1\}$

整体而言，注意力机制能够聚焦于关键信息，多头注意力机制能够捕获不同子空间的信息，协注意力机制则能融合不同表示，因此，本研究拟利用这些机制进行相互作用信息的挖掘与多模态信息的融合，从而改善模型的预测表现。

2.3.3 基于图注意力网络的编码模型

在社交网络与生物系统等复杂的系统中，相关的数据与关联（例如分子内不同原子的关系）通常以图或网络的方式呈现。对于图结构的数据，常规的机器学习任务往往表现不佳，而图神经网络则能更好地挖掘节点之间的关联^[103]。常见的图神经网络包括图卷积网络^[104]、图同构网络以及图注意力网络（Graph

Attention Network, GAT) 等。2017 年, Velivckovic 等人^[105]将注意力机制引入图神经网络并提出 GAT。与图卷积网络等模型相比, GAT 在聚合过程中动态权衡相邻节点的重要性, 能够同时学习局部模式与全局图结构信息, 因此在论文引用识别等多项任务中具有更好的表现^[105]。

GAT 包括注意力系数计算与信息聚合两项步骤, 其中前者根据每个节点的嵌入表示计算相邻节点之间的影响, 从而识别最相关的节点。

以任意相邻节点 i 与 j 为例, 记节点嵌入为 $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^d$, 边嵌入为 $\mathbf{h}_e \in \mathbb{R}^{d'}$, 节点 i 与 j 之间的注意力系数为 α_{ij} , 则其计算过程如 (2.5) 与 (2.6) 所示。

$$e_{ij} = \mathbf{a}^\top (\mathbf{W}\mathbf{h}_i + \mathbf{W}\mathbf{h}_j + \mathbf{W}'\mathbf{h}_e) \quad (2.5)$$

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(e_{ij}))}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp(\text{LeakyReLU}(e_{ik}))} \quad (2.6)$$

其中 $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{W} \in \mathbb{R}^{d \times d}$ 以及 $\mathbf{W}' \in \mathbb{R}^{d \times d'}$ 为可学习的权重, $\mathcal{N}_i$ 表示节点 i 的所有邻接节点, LeakyReLU 为激活函数^[106]。

随后, GAT 根据注意力系数对邻接节点的信息进行聚合。以节点 i 为例, 记该节点聚合后的表示为 $\mathbf{h}'_i$, 则其计算过程如式 (2.7) 所示。

$$\mathbf{h}'_i = \text{ELU} \left(\sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij} \mathbf{W}\mathbf{h}_j + \mathbf{b} \right) \quad (2.7)$$

其中 $\mathbf{b} \in \mathbb{R}^d$ 为可学习的偏置, ELU 为激活函数。

通过上述运算, GAT 能够动态权衡节点的相关性并聚合局部信息。在实际应用中, GAT 通常会利用多头注意力机制分别计算注意力系数并聚合, 最终输出不同聚合结果的平均值, 从而增强模型的稳定性^[105]。相比于 GCN 等模型, GAT 具备更强的学习能力与更好的表现, 因此本研究拟利用 GAT 对药物的层次化分子图与靶点的二维接触图进行编码, 从而改善模型的预测性能。

2.4 本章小结

本研究涉及分子生物学与深度学习两个不同的领域, 为便于读者理解, 本章系统介绍了相关知识与方法, 包含常见的药物表示形式、常见的靶点表示形式以及各类深度学习模型。整体而言, 药物与靶点的不同表示形式各有优势并具有融合潜力, 例如一维序列信息表征效率更高, 而二维图信息表征总量更高, 然而现阶段的药物相互作用预测方法通常仅利用一种表示形式, 而未实现多模态信息的融合。在此背景下, 本研究拟利用深度学习方法, 从相互作用信息挖掘与多模态信息融合等多个角度, 对 DDI 预测与 DTI 预测进行研究, 以期药物研发与药物再利用等下游任务提供更准确的底层计算模型。

第3章 基于图表示学习的药物-药物相互作用预测方法

3.1 引言

作为药物相互作用预测领域的代表性问题，DDI 预测对于识别潜在的药物-药物相互作用以及提升联合用药的安全性具有重要意义。就问题定义而言，现有研究通常将 DDI 预测视为二分类任务，即预测两个药物之间是否存在某种指定类型的相互作用^[54-56,58]。

在 DDI 预测任务中，药物的表示形式例如一维表示以及分子图等对模型的表现具有较大影响。早期的方法通常基于药物的 SMILES 序列或 ECFP 等一维表示进行编码，但序列无法反映药物分子的拓扑结构，因此对应方法的识别准确率受到一定限制^[19,32,107]。随着图神经网络等图编码方法的发展，基于分子图的 DDI 预测方法^[47-50]取得了更好的表现。

药物通常由多个官能团或化学子结构组成，它们决定了药物的药效学和药代动力学特性，并直接影响药物之间的相互作用^[52-53]。基于此，近年来基于子结构的 DDI 预测^[54-56,58]逐渐成为研究热点，然而现有大多数方法主要以模型感知的方式挖掘子结构，例如 SSI-DDI 等模型以分子图中节点的表示作为子结构表示，而 GMPNN-CS 等模型以池化后的全局表示作为子结构表示。这些由模型感知到的子结构缺乏化学可解释性，往往需要额外的人为分析方能验证其合理性。此外，部分方法^[12]利用可解释算法提取化学子结构，但其对同一分子内部的不同子结构进行独立编码，模型无法捕获子结构之间的相互作用信息。

在分子属性预测领域，基于子结构的方法也受到广泛关注。2021 年，Zhang 等人^[86]提出了基于分子模式的自监督学习方法 MGSSL，该方法利用 BRICS 算法与细化规则将分子分解为多个子结构并构建子结构树。2023 年，Zang 等人^[87]提出了一种层次化分子自监督学习框架 HiMol，该方法通过分解分子构建出包含节点、子结构和图的层次化分子表示。实验结果表明，层次化分子表示能够在分子编码过程中实现不同层次之间的交互学习，从而有效提升模型的预测性能^[86-87]，然而这类表示目前在 DDI 预测领域尚未受到充分关注。

为解决现有 DDI 预测方法存在的模型感知子结构可解释性不足、实际化学子结构相互孤立的问题，本章提出了一种基于层次化分子图与双视图表示学习的 DDI 预测模型，即 HDN-DDI。该模型的主要贡献如下：（1）可解释的子结构提取模块。HDN-DDI 通过可解释的分解算法将药物分子分解为多个化学子结构，从根本上改善了子结构的可解释性；（2）层次化的分子表示方式。HDN-DDI 引入了“原子-子结构-分子”层次化分子图对分子进行表征，在编码过程中充分整合分子的全局拓扑信息和局部子结构信息，从而提升了模型的编码效率；（3）增

强型双视图表示学习。HDN-DDI 设计了一种增强型双视图表示学习方法,使得子结构信息能够在分子内部与分子之间传递,同时该方法能够捕获相互作用过程中的关键子结构,从而充分挖掘药物分子的结构信息与相互作用信息。

3.2 方法模型

本节将依次介绍 HDN-DDI 的整体框架、编码模块的输入、HDN 编码器、HDN 解码器以及模型训练过程中所利用的损失函数。

3.2.1 HDN-DDI 整体框架

HDN-DDI 首先使用可解释的子结构提取模块分解药物分子,随后通过 HDN 编码器获取药物分子的层次化表示、相互作用表示以及全局表示,最后通过 HDN 解码器聚合不同深度的全局表示并进行 DDI 预测,其整体框架如图3.1所示。

具体而言,针对给定的两种药物 D_x 和 D_y , HDN-DDI 采用改进后的 BRICS 算法^[87]对药物分子进行分解,提取其化学子结构并构建层次化分子图 G_x 和 G_y 。HDN 编码器由多个 HDN 编码块组成,每个编码块以初始的层次化分子图或前一编码块更新后的图为输入。HDN 编码块使用 GAT^[105]在分子内部学习层次化表示,在两个分子之间学习相互作用表示,最终对层次化分子图进行更新并输出两个药物分子的全局表示。HDN 解码器通过协注意力机制整合不同 HDN 编码块的输出并实现 DDI 预测。

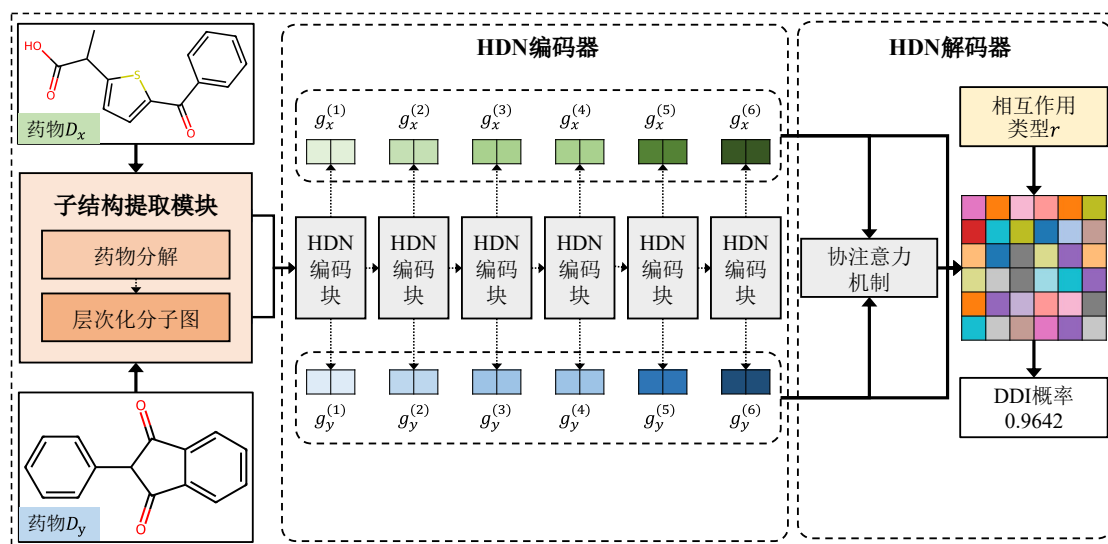


图 3.1 HDN-DDI 总体框架图

3.2.2 模型输入

HDN-DDI 模型以药物的 SMILES 序列为输入,通过 RDKit 库^[91]将序列解析为分子图,在此基础上进一步构建层次化分子图 G 。记 V 为节点集合,其包

含原子级节点、子结构级节点和分子级节点, $E \subset V \times V$ 为边集合, 其表示节点之间的拓扑连接关系, 则有 $G = (V, E)$ 。

在图 G 中, 每个节点均具有 66 维初始特征, 具体含义如表A.1所示。就特征的获取方式而言, 原子级节点的特征由原子的理化特性决定, 而子结构级与分子级节点则分别具有独立的节点标识, 其中子结构级节点除标识外的剩余 65 维特征通过对子结构中所有原子级节点的特征求和而得出, 分子级节点的剩余 65 维特征通过对子结构级节点的特征进行求和而得出。此外, 依照基线模型^[52,55,58]的设定, 图 G 中的边均为无属性的双向边, 其仅表示节点之间的邻接关系。

3.2.3 子结构提取模块

子结构提取模块利用改进后的 BRICS 算法^[87]分解药物分子, 并采用层次化分子图对分子结构进行表征, 其处理流程如图3.2与图3.3所示。具体而言, 给定药物的 SMILES 序列, 该模块通过 RDKit 库^[91]将其转换为分子图, 并利用 BRICS 算法^[108]将分子图初步分解为若干子结构。对于 BRICS 算法无法分解的大型片段, 模块依据细化分解规则^[87]将对其作进一步分解。最终, 该模块能够将完整的药物分子分解为多个具有化学意义的子结构, 如图3.2所示。

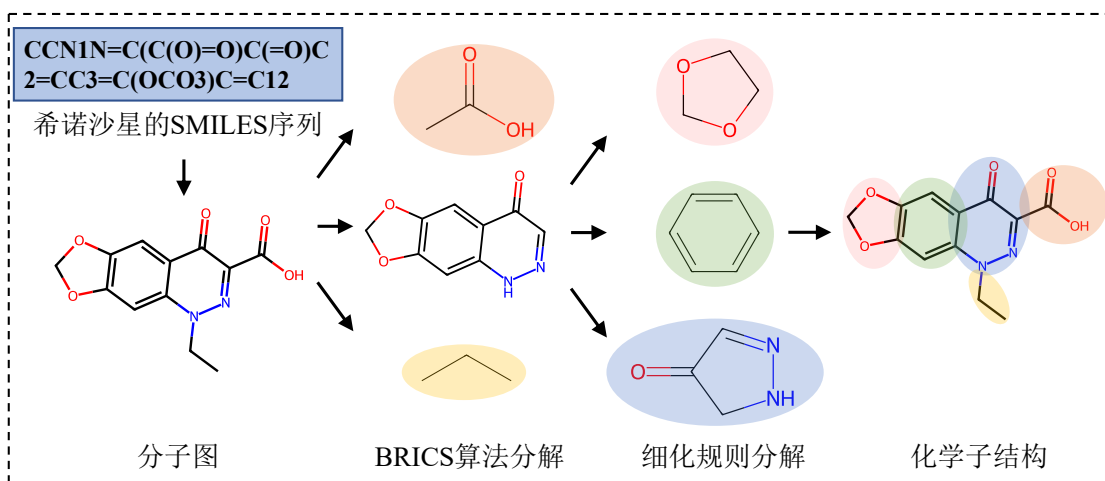


图 3.2 子结构提取模块中药物分子的分解过程

在基于算法分解药物分子的基础上, 子结构提取模块进一步将初始分子图转化为层次化分子图, 如图3.3所示。对于每个分解得到的化学子结构, 模块在分子图中引入相应的子结构级节点, 并根据子结构与其所包含的原子之间的对应关系, 在子结构级节点和原子级节点之间建立双向连接。例如, 如果子结构 S_i 包含原子 A_j , 则模块在子结构级节点 v_{S_i} 与原子级节点 v_{A_j} 之间添加一条双向边。该过程对所有子结构级节点和原子级节点重复执行, 如图3.3中间部分所示。通过上述步骤, 转换后的分子图能够清晰地反映原子和子结构之间的层级关系, 从而为后续编码过程提供更完整的结构信息。

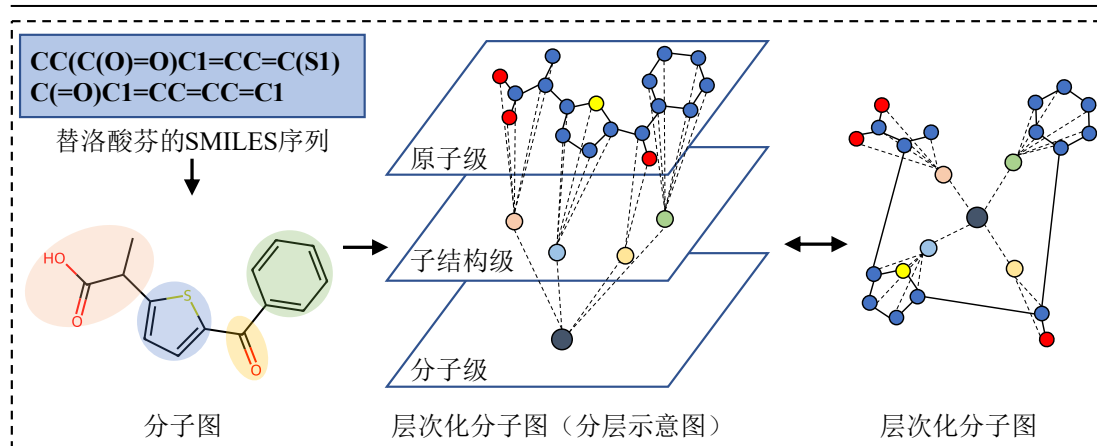


图 3.3 子结构提取模块中层次化分子图的构建过程

此外,该模块还会引入一个分子级节点 v_M ,并在该节点与所有子结构级节点之间建立双向连接。具体而言,对于每个子结构级节点 v_{S_i} ,模块均在其与分子级节点 v_M 之间添加一条双向边 $v_M \leftrightarrow v_{S_i}$,从而促进局部结构信息的向上聚合和全局信息的向下传播,增强层次化分子图的表征能力。

最终,子结构提取模块构建出完整的层次化分子图,其包含原子级节点、子结构级节点以及分子级节点,分别代表原子、分解得到的化学子结构以及整个药物分子。层次化分子图不仅保留了原子级节点之间的原始拓扑结构,还引入了上下层级之间的双向连接,从而增强层级间的信息交互。

在层次化分子图中,各层级的节点均可感知并整合其它层级的信息。具体而言,原子级节点侧重于局部特征的学习,分子级节点汇聚全局信息,而子结构级节点则捕获官能团或化学子结构的特征。通过这种层次化结构,原子级节点可以感知分子的整体信息,分子级节点可以整合所有原子级特征,而子结构级节点则在两者之间发挥中介作用,从而促进不同层级的信息流动与聚合。

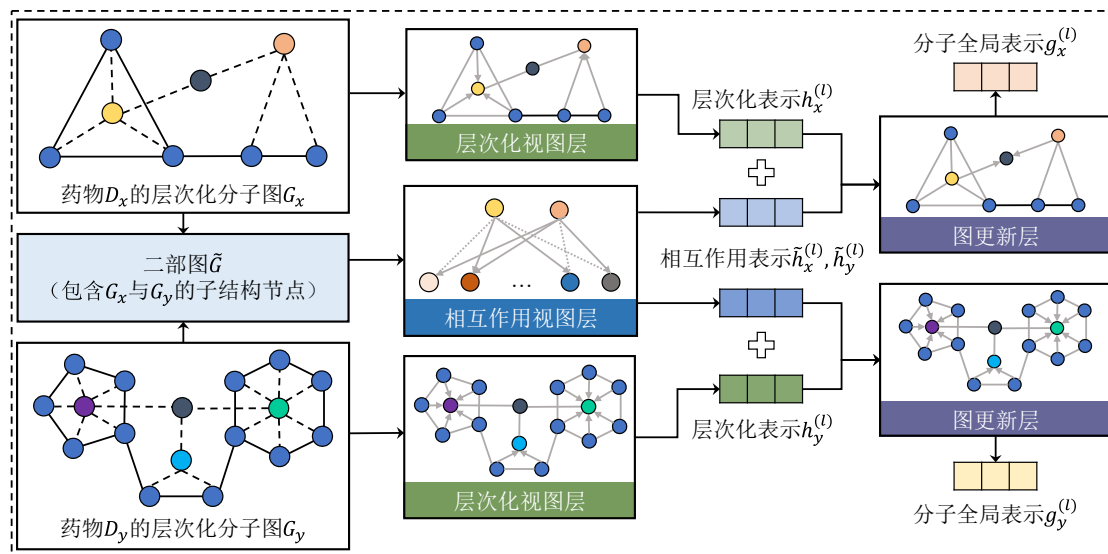


图 3.4 HDN 编码块内部结构示意图

3.2.4 HDN 编码器

如图3.4所示, HDN 编码器由多个 HDN 编码块组成, 每个 HDN 编码块包含三个编码层: 面向单个药物的层次化视图层、面向药物对的相互作用视图层以及用于整合信息的更新层。这些编码层均采用 GAT 进行图编码, 然而其编码对象与权重均不同, 因此能够差异化捕获信息, 即层次化视图层捕获结构信息, 相互作用视图层捕获相互作用信息, 而更新层则融合两类信息并对分子图进行更新。

在层次化视图层和更新层中, 节点可以聚合来自所有邻接节点的信息。例如, 原子级节点不仅能够从相邻的子结构级节点获取子结构信息, 还可以整合相邻原子级节点的编码表示。在具体实现上, 层次化视图层和更新层均采用单层 GAT 来执行信息传输。记第 $l+1$ 个编码块中节点 i 的编码表示为 $\mathbf{h}_i^{(l+1)}$, 则层次化视图层与更新层的计算过程均如式 (3.1) 至式 (3.3) 所述。

$$\mathbf{h}_i^{(l+1)} = \sigma \left(\sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij} \mathbf{W}^{(l+1)} \mathbf{h}_j^{(l)} + \mathbf{b}^{(l+1)} \right) \quad (3.1)$$

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(e_{ij}))}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp(\text{LeakyReLU}(e_{ik}))} \quad (3.2)$$

$$e_{ij} = \mathbf{a}^{(l+1)\top} \left(\mathbf{W}^{(l+1)} \mathbf{h}_i^{(l)} + \mathbf{W}^{(l+1)} \mathbf{h}_j^{(l)} \right) \quad (3.3)$$

其中 $\mathbf{W}^{(l+1)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ 和 $\mathbf{a}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 为第 $l+1$ 个编码块的层次化视图层或更新层中 GAT 的可训练权重, 而 $\mathbf{b}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 为可训练偏差。 $\top$ 为转置操作。LeakyReLU($\cdot$)^[106] 和 $\sigma(\cdot)$ 为激活函数。 $d^{(l)}$ 和 $d^{(l+1)}$ 分别是第 $l+1$ 个编码块中层次化视图层或更新层的输入和输出特征维度。 $\mathcal{N}_i$ 为节点 i 的所有邻接节点的集合。 $\mathbf{h}_j^{(l)}$ 为节点 j 在第 l 个编码块的表示, α_{ij} 为节点 i 和 j 之间的注意力系数。

相互作用视图层以二部图 $\tilde{\mathbf{G}} = \mathbf{V}_s^{(x)} \times \mathbf{V}_s^{(y)}$ 为输入并利用 GAT 对其编码, 其中 $\mathbf{V}_s^{(x)}$ 和 $\mathbf{V}_s^{(y)}$ 分别表示层次化分子图 G_x 和 G_y 中的子结构级节点的集合。相关研究^[52,54]表明, DDI 通常由有限数量的关键子结构所触发。受此启发, 相互作用视图层在编码过程中对二部图中的连接进行随机剪枝, 即对图中的边进行 Dropout 处理, 从而引导模型挖掘 DDI 中的关键子结构。记第 $l+1$ 个编码块中节点 i 的相互作用表示为 $\tilde{\mathbf{h}}_i^{(l+1)}$, 则其计算过程如式 (3.4) 至式 (3.6) 所述。

$$\tilde{\mathbf{h}}_i^{(l+1)} = \sigma \left(\sum_{j \in \tilde{\mathcal{N}}_i \cup \{i\}} \beta_{ij} \tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_j^{(l)} + \tilde{\mathbf{b}}^{(l+1)} \right) \quad (3.4)$$

$$\beta_{ij} = \frac{\exp(\text{LeakyReLU}(\tilde{e}_{ij}))}{\sum_{k \in \tilde{\mathcal{N}}_i \cup \{i\}} \exp(\text{LeakyReLU}(\tilde{e}_{ik}))} \quad (3.5)$$

$$\tilde{e}_{ij} = \tilde{\mathbf{a}}^{(l+1)\top} \left(\tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_i^{(l)} + \tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_j^{(l)} \right) \quad (3.6)$$

其中 $\widetilde{\mathbf{W}}^{(l+1)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ 和 $\widetilde{\mathbf{a}}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 为第 $l+1$ 个编码块中相互作用视图层中 GAT 的可训练权重, 而 $\widetilde{\mathbf{b}}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 为可训练偏差。 $\widetilde{\mathcal{N}}_i$ 为随机剪枝后节点 i 在二部图中的邻接节点集合。具体而言, 若子结构级节点 i 来源于药物 D_x , 则 $\widetilde{\mathcal{N}}_i$ 为药物 D_y 中子结构级节点的随机子集, 反之亦然。此外, β_{ij} 表示节点 i 和 j 之间的注意力系数。

为了保持节点表示维度的一致性, 相互作用视图层对所有原子级节点和分子级节点应用非线性变换, 该变换与相互作用视图层中的 GAT 共享权重, 其计算过程如式 (3.7) 所示

$$\widetilde{\mathbf{h}}_k^{(l+1)} = \sigma \left(\widetilde{\mathbf{W}}^{(l+1)} \mathbf{h}_k^{(l)} + \widetilde{\mathbf{b}}^{(l+1)} \right) \quad (3.7)$$

其中 $\widetilde{\mathbf{h}}_k^{(l+1)}$ 为节点 k 的相互作用表示。

针对节点 i 的层次表示 $\mathbf{h}_i^{(l+1)}$ 和相互作用表示 $\widetilde{\mathbf{h}}_i^{(l+1)}$, 更新层通过单个线性层与指数线性单元激活函数 $\text{ELU}(\cdot)^{[109]}$ 实现信息聚合, 如式 (3.8) 所示。随后, 更新层通过单层 GAT 更新整个层次化分子图, 其计算原理与式 (3.1) 至式 (3.3) 相同。所有节点的更新表示作为下一个 HDN 编码块的输入。

$$\mathbf{h}_i^{(l+1)} = \text{ELU} \left(\text{Linear} \left(\mathbf{h}_i^{(l+1)} \parallel \widetilde{\mathbf{h}}_i^{(l+1)} \right) \right) \quad (3.8)$$

如3.2.3小节所述, 分子级节点能够在编码过程中整合所有子结构级节点的信息, 因此编码块可以通过分子级节点获取分子的全局表示, 如式 (3.9) 所示。

$$\mathbf{g}_x^{(l+1)} = \mathbf{h}_{x_mol}^{(l+1)}, \mathbf{g}_y^{(l+1)} = \mathbf{h}_{y_mol}^{(l+1)} \quad (3.9)$$

其中 $\mathbf{h}_{x_mol}^{(l+1)}$ 和 $\mathbf{h}_{y_mol}^{(l+1)}$ 分别是层次化分子图 G_x 和 G_y 中分子级节点的表示。此外, $\mathbf{g}_x^{(l+1)}$ 和 $\mathbf{g}_y^{(l+1)}$ 表示第 $l+1$ 个编码块中药物 D_x 和 D_y 的全局表示。与传统分子图相比, 层次化分子图无需额外的池化操作, 从而有效降低了计算成本。

3.2.5 HDN 解码器

通过利用多个 HDN 编码块进行迭代式编码, HDN 编码器能够捕获药物的分子内层次结构信息和分子间相互作用信息, 并输出不同深度的全局表示。HDN 解码器则利用协注意力机制^[52,55]对这些表示进行聚合。具体而言, HDN 解码器首先逐层计算两种药物全局表示的关联强度, 如式 (3.10) 所示。

$$\gamma_{ll'} = \boldsymbol{\alpha}^\top \tanh \left(\mathbf{W}_x \mathbf{g}_x^{(l)} + \mathbf{W}_y \mathbf{g}_y^{(l')} \right) \quad (3.10)$$

其中 $\boldsymbol{\alpha} \in \mathbb{R}^d$ 为可学习的权重向量, $\mathbf{W}_x, \mathbf{W}_y \in \mathbb{R}^{d \times d}$ 为可学习的权重矩阵。 d 为分子表示的维度, $\mathbf{g}_x^{(l)}, \mathbf{g}_y^{(l)} \in \mathbb{R}^d$ 为第 l 个编码块中药物 D_x 和 D_y 的全局表示。

随后, HDN 解码器学习相互作用类型的嵌入表征并计算不同深度下药物-药物相互作用分数, 最终对其求和以计算 DDI 概率, 如式 (3.11) 所示。

$$s(D_x, r, D_y) = \sigma \left(\sum_{l=1}^L \sum_{l'=1}^L \gamma_{ll'} \mathbf{g}_x^{(l)} \mathbf{M}_r \mathbf{g}_y^{(l')} \right) \quad (3.11)$$

其中 $\mathbf{M}_r \in \mathbb{R}^{d \times d}$ 为相互作用 r 的嵌入矩阵, L 为 HDN 编码块的数量, σ 为 sigmoid 激活函数, $s(D_x, r, D_y)$ 为药物 D_x 和 D_y 产生相互作用 r 的概率。

3.2.6 损失函数

如3.1节所述, DDI 预测可被视为二分类任务, 即预测两个药物之间是否存在相互作用, 形式化而言, 即判断给定的三元组 (D_x, r, D_y) 是否构成实际存在的药物-药物相互作用。因此, HDN-DDI 使用二元交叉熵损失函数进行端到端训练, 如式 (3.12) 所示。

$$\text{Loss} = -\frac{1}{|\mathcal{T}|} \sum_{t=(D_x, r, D_y) \in \mathcal{T}} [y_t \log p_t + (1 - y_t) \log(1 - p_t)] \quad (3.12)$$

其中 $y_t \in \{0, 1\}$ 为样本 t 的真实类别, $p_t \in [0, 1]$ 为模型预测的相互作用概率, $\mathcal{T}$ 为所有样本的集合。

3.3 实验及结果分析

3.3.1 实验设置

本节将依次介绍实验的设置情况, 即数据集、评估场景、实验环境、模型架构与超参数、评估指标以及基线模型。

3.3.1.1 数据集

实验采用两个被广泛使用的数据集 DrugBank^[6] 和 Twosides^[110]。DrugBank 是一个综合性药物数据集, 其涵盖药物-药物相互作用与药物-靶点相互作用等多类数据, 并且记录了 1706 种药物和 86 种药物-药物相互作用, 共产生 191 808 个 DDI 实例。Twosides 由 Zitnik 等人^[110] 对原始数据集^[111] 进一步筛选得出, 其包含 645 种药物和 963 种相互作用, 共产生 4 576 287 个 DDI 实例。

除了数据集规模的差异外, 这两个数据集在药物相互作用类型的定义和数量上也有所不同。DrugBank 记录了两种药物之间的单一相互作用, 其描述了一种药物如何影响另一种药物的代谢过程^[6]。例如, 药物 A 会增加药物 B 的镇静作用, 或药物 C 会减弱药物 D 的疗效。相比之下, Twosides 记录了两种药物之间的多种相互作用, 其通常以潜在副作用的形式呈现^[54]。例如, 药物 C 和药物 D 共同给药会导致患者头痛、咳嗽和发烧。

由于两个数据集均未包含 DDI 负样本, 即未记录不存在相互作用的药物对, 故需要对负样本进行生成。本节的生成方式与先前研究^[54-56] 相一致, 即对于给定的正样本 $t = (D_x, r, D_y)$, 通过随机替换 D_x 为 D'_x 或 D_y 为 D'_y , 从而构造负样本 (D_x, r, D'_y) 或 (D'_x, r, D_y) 。为确保负样本的合理性, 只有当原始数据集中不存

在 (D'_x, D_y) 之间或 (D_x, D'_y) 之间的相互作用记录时, 才会生成对应的负样本。此外, 为确保正负样本数量均衡, 每条正样本仅生成一条负样本^[55-56]。

3.3.1.2 评估场景

实验在热启动场景与冷启动场景中进行: 在热启动场景中, 一种药物可能同时出现在训练集和测试集; 在冷启动场景中, 每种药物仅会被分配到训练集或测试集, 从而确保训练集和测试集中的药物没有重叠。实验在两种场景中的评估数据集以及划分比例同现有研究^[54-56]保持一致。在热启动场景下, 使用 DrugBank 与 Twosides 两个数据集评估模型, 且训练集、验证集和测试集的划分比例为 6:2:2。在冷启动场景下, 仅使用 DrugBank 数据集评估模型, 并随机选择 20% 的药物构建测试集, 剩余 80% 的药物构建训练集和验证集。

3.3.1.3 实验环境

HDN-DDI 基于 Pytorch 框架构建, 其依赖的工具库包括 torch_geometric 和 RDKit 等。此外, 模型的训练、评估以及对比均在 Ubuntu 22 操作系统中进行, 中央处理器 (Central Processing Unit, CPU) 型号为 Intel Core i9-1300K 3.00GHz, 图像处理器 (Graphics Processing Unit, GPU) 型号为 NVIDIA GeForce RTX 4090, 内存容量为 128GB, 显存容量为 24GB。

3.3.1.4 模型架构与超参数

如图3.1所示, HDN-DDI 包含多个 HDN 编码块, 每个编码块由一个层次化视图层、一个相互作用视图层和一个更新层组成, 所有层均采用单层多头 GAT 进行消息传递。最优超参数通过在验证集上随机搜索而确定, 搜索空间如表A.2所示。具体而言, HDN-DDI 包含 6 个 HDN 编码块。在每个编码块中, 层次化视图层和相互作用视图层均产生 64 维表示, 而更新层产生 128 维表示。此外, HDN-DDI 中的所有 GAT 均使用两个注意力头。在模型训练过程中, 批次样本数量设置为 512, 学习率设置为 1.0×10^{-3} 。

3.3.1.5 评估指标

由于单个评估指标难以全面地衡量模型性能^[18,95,112], 实验采用四个常见的指标来评估 HDN-DDI 以及基线模型, 具体如下:

(1) 准确率 (Accuracy, ACC), 即正确预测的样本占有所有预测样本的比例, 该指标能直观反映模型的预测能力;

(2) 受试者工作特征曲线下面积 (Area Under the Receiver Operating Characteristic Curve, AUROC)。AUROC 在数值上等于模型对随机选择的正样本的评分

高于随机选择负样本的评分的概率^[55]，其同样能衡量模型的分类能力；

(3) 精确率-召回率曲线下面积 (Area Under the Precision-Recall Curve, AUPR)。该指标能够衡量模型维持高精确率和高召回率的能力；

(4) F1 值 (F1 Score)，即精确率和召回率的调和平均值。在样本不均衡或正样本更重要的场景中，AUPR 与 F1 值通常是更合适的评估指标^[112-113]。

3.3.1.6 基线模型

实验对 HDN-DDI 与 DDI 预测领域的主流方法以及近期工作进行了比较。参与对比的基线模型包括基于子结构的模型和基于双视图表示学习的模型，具体模型及其主要贡献如下：

(1) GMPNN-CS^[54]：通过门控消息传递网络学习多尺度的子结构信息，并对子结构之间的相互作用进行建模，从而实现 DDI 预测；

(2) DSN-DDI^[55]：迭代式进行局部表示学习和全局表示学习，并通过双视图表示学习从单个药物和药物对中挖掘关键子结构；

(3) BDN-DDI^[56]：通过双线性层以及交互式图神经网络捕获成对的原子相互作用，并利用双视图表示学习从药物分子中提取子结构；

(4) SRR-DDI^[114]：使用子结构增强机制完善药物的子结构特征，同时引入药物相似性编码，从而提升 DDI 预测准确性并改善模型的泛化能力。

(5) PEB-DDI^[57]：在分子编码过程中整合化学键信息，并且针对热启动与冷启动场景分别设计了分子指纹-分子图以及二部图-分子图的双视图策略。

表 3.1 热启动场景下 HDN-DDI 等模型在不同数据集上的表现 (%)

模型	DrugBank				Twosides			
	ACC	AUROC	AUPR	F1 值	ACC	AUROC	AUPR	F1 值
GMPNN-CS	95.30	98.46	97.94	95.39	82.83	90.07	87.24	84.08
DSN-DDI	96.94	99.47	99.37	96.93	98.83	99.90	99.89	98.83
BDN-DDI	96.33	99.34	99.28	94.40	99.07	99.92	99.90	99.07
SRR-DDI	96.67	99.05	98.74	96.72	87.52	93.20	90.37	88.32
PEB-DDI	98.18	99.65	99.63	98.19	79.60	86.00	81.28	81.37
HDN-DDI	97.93	99.73	99.72	98.03	99.38	99.97	99.97	99.38

注：每项指标的最高值用加粗字体显示。

3.3.2 热启动场景下的实验结果

在热启动场景中，本节进行三次独立实验并取平均值作为模型最终表现。每次实验根据相互作用类型对数据集进行随机分层划分，以确保训练集、验证集和

测试集中各类相互作用的占比相同。此外，负样本生成和数据集划分在模型训练之前完成，从而使所有模型共享相同的训练集、验证集与测试集。表3.1显示了各模型在 DrugBank 与 Twosides 数据集上的表现。

如3.1所示，在 DrugBank 数据集上，HDN-DDI 的表现与 PEB-DDI 相近，并且两者的表现均显著优于 SRR-DDI 以及更早的基线模型。在 Twosides 数据集上，HDN-DDI 的表现优于所有基线模型。此外，HDN-DDI 在两个数据集上的 AUPR 分别为 99.72% 与 99.97%，这表明 HDN-DDI 具有良好的正样本预测即识别潜在 DDI 的能力。与此同时，HDN-DDI 与基线模型在 AUROC 与 AUPR 指标上均接近 99%，这表明热启动场景下的 DDI 预测问题已取得十分显著的进展。

3.3.3 冷启动场景下的实验结果

在冷启动场景中，训练集和测试集是基于药物划分的。模型在训练过程中仅能利用训练集中的药物信息进行学习，而在测试过程中需要对包含新型药物的 DDI 进行预测，因此冷启动场景更考验模型对新型药物的泛化能力^[55,115]。记所有药物的集合为 $\mathcal{G}$ ，用于训练的药物集合为 $\mathcal{G}_{old}$ ，用于测试的药物集合为 $\mathcal{G}_{new}$ ，则冷启动场景的要求为 $\mathcal{G}_{new} \cup \mathcal{G}_{old} = \mathcal{G}$ 且 $\mathcal{G}_{new} \cap \mathcal{G}_{old} = \emptyset$ 。在冷启动场景中，数据集被划分为以下三个部分：

$$\mathcal{D}_{train} = \{(D_x, r, D_y) | D_x \in \mathcal{G}_{old} \wedge D_y \in \mathcal{G}_{old}\} \quad (3.13)$$

$$\mathcal{D}_{s1} = \{(D_x, r, D_y) | D_x \in \mathcal{G}_{new} \wedge D_y \in \mathcal{G}_{new}\} \quad (3.14)$$

$$\begin{aligned} \mathcal{D}_{s2} = & \{(D_x, r, D_y) | D_x \in \mathcal{G}_{new} \wedge D_y \in \mathcal{G}_{old}\} \\ & \cup \{(D_x, r, D_y) | D_x \in \mathcal{G}_{old} \wedge D_y \in \mathcal{G}_{new}\} \end{aligned} \quad (3.15)$$

其中 $\mathcal{D}_{train}$ 为训练集，即仅由训练药物构成的样本的集合； $\mathcal{D}_{s1}$ 为测试集的 S1 划分，其包含仅由测试药物构成的样本的集合； $\mathcal{D}_{s2}$ 为测试集的 S2 划分，其包含由一种测试药物和一种训练药物组成的样本的集合。

本节进行三次独立实验并计算平均表现。在每次实验中，随机抽取 20% 的药物作为新药物，从而在三次实验中构建不同的测试集。此外，为确保生成的样本符合冷启动场景的条件，训练集中的负样本基于训练药物生成，测试集中的负样本基于测试药物生成。为确保所有模型使用相同的训练集、验证集和测试集，新药物选择和负样本生成均在模型训练之前进行。表3.2显示了 HDN-DDI 等模型在冷启动场景中 S1 划分与 S2 划分上的表现。

如表3.2所示，冷启动场景下所有模型的表现均出现大幅度下降，然而 HDN-DDI 的表现仍然优于基线模型。在 S1 划分上，相比于基线模型 PEB-DDI，HDN-DDI 的准确率提高了 4.19%，F1 值提升了 1.54%。相比于基线模型 BDN-DDI，HDN-DDI 的 AUROC 提升了 3.38%，AUPR 提升了 3.81%。在 S2 划分上，相比于

表 3.2 冷启动场景下 HDN-DDI 等模型在 S1 和 S2 划分上的表现 (%)

模型	S1 划分				S2 划分			
	ACC	AUROC	AUPR	F1 值	ACC	AUROC	AUPR	F1 值
GMPNN-CS	68.57	74.96	75.44	65.32	77.72	84.84	84.87	78.29
DSN-DDI	73.42	81.79	81.82	70.34	81.92	91.01	91.09	80.18
BDN-DDI	74.88	86.10	85.84	70.21	85.35	93.68	93.64	84.34
SRR-DDI	67.16	73.72	73.35	65.56	75.88	83.19	83.32	73.24
PEB-DDI	75.65	83.08	82.60	75.88	83.01	91.94	92.05	81.52
HDN-DDI	79.84	89.48	89.65	77.42	89.43	95.92	95.95	89.34

注：每项指标的最高值用加粗字体显示。

基线模型 BDN-DDI，HDN-DDI 的准确率提升了 4.08%，AUROC 提升了 2.24%，AUPR 提升了 2.31%，F1 值提高了 5.00%。这表明 HDN-DDI 能够更准确地预测涉及新药物的 DDI。尽管模型训练过程中的负样本源于随机生成，HDN-DDI 仍然能够有效学习并展现出对原始正样本较好的预测能力。

综合以上结果可以得出，HDN-DDI 在热启动场景中接近或优于最优基线模型，在冷启动场景中能够达到最优预测表现。

3.3.4 消融实验

为了进一步分析 HDN-DDI 中各模块的贡献，本节在任务难度更大的冷启动场景下进行了消融实验，以明确不同模块对模型性能的影响。消融实验中所使用的 HDN-DDI 变体的含义如下：

(1) wo-SEM：移除子结构提取模块，并使用传统分子图替代层次化分子图，以验证对应模块和层次化分子图在分子表示学习和 DDI 预测中的有效性；

(2) wo-EDV：使用原始的双视图表示学习方法^[55]替代增强型双视图表示学习方法，从而评估增强型方法在捕获相互作用信息方面的重要性；

(3) with-pool：采用常见的求和函数计算全局表示，不区分分子结构的贡献，从而评估识别关键子结构这一方式对模型预测性能的影响；

(4) with-SAG：采用 SAGPooling 方法^[116]生成全局表示，而非基于分子级节点生成，从而验证后者在全局表示学习中的有效性。

表3.3显示了 HDN-DDI 及其变体在 DrugBank 数据集 S1 和 S2 划分上的表现。实验结果表明，HDN-DDI 在所有指标上均优于其变体，这表明 HDN-DDI 中的层次化分子图和增强型双视图表示学习方法在提高模型性能方面均发挥了重要作用。针对消融实验结果的详细分析如下：

表 3.3 冷启动场景下 HDN-DDI 及其变体在 S1 和 S2 划分上的表现 (%)

模型	S1 划分				S2 划分			
	ACC	AUROC	AUPR	F1 值	ACC	AUROC	AUPR	F1 值
wo-SEM	69.96	77.23	77.62	66.83	78.01	85.80	85.79	78.15
wo-EDV	76.15	85.44	84.91	74.00	84.98	92.60	92.17	84.97
with-pool	78.56	87.56	87.33	76.89	87.18	94.42	94.22	87.24
with-SAG	78.20	88.05	88.15	75.41	87.67	94.71	94.70	87.60
HDN-DDI	79.84	89.48	89.65	77.42	89.43	95.92	95.95	89.34

注：每项指标的最高值用加粗字体显示。

(1) 子结构提取模块在 HDN-DDI 中具有最重要的作用。移除该模块后，wo-SEM 变体的性能显著下降，而保留该模块的 HDN-DDI 及其它变体仍然具有较好的表现。这一结果表明，子结构提取模块能够有效整合分子中的化学子结构信息以及层次结构信息，进而可以改善后续的药物分子表征学习。

(2) 增强型双视图表示学习同样能够提升模型表现。相较于采用原始双视图表示学习方法^[55]的 wo-EDV，采用了增强型方法的 HDN-DDI 在 S1 划分与 S2 划分上的准确率分别提高了 3.69% 和 4.45%。这表明增强型方法在捕获药物结构信息和相互作用信息方面更有优势。

(3) 层次化分子图能够有效学习分子的全局表示。如表3.3所示，基于分子级节点提取全局表示的 HDN-DDI 的表现优于基于求和池化的 with-pool 以及基于 SAGPooling 池化的 with-SAG。此外，相比于上述变体模型，HDN-DDI 的计算成本更低，这进一步表明层次化分子图在分子表示学习中的有效性。

3.3.5 实际应用场景评估

为了评估 HDN-DDI 的实际应用价值，本节利用美国食品药品监督管理局 (Food and Drug Administration, FDA) 早期批准的药物^[117]构建训练集以及验证集，利用 FDA 在近十年内批准的药物构建测试集，从而模拟实际应用场景，并对各模型的预测性能进行评估。

数据集划分依据如下：对于 DrugBank 中的每个药物分子，收集所有以该药物分子为活性成分的 FDA 批准药物，并将这些药物中最早的批准日期确定为该药物分子的批准日期，最后根据药物的批准日期将其分为两组，即现有药物（2015 年之前批准，包含 1585 种药物）和新药物（2015 年及以后批准，包含 121 种药物）。最终，训练集和验证集包含 166 342 个仅涉及现有药物 DDI 元组，而测试数据集包含 25 466 个至少涉及一种新批准药物的 DDI 元组。此外，采用与^[7]小节相同的方式生成负样本，即不存在相互作用的药物对。

由于 DSN-DDI^[55] 和 BDN-DDI^[56] 在基线模型中表现较好, 本节重点对 HDN-DDI 与这两个模型进行比较。此外, DSN-DDI 仅具有 4 个编码块, 而 BDN-DDI 与 HDN-DDI 均具有 6 个编码块, 为保证对比的公平性, 本节同时评估了 DSN-DDI 原始模型及其具有 6 个编码块的变体。所有模型的超参数均根据对应论文中指定的数值进行设定。表3.4显示了 HDN-DDI 等模型在 FDA 新批准药物上的表现, 图3.5进一步显示了各模型的受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线与精确率-召回率 (Precision-Recall, PR) 曲线。

如表3.4所示, HDN-DDI 在四项评估指标上均具有显著优势。相较于基线模型 BDN-DDI, HDN-DDI 的准确率提升了 5.79%, AUROC 提升了 4.84%, AUPR 提升了 4.62%, F1 值提高了 6.09%。

表 3.4 HDN-DDI 等模型在 FDA 新批准药物上的表现 (%)

模型	ACC	AUROC	AUPR	F1 值
DSN-DDI	74.34	81.86	81.12	74.54
DSN-DDI (6 个编码块)	76.97	84.60	84.37	76.98
BDN-DDI	82.35	90.00	90.11	82.12
HDN-DDI	88.14	94.84	94.73	88.21

注: 每项指标的最高值用加粗字体显示。

如图3.5所示, 在召回率较低的区间内, HDN-DDI 的 PR 曲线更为稳定, 这表明 HDN-DDI 不仅具有更好的综合预测性能, 其在高置信度条件下也能保持更高的精确率, 因此能够在一定程度上减少不必要的药物理化实验验证。

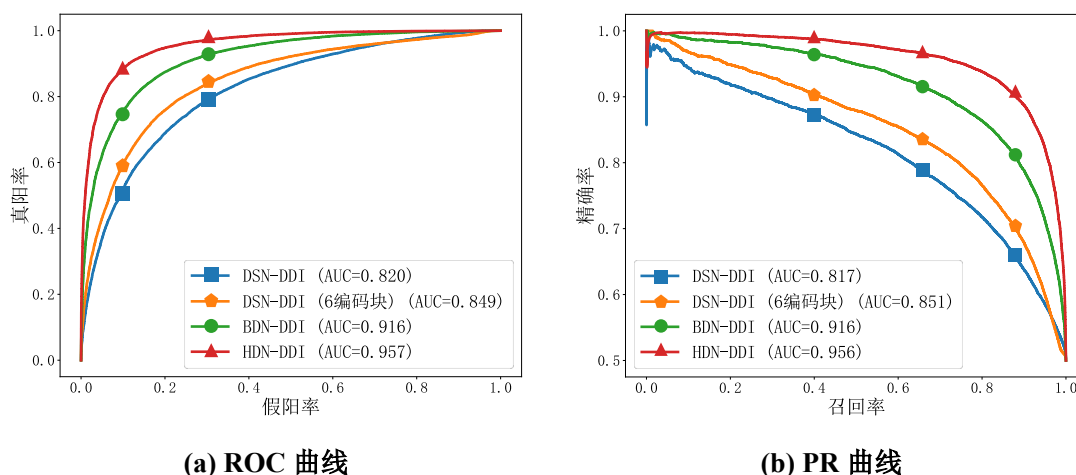


图 3.5 HDN-DDI 及基线模型的预测曲线

综上所述, HDN-DDI 在实际应用场景下同各模型的对比结果与其在冷启动场景下的对比结果一致, 即 HDN-DDI 能够达到最优表现, 这表明 HDN-DDI 对于新药物具有较好的泛化能力, 因而具备在实际场景中的应用潜力。

3.3.6 可视化分析

为了验证 HDN-DDI 识别出的关键子结构的合理性,本节结合具体的 DDI 案例对 HDN-DDI 的预测过程进行可视化分析。

由于 HDN-DDI 具有多个 HDN 编码块,并且每个 HDN 编码块的更新层能够计算并输出各个子结构的贡献值,故 HDN-DDI 最终会输出子结构在多个深度下的贡献值。基于此,本节对所有更新层的输出结果计算平均值,从而得到子结构的综合贡献值,最终选取贡献值排名前 40% 的子结构作为模型识别的关键子结构。图3.6显示了 HDN-DDI 对 3 组 DDI 的子结构识别情况。

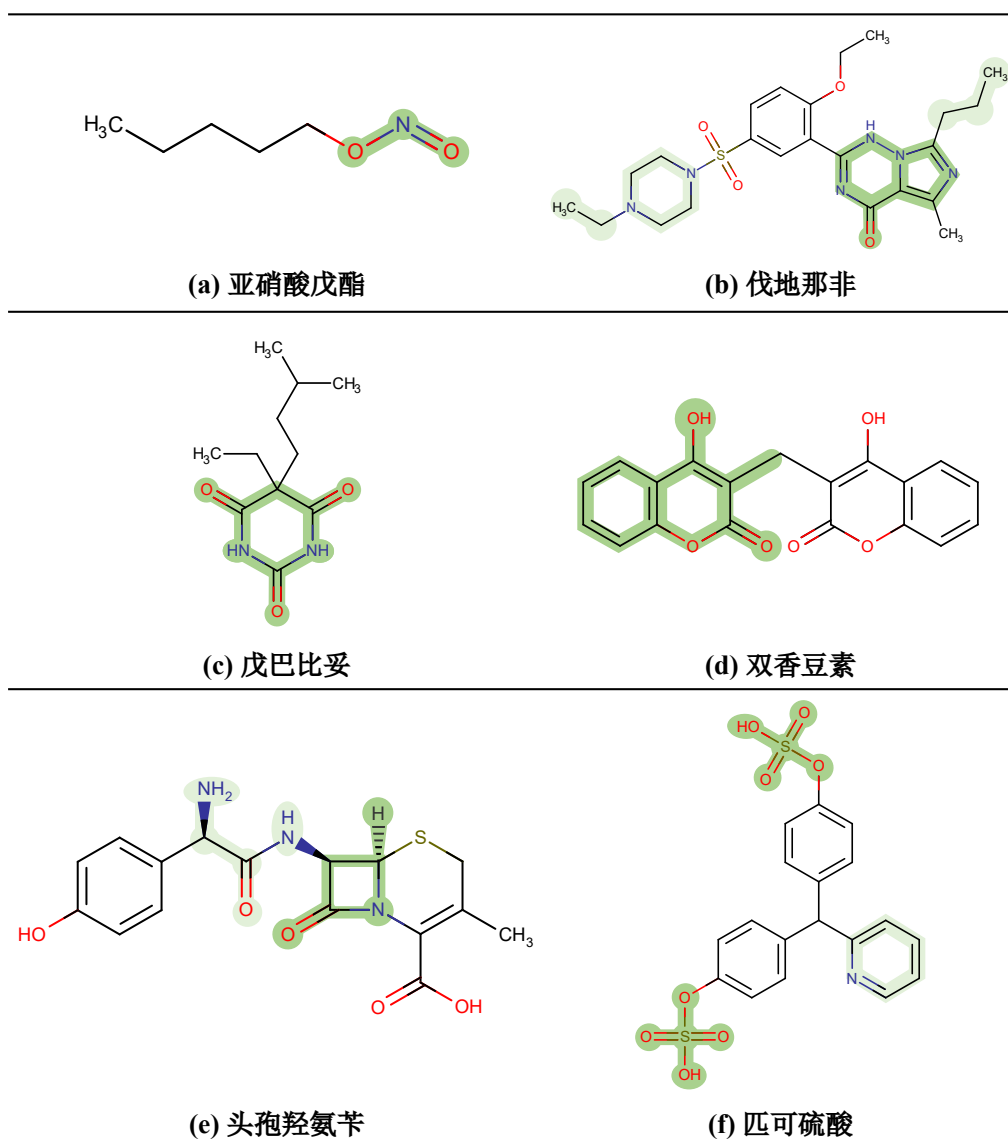


图 3.6 HDN-DDI 针对 3 组 DDI 的子结构识别情况

注: 贡献排名前 40% 的子结构被突出显示, 子结构颜色越深表示其贡献越大。

如图3.6所示, HDN-DDI 能够有效识别亚硝酸戊酯与伐地那非等药物中的关键子结构。结合药物的特性以及相互作用机制, 可以对 HDN-DDI 的识别情况进行深入分析, 具体如下:

(1) 亚硝酸戊酯和伐地那非。前者是一种用于缓解心绞痛的有机硝酸盐,后者是一种用于治疗勃起功能障碍的磷酸二酯酶-5 抑制剂。这两种药物都有抗高血压的作用,其联合使用可以迅速降低血压^[3,6]。具体而言,伐地那非是 P-糖蛋白(P-glycoprotein, P-gp)的底物,而亚硝酸戊酯会抑制 P-gp 的活性。两者同时服用可以提高伐地那非在患者体内的浓度,从而增强其治疗效果^[3,56]。亚硝酸戊酯的治疗效果主要取决于亚硝基官能团,而伐地那非的主要官能团包括吡啶环和酰胺键^[56]。赤藓醇四硝酸酯和他达拉非^[6]也会产生类似的相互作用。如图3.6的(a)和(b)所示,HDN-DDI 准确识别了这些关键官能团。

(2) 戊巴比妥和双香豆素。前者是一种用于催眠和镇静的巴比妥类衍生物,后者是一种抗凝血剂,两者联合使用则会降低双香豆素的凝血作用^[6,118]。具体而言,戊巴比妥可以增强肝微粒体酶的活性,从而加速双香豆素的代谢并削弱其治疗效果。巴比妥酸是巴比妥类衍生物中的关键官能团,而在双香豆素中苯环和呋喃环都发挥着十分重要的作用^[55,118]。如图3.6中的(c)和(d)所示,HDN-DDI 准确地识别出了对应药物内的关键官能团。

(3) 头孢羟氨苄和匹可硫酸。前者是一种头孢菌素类抗生素,用于治疗细菌感染,其主要官能团为头孢菌素 β -内酰胺环^[119]。后者是一种结肠清洁泻药,其主要子结构是磺酸基团^[52]。两种药物共同给药会导致匹可硫酸的疗效降低^[6,52]。如图3.6的(e)和(f)所示,HDN-DDI 准确识别了 β -内酰胺环和磺酸基团。这些 DDI 案例进一步验证了 HDN-DDI 能够精确识别药物分子内的关键子结构。

3.4 讨论

本章提出了一种基于层次化分子图与增强型双视图表示学习的 DDI 预测模型 HDN-DDI。本节将深入分析 HDN-DDI 的技术贡献,并结合 HDN-DDI 的局限性为后续研究提供改进方向。

首先,HDN-DDI 采用了一种更有效的分子表示形式,即层次化分子图。从理论层面分析,层次化分子图的构建依赖于可解释的化学子结构分解算法,因此其能够有效整合具备实际化学意义的子结构信息。同时,层次化分子图在传统分子图的基础上进一步引入了子结构级与分子级节点,能够更全面地表征分子的拓扑结构与官能团信息。此外,层次化分子图能够通过分子级节点提取分子的全局表示,无需额外的池化操作,从而降低了计算需求。消融实验结果进一步表明,层次化分子图能够有效提升模型的预测表现。

其次,HDN-DDI 设计了一种与 DDI 底层作用机制更契合的增强型双视图表示学习方法。从理论层面分析,现有的双视图表示学习方法^[55-56]通常对药物分子中的所有原子进行编码,而 HDN-DDI 中的增强型方法仅对子结构级二部图进

行编码,从而捕获子结构级别的相互作用信息,并且有效滤除了原子级节点产生的噪声。同时,该方法在相互作用视图层对二部图进行随机剪枝,使得同一分子内的子结构级节点能够从不同的邻接范围感知信息,从而有效防止节点表示同质化。此外,由于随机剪枝操作在每个 HDN 编码块中独立执行,同一个子结构节点在不同编码块中具有不同的邻接范围,从而进一步扩展了节点的关注范围并改善了编码器的学习能力。消融实验结果表明,增强型双视图表示学习方法能够有效提升模型表现,同时改善了模型识别关键化学子结构的能力。

然而,HDN-DDI 仍存在一定局限性,例如 HDN-DDI 模型通过 BRICS 算法及附加规则将药物分解为若干子结构,然而当前分解算法的适用范围仍然有限,部分药物分子无法得到充分分解,即分解结果中仍存在一些较大片段。此外,由于原始的 GAT 方法以及 DSN-DDI 等基线模型^[52,55]未考虑边属性,且对边进行编码需要更复杂的网络结构,故 HDN-DDI 未对三键与芳香键等化学键进行充分编码。因此,后续工作可围绕子结构分解算法优化与化学键编码等方面展开。

3.5 本章小结

本章提出了一种基于层次化分子图与增强型双视图表示学习的 DDI 预测模型,即 HDN-DDI,该模型解决了已有研究中存在的模型感知子结构可解释性不足、实际化学子结构相互孤立的问题。实验结果表明,HDN-DDI 具有更好的预测性能,其在热启动场景下达到了 97.93% 和 99.38% 的预测准确率,在冷启动场景和实际场景应用中各项指标均产生了大幅度提升,其中准确率提升约 4%,F1 值提高约 6%,这表明 HDN-DDI 对新型药物具备良好的泛化能力。此外,消融研究验证了 HDN-DDI 中各模块的重要性。通过使用可解释的子结构提取模块分解药物分子,HDN-DDI 可以有效利用化学子结构的信息。通过使用“原子-子结构-分子”层次化分子图表示药物分子,HDN-DDI 整合了层次结构并显著改善了后续分子表示学习。此外,通过增强型双视图表示学习,HDN-DDI 能够有效捕获单个药物内部的层次结构和两个药物之间的相互作用信息。综上所述,通过多类场景中实现更准确的预测,HDN-DDI 为识别潜在 DDI 提供了一种有价值的预测方法,从而促进更安全且更有效的新药研发和多药物联合治疗。

第4章 基于二维图结构的药物-靶点相互作用预测方法

4.1 引言

药物-靶点相互作用预测旨在识别药物与靶点（如蛋白质与基因）之间的相互作用，其对于药物副作用识别与药物再利用等下游任务具有重要意义。如1.2.2小节所述，本文重点研究药物与蛋白质靶点之间的相互作用预测问题。具体而言，针对给定的药物和蛋白质靶点，本文的研究目标为预测两者之间是否存在相互作用（即是否能够形成药靶复合体）。

近年来，基于深度学习的预测方法已成为 DTI 预测领域的主流方法，多项研究^[59,73-74,76-78]利用 CNN 与注意力机制等深度学习技术在 DTI 预测任务中取得了优秀的表现，然而该类工作主要依赖于一维序列进行特征编码，未能充分利用药物与靶点的二维图结构以及三维空间构象。由于一维序列包含的信息有限，深度学习模型难以从中挖掘出原子之间以及氨基酸之间的拓扑邻接关系以及三维空间结构信息，模型的预测性能有待进一步提升。

针对该问题，部分研究^[79-80,82]尝试利用药物与靶点的高维结构进行编码，从而实现了更准确的 DTI 预测，然而现有方法主要关注药物与靶点的内部结构信息，对两者之间的相互作用信息挖掘仍不充分。综上所述，现有的大多数 DTI 预测方法未充分利用二维图或三维结构等高维结构信息，而部分基于图的预测方法则未深入挖掘药物与靶点之间的相互作用信息。

随着分子表示学习方法的持续改进，药物与靶点的表示形式也在不断丰富完善。在 DDI 预测领域，药物的分子图已经成为常见的药物表示形式。在蛋白质表示学习领域，蛋白质的二维图（如氨基酸距离图与接触图等）等近年来逐渐受到关注。蛋白质的距离图是一种紧凑的空间结构表示，其表征了氨基酸之间的成对接触情况。图中的节点代表氨基酸，边则表示氨基酸之间的邻接关系。当两个氨基酸的空间距离小于某一阈值（例如 6 Å）时，对应节点将被视为邻接^[80]。2023 年，Lin 等人^[81]提出蛋白质语言模型 ESM-2，该模型能够根据氨基酸序列推断蛋白质的三维结构，并在预测过程中生成二维接触图。与氨基酸距离图相比，ESM-2 生成的接触图具备相似的空间构象表征能力，但获取难度更低，因此其更适合作为靶点的表示形式^[79,81]。相关研究表明，蛋白质的二维图能够以更高的效率表征空间结构信息，从而有效改善模型在蛋白质结构预测与 DTI 预测等任务上的表现^[81,120]，然而现阶段基于二维图的 DTI 预测研究仍然较少。

针对主流 DTI 预测方法存在的高维结构信息利用不充分、相互作用信息挖掘不足等问题，本章给出了一种基于二维图结构与双视图表示学习的 DTI 预测方法，即 GDN-DTI。该方法的主要贡献如下：（1）二维图结构编码。GDN-DTI 通

过改进后的 BRICS 算法分解药物分子并构建药物的层次化分子图，通过 ESM-2 预训练模型预测靶点的二维接触图，从而充分利用药物与靶点的二维拓扑关系，弥补了一维序列表征信息简单的不足；（2）双视图表示学习。通过引入 HDN-DDI 中的增强型双视图表示学习方法，GDN-DTI 实现了药物内部的子结构之间、靶点内部的氨基酸之间以及药物的子结构与靶点的氨基酸之间的信息传递，从而充分挖掘药物与靶点的结构信息与相互作用信息；（3）相互作用得分矩阵。GDN-DTI 能够计算不同深度的药物与靶点的相互作用得分并形成相互作用得分矩阵，从而有效聚合不同深度的全局表示，同时在一定程度上改善模型预测的可解释性。

4.2 方法模型

4.2.1 GDN-DTI 整体框架

GDN-DTI 包含二维图结构生成模块、GDN 编码器以及 GDN 解码器，如图4.1所示。其中，GDN 解码器由多个 GDN 编码块及其之间的残差连接组成。具体而言，针对药物与靶点的一维序列，GDN-DTI 模型通过二维图结构生成模块获取药物的层次化分子图与靶点的接触图，随后通过 GDN 编码器获取药物与靶点的结构表示、相互作用表示以及全局表示，最终通过 GDN 解码器聚合不同深度的药物与靶点的全局表示并进行 DTI 预测。

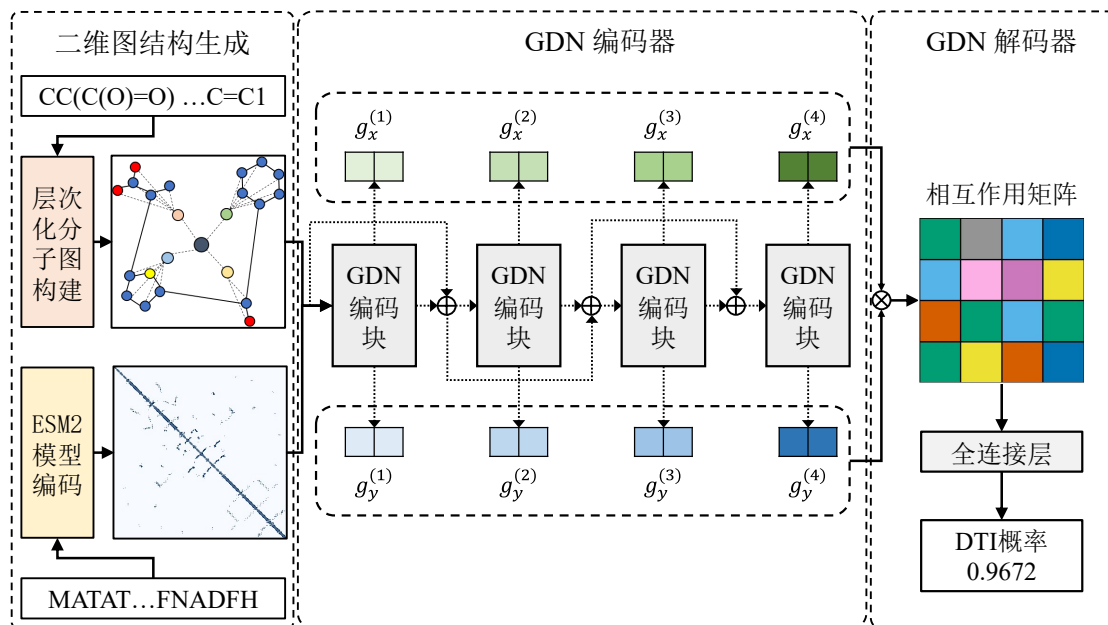


图 4.1 GDN-DTI 总体框架图

4.2.2 二维图结构生成模块

如图4.1中左部分所示，二维图结构生成模块的主要功能为通过相关算法及模型将药物与靶点的一维序列转换为二维图结构。

针对药物的 SMILES 序列, 该模块通过 RDKit 库^[91]获取对应的分子图, 随后采用3.2.3小节中的子结构提取方法将其分解为若干子结构, 进而构建出包含原子级节点、子结构级节点以及分子级节点的层次化分子图。

针对靶点的氨基酸序列, 该模块通过 ESM-2 预训练模型^[81]获取靶点的二维接触图, 其包含氨基酸节点之间的邻接矩阵以及每个节点的高维度表示。由于 ESM-2 模型输出的二维接触图包含所有氨基酸之间的成对接触信息, 即所有氨基酸节点之间均存在边, 为避免不同节点的表示随着层数加深而趋于同质化, 本文遵循 Koh 等人^[79]的处理方法, 仅保留接触值大于 0.5 的边, 并剔除其余较弱的接触边, 从而对处于全连接状态的原始接触图进行合理剪枝。

4.2.3 编码模块输入

GDN-DTI 以一维序列即药物的 SMILES 序列与靶点的氨基酸序列为输入, 通过二维图结构生成模块获取药物的层次化分子图 G_x 以及靶点的二维接触图 G_y , 并将其进一步输入到 GDN 编码器。其中, G_x 与 G_y 均为二维图, 两者均由节点与边组成。

在药物的层次化分子图 G_x 中, 节点可以表示原子、子结构或分子, 边可以表示化学键、原子与子结构以及子结构与分子之间的对应关系。节点特征含义如表A.1所示, 边的特征为 1 至 6 的数字编码, 分别表示单键、双键、三键、芳香键以及不同层次之间的边。

在靶点的二维接触图 G_y 中, 每个节点代表一个氨基酸, 节点之间的边则表示氨基酸之间的邻接关系。节点表示包含两部分: ESM-2 模型输出的 1280 维嵌入表示, 以及与氨基酸自身相关的 33 维理化特征表示, 其含义如表A.3所示。边的特征值为两个氨基酸之间的接触强度, 该强度为离散值, 区间为 [0, 1], 其与两者的空间距离负相关, 即两者距离越近, 接触强度越大。

4.2.4 GDN 编码器

GDN 编码器以药物与靶点的二维图结构为输入, 通过多个 GDN 编码块学习包含结构信息与相互作用信息的药物表示和靶点表示。为充分发挥深层网络的优势并防止模型过拟合, GDN 编码器在不同 GDN 编码块之间引入了残差连接, 如图4.1所示。

针对药物与靶点的二维图结构, GDN 编码器通过 Embedding 嵌入层以及多层感知机对节点及边的特征进行编码。具体而言, 针对药物层次化分子图 G_x 中的任一节点 i , GDN 编码器根据节点标识计算其嵌入表示, 同时利用多层感知机对节点 i 的理化性质进行编码, 最后将两类表示相加, 从而得到节点 i 的初始表示 $h_i^{(0)}$ 。针对 G_x 中的边, GDN 编码器通过嵌入层对不同类型的边进行编码。针

对蛋白质的二维接触图 G_y 中的任一节点 j , GDN 编码器利用多层感知机将 1280 维嵌入表示和 33 维理化特征分别转化为同维度嵌入表示, 并将两类表示求和, 从而得到节点 j 的初始表示 $h_j^{(0)}$ 。针对 G_y 中的边, GDN 编码器采用羟基径向函数将边的单维数值转换为高维表示。

在获取节点表示和边表示后, GDN 编码器通过一系列 GDN 编码块进行迭代式双视图表示学习。GDN 编码块的内部结构如图4.2所示, 其包含结构视图层、相互作用视图层和图更新层。由于药物通常为包含数十个原子的小分子, 而靶点为包含数百个甚至数千个氨基酸的大分子, 因此它们需要通过不同的图编码器进行学习。基于此, GDN 编码块在结构视图层与图更新层中分别利用不同权重的 GAT 对药物与靶点进行学习。

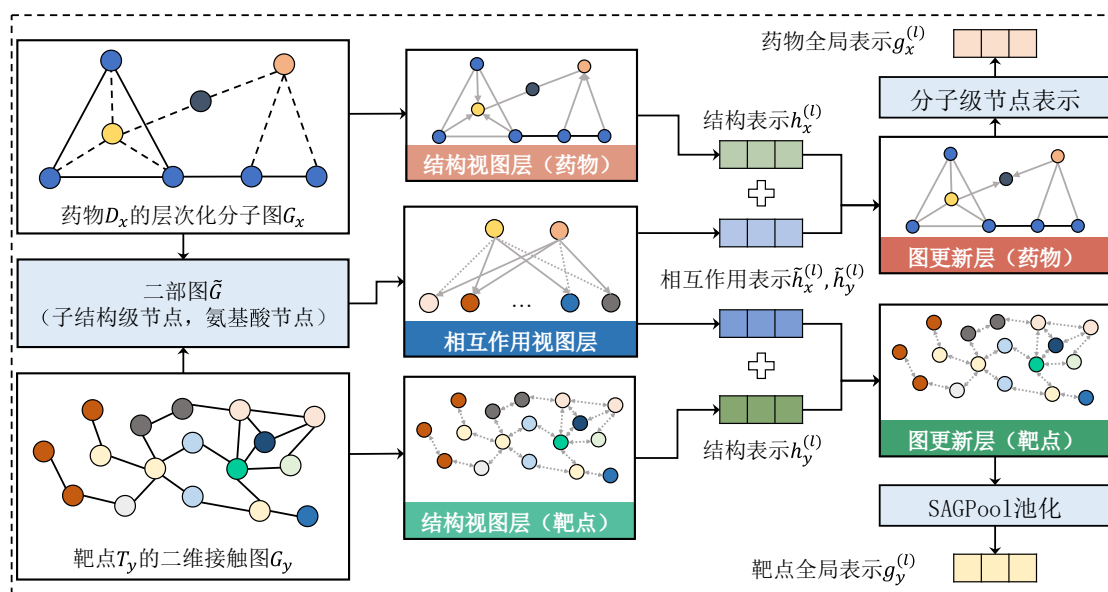


图 4.2 GDN 编码块内部结构示意图

在结构视图层, GDN 编码块采用两个相互独立的 GAT 分别对药物层次化分子图 G_x 与靶点的二维接触图 G_y 进行编码, 从而获取两者内部的结构信息。记第 $l+1$ 个 GDN 编码块中 G_x 或 G_y 的任一节点 i 的编码表示为 $h_i^{(l+1)}$, 则其计算过程如式 (4.1) 至式 (4.3) 所述。

$$h_i^{(l+1)} = \sigma \left(\sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij} \mathbf{W}^{(l+1)} h_j^{(l)} + \mathbf{b}^{(l+1)} \right) \quad (4.1)$$

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(e_{ij}))}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp(\text{LeakyReLU}(e_{ik}))} \quad (4.2)$$

$$e_{ij} = \mathbf{a}^{(l+1)\top} \left(\mathbf{W}^{(l+1)} h_i^{(l)} + \mathbf{W}^{(l+1)} h_j^{(l)} + \mathbf{W}_e^{(l+1)} h_{ij}^{(l)} \right) \quad (4.3)$$

其中 $\mathbf{W}^{(l+1)}, \mathbf{W}_e^{(l+1)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ 和 $\mathbf{a}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 表示第 $l+1$ 个 GDN 编码块中 GAT 的可训练权重, 而 $\mathbf{b}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 表示可训练偏差。T 为转置操作。

LeakyReLU($\cdot$)^[106]和 $\sigma(\cdot)$ 为激活函数。 $d^{(l)}$ 和 $d^{(l+1)}$ 分别为第 $l+1$ 个GDN编码块中结构视图层及图更新层的输入和输出特征维度。 $\mathcal{N}_i$ 表示节点 i 的所有邻接节点的集合。 $\mathbf{h}_j^{(l)}$ 表示节点 j 在第 l 个GDN编码块的表示, $\mathbf{h}_{ij}^{(l)}$ 表示节点 i 与 j 之间的边在第 l 个GDN编码块的表示, α_{ij} 表示节点 i 和 j 之间的注意力系数。

在相互作用视图层,GDN编码块抽取 G_x 中的子结构级节点 $V_s^{(x)}$ 以及 G_y 中的所有节点 $V^{(y)}$ 构建“子结构-氨基酸”二部图 $\tilde{G}$,即 $\tilde{G} = V_s^{(x)} \times V^{(y)}$,随后利用单个GAT对 $\tilde{G}$ 进行编码,从而捕获药物与靶点之间的相互作用信息。

相互作用视图层与结构视图层具有相似的计算原理,但前者主要从二部图 $\tilde{G}$ 中捕获相互作用信息,而后者则从二维图 G_x 与 G_y 中捕获内部结构信息,即两个视图层的编码对象不同,因此采用不同的权重。此外,在二部图 $\tilde{G}$ 中,节点具备嵌入表示,边仅用于表示节点之间的连接关系,无嵌入表示,因此相互作用视图层的GAT无需对边进行编码。

与DDI的发生机制类似,DTI通常由少量的关键子结构与关键肽段所触发^[59,82]。受此启发,GDN编码块在编码过程中随机删除二部图 $\tilde{G}$ 内的边,从而促使模型识别关键的子结构及氨基酸。记第 $l+1$ 个GDN编码块中节点 i 的相互作用表示 $\tilde{\mathbf{h}}_i^{(l+1)}$,则其计算步骤如式(4.4)至式(4.6)所述。

$$\tilde{\mathbf{h}}_i^{(l+1)} = \sigma \left(\sum_{j \in \tilde{\mathcal{N}}_i \cup \{i\}} \beta_{ij} \tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_j^{(l)} + \tilde{\mathbf{b}}^{(l+1)} \right) \quad (4.4)$$

$$\beta_{ij} = \frac{\exp(\text{LeakyReLU}(\tilde{e}_{ij}))}{\sum_{k \in \tilde{\mathcal{N}}_i \cup \{i\}} \exp(\text{LeakyReLU}(\tilde{e}_{ik}))} \quad (4.5)$$

$$\tilde{e}_{ij} = \tilde{\mathbf{a}}^{(l+1)\top} \left(\tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_i^{(l)} + \tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_j^{(l)} \right) \quad (4.6)$$

其中 $\tilde{\mathbf{W}}^{(l+1)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ 和 $\tilde{\mathbf{a}}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 表示第 $l+1$ 个GDN编码块中相互作用视图层GAT的可训练权重, $\tilde{\mathbf{b}}^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ 表示可训练偏差。 $\tilde{\mathcal{N}}_i$ 为节点 i 在修剪后二部图中的邻接节点集合。具体来说,如果子结构级节点 i 来自层次化分子图 G_x ,则 $\tilde{\mathcal{N}}_i$ 表示靶点接触图 G_y 中节点的随机子集,反之亦然。 β_{ij} 表示节点 i 和 j 之间的注意力系数。

针对层次化分子图 G_x 中的原子级与分子级节点,GDN编码块对其进行非线性编码,以确保所有节点具有相同维度的表示。非线性编码与上述GAT共享权重,以节点 k 的相互作用表示 $\tilde{\mathbf{h}}_k^{(l+1)}$ 为例,其计算过程如式(4.7)所示。

$$\tilde{\mathbf{h}}_k^{(l+1)} = \sigma \left(\tilde{\mathbf{W}}^{(l+1)} \mathbf{h}_k^{(l)} + \tilde{\mathbf{b}}^{(l+1)} \right) \quad (4.7)$$

在图更新层,针对 G_x 或 G_y 中任一节点 i 的结构表示 $\mathbf{h}_i^{(l+1)}$ 与相互作用表示 $\tilde{\mathbf{h}}_i^{(l+1)}$,GDN编码块将两者进行拼接并进行LayerNorm归一化^[121],随后利用

ELU 激活函数^[109]进行非线性转换, 如式 (4.8) 所示。

$$\mathbf{h}_i^{(l+1)} = \text{ELU} \left(\text{LayerNorm} \left(\mathbf{h}_i^{(l+1)} \parallel \tilde{\mathbf{h}}_i^{(l+1)} \right) \right) \quad (4.8)$$

在此基础上, GDN 编码块采用两个相互独立的 GAT 分别更新 G_x 与 G_y , 计算原理与式 (4.1) 至式 (4.3) 相同。

通过结构视图层、相互作用视图层和图更新层的编码, GDN 编码块能够获取更新后的二维图。针对层次化分子图 G_x , GDN 编码块提取分子级节点 v_{x_mol} 的更新后表示 $\mathbf{h}_{x_mol}^{(l+1)}$ 作为药物分子的全局表示 $\mathbf{g}_x^{(l+1)}$, 如式 (4.9) 所示。

$$\mathbf{g}_x^{(l+1)} = \mathbf{h}_{x_mol}^{(l+1)} \quad (4.9)$$

针对二维接触图 G_y , GDN 编码块通过自注意力图池化方法 SAGPooling^[116]提取靶点大分子的全局表示 $\mathbf{g}_y^{(l+1)}$, 如式 (4.10) 与 (4.11) 所示。

$$\mathbf{g}_y^{(l+1)} = \sigma \left(\sum_{i=1}^{N_y} \theta_i \mathbf{h}_i^{(l)} \right) \quad (4.10)$$

$$\theta_i = \text{softmax} \left(\mathbf{A}_y \mathbf{H}_y^{(l)} \mathbf{W}_s^{(l)} \right) \quad (4.11)$$

其中 $\sigma(\cdot)$ 为非线性激活函数, N_y 为接触图 G_y 中的节点数量, $\mathbf{A}_y \in \mathbb{R}^{N_y \times N_y}$ 为 G_y 的邻接矩阵, $\mathbf{H}_y^{(l+1)} \in \mathbb{R}^{N_y \times d^{(l+1)}}$ 为 G_y 中所有节点在第 $l+1$ 个编码块中的更新后表示, $\mathbf{W}_s^{(l+1)} \in \mathbb{R}^{d^{(l+1)} \times 1}$ 为可训练权重, $\theta_i \in \mathbb{R}$ 为各节点的权重系数。

此外, 针对任意节点 i , GDN 编码块通过 ELU 激活函数对该节点的表示 $\mathbf{h}_i^{(l+1)}$ 进行非线性转换, 随后通过残差连接将节点 i 的原始表示 $\mathbf{h}_i^{(l)}$ 与激活后的输出表示相加, 并采用 Dropout 方法进行正则化处理, 从而获得当前编码块的最终输出, 如式 (4.12) 所示。

$$\mathbf{h}_i^{(l+1)} = \text{Dropout} \left(\mathbf{h}_i^{(l)} + \text{ELU} \left(\mathbf{h}_i^{(l+1)} \right) \right) \quad (4.12)$$

4.2.5 GDN 解码器

通过多个 GDN 编码块的迭代编码, GDN 编码器生成多个深度的药物全局表示和靶点全局表示。针对这些全局表示, GDN 解码器采用矩阵乘法计算不同深度之间的药物与靶点的相互作用分值, 从而得到相互作用得分矩阵 $\mathbf{S}$ 。记 GDN 编码块的数量为 L , 全局表示的编码维度为 d , 则得分矩阵 $\mathbf{S}$ 的计算过程如式 (4.13) 所示。

$$\mathbf{S} = \mathbf{g}_x \mathbf{g}_y^T \quad (4.13)$$

其中 $\mathbf{g}_x \in \mathbb{R}^{L \times d}$ 为层次化分子图 G_x 的全局表示矩阵, $\mathbf{g}_y \in \mathbb{R}^{L \times d}$ 为靶点接触图的 G_y 的全局表示矩阵, $\mathbf{S} \in \mathbb{R}^{L \times L}$, T 为转置运算。

在获取得分矩阵 $\mathbf{S}$ 后, GDN 解码块通过线性层对 $\mathbf{S}$ 进行汇总, 最终通过 sigmoid 激活函数计算药物与靶点之间存在相互作用的概率值 p , 计算过程如式 (4.14) 所示。

$$p = \text{sigmoid}(\mathbf{W} \cdot \text{Flatten}(\mathbf{S}) + b) \quad (4.14)$$

其中 $\text{Flatten}(\cdot)$ 为矩阵扁平化操作, $\text{Flatten}(\mathbf{S}) \in \mathbb{R}^{(L \times L) \times 1}$ 。 $\mathbf{W} \in \mathbb{R}^{1 \times (L \times L)}$ 为可学习的权重向量, $b \in \mathbb{R}$ 为可学习的偏差值。

4.2.6 损失函数

本章所研究的 DTI 预测旨在预测元组 (D_x, P_y) 是否存在, 即药物 D_x 与靶点 P_y 之间是否存在相互作用, 因此可视为二分类问题。在二分类任务中, 交叉熵损失函数最为常见。相较于其它损失函数, 交叉熵损失函数易于优化、可解释性较强并且能够提高数值计算的稳定性, 故在 DTI 预测任务中^[59,76,78]得到广泛应用。因此, 本文采取二元交叉熵损失函数对模型进行端到端训练, 具体计算方式如式 (4.15) 所示。

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \quad (4.15)$$

其中 N 为训练集样本的总数量, p_i 为模型输出的第 i 个样本为真的概率值, y_i 为第 i 个样本的实际标签, 即第 i 个样本存在 DTI 时 $y_i = 1$, 否则 $y_i = 0$ 。

4.3 实验及结果分析

4.3.1 实验设置

本节将依次介绍实验的设置情况, 即数据集、实验环境、模型超参数、评估指标以及基线模型。

4.3.1.1 数据集

针对模型的训练与评估, 实验采用两个免费开放的公共数据集: DrugBank^[6]和 BioSNAP^[94]。DrugBank 和 BioSNAP 数据集均为综合数据集, 两者均记录了药物-药物相互作用与药物-靶点相互作用等多项药物相关的数据, 本章仅利用其中与 DTI 预测任务相关的数据。

就数据集规模而言, BioSNAP 数据集包含 4510 类药物与 2181 类靶点, 共计 13 830 个正样本 (存在相互作用的药物-靶点对) 和 13 634 个负样本 (不存在相互作用的药物-靶点对)。DrugBank 数据集包含 6655 种药物与 4294 种靶点。由于该数据集仅包含正样本, 本章采取与 HyperAttentionDTI^[77]相同的药物筛选标准及负样本生成方法, 即排除过于微小的以及无法被 RDKit 工具包识别的药物

分子,并随机选择数据集中无相互作用记录的药物与靶点对作为负样本。负样本的生成数量与数据集中原有的正样本数量相同,处理后的 DrugBank 数据集包含 17 511 个正样本与 17 511 个负样本。

4.3.1.2 实验环境与模型超参数

GDN-DTI 同样基于 Pytorch 框架构建,其依赖的工具库包括 torch_geometric 和 ESM 等,模型训练与评估的硬件环境则与第 3 章相同。

在模型参数方面,如图4.1所示,GDN-DTI 由 GDN 编码器和 GDN 解码器组成。针对 GDN-DTI 超参数设置,本章采取交叉验证的方式确定 GDN 编码块的数量以及模型内部的超参数,训练过程中的超参数则参照 MCANet 设定。具体而言,GDN-DTI 包含 4 个 GDN 编码块,每个 GDN 编码块包含两个结构视图层、一个相互作用视图层和两个图更新层,所有层均使用单层多头 GAT 进行图编码。其中,结构视图层和相互作用视图层输入 200 维表示,输出 100 维表示,而更新层输入与输出均为 200 维表示。所有 GAT 均使用四个注意力头。此外,模型训练期间使用的批次大小为 64,学习率通过 CyclicLR^[122]动态调整,基础学习率为 1.0×10^{-4} ,最大调整倍率为 10 倍,权重衰减率为 1.0×10^{-4} 。

4.3.1.3 评估指标

在模型评估环节,本章遵循已有工作^[63,76-77]的设定,采用准确率、精确率、召回率、AUROC 和 AUPR 共五项指标对不同模型进行对比,其中准确率和 AUROC 用于评估模型预测的准确程度,而精确率、召回率和 AUPR 则用于评估模型对正样本的预测能力。

4.3.1.4 基线模型

在基线模型方面,本章选取了 DTI 领域近年来的代表性模型作为基线模型,并依照其原始论文中的值来设定超参数。各模型的名称及主要贡献如下:

(1) MolTrans^[76]: 利用无标签数据库构建片段分解算法,将 SMILES 序列与氨基酸序列分解为若干片段,利用 Transformer 分别对两个片段序列编码并计算相互作用得分矩阵,最终利用 CNN 进行再次编码并实现 DTI 预测;

(2) HyperAttentionDTI^[77]: 利用 CNN 对药物与靶点的序列进行编码,并通过超注意力机制捕获两者之间的相互作用信息,从而实现 DTI 预测;

(3) MCANet^[63]: 利用一维 CNN 对 SMILES 序列与氨基酸序列编码,在此基础上通过交叉注意力机制捕获药物与靶点之间的双向影响,最终拼接药物表示与靶点表示并通过全连接层实现 DTI 预测。

(4) BINDTI^[60]: 采用图神经网络与 ACMix 模型对药物的分子图与靶点的

氨基酸序列进行编码,并通过双向意图网络融合药物和靶点的特征,最终通过全连接层实现 DTI 预测。

(5) PSICHIC^[79]: 采用连接树分解算法将药物分子分解为多个官能团,采用 ESM-2 模型生成靶点的接触图并利用图聚类算法将其分解为多个功能区,最终通过图神经网络学习两者的结构信息与相互作用信息,从而实现 DTI 预测。

4.3.2 实验结果

本章主要在热启动场景中对 GDN-DTI 与各基线模型进行对比,即通过对全部样本进行随机划分获得训练集与测试集,两者的比例为 8:2,该比例与现有研究^[63]中的实验保持一致。模型训练及选择采用 5 折交叉验证,即在训练过程中,将训练集会进一步划分成 5 个子集,其中 4 个子集用于模型训练,剩余 1 个子集用于模型验证。此外,为确保对比的公平性,GDN-DTI 在不同数据集上使用相同的超参数,例如学习率、编码维度以及注意力头数等。

表4.1和表4.2分别显示了 GDN-DTI 与各基线模型在 DrugBank 数据集和 BioSNAP 数据集上的表现。实验结果表明,GDN-DTI 在多个评估指标上接近或优于现阶段领域最优模型。

如表4.1所示,在 DrugBank 数据集上,GDN-DTI 的召回率、AUROC 和 AUPR 优于基线模型。与此前最优模型 PSICHIC 相比,GDN-DTI 的对应指标分别提升了 1.3%、0.18% 以及 0.58%,而准确率与精确率分别下降了 0.24% 与 0.77%。与基线模型 MCANet 相比,GDN-DTI 的准确率、召回率和 AUROC 分别提升了 0.75%、1.89% 和 0.94%。整体而言,GDN-DTI 在 DrugBank 数据集上的综合表现优于 MCANet,并且接近此前最优模型 PSICHIC。

表 4.1 GDN-DTI 等模型在 DrugBank 数据集上的表现 (%)

模型	准确率	精确率	召回率	AUROC	AUPR
MolTrans	78.48	77.01	81.97	85.98	86.67
HyperAttentionDTI	81.00	79.90	82.90	88.90	88.44
MCANet	82.60	82.42	82.74	89.71	90.45
BINDTI	80.43	79.10	81.33	87.42	86.95
PSICHIC	83.59	83.05	83.33	90.47	89.94
GDN-DTI	83.35	82.28	84.63	90.65	90.52

注: 每项指标的最高值用加粗字体显示。

如表4.2所示,在 BioSNAP 数据集上,GDN-DTI 的准确率、精确率和 AUROC 优于基线模型,其中精确率与 PSICHIC 持平。相比于 PSICHIC, GDN-DTI 的其余四项指标分别提升了 0.59% 至 1.50%。此外, GDN-DTI 的召回率较 HyperAtten-

tionDTI 低 0.66%，AUPR 较 MCANet 低 0.37%。整体而言，GDN-DTI 在 BioSNAP 数据集上的综合表现优于 PSICHIC，并且接近 MCANet 及 HyperAttentionDTI。

表 4.2 GDN-DTI 等模型在 BioSNAP 数据集上的表现 (%)

模型	准确率	精确率	召回率	AUROC	AUPR
MolTrans	79.74	79.32	80.87	86.06	86.34
HyperAttentionDTI	84.20	82.80	86.61	91.10	91.89
MCANet	84.58	83.57	85.76	91.74	92.24
BINDTI	81.39	80.14	83.06	88.25	87.84
PSICHIC	84.14	83.72	84.45	91.21	90.60
GDN-DTI	84.78	83.72	85.95	91.80	91.87

注：每项指标的最高值用加粗字体显示。

4.3.3 消融实验

为了验证 GDN-DTI 模型中各模块的有效性，本章在 DrugBank 数据集上进行了消融实验，实验过程中使用的 GDN-DTI 变体如下：

(1) **wo-drug-graph**：针对药物分子，使用 SMILES 序列替代层次化分子图，并采用注意力机制替代 GAT，从而验证层次化分子图在 DTI 预测中的有效性；

(2) **wo-target-graph**：针对靶点，采用氨基酸序列替代靶点的接触图，采用注意力机制替代 GAT，即通过双视图表示学习对药物的层次化分子图与靶点的氨基酸序列进行编码，从而验证靶点接触图在 DTI 预测中的有效性；

(3) **wo-dual-view**：移除 GDN 解码块中的双视图表示学习机制，即每个编码块内仅通过单层 GAT 对药物与靶点的二维图进行编码，从而评估双视图表示学习对 GDN-DTI 模型表现的影响。

(4) **wo-matrix**：移除 GDN 解码块中的相互作用得分矩阵，采用多层感知机对不同深度的药物与靶点表示进行汇总并实现 DTI 预测，从而验证相互作用得分矩阵在融合不同深度信息时的有效性。

(5) **wo-residual**：移除 GDN 编码块之间的残差连接，从而评估残差连接对于 GDN-DTI 模型表现的影响。

表4.3显示了 GDN-DTI 及其变体在 DrugBank 数据集上的表现。实验结果表明，GDN-DTI 的表现优于所有变体模型，这进一步验证了各模块对于提高模型性能的重要性。结合表4.3，可以对消融实验结果以及 GDN-DTI 内二维图表示以及相互作用得分矩阵等模块的有效性进行深入分析，具体如下：

(1) 二维图表示形式在 GDN-DTI 中发挥了重要作用。相比于完整模型，移除了层次化分子图的 **wo-drug-graph** 的准确率、AUROC 和 AUPR 分别降低了

表 4.3 GDN-DTI 及其变体在 DrugBank 数据集上的表现 (%)

模型	准确率	精确率	召回率	AUROC	AUPR
wo-drug-graph	79.40	79.72	78.37	86.29	86.16
wo-target-graph	80.95	82.24	78.48	87.85	86.75
wo-dual-view	81.86	82.28	80.82	89.23	89.75
wo-matrix	79.64	79.17	79.97	86.75	85.37
wo-residual	82.07	80.82	83.69	89.24	89.10
GDN-DTI	83.35	82.28	84.63	90.65	90.52

注：每项指标的最高值用加粗字体显示。

3.95%、4.36% 以及 4.36%，移除了接触图的 wo-prot-graph 的对应指标分别降低了 2.40%、2.80% 以及 3.77%。结合图4.3可以看出，GDN-DTI 在训练过程中表现更高且表现曲线更稳定，同时模型训练所需的迭代轮数更少。这表明二维图是一种更有效的表示形式，其更有利于模型训练并且能够提升模型的预测性能。

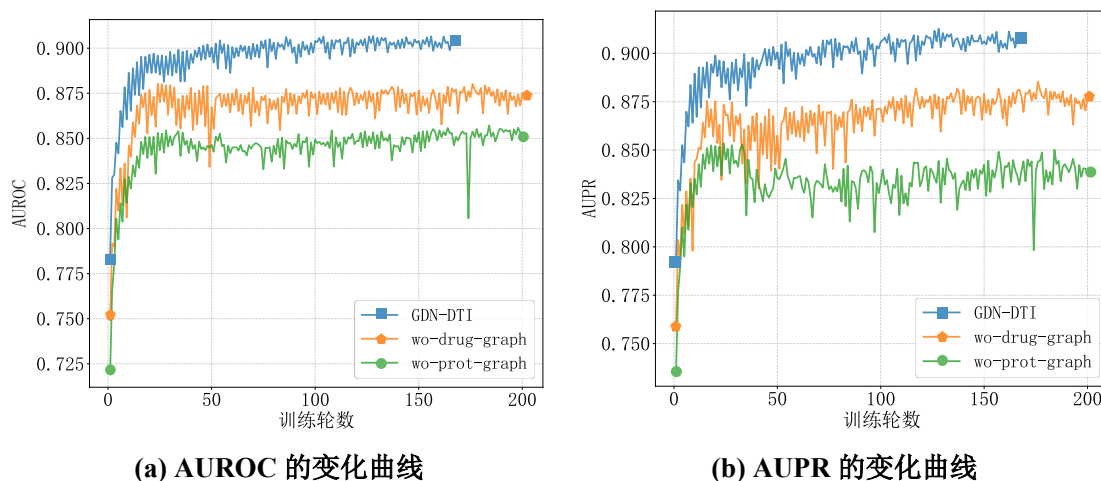


图 4.3 GDN-DTI 及其变体在训练过程中 AUROC 与 AUPR 的变化曲线

(2) 相互作用得分矩阵及残差连接在 GDN-DTI 中发挥了较大作用。与变体模型 wo-matrix 相比，GDN-DTI 的准确率、AUROC 和 AUPR 分别提升了 3.71%、3.90% 以及 5.15%，这表明相互作用得分矩阵能够有效融合不同深度的药物表示与靶点表示。与变体模型 wo-residual 相比，完整模型 GDN-DTI 的相关指标分别提升了 1.28%、1.41% 以及 1.42%。通过引入残差连接，GDN-DTI 能够充分发挥深层网络的优势，同时避免了编码过程中节点表示同质化和原始特征丢失等问题。

(3) 双视图表示学习方法对 GDN-DTI 的表现也具有一定影响。结合表4.3可以看出，相比于完整模型，变体模型 wo-dual-view 的预测表现同样出现明显下降，这进一步验证了双视图表示学习能够有效捕获结构信息以及相互作用信息，从而为最终的 DTI 预测提供支撑。

4.4 讨论

本章给出了一种基于二维图表示与双视图表示学习的 DTI 预测模型 GDN-DTI。相较于一维序列，药物的分子图以及层次化分子图等二维图能够表征更多的拓扑结构信息，而靶点的距离图与接触图等二维图能够在一定程度上反映其三维空间构象。实验结果表明，本章所构建的 GDN-DTI 与 MCANet 等序列编码模型^[63,77]整体表现相近，本节将对该现象进行分析讨论。

首先，GDN-DTI 使用了二维表示形式，即层次化分子图与接触图，因此需要对这两类表示形式进行分析。

针对药物的层次化分子图，结合3.3.4小节中 wo-SEM 变体的表现可以看出，相较于传统的分子图，层次化分子图能够有效改善 DDI 预测模型对药物分子的编码能力。在4.3.3小节的消融实验中，在药物的层次化分子图被替换为 SMILES 序列后，wo-drug-graph 变体模型的表现同样出现大幅度下降，这进一步表明药物的层次化分子图是一种有效的表示形式。

针对靶点的接触图，本节随机选取了 P45059 与 P00488 两种靶点并分别绘制了两者的预测接触图与实际距离图的邻接矩阵，如图4.4与图A.1所示。其中，接触图由 ESM-2 模型^[81]根据氨基酸序列预测生成，而距离图则基于靶点的实际三维结构计算得到，即依次计算两个氨基酸在三维结构中的空间距离，如果距离小于 6 Å，则视为两者之间存在邻接关系^[123]。

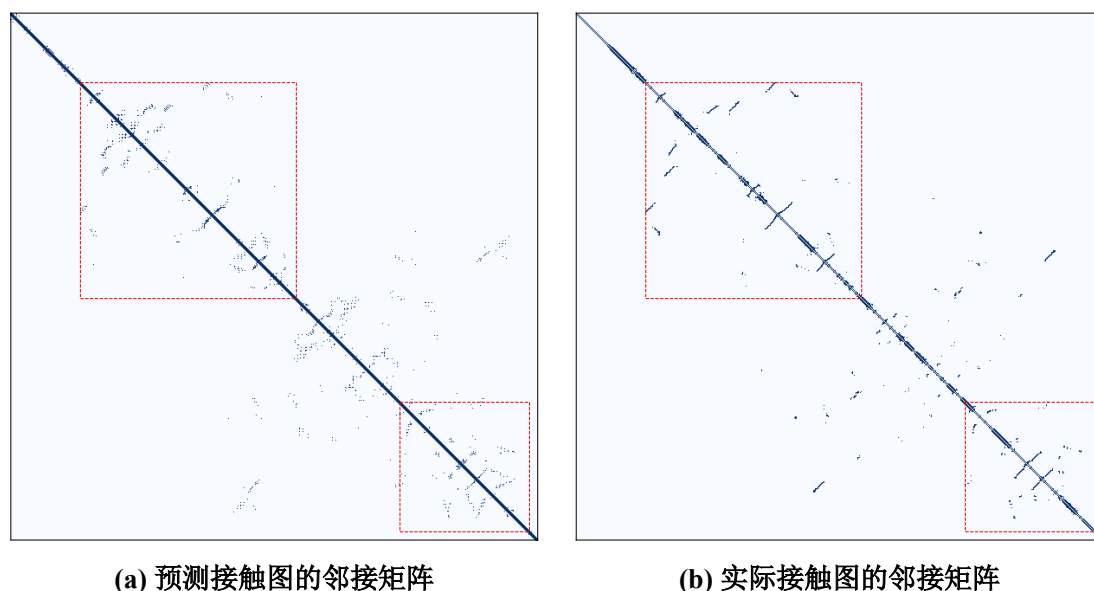


图 4.4 靶点 P45059 的预测接触图与实际距离图的邻接矩阵

结合图4.4与图A.1可以看出：（1）ESM-2 模型的预测接触图的邻接矩阵与实际距离图的邻接矩阵高度相似，即预测接触图能够反映氨基酸的实际邻接关系，该结论亦得到其原始研究^[81]的验证。此外，如4.1节所述，相比于实际距离图，接触图具有更好的信息表征能力，并且其获取难度更低。因此，接触图更适合作为

靶点的二维表示。(2) 无论是预测接触图还是实际距离图,邻接关系最密集的区域均位于邻接矩阵对角线附近。氨基酸节点在邻接矩阵中从左至右与从上至下的排列方式均遵循其在一维序列中的顺序,即邻接矩阵的对角线反映了氨基酸在一维序列中的邻接关系。同时,邻接矩阵又反映了氨基酸在三维空间中的邻接关系。因此,氨基酸之间的三维空间距离及其邻接关系与其在一维序列中的排列顺序密切相关。换言之,尽管靶点的二维表示形式能够在一定程度上作为三维空间结构的紧凑表示,但由于其整体规模较大,因此其信息表征效率弱于一维序列,这可能是 GDN-DTI 与基于序列编码的模型表现相近的原因之一。

其次,GDN-DTI 使用了图注意力网络作为图编码器。相比于 MCANet 等模型中所使用的一维 CNN,图神经网络及图注意力网络具有更复杂的计算机制,因此对训练数据的数量与质量要求更高。在 DTI 预测任务中,现有数据集通常包含数千种靶点,例如 BioSNAP 数据集包含 2181 种靶点,DrugBank 数据集包含 4294 种靶点,而这些靶点的平均约包含 470 个氨基酸,即数据集所包含的图的数量相对较少,而单个图的节点数量较多,这便对图编码器的学习能力提出了更高要求。因此,现有的图编码器难以有效处理靶点的接触图与距离图等复杂图结构,该问题可能是 GDN-DTI 未呈现显著优势的另一重要原因。

通过从二维图表示形式和图编码器这两个角度进行分析,可以得出以下初步结论:(1) 靶点的一维序列与二维图各有优势。相较于一维序列表示,靶点的距离图与接触图等二维图能够更好地表征空间结构信息,而一维序列由于其表示形式更为紧凑,信息表征效率更高;(2) 靶点的二维图作为一种具有大量节点的稀疏图,现有图编码器在对其进行学习时中仍存在一定局限性。因此,与已有的序列编码模型相比,GDN-DTI 并未呈现显著优势。

基于以上分析,本节为今后基于二维图的 DTI 预测工作提供了一些改进方向:首先,应进一步优化靶点的二维表示形式,例如提高接触图的信息表征效率,或者结合可解释算法为蛋白质划分功能区域,构建与“原子-子结构-分子”形式相类似的“氨基酸-肽段-蛋白质”层次化大分子图;其次,应进一步改善图编码器的学习能力,例如针对靶点接触图以及距离图设计专用的图编码器,或者结合大规模蛋白质数据库构建预训练图编码模型,从而构建出更适用于靶点二维图编码任务的图编码器。

4.5 本章小结

现阶段基于深度学习的 DTI 预测已逐渐成为该领域的主流方法,近年来,基于卷积神经网络与注意力机制的 DTI 预测模型不断刷新领域最佳表现。然而,现有工作仍存在着高维结构信息利用不足、相互作用信息挖掘不充分等挑战。针对

这些问题，本章给出了一种基于二维图表示与双视图表示学习的 DTI 预测模型 GDN-DTI。实验结果表明，GDN-DTI 在两个数据集上的 AUROC 与 AUPR 均达到了 90%，并且在多个指标上的表现接近或优于现阶段领域最优模型 PSICHIC 及 MCANet。在 DrugBank 数据集上，GDN-DTI 的召回率提升了 1.3%，AUPR 提升了 0.58%，整体表现优于 MCANet 并与 PSICHIC 接近，在 BioSNAP 数据集上则优于 PSICHIC 并与 MCANet 表现接近。此外，消融实验进一步验证了 GDN-DTI 中二维图表示、双视图表示学习、残差连接以及相互作用得分矩阵的有效性。最后，本章针对实验结果进行了分析，深入讨论了药物层次化分子图的和靶点接触图的有效性，并探讨了 DTI 预测任务中图编码器面临的挑战，同时为今后基于图的 DTI 预测提供了多个改进方向。

第5章 基于多模态信息融合的药物-靶点相互作用预测方法

5.1 引言

如1.2节与4.1节所述,近年来,基于一维序列和基于二维图的预测方法在DTI预测领域均取得了显著进展,然而现有大多数方法仅基于单种表示形式进行研究,即仅利用序列或二维图进行编码,对于不同表示形式的信息融合仍不充分。

本文的4.4节对靶点的二维接触图与一维序列的关系进行了探讨。为进一步比较一维序列与二维图的优势,本节以靶点 P35228 为例,绘制了该靶点距离图的邻接矩阵以及其局部放大图,如图5.1所示。

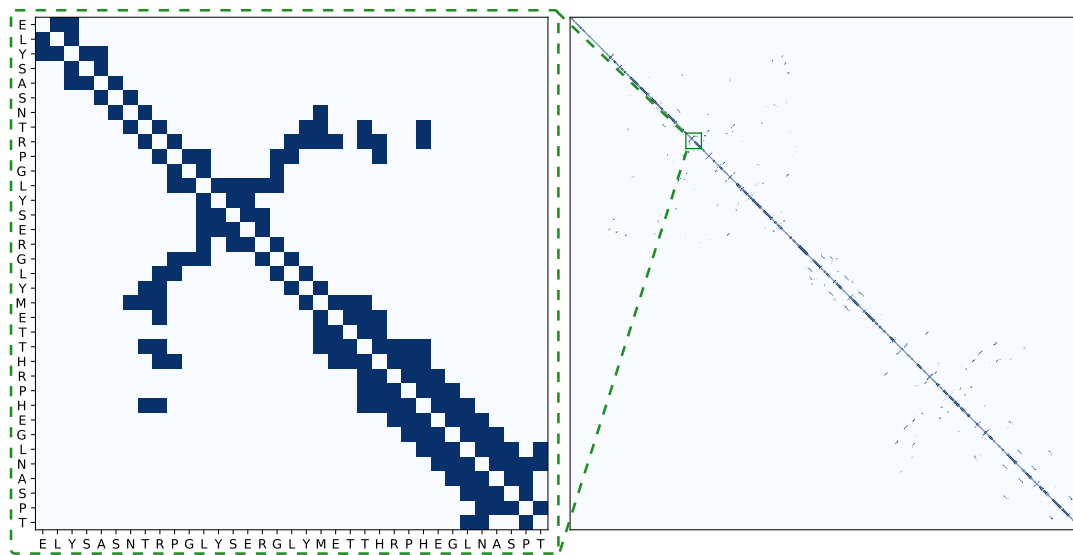


图 5.1 靶点 P35228 的距离图邻接矩阵及其局部放大图

结合图5.1可以观察出:邻接矩阵中的点主要集中在矩阵对角线周围,即氨基酸在三维空间中的距离和邻接关系与其在一维序列中的排列密切相关,换言之,氨基酸序列能够以更紧凑的方式反映靶点的三维结构信息。相比于矩阵对角线,完整的邻接矩阵则具有更多信息,即与一维序列相比,接触图和距离图等二维图能够表征更丰富的空间结构信息。因此,靶点的一维序列与二维图在表征效率与信息总量方面各有优势。

在此基础上,可以对相关的DIT预测方法进行分析。一方面,由于一维序列能够更紧凑地反映氨基酸之间的邻接信息,且序列可以通过卷积神经网络和注意力机制等低复杂度模型进行编码,基于序列的DTI预测模型^[63,77,85]具备训练难度低且计算速度快等优势。然而,氨基酸序列所包含的信息有限,故此类方法的预测准确性仍有待提升。另一方面,由于二维图具备更多结构信息,基于图的DTI预测往往具有更好的表现^[79-80,82]。然而,靶点的二维图作为一种节点众多

且邻接关系稀疏的图，其对编码器的学习能力具有更高的要求。此外，相比于一维卷积神经网络等序列编码器，图编码器参数更多且计算资源需求更高。综上所述，一维序列与二维图以及对应的 DTI 预测方法各有优势，但现有研究通常仅利用一种表示方式进行预测。

针对当前 DTI 预测方法主要依赖单一模态表示、多模态信息融合不足的问题，本章给出了一种基于双视图表示学习与多模态信息融合的 DTI 预测方法 MIF-DTI。该方法的主要贡献如下：（1）多模态特征编码。MIF-DTI 从药物的 SMILES 序列与靶点的氨基酸序列中捕获低层次邻接信息，从药物的层次化分子图与靶点的接触图中捕获高层次拓扑结构信息，从而实现从单一信息源到双信息源的改进，弥补了单模态信息表征能力的不足；（2）双视图表示学习。在序列编码模块与图编码模块中，MIF-DTI 通过双视图表示学习捕获药物与靶点的分子内结构信息与分子间相互作用信息，从而实现更准确的 DTI 预测；（3）多模态融合。MIF-DTI 利用协注意力机制实现序列模态与图模态的融合，并通过相互作用得分矩阵计算 DTI 概率，从而充分利用不同模态的信息。

5.2 方法模型

5.2.1 MIF-DTI 整体框架

MIF-DTI 以药物的 SMILES 序列和靶点的氨基酸序列为输入，其包含序列编码模块、图编码模块以及 MIF 解码模块，整体结构如图5.2所示。

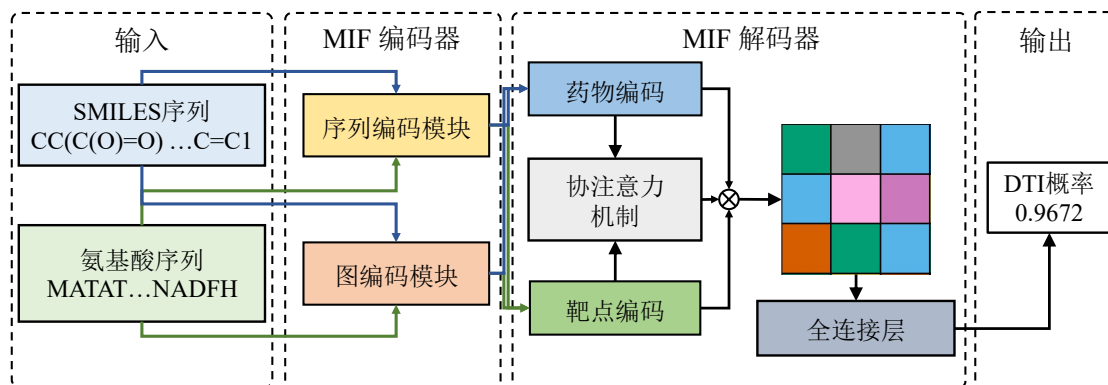


图 5.2 MIF-DTI 总体框架图

具体而言，序列编码模块用于处理一维序列，该模块通过卷积神经网络提取药物与靶点的序列信息，通过注意力机制捕获一维相互作用信息。图编码模块用于处理二维图结构，该模块将 SMILES 序列转换为层次化分子图，并将氨基酸序列转换为接触图，随后通过图注意力网络与双视图表示学习方法捕获结构信息与二维相互作用信息。MIF 解码模块通过协注意力机制计算多模态融合系数以及相互作用得分矩阵，最终通过全连接层聚合得分矩阵并计算 DTI 概率。

5.2.2 序列编码模块

MIF 序列编码模块由嵌入部分和编码部分组成,其中编码部分包含多个 MIF-1D 编码块。模块最终输出多个深度的药物表示与靶点表示,如图5.3所示。

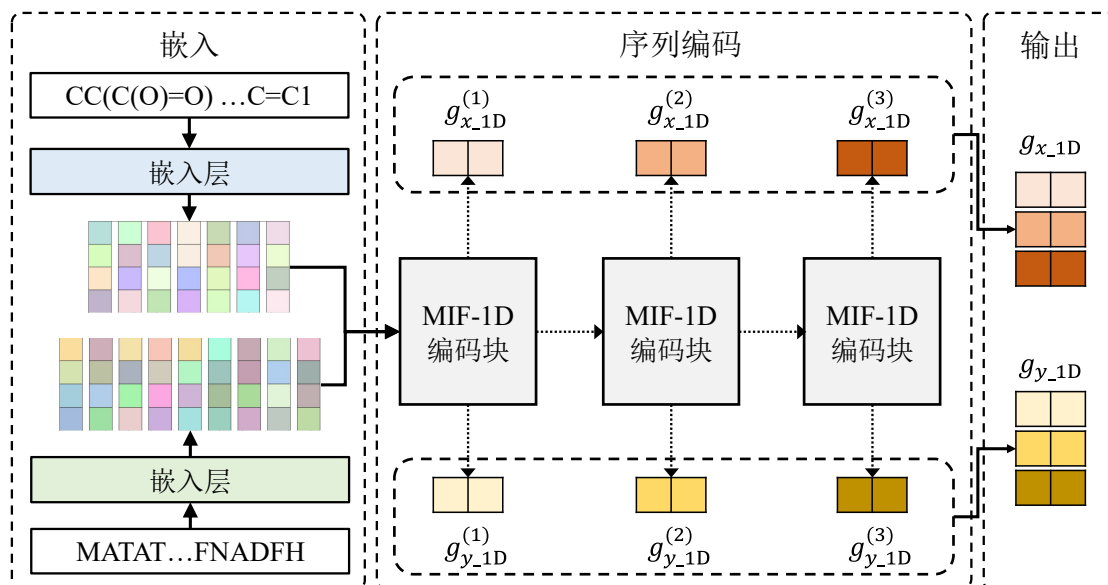


图 5.3 MIF-DTI 中序列编码模块的示意图

针对药物的 SMILES 序列和靶点的氨基酸序列,该模块首先对序列中的字符进行编码。具体而言,SMILES 序列中的原子与化学键等字符被依次映射为整数 1 至 64,氨基酸序列中的字符依次被映射为整数 1 至 24,该编码方式与 MCANet^[63] 相同。对于批次数据,模块依照 MCANet^[63] 的定值处理方式,即设定药物序列的对齐长度为 200,氨基酸序列的对齐长度为 1500。对于小于对齐长度的序列,模块采用 0 进行补齐。对于大于对齐长度的序列,模块对其进行截取,从而确保所有序列长度一致。在此基础上,模块采用不同的嵌入层分别为 SMILES 序列与氨基酸序列中的每一项数字编码生成高维表示。

MIF-1D 编码块的整体处理流程如图5.4所示。对于 SMILES 序列与氨基酸序列的嵌入表示,MIF-1D 编码块利用多层一维 CNN 进行编码。具体而言,每层 CNN 通过在序列上滑动固定大小的卷积核来捕获相邻单位之间的局部依赖。通过对多层 CNN 进行堆叠,MIF-1D 编码块可以逐层提取并集成不同区域的特征,如图5.5所示。

给定 SMILES 序列或氨基酸序列的嵌入表示 $\mathbf{h} \in \mathbb{R}^{L \times d}$,其中 L 为序列初始长度, d 为嵌入维度,则单层一维 CNN 的计算过程如式 (5.1) 所示。

$$\mathbf{h}' = \text{ReLU}(\text{CNN}(\mathbf{h})) \quad (5.1)$$

其中 $\mathbf{h}' \in \mathbb{R}^{L' \times d'}$ 为输出表示, L' 为卷积后序列的长度, d' 为信道数量,即输出维度。记卷积核大小为 k ,步长为 s ,则 L' 的计算方式如 (5.2) 所示。

$$L' = \left\lfloor \frac{L - k + 1}{s} \right\rfloor \quad (5.2)$$

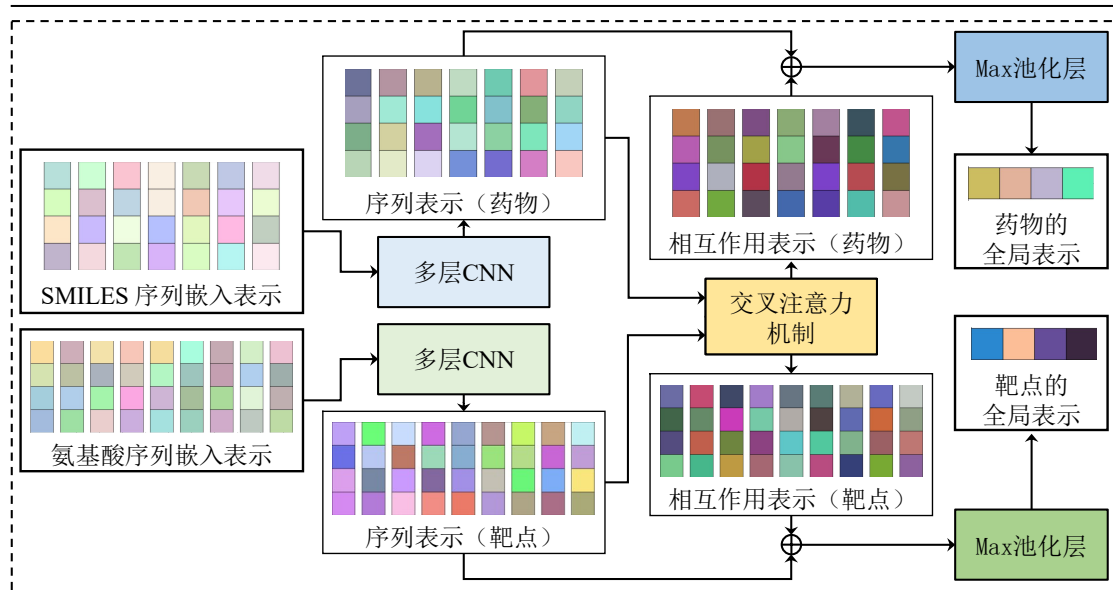


图 5.4 MIF-1D 编码块的结构示意图

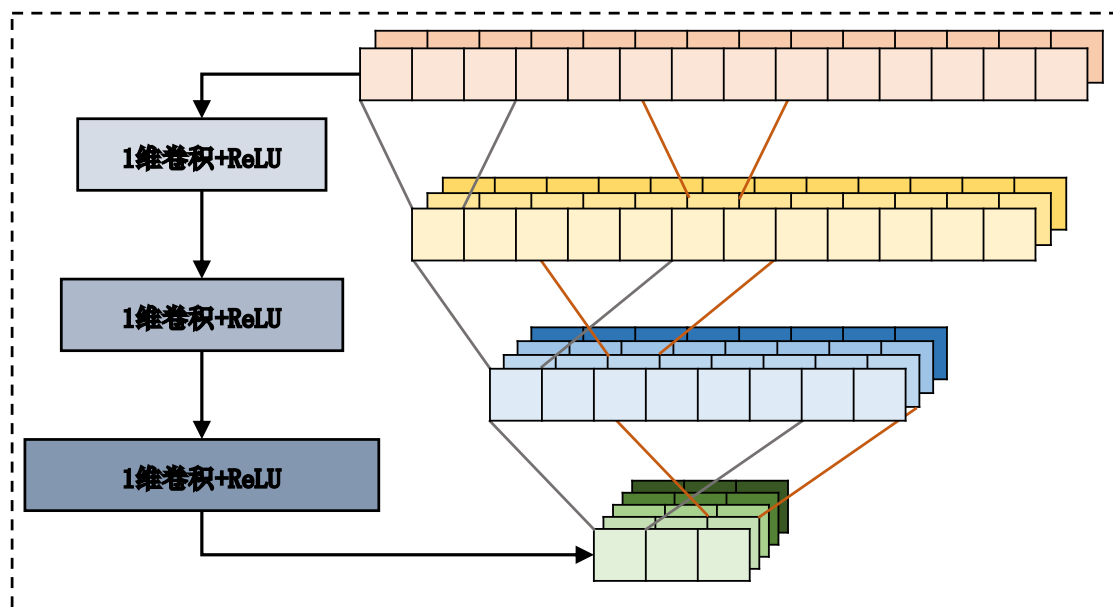


图 5.5 MIF-1D 编码块内多层卷积神经网络示意图

在多层 CNN 提取特征后，MIF-1D 编码块通过共享权重的交叉注意力机制进一步编码。具体而言，MIF-1D 编码块以药物的序列表示为查询值，以靶点的序列表示为键值对，通过多头注意力机制捕获对药物具有重要影响的靶点片段，从而生成药物的相互作用表示。

与此同时，MIF-1D 编码块以靶点的序列表示为查询值，以药物的序列表示为键值对，通过相同方式获取靶点的相互作用表示。在药物及靶点的查询过程中，MIF-1D 编码块通过单个多头注意力模块进行实现，即药物与靶点的查询、键和值的投影层共享权重。

记 SMILES 序列与氨基酸序列经多层 CNN 处理后的长度分别为 L_D 和 L_T ，编码维度为 d_C ，则其编码表示分别为 $\mathbf{h}_x \in \mathbb{R}^{L_D \times d_C}$ 与 $\mathbf{h}_y \in \mathbb{R}^{L_T \times d_C}$ 。注意力模块

通过线性层计算药物的查询向量 $q_x^i \in \mathbb{R}^{L_D \times d_H}$ 、靶点的键向量 $k_y^i \in \mathbb{R}^{L_T \times d_H}$ 与靶点的值向量 $v_y^i \in \mathbb{R}^{L_T \times d_H}$ ，如式 (5.3) 所示。

$$\begin{cases} q_x^i = h_x W_q^i \\ k_y^i = h_y W_k^i \\ v_y^i = h_y W_v^i \end{cases} \quad i = 1, 2, \dots, N_H \quad (5.3)$$

其中 N_H 为注意力头数， $W_q^i, W_k^i, W_v^i \in \mathbb{R}^{d_C \times d_H}$ 分别为第 i 个注意力头中查询、键和值的投影层权重， d_H 为第 i 个注意力头的输出维度并且有 $d_C = d_H \times N_H$ 。

与此同时，注意力模块利用共享权重计算靶点的查询向量 $q_y^i \in \mathbb{R}^{L_T \times d_H}$ 、药物的键向量 $k_x^i \in \mathbb{R}^{L_D \times d_H}$ 以及药物的值向量 $v_x^i \in \mathbb{R}^{L_D \times d_H}$ ，如式 (5.4) 所示。

$$\begin{cases} q_y^i = h_y W_q^i \\ k_x^i = h_x W_k^i \\ v_x^i = h_x W_v^i \end{cases} \quad i = 1, 2, \dots, N_H \quad (5.4)$$

在实现对药物与靶点的查询、键和值进行线性投影的基础上，注意力模块通过点积运算以及 softmax 函数计算药物在第 i 个注意力头的特征矩阵 f_x^i 和靶点在第 i 个注意力头的特征矩阵 f_y^i ，如式 (5.5) 所示。

$$\begin{cases} f_x^i = \text{softmax} \left(\frac{q_x^i k_y^{i\top}}{\sqrt{d_H}} \right) v_y^i = A_x^i v_y^i \\ f_y^i = \text{softmax} \left(\frac{q_y^i k_x^{i\top}}{\sqrt{d_H}} \right) v_x^i = A_y^i v_x^i \end{cases} \quad i = 1, 2, \dots, N_H \quad (5.5)$$

其中， $A_x^i \in \mathbb{R}^{L_D \times L_T}$ 为药物的标准化注意力得分矩阵， $A_y^i \in \mathbb{R}^{L_T \times L_D}$ 为靶点的标准化注意力得分矩阵，最终结果满足 $f_x^i \in \mathbb{R}^{L_D \times d_H}$ 以及 $f_y^i \in \mathbb{R}^{L_T \times d_H}$ 。

在此基础上，注意力模块对各注意力头的输出特征进行拼接并作线性转换，从而得到药物的相互作用表示 h'_x 与靶点的相互作用表示 h'_y ，如式 (5.6) 所示。

$$\begin{cases} h'_x = \text{concat}(f_x^1, f_x^2, \dots, f_x^{N_H}) W_c \\ h'_y = \text{concat}(f_y^1, f_y^2, \dots, f_y^{N_H}) W_c \end{cases} \quad (5.6)$$

其中， $W_c \in \mathbb{R}^{d_C \times d_C}$ 为可学习权重， $h'_x \in \mathbb{R}^{L_D \times d_C}$ ， $h'_y \in \mathbb{R}^{L_T \times d_C}$ 。

通过上述运算，MIF-1D 编码块通过多层 CNN 分别计算出药物与靶点的序列表示 h_x 与 h_y ，随后通过交叉注意力机制进一步计算出药物与靶点的相互作用表示 h'_x 与 h'_y ，最后将药物的两类表示与靶点的两类表示分别进行融合。记当前编码块为第 l 个编码块，则药物与靶点最终的编码表示 $h_x^{(l)} \in \mathbb{R}^{L_D \times d_C}$ 与 $h_y^{(l)} \in \mathbb{R}^{L_T \times d_C}$ 的计算过程如式 (5.7) 所示。

$$\begin{cases} h_x^{(l)} = 0.5 \cdot h_x + 0.5 \cdot h'_x \\ h_y^{(l)} = 0.5 \cdot h_y + 0.5 \cdot h'_y \end{cases} \quad (5.7)$$

最终, 编码块利用 MaxPool 方法进行池化, 即逐个维度计算序列中的最大值, 从而获取两者的全局表示 $\mathbf{g}_{x_1D}^{(l)}, \mathbf{g}_{y_1D}^{(l)} \in \mathbb{R}^{1 \times d_c}$, 计算过程如式 (5.8) 所示。

$$\begin{cases} \mathbf{g}_{x_1D}^{(l)} = \text{MaxPool}(\mathbf{h}_x^{(l)}) \\ \mathbf{g}_{y_1D}^{(l)} = \text{MaxPool}(\mathbf{h}_y^{(l)}) \end{cases} \quad (5.8)$$

综上所述, 给定药物与靶点的嵌入表示, MIF-1D 编码块通过多层一维 CNN 捕获两者的序列结构信息, 通过多头交叉注意力机制捕获两者的相互作用信息, 最终将两类信息进行结合, 从而得到药物与靶点的输出特征并将其输入到下一个 MIF-1D 编码块, 此外, MIF-1D 编码块会对输出特征进行池化, 计算药物与靶点的全局表示, 从而为最终的 MIF 解码模块提供输入。

5.2.3 图编码模块

针对药物的 SMILES 序列与靶点的氨基酸序列, 图编码模块依次对其进行二维图结构生成以及双视图表示编码, 其整体处理流程与本文第 4 章所提出的 GDN-DTI 模型相似。

首先, 图编码模块通过二维图结构生成模块根据 SMILES 序列构建层次化分子图, 根据氨基酸序列构建靶点接触图, 处理流程如 4.2.2 小节所述。随后, 图编码模块通过多个 GDN 编码器对药物的层次化分子图与靶点的接触图进行编码。如 4.2.4 小节所述, GDN 编码器通过两个相互独立的结构视图层分别学习药物的二维结构表示与靶点的二维结构表示, 通过一个相互作用视图层学习两者的二维相互作用表示, 随后通过两个相互独立的图更新层分别对药物的两类表示和靶点的两类表示进行融合并对二维图进行更新。最后, 针对药物的层次化分子图, GDN 编码块提取分子级节点的表示作为全局表示; 针对靶点的二维接触图, GDN 编码块通过 SAGPooling 池化方法提取靶点的全局表示。

通过多个 GDN 编码块对层次化分子图 G_x 和靶点接触图 G_y 进行迭代式编码, 即上一编码块的输出作为邻接的下一编码块的输入, 图编码模块能够获得不同深度的药物全局表示和靶点全局表示, 如式 (5.9) 所示。

$$\mathbf{g}_{x_2D}^{(l)}, \mathbf{g}_{y_2D}^{(l)} = \text{GDNBlock}^{(l)}(G_x, G_y) \quad (5.9)$$

其中 $\text{GDNBlock}^{(l)}$ 表示第 l 个 GDN 编码块, G_x 与 G_y 为初始的二维图或者上一个 GDN 编码块的输出, 两者均包含节点的嵌入表示以及边的嵌入表示。记 GDN 编码块的输出维度为 d_{2D} , 则有 $\mathbf{g}_{x_2D}^{(l)}, \mathbf{g}_{y_2D}^{(l)} \in \mathbb{R}^{1 \times d_{2D}}$ 。

5.2.4 MIF 解码模块

通过多个编码块的迭代编码, 序列编码模块与图编码模块分别提取多个深度的序列全局表示及图全局表示。针对这些不同深度与不同模态的全局表示, MIF

解码模块利用协注意力机制^[52,55]计算多模态融合系数,并通过矩阵乘法计算相互作用得分矩阵,最后采用全连接层对得分矩阵进行聚合并实现 DTI 预测。

首先, MIF 解码模块对药物与靶点的全局表示分别进行堆叠,从而获取药物的全局表示矩阵 $\mathbf{g}_x$ 和靶点的全局表示矩阵 $\mathbf{g}_y$ 。记 MIF-1D 编码块的数量为 N_1 , GDN 编码块的数量为 N_2 , 全局表示维度为 d , 则合并过程如式 (5.10) 所示。

$$\begin{cases} \mathbf{g}_x = \text{Stack}(\mathbf{g}_{x_1D}^{(1)}, \dots, \mathbf{g}_{x_1D}^{(N_1)}, \mathbf{g}_{x_2D}^{(1)}, \dots, \mathbf{g}_{x_2D}^{(N_2)}), \\ \mathbf{g}_y = \text{Stack}(\mathbf{g}_{y_1D}^{(1)}, \dots, \mathbf{g}_{y_1D}^{(N_1)}, \mathbf{g}_{y_2D}^{(1)}, \dots, \mathbf{g}_{y_2D}^{(N_2)}) \end{cases} \quad (5.10)$$

其中 Stack 表示堆叠操作, $\mathbf{g}_x, \mathbf{g}_y \in \mathbb{R}^{(N_1+N_2) \times d}$ 。

随后, MIF 解码模块利用协注意力机制逐层计算多模态融合系数,从而得到融合系数矩阵 $\mathbf{M}$, 计算过程如式 (5.11) 所示。在此基础上, 模块对融合系数矩阵 $\mathbf{M}$ 、药物的全局表示矩阵 $\mathbf{g}_x$ 以及靶点的全局表示矩阵 $\mathbf{g}_y$ 执行矩阵乘法运算, 从而得到相互作用得分矩阵 $\mathbf{S}$, 计算过程如式 (5.12) 所示。

$$M_{ij} = \alpha^\top \tanh(\mathbf{W}_x \mathbf{g}_x^{(i)} + \mathbf{W}_y \mathbf{g}_y^{(j)}) \quad (5.11)$$

$$\mathbf{S} = \mathbf{M} \circ (\mathbf{g}_x \mathbf{g}_y^\top) \quad (5.12)$$

其中 $\mathbf{g}_x^{(i)}, \mathbf{g}_y^{(j)} \in \mathbb{R}^d$ 分别为表示矩阵 $\mathbf{g}_x$ 的第 i 行与表示矩阵 $\mathbf{g}_y$ 的第 j 行, $M_{ij} \in \mathbb{R}$ 表示系数矩阵 $\mathbf{M}$ 第 i 行第 j 列的元素值并且 $1 \leq i, j \leq (N_1 + N_2)$ 。 $\mathbf{W}_x, \mathbf{W}_y \in \mathbb{R}^{d \times d}$ 为可学习的权重矩阵, $\alpha \in \mathbb{R}^d$ 为可学习的权重向量, $\top$ 为转置操作, $\alpha^\top \in \mathbb{R}^{1 \times d}$ 。此外, $\circ$ 表示逐元素乘法, $\mathbf{M} \in \mathbb{R}^{(N_1+N_2) \times (N_1+N_2)}$ 。

最终, MIF 解码模块通过线性层对得分矩阵 $\mathbf{S}$ 进行汇总求和, 并通过 sigmoid 激活函数计算 DTI 的存在概率 p , 计算过程如式 (5.13) 所示。

$$p = \text{sigmoid}(\mathbf{W} \cdot \text{Flatten}(\mathbf{S}) + b) \quad (5.13)$$

其中 Flatten($\cdot$) 为矩阵扁平化操作, $\mathbf{W} \in \mathbb{R}^{1 \times [(N_1+N_2) \times (N_1+N_2)]}$ 为可学习的权重向量, $b \in \mathbb{R}$ 为可学习的偏置项。

5.2.5 损失函数

本章的任务目标与第 4 章相同, 均为预测药物与靶点之间是否存在相互作用, 因此本章同样采用二元交叉熵损失函数进行模型训练, 具体计算方式如式 (5.14) 所示。

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \quad (5.14)$$

其中 N 为训练集样本的总数量, $p_i \in [0, 1]$ 为模型输出的第 i 个样本为真的概率值, $y_i \in 0, 1$ 为第 i 个样本的真实标签。

5.3 实验及结果分析

5.3.1 实验设置

本节将依次介绍实验的设置情况，即数据集、实验环境与评估指标、模型架构与超参数以及基线模型。

5.3.1.1 数据集

在数据集方面，本章采用3个免费开放的公共数据集：DrugBank^[6]、BioSNAP^[94]和Davis^[124]。DrugBank和BioSNAP为综合数据集，本文4.3.1小节详细介绍了两者的内容及数据量。Davis数据集主要记录了与激酶相关的药物-靶点相互作用，就数据集规模而言，Davis包含68类药物与379类靶点，共计7320个正样本（即存在相互作用的药物-靶点对）和18452个负样本（即不存在相互作用的药物-靶点对）。

此外，在模型训练过程中，本章从全部样本中随机划分出训练集与测试集，两者的比例为8:2，该比例与现有实验^[63]保持一致。同时，通过5折交叉验证的方式进行模型选择与评估，即将训练集划分为5个子集，依次利用其中1个子集选择最优模型，其余4个子集则用于模型训练，选择出的最优模型在测试集上进行评估，最终以5次评估的平均值作为模型表现。

5.3.1.2 实验环境与评估指标

在实验环境与评估指标方面，本章与第4章保持一致，即采用Pytorch构建MIF-DTI，并采用准确率、AUROC与AUPR等五项指标对各模型进行对比，其中准确率与AUROC着重反映模型预测的准确程度，精确率、召回率和AUPR则衡量模型对真实DTI的预测能力。

5.3.1.3 模型架构与超参数

在模型架构与超参数方面，MIF-DTI由序列编码模块、图编码模块和MIF解码模块组成，其超参数通过随机搜索以及参考GDN-DTI与MCANet等相关工作而确定，具体如下：

序列编码模块包含3个MIF-1D编码块，每个MIF-1D编码块包含两个CNN子模块和一个注意力子模块。两个CNN子模块均包含3层CNN，针对药物的卷积核大小依次为4、6和8，针对靶点的卷积核大小依次为4、8和12。注意力子模块包含5个注意力头。此外，所有子模块均产生200维表示。

图编码模块包含3个GDN编码块，每个GDN编码块包含两个结构视图层、一个相互作用视图层和两个图更新层，所有层采用单层2头GAT。在每个块中，

结构视图层和相互作用视图层均产生 100 维表示,而更新层产生 200 维表示。此外,模型训练期间使用的批次大小为 64,学习率由 CyclicLR^[122]动态调整,基础学习率与权重衰减率均为 1.0×10^{-4} ,最大调整倍率为 10 倍。

5.3.1.4 基线模型

在基线模型方面,本章选取基于序列的代表性方法 MCANet^[63]、基于图的代表性方法 PSICHIC^[79]、基于序列与图的代表性方法 BINDTI^[60]以及本文第 4 章给出的 GDN-DTI 作为基线模型,各模型的名称及主要贡献如 4.3.1 小节所述。

5.3.2 实验结果

本章在热启动场景中对 MIF-DTI 与各基线模型进行对比,即随机划分出训练集、验证集与测试集。表 5.1、表 5.2 和表 5.3 分别显示了 MIF-DTI 和各基线模型在 DrugBank、BioSNAP 和 Davis 数据集上的表现。实验结果表明,MIF-DTI 在三个数据集上的多项评估指标均优于此前最优的序列编码与图编码模型。

表 5.1 MIF-DTI 等模型在 DrugBank 数据集上的表现 (%)

模型	表示形式	准确率	精确率	召回率	AUROC	AUPR
MCANet	序列	82.60	82.42	82.74	89.71	90.45
BINDTI	序列与图	80.43	79.10	81.33	87.42	86.95
PSICHIC	图	83.59	83.05	83.33	90.47	89.94
GDN-DTI	图	83.35	82.28	84.63	90.65	90.52
MIF-DTI	序列与图	85.46	85.14	85.60	92.77	92.95

注:每项指标的最高值用加粗字体显示。

结合表 5.1 可以看出,在 DrugBank 数据集上,MIF-DTI 的各项指标均优于此前的最优模型。与序列编码模型 MCANet 相比,MIF-DTI 的准确率、AUROC 和 AUPR 分别提升了 2.86%、3.06% 以及 2.50%。与图编码模型 PSICHIC 相比,MIF-DTI 的对应指标分别提升了 1.87%、2.30% 以及 3.01%。此外,与本文第 4 章提出的图编码模型 GDN-DTI 相比,MIF-DTI 的三项对应指标具有约 2% 的提升。

结合表 5.2 可以看出,在 BioSNAP 数据集上,MIF-DTI 的各项指标均优于基线模型。相比于 MCANet,MIF-DTI 的准确率、AUROC 和 AUPR 分别提升了 2.06%、1.91% 以及 1.77%。相比于 PSICHIC,MIF-DTI 的对应指标分别提升了 2.50%、2.44% 以及 3.41%。与本文第 4 章提出的 GDN-DTI 相比,MIF-DTI 的对应指标分别提升了 1.86%、1.85% 以及 2.14%。

结合表 5.3 可以看出,在 Davis 数据集上,与序列编码模型 MCANet 相比,MIF-DTI 的准确率提升了 0.37%,AUROC 提升了 0.25%,AUPR 提升了 0.40%,

表 5.2 MIF-DTI 等模型在 BioSNAP 数据集上的表现 (%)

模型	表示形式	准确率	精确率	召回率	AUROC	AUPR
MCANet	序列	84.58	83.57	85.76	91.74	92.24
BINDTI	序列与图	81.39	80.14	83.06	88.25	87.84
PSICHIC	图	84.14	83.72	84.45	91.21	90.60
GDN-DTI	图	84.78	83.72	85.95	91.80	91.87
MIF-DTI	序列与图	86.64	86.10	87.10	93.65	94.01

注：每项指标的最高值用加粗字体显示。

精确率提升了 0.93%，召回率与 MCANet 基本持平。与图编码模型 PSICHIC 与 GDN-DTI 相比，MIF-DTI 的多项指标具有显著提升，其中准确率、AUROC 和 AUPR 较 PSICHIC 分别提升了 1.22%、1.59% 以及 1.85%，相比于 GDN-DTI 分别提升了 1.33%、1.47% 以及 2.65%。

表 5.3 MIF-DTI 等模型在 Davis 数据集上的表现 (%)

模型	表示形式	准确率	精确率	召回率	AUROC	AUPR
MCANet	序列	87.05	77.54	76.67	92.56	85.11
BINDTI	序列与图	75.03	54.38	87.45	87.51	77.71
PSICHIC	图	86.20	78.91	71.70	91.22	83.66
GDN-DTI	图	86.09	75.73	75.05	91.34	82.86
MIF-DTI	序列与图	87.42	78.47	76.66	92.81	85.51

注：每项指标的最高值用加粗字体显示。

综上所述，MIF-DTI 在三个数据集上的表现均优于现有的各类方法，这表明 MIF-DTI 能够有效融合不同模态的信息，从而实现更准确的 DTI 预测。

5.3.3 消融实验

为了评估 MIF-DTI 中各模块对模型预测表现的影响，本章在 DrugBank 数据集上进行消融实验。消融实验所使用的 MIF-DTI 变体如下：

(1) wo-1D-encoder：移除序列编码模块，仅保留图编码模块以及 MIF 解码模块，从而表明序列编码在多模态 DTI 预测中的有效性；

(2) wo-2D-encoder：移除图编码模块，仅保留序列编码模块以及 MIF 解码模块，从而表明图编码在多模态 DTI 预测中的有效性；

(3) wo-2D-with-attn：移除图编码模块，仅保留序列编码模块以及 MIF 解码模块，并采用多头自注意力机制替代序列编码模块中的多层 CNN，从而表明多层 CNN 在序列编码方面的有效性；

(4) with-attention: 移除 MIF 解码模块中的协注意力机制以及相互作用得分矩阵, 通过交叉注意力机制进行多模态融合, 并采用最大值池化方法获取药物表示与靶点表示, 最终通过全连接层实现 DTI 预测, 从而验证协注意力机制以及相互作用得分矩阵在多模态信息融合过程中的有效性。

如表5.4所示, MIF-DTI 的表现优于其所有变体, 这一结果表明 MIF-DTI 内的序列编码模块、图编码模块及 MIF 解码模块均能有效改善模型的预测性能。针对各变体表现的具体分析如下:

表 5.4 MIF-DTI 及其变体在 DrugBank 数据集上的表现 (%)

模型	准确率	精确率	召回率	AUROC	AUPR
wo-1D-encoder	83.26	82.22	84.52	90.57	90.36
wo-2D-encoder	79.40	80.27	77.69	87.25	87.12
wo-2D-with-attn	75.62	75.07	76.21	83.83	84.41
with-attention	83.70	83.40	83.87	91.08	91.20
MIF-DTI	85.46	85.14	85.60	92.77	92.95

注: 每项指标的最高值用加粗字体显示。

(1) 序列编码模块与图编码模块在 MIF-DTI 中发挥了重要作用。结合表5.4可以看出, 相比于完整模型 MIF-DTI, 移除了序列编码模块的 wo-1D-encoder 的准确率、AUROC 和 AUPR 分别下降了 2.20%、2.20% 以及 2.59%, 移除了图编码模块的 wo-2D-encoder 则出现了更显著的下降, 其准确率、AUROC 和 AUPR 分别下降了 6.06%、5.52% 以及 5.83%。这一结果表明, 序列编码和图编码均对 DTI 预测均具有重要影响, 其中图编码对完整模型的贡献更大。此外, 结合图5.6可以观察到, 相比于单模态预测方法 wo-1D-encoder 和 wo-2D-encoder, 多模态预测方法 MIF-DTI 所需训练轮数更少且预测性能更优, 因此可以确定多模态预测方法更适用于 DTI 预测任务。

(2) 多层卷积神经网络更适用于药物与靶点的序列编码学习。相较于使用注意力机制对序列进行编码的 wo-2D-with-attn, 使用多层 CNN 的 wo-2D-encoder 具有更好的表现, 其准确率提升了 3.78%, AUROC 提升了 3.42%, AUPR 提升了 2.71%, 这一结果表明多层 CNN 能够更好地从序列中提取有效信息。从理论层面分析, 注意力机制能够捕获长序列中的远程依赖信息, 而多层 CNN 则在捕获局部信息时更有优势。在 DTI 预测任务中, 药物与靶点虽然具有较长的序列, 但两者的空间结构往往与序列中原子或氨基酸的局部排列紧密相关, 即关键信息通常蕴含于相邻的局部区域, 因此在该场景中多层 CNN 能够产生更好的表现。

(3) 协注意力机制及相互作用得分矩阵在多模态融合过程中具有重要作用。结合表5.4可以看出, 相较于使用协注意力机制及相互作用得分矩阵进行信息融

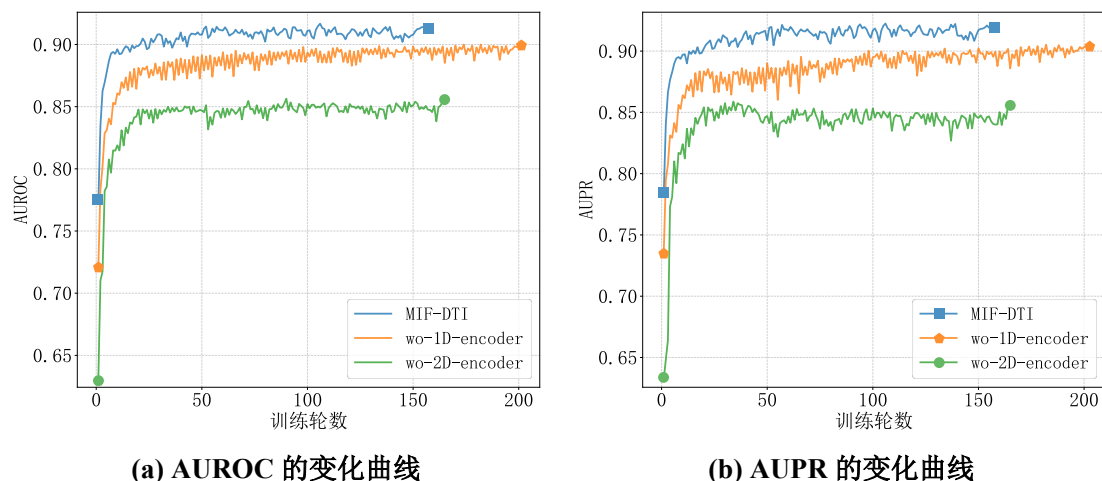


图 5.6 MIF-DTI 及其变体在训练过程中 AUROC 与 AUPR 的变化曲线

合的 MIF-DTI，使用传统注意力机制及池化方法的变体模型 with-attention 的准确率下降了 1.76%，AUROC 降低了 1.69%，AUPR 降低了 1.75%。这一结果表明，协注意力机制能够有效融合不同模态与不同深度的药物表示与靶点表示。

5.4 本章小结

针对现有 DTI 预测方法存在的编码信息源单一、多模态信息融合不充分的问题，本章给出了一种基于双视图表示学习与多模态信息融合的 DTI 预测方法 MIF-DTI。该方法通过序列编码模块对药物的 SMILES 序列与靶点的氨基酸序列进行双视图学习，通过图编码模块对药物的层次化分子图与靶点的接触图进行双视图学习，从而捕获药物与靶点的序列信息、拓扑结构信息与相互作用信息。最终，MIF-DTI 采用协注意力机制计算多模态融合系数与相互作用得分矩阵，实现了不同模态全局表示的融合，同时提升了 DTI 预测的准确性。实验结果表明，MIF-DTI 三项数据集上均取得了更好的表现。消融实验进一步验证了 MIF-DTI 中序列编码模块、图编码模块及 MIF 解码模块的有效性。综上所述，通过充分挖掘并融合多模态信息，MIF-DTI 能够更准确地识别潜在的 DTI，从而为药物作用机理研究等任务提供了更可靠的 DTI 预测方法。

第6章 总结与展望

6.1 工作总结

药物相互作用预测是生物信息学和药物学领域中的重要研究方向，其旨在预测药物与药物、药物与蛋白质以及药物与基因之间的相互作用。该研究对新药研发、药物作用机理探究以及已有药物再利用等下游任务具有重要意义，同时能够为安全有效的联合用药治疗提供理论支持。

作为药物相互作用预测任务的代表性子课题，药物-药物相互作用预测旨在预测两项或多项药物共同给药时所产生的协同或拮抗等相互作用，而药物-靶点相互作用预测旨在预测药物与靶点（如蛋白质与基因）的结合情况。两者尽管在预测目标与编码难度上存在差异，但均涉及分子表示学习与相互作用信息挖掘，因此可采用统一的框架进行研究。

随着大规模数据库的陆续建立以及深度学习方法的不断改进，深度学习技术已成为药物相互作用预测的重要手段。然而，现有基于深度学习的预测方法仍存在诸多局限，例如药物子结构可解释性不足、靶点信息模态单一以及相互作用信息挖掘不充分等问题。针对这些挑战，本文提出了一种药物-药物相互作用预测方法，给出了两种药物-靶点相互作用预测方法，主要工作包括：

（1）基于图表示学习的药物-药物相互作用预测方法。药物由各种官能团或化学子结构组成，它们决定了药物的药效学和药代动力学特性，并最终影响药物之间的相互作用，因此基于子结构的 DDI 预测逐渐成为主流方法。然而，现有方法存在感知子结构可解释性不佳、化学子结构相互孤立的问题。针对这些不足，本文提出了一种基于层次化分子图与双视图表示学习的 DDI 预测方法，即 HDN-DDI。该方法首先使用化学领域的可解释性算法将药物分解为多个实际存在的化学子结构，在此基础上构建具有原子级节点、子结构级节点以及分子级节点的层次化分子图。随后，HDN-DDI 通过双视图表示学习提取分子内部的层次化信息以及分子之间的相互作用信息，最终利用协注意力机制对不同深度的分子表示进行聚合并实现 DDI 预测。多项实验结果表明，HDN-DDI 在热启动、冷启动以及实际应用场景下均表现出优异的性能。在热启动场景中，HDN-DDI 在 DrugBank 数据集上的准确率达到 97.93%，AUPR 达到 99.72%，在 Twosides 数据集上的准确率达到 99.38%，AUPR 达到 99.97%。在冷启动场景中，HDN-DDI 在 S1 与 S2 划分上的准确率分别为 79.84% 和 89.43%，相较于此前的领域最佳模型提升了 4.19% 和 4.08%，在 AUROC 和 AUPR 等评估指标上也表现出较大优势。此外，HDN-DDI 针对 FDA 新认证药物的识别准确率较此前最优方法提升了 5.79%，进一步验证了其在预测与新药物相关的 DDI 时的有效性。

(2) 基于二维图结构的药物-靶点相互作用预测方法。当前主流的 DTI 预测方法通常仅对药物与靶点的一维序列进行编码,然而序列包含的信息有限,深度学习模型无法有效捕获分子的拓扑结构信息。近年来,部分工作尝试针对二维图结构进行编码,但该类方法通常仅关注药物与靶点的自身结构,对两者之间的相互作用信息挖掘仍不充分。为解决这些问题,本文给出了一种基于二维图结构与双视图表示学习的 DTI 预测模型 GDN-DTI。该方法以药物的 SMILES 序列与靶点的氨基酸序列为输入,首先通过二维图结构生成模块构建药物的层次化分子图以及靶点的接触图,随后通过双视图表示学习提取分子内结构信息以及分子间相互作用信息,最后通过药物-靶点相互作用得分矩阵实现多尺度表示融合并通过全连接层实现 DTI 预测。实验结果表明,GDN-DTI 在 DrugBank 与 BioSNAP 数据集上的 AUROC 与 AUPR 值均达到了 90%,并且在多个评估指标上优于基线模型。消融实验进一步验证了 GDN-DTI 中二维图结构、双视图表示学习以及相互作用得分矩阵的有效性。此外,本文也对靶点的三维结构与一维序列的关系进行了深入剖析,同时指出了 GDN-DTI 的局限性,并为今后基于二维图的 DTI 预测研究提供了改进方向。

(3) 基于多模态信息融合的药物-靶点相互作用预测方法。当前大多数 DTI 预测方法针对一维序列进行编码,部分方法从药物与靶点的二维图或三维空间结构中提取信息,然而这些方法往往仅关注单一模态信息,缺乏对多模态信息的有效融合。针对该问题,本文给出了一种基于双视图表示学习与多模态信息融合的 DTI 预测模型 MIF-DTI。该方法通过序列编码模块对药物的 SMILES 序列与靶点的氨基酸序列进行编码,从而学习序列信息与一维相互作用信息;通过图编码模块获取药物的层次化分子图与靶点的接触图并对其进行编码,从而捕获拓扑结构信息与二维相互作用信息。最终,MIF-DTI 通过解码模块实现不同模态与不同深度的信息融合,从而改善 DTI 预测表现。实验结果表明,MIF-DTI 在多个数据集上均优于此前最佳方法,其在 DrugBank 与 BioSNAP 数据集上的准确率分别提升了 1.87% 与 1.86%,AUROC 分别提升了 2.12% 与 1.85%,AUPR 分别提升了 2.43% 与 1.77%。此外,消融实验进一步验证了 MIF-DTI 中序列编码模块、图编码模块以及 MIF 解码模块的重要性,同时也表明多模态信息融合有助于提升 DTI 预测模型的准确性。

综上所述,本文针对药物相互作用预测领域中存在的问题与不足,提出了一种基于层次化分子图与双视图表示学习的药物-药物相互作用预测方法,即 HDN-DDI。在此基础上,进一步给出了两种药物-靶点相互作用预测方法,即 GDN-DTI 与 MIF-DTI。实验结果表明,HDN-DDI、GDN-DTI 以及 MIF-DTI 在多个数据集上具有更优的预测性能,并且具备更好的泛化能力,从而为新药研发、药物作用机理探究以及药物再利用等下游任务提供了有效的计算工具。

6.2 未来展望

本文的改进方向与未来展望如下：

(1) 进一步优化 HDN-DDI 模型中的子结构分解算法。HDN-DDI 模型通过 BRICS 算法及附加规则将药物分解为若干子结构,然而当前分解算法的适用范围仍然有限,部分药物分子无法得到充分分解。与此同时,现阶段已经存在 PubChem 与 ZINC 等包含数亿种化合物的大规模分子库,并且已有 MolTrans 和 FMCANet 等工作尝试以数据驱动的方式进行药物分解,但这些方法的分解结果可解释性仍然较低。因此,今后工作可以尝试结合化学领域专业知识以及大规模数据,构建出适用范围更广且可解释性更强的化学子结构分解算法,从而进一步提升 DDI 预测模型的准确性与泛化性。

(2) 构建适用于靶点二维图的图编码器。在 GDN-DTI 模型中,由于靶点的预测接触图节点数量多、邻接关系稀疏且总体样本有限,传统的图神经网络与图注意力网络等图编码器难以有效训练并提取信息,这在一定程度上影响了模型的预测表现。近年来,在分子表示学习领域,以 MolCLR 和 HiMol 为代表的预训练方法在分子属性预测等任务中表现突出,其核心在于结合大规模分子库与图神经网络构建预训练分子编码器。在蛋白质表示学习领域,针对一维序列的蛋白质语言模型如 ESM 和 ProGen 等取得了良好表现,但现阶段仍缺乏针对二维图的预训练编码器。因此,今后的研究可以考虑设计适用于靶点接触图的图神经网络,或者利用大规模数据库构建预训练的靶点编码器,从而改善蛋白质表示学习及 DTI 预测等下游任务的表现。

(3) 实现更多模态的信息融合。本文所提出的 MIF-DTI 融合了药物与靶点的一维序列以及二维图结构,实验结果表明,双模态融合的 DTI 预测方法相较单模态 DTI 预测方法有显著的性能提升。近年来,随着生物学数据不断积累以及 AlphaFold 等深度学习模型的逐步应用,大量药物及靶点的三维结构被解析并收录至相关数据库。与此同时,药物与靶点的功能数据库和同源知识数据库也在不断完善。因此,未来的研究可以探索融合更多模态的信息,例如引入三维空间构象与结合位点等结构信息以及作用途径等同源知识,以实现更准确的药物相互作用预测。

参考文献

- [1] KATZUNG B G, MASTERS S B, TREVOR A J. Basic and clinical pharmacology[M]. 14th ed. New York: McGraw-Hill Education, 2018.
- [2] 孙立, 刘荣, 李勇, 等. 阿司匹林和低分子肝素预防手足创伤骨折患者术后静脉血栓栓塞的临床研究[J]. 中国临床药理学杂志, 2024, 40(12): 1718-1722.
- [3] SLAVIN S. Recreational use of amyl nitrite[J]. Venereology, 2001, 14(2): 81-82.
- [4] IQBAL A B, SHAH I A, INJILA, et al. A review of deep learning algorithms for modeling drug interactions[J]. Multimedia Systems, 2024, 30(3): 124.
- [5] CHEN X, YAN C C, ZHANG X, et al. Drug–target interaction prediction: databases, web servers and computational models[J]. Briefings in Bioinformatics, 2016, 17(4): 696-712.
- [6] WISHART D S, FEUNANG Y D, GUO A C, et al. Drugbank 5.0: a major update to the drugbank database for 2018[J]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082.
- [7] JOURDAN J P, BUREAU R, ROCHAIS C, et al. Drug repositioning: a brief overview[J]. Journal of Pharmacy and Pharmacology, 2020, 72(9): 1145-1151.
- [8] MEI S, ZHANG K. A machine learning framework for predicting drug–drug interactions[J]. Scientific Reports, 2021, 11(1): 17619.
- [9] 刘畅, 袁恒杰. 基于 OpenFDA 平台挖掘托法替布药物不良反应的临床分析[J]. 中国临床药理学杂志, 2023, 39(14): 2099-2102.
- [10] LEE J S, KIM E S, CHO K B, et al. Pain intensity at injection site during esophagogastroduodenoscopy using long-and medium-chain versus long-chain triglyceride propofol: a randomized controlled double-blind study[J]. Gut and Liver, 2020, 15(4): 562-568.
- [11] PANG K, WAN Y W, CHOI W T, et al. Combinatorial therapy discovery using mixed integer linear programming[J]. Bioinformatics, 2014, 30(10): 1456-1463.
- [12] FAN Z, ZHENG H. Pretraining molecular and substructural encoders for predicting drug-drug interactions in cold-start scenarios[C]//Proceedings of the 2nd International Conference on Biomedical and Intelligent Systems (IC-BIS): Vol. 12724. 2023: 202-210.
- [13] LAJOIE A C, LAUZIÈRE G, LEGA J C, et al. Combination therapy versus monotherapy for pulmonary arterial hypertension: a meta-analysis[J]. The Lancet Respiratory Medicine, 2016, 4(4): 291-305.
- [14] MOKHTARI R B, HOMAYOUNI T S, BALUCH N, et al. Combination therapy in combating cancer[J]. Oncotarget, 2017, 8(23): 38022-38043.
- [15] KHAWCHAROENPORN T, CHUNCHARUNEE A, MALUANGNON C, et al. Active monotherapy and combination therapy for extensively drug-resistant pseudomonas aeruginosa

- nosa pneumonia[J]. International Journal of Antimicrobial Agents, 2018, 52(6): 828-834.
- [16] POLEKSIC A, XIE L. Database of adverse events associated with drugs and drug combinations[J]. Scientific Reports, 2019, 9: 20025.
- [17] MALKI M A, PEARSON E R. Drug–drug–gene interactions and adverse drug reactions[J]. The Pharmacogenomics Journal, 2020, 20(3): 355-366.
- [18] WU L, WEN Y, LENG D, et al. Machine learning methods, databases and tools for drug combination prediction[J]. Briefings in Bioinformatics, 2022, 23(1): bbab355.
- [19] LIN X, DAI L, ZHOU Y, et al. Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction[J]. Briefings in Bioinformatics, 2023, 24(4): bbad235.
- [20] VILAR S, HARPAZ R, URIARTE E, et al. Drug–drug interaction through molecular structure similarity analysis[J]. Journal of the American Medical Informatics Association, 2012, 19(6): 1066-1074.
- [21] PORAT D, DAHAN A. Pharmacokinetics after bariatric surgery: adverse effects and drug safety issues in bariatric patients[J]. Expert Review of Clinical Pharmacology, 2025, 18(3): 101-108.
- [22] THAFAR M A, OLAYAN R S, ASHOOR H, et al. Dtigems+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques[J]. Journal of Cheminformatics, 2020, 12(44): 1-17.
- [23] BAI P, MILJKOVIĆ F, JOHN B, et al. Interpretable bilinear attention network with domain adaptation improves drug–target prediction[J]. Nature Machine Intelligence, 2023, 5(2): 126-136.
- [24] MAHMUD S H, CHEN W, LIU Y, et al. Predtis: prediction of drug–target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques[J]. Briefings in Bioinformatics, 2021, 22(5): bbab046.
- [25] RONG Y, BIAN Y, XU T, et al. Self-supervised graph transformer on large-scale molecular data[C]//Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS): Vol. 33. 2020: 12559-12571.
- [26] YU L, SU Y, LIU Y, et al. Review of unsupervised pretraining strategies for molecular representation[J]. Briefings in Functional Genomics, 2021, 20(5): 323-332.
- [27] WANG M, WANG J, WENG G, et al. Remode: a deep learning-based web server for target-specific drug design[J]. Journal of Cheminformatics, 2022, 14(84): 1-11.
- [28] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with alphafold[J]. Nature, 2021, 596(7873): 583-589.
- [29] CHANG Y, PARK H, YANG H J, et al. Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature[J]. Scientific

- Reports, 2018, 8(8857): 1-11.
- [30] ZHU J, WANG J, WANG X, et al. Prediction of drug efficacy from transcriptional profiles with deep learning[J]. Nature Biotechnology, 2021, 39(11): 1444-1452.
- [31] CHENG Z, ZHAO Q, LI Y, et al. lifdti: predicting drug–target interactions through interactive and independent features based on attention mechanism[J]. Bioinformatics, 2022, 38(17): 4153-4161.
- [32] LIN J, WU L, ZHU J, et al. R2-ddi: relation-aware feature refinement for drug–drug interaction prediction[J]. Briefings in Bioinformatics, 2023, 24(1): bbac576.
- [33] KASTRIN A, FERK P, LESKOŠEK B. Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning[J]. PLOS ONE, 2018, 13(5): e0196865.
- [34] SHI J Y, MAO K T, YU H, et al. Detecting drug communities and predicting comprehensive drug–drug interactions via balance regularized semi-nonnegative matrix factorization[J]. Journal of Cheminformatics, 2019, 11(28): 1-16.
- [35] DENG Y, XU X, QIU Y, et al. A multimodal deep learning framework for predicting drug–drug interaction events[J]. Bioinformatics, 2020, 36(15): 4316-4322.
- [36] FENG Y H, ZHANG S W. Prediction of drug-drug interaction using an attention-based graph neural network on drug molecular graphs[J]. Molecules, 2022, 27(9): 3004.
- [37] YIN Q, FAN R, CAO X, et al. Deepdrug: a general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction[J]. Quantitative Biology, 2023, 11(3): 260-274.
- [38] GUO Z, GUO K, NAN B, et al. Graph-based molecular representation learning[C]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI). 2023: 6638-6646.
- [39] KARIM M R, COCHEZ M, JARES J B, et al. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network[C]//Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB). 2019: 113-123.
- [40] YU Y, HUANG K, ZHANG C, et al. Sumgnn: multi-typed drug interaction prediction via efficient knowledge graph summarization[J]. Bioinformatics, 2021, 37(18): 2988-2995.
- [41] SU X, HU L, YOU Z, et al. Attention-based knowledge graph representation learning for predicting drug-drug interactions[J]. Briefings in Bioinformatics, 2022, 23(3): bbac140.
- [42] LEE J, YOON W, KIM S, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
- [43] HE H, CHEN G, YU-CHIAN CHEN C. 3dgt-ddi: 3d graph and text based neural network

- for drug–drug interaction prediction[J]. Briefings in Bioinformatics, 2022, 23(3): bbac134.
- [44] RYU J Y, KIM H U, LEE S Y. Deep learning improves prediction of drug–drug and drug–food interactions[J]. Proceedings of the National Academy of Sciences, 2018, 115(18): E4304-E4311.
- [45] PREUER K, LEWIS R P, HOCHREITER S, et al. Deepsynergy: predicting anti-cancer drug synergy with deep learning[J]. Bioinformatics, 2018, 34(9): 1538-1546.
- [46] YANG Z, TONG K, JIN S, et al. Cnn-siam: multimodal siamese cnn-based deep learning approach for drug–drug interaction prediction[J]. BMC Bioinformatics, 2023, 24(1): 110.
- [47] YU H, ZHAO S, SHI J. Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions[J]. Briefings in Bioinformatics, 2022, 23(4): bbac209.
- [48] TANG Z, CHEN G, YANG H, et al. Dsil-ddi: A domain-invariant substructure interaction learning for generalizable drug–drug interaction prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(8): 10552-10560.
- [49] XU N, WANG P, CHEN L, et al. Mr-gnn: multi-resolution and dual graph neural network for predicting structured entity interactions[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). 2019: 3968-3974.
- [50] DEAC A, HUANG Y H, VELIĆKOVIĆ P, et al. Drug-drug adverse effect prediction with graph co-attention[A]. 2019. arXiv: 1905.00534.
- [51] LI G, XIONG C, THABET A, et al. Deepergcn: All you need to train deeper gcns[A]. 2020. arXiv: 2006.07739.
- [52] NYAMABO A K, YU H, SHI J Y. Ssi-ddi: substructure–substructure interactions for drug–drug interaction prediction[J]. Briefings in Bioinformatics, 2021, 22(6): bbab133.
- [53] HARROLD M, ZAVOD R. Basic concepts in medicinal chemistry[M]. American Society of Health-System Pharmacists (ASHP), 2013.
- [54] NYAMABO A K, YU H, LIU Z, et al. Drug–drug interaction prediction with learnable size-adaptive molecular substructures[J]. Briefings in Bioinformatics, 2022, 23(1): bbab441.
- [55] LI Z, ZHU S, SHAO B, et al. Dsn-ddi: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning[J]. Briefings in Bioinformatics, 2023, 24(1): bbac597.
- [56] NING G, SUN Y, LING J, et al. Bdn-ddi: A bilinear dual-view representation learning framework for drug–drug interaction prediction[J]. Computers in Biology and Medicine, 2023, 165 (1): 107340.
- [57] SHEN X, LI Z, LIU Y, et al. Peb-ddi: A task-specific dual-view substructural learning framework for drug-drug interaction prediction[J]. IEEE Journal of Biomedical and Health Informatics, 2024, 28(1): 569-579.

- [58] YANG Z, ZHONG W, LV Q, et al. Learning size-adaptive molecular substructures for explainable drug–drug interaction prediction by substructure-aware graph neural network[J]. *Chemical Science*, 2022, 13(29): 8693-8703.
- [59] ZHANG Q, ZUO L, REN Y, et al. Fmca-dti: A fragment-oriented method based on a multihead cross attention mechanism to improve drug-target interaction prediction[J]. *Bioinformatics*, 2024, 40(6): btae347.
- [60] PENG L, LIU X, YANG L, et al. Bindti: a bi-directional intention network for drug-target interaction identification based on attention mechanisms[J]. *IEEE Journal of Biomedical and Health Informatics*, 2024, 29(3): 1602-1612.
- [61] TÜRK D, FUHR L M, MAROK F Z, et al. Novel models for the prediction of drug–gene interactions[J]. *Expert Opinion on Drug Metabolism & Toxicology*, 2021, 17(11): 1293-1310.
- [62] AISOPOS F, PALIOURAS G. Comparing methods for drug–gene interaction prediction on the biomedical literature knowledge graph: performance versus explainability[J]. *BMC Bioinformatics*, 2023, 24(1): 272.
- [63] BIAN J, ZHANG X, ZHANG X, et al. Mcanet: shared-weight-based multiheadcrossattention network for drug–target interaction prediction[J]. *Briefings in Bioinformatics*, 2023, 24(2): bbad082.
- [64] CHEN P, ZHENG H. Drug-target interaction prediction based on spatial consistency constraint and graph convolutional autoencoder[J]. *BMC Bioinformatics*, 2023, 24(1): 151.
- [65] NAPOLITANO F, ZHAO Y, MOREIRA V M, et al. Drug repositioning: a machine-learning approach through data integration[J]. *Journal of Cheminformatics*, 2013, 5(30): 1-9.
- [66] KEISER M J, ROTH B L, ARMBRUSTER B N, et al. Relating protein pharmacology by ligand chemistry[J]. *Nature Biotechnology*, 2007, 25(2): 197-206.
- [67] FAN J, FU A, ZHANG L. Progress in molecular docking[J]. *Quantitative Biology*, 2019, 7(1): 83-89.
- [68] KUNTZ I D, BLANEY J M, OATLEY S J, et al. A geometric approach to macromolecule–ligand interactions[J]. *Journal of Molecular Biology*, 1982, 161(2): 269-288.
- [69] MATOS G D R, PAK S, RIZZO R C. Descriptor-driven de novo design algorithms for dock6 using rdkit[J]. *Journal of Chemical Information and Modeling*, 2023, 63(18): 5803-5822.
- [70] MILON T I, WANG Y, FONTENOT R L, et al. Development of a novel representation of drug 3d structures and enhancement of the tsr-based method for probing drug and target interactions[J]. *Computational Biology and Chemistry*, 2024, 112(1): 108117.
- [71] BAGHERIAN M, SABETI E, WANG K, et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper[J]. *Briefings in Bioinformatics*,

- 2021, 22(1): 247-269.
- [72] GENTILE F, AGRAWAL V, HSING M, et al. Deep docking: A deep learning platform for augmentation of structure-based drug discovery[J]. ACS Central Science, 2020, 6(6): 939-949.
- [73] ÖZTÜRK H, ÖZGÜR A, OZKIRIMLI E. Deepdta: deep drug–target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.
- [74] LEE I, KEUM J, NAM H. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences[J]. PLOS Computational Biology, 2019, 15(6): e1007129.
- [75] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS): Vol. 30. 2017: 5998-6008.
- [76] HUANG K, XIAO C, GLASS L M, et al. Moltrans: molecular interaction transformer for drug–target interaction prediction[J]. Bioinformatics, 2021, 37(6): 830-836.
- [77] ZHAO Q, ZHAO H, ZHENG K, et al. Hyperattentiondti: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism[J]. Bioinformatics, 2022, 38(3): 655-662.
- [78] DOU L, ZHANG Z, QIAN Y, et al. Bcm-dti: A fragment-oriented method for drug–target interaction prediction using deep learning[J]. Computational Biology and Chemistry, 2023, 104(1): 107844.
- [79] KOH H Y, NGUYEN A T, PAN S, et al. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data[J]. Nature Machine Intelligence, 2024, 6(1): 673-687.
- [80] ZHENG S, LI Y, CHEN S, et al. Predicting drug–protein interaction using quasi-visual question answering system[J]. Nature Machine Intelligence, 2020, 2(2): 134-140.
- [81] LIN Z, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.
- [82] LI F, ZHANG Z, GUAN J, et al. Effective drug–target interaction prediction with mutual interaction neural network[J]. Bioinformatics, 2022, 38(14): 3582-3589.
- [83] MYSINGER M M, CARCHIA M, IRWIN J J, et al. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking[J]. Journal of Medicinal Chemistry, 2012, 55(14): 6582-6594.
- [84] ZHANG Q, WEI Y, LIAO B, et al. Mmd-dta: A multi-modal deep learning framework for drug-target binding affinity and binding region prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024, 21(6): 2200-2211.

- [85] CHEN L, TAN X, WANG D, et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments[J]. *Bioinformatics*, 2020, 36(16): 4406-4414.
- [86] ZHANG Z, LIU Q, WANG H, et al. Motif-based graph self-supervised learning for molecular property prediction[C]//*Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*: Vol. 34. 2021: 15870-15882.
- [87] ZANG X, ZHAO X, TANG B. Hierarchical molecular graph self-supervised learning for property prediction[J]. *Communications Chemistry*, 2023, 6(1): 34.
- [88] BOYD D B, LIPKOWITZ K B, CUNDARI T R. *Reviews in computational chemistry*[M]. Wiley-VCH, 2008.
- [89] KRENN M, HÄSE F, NIGAM A, et al. Self-referencing embedded strings (selfies): A 100% robust molecular string representation[J]. *Machine Learning: Science and Technology*, 2020, 1(4): 045024.
- [90] HANSON R M. Jmol smiles and jmol smarts: specifications and applications[J]. *Journal of Cheminformatics*, 2016, 8(1): 1-20.
- [91] LANDRUM G. RDKit: Open-source cheminformatics[EB/OL]. 2013. <http://www.rdkit.org>.
- [92] YUE Z X, YAN T C, XU H Q, et al. A systematic review on the state-of-the-art strategies for protein representation[J]. *Computers in Biology and Medicine*, 2023, 152(1): 106440.
- [93] ALBERTS B, HEALD R, JOHNSON A, et al. *Molecular biology of the cell: seventh international student edition with registration card*[M]. WW Norton & Company, 2022.
- [94] LESKOVEC J, SOSIČ R. Snap: A general-purpose network analysis and graph-mining library[J]. *ACM Transactions on Intelligent Systems and Technology*, 2016, 8(1): 1-20.
- [95] DAVIS J, GOADRICH M. The relationship between precision-recall and roc curves[C]//*Proceedings of the 23rd International Conference on Machine Learning (ICML)*. 2006: 233-240.
- [96] ADAM G, RAMPÁŠEK L, SAFIKHANI Z, et al. Machine learning approaches to drug response prediction: challenges and recent progress[J]. *NPJ Precision Oncology*, 2020, 4(19): 1-10.
- [97] 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.
- [98] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge [J]. *International Journal of Computer Vision*, 2015, 115(1): 211-252.
- [99] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//*Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016: 770-778.

- [100] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. *Neurocomputing*, 2021, 452(1): 48-62.
- [101] BAHDANAU D, CHO K H, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//*Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 2015.
- [102] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]//*Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*: Vol. 29. 2016: 289-297.
- [103] VRAHATIS A G, LAZAROS K, KOTSIANTIS S. Graph attention networks: a comprehensive review of methods and applications[J]. *Future Internet*, 2024, 16(9): 318.
- [104] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [C]//*Proceedings of the 5th International Conference on Learning Representations (ICLR)*. 2017.
- [105] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//*Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018.
- [106] MAAS A L, HANNUN A Y, NG A Y, et al. Rectifier nonlinearities improve neural network acoustic models[C]//*Proceedings of Machine Learning Research*: Vol. 28 *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Atlanta, GA, USA: PMLR, 2013: 3-9.
- [107] VO T H, NGUYEN N T K, KHA Q H, et al. On the road to explainable ai in drug-drug interactions prediction: A systematic review[J]. *Computational and Structural Biotechnology Journal*, 2022, 20: 2112-2123.
- [108] DEGEN J, WEGSCHEID-GERLACH C, ZALIANI A, et al. On the art of compiling and using 'drug-like' chemical fragment spaces[J]. *ChemMedChem*, 2008, 3(10): 1503.
- [109] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (elus)[A]. 2016. arXiv: 1511.07289.
- [110] ZITNIK M, AGRAWAL M, LESKOVEC J. Modeling polypharmacy side effects with graph convolutional networks[J]. *Bioinformatics*, 2018, 34(13): i457-i466.
- [111] TATONETTIN P, YE P P, DANESHJOU R, et al. Data-driven prediction of drug effects and interactions[J]. *Science Translational Medicine*, 2012, 4(125): 125ra31-125ra31.
- [112] MCDERMOTT M, ZHANG H, HANSEN L, et al. A closer look at auROC and aupRC under class imbalance[C]//*Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*: Vol. 37. 2024: 44102-44163.
- [113] YANG Y, LICHTENWALTER R N, CHAWLA N V. Evaluating link prediction methods[J]. *Knowledge and Information Systems*, 2015, 45: 751-782.

- [114] NIU D, XU L, PAN S, et al. Srr-ddi: a drug–drug interaction prediction model with substructure refined representation learning based on self-attention mechanism[J]. Knowledge-Based Systems, 2024, 285(1): 111337-111347.
- [115] DEWULF P, STOCK M, DE BAETS B. Cold-start problems in data-driven prediction of drug–drug interaction effects[J]. Pharmaceuticals, 2021, 14(5): 429.
- [116] LEE J, LEE I, KANG J. Self-attention graph pooling[C]//Proceedings of the 36th International Conference on Machine Learning (ICML). 2019: 3734-3743.
- [117] MULLARD A. 2021 fda approvals[J]. Nature Reviews Drug Discovery, 2022, 21(2): 83-88.
- [118] COUPEY S M. Barbiturates[J]. Pediatrics in Review, 1997, 18(8): 260-265.
- [119] DE MARCO B A, SALGADO H R N. Characteristics, properties and analytical methods of cefadroxil: a review[J]. Critical Reviews in Analytical Chemistry, 2017, 47(2): 93-98.
- [120] ANAND N, HUANG P. Generative modeling for protein structures[C]//Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS): Vol. 31. 2018: 7494-7505.
- [121] BA J L, KIROU J R, HINTON G E. Layer normalization[A]. 2016. arXiv: 1607.06450.
- [122] SMITH L N. Cyclical learning rates for training neural networks[C]//Proceedings of the 10th IEEE Winter Conference on Applications of Computer Vision (WACV 2017). 2017: 464-472.
- [123] WU H, LIU J, JIANG T, et al. Attentionmgt–dta: A multi–modal drug–target affinity prediction using graph transformer and attention mechanism[J]. Neural Networks, 2024, 169(1): 623-636.
- [124] DAVIS M I, HUNT J P, HERRGARD S, et al. Comprehensive analysis of kinase inhibitor selectivity[J]. Nature Biotechnology, 2011, 29(11): 1046-1051.

附录 A 补充材料

表 A.1 层次化节点的特征含义

特征含义	维度
节点标识	53
节点度数	1
隐含价	1
形式电荷	1
自由基电子数	1
杂化类型	7
原子是否处于芳香环	1
原子上氢原子的总数	1
总计	66

表 A.2 HDN-DDI 超参数搜索空间

超参数	搜索空间
HDN 编码块的数量	[3, 4, 5, 6, 7]
隐藏层的维度	[32, 64, 128, 256]
GAT 中的头数	[1, 2, 4]
批量大小	[256, 512, 1024]
学习率	[1e-2, 1e-3, 1e-4]

表 A.3 氨基酸节点的理化特征

特征含义	维度
氨基酸标识	21
是否为脂肪族氨基酸、是否为芳香族氨基酸	2
是否为极性中性氨基酸、是否为酸性带电氨基酸、是否为碱性带电氨基酸	3
分子量	1
酸性解析常数、碱性解离常数、其他解离常数	3
等电点	1
pH 值为 2 与 pH 为 7 时的疏水性指数	2
总计	33

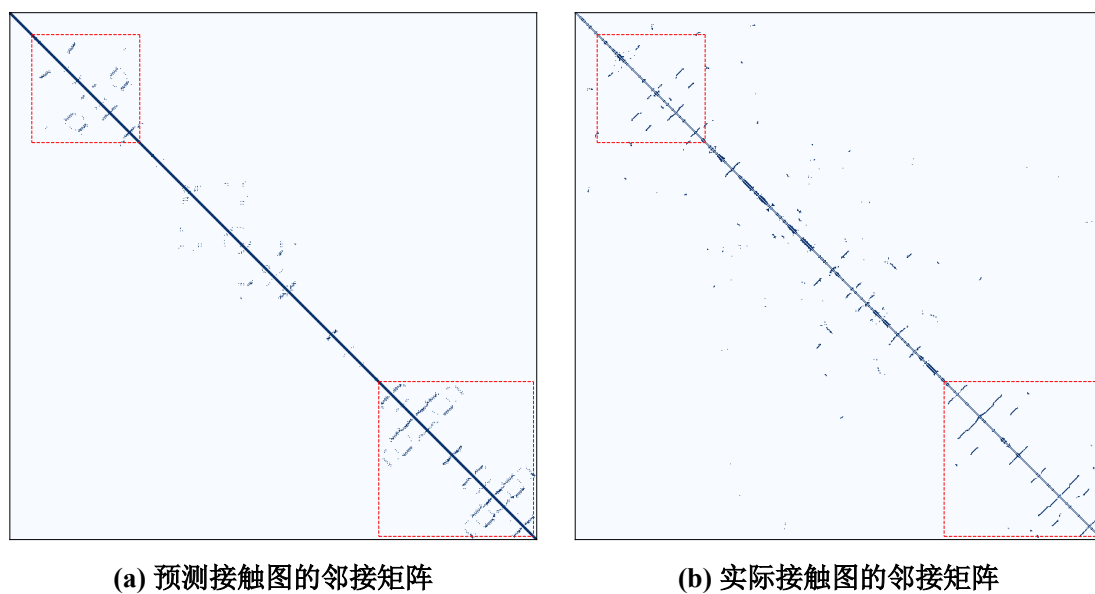


图 A.1 靶点 P00488 的预测接触图与实际距离图的邻接矩阵