分类号：O1　　学校代号：10150
U D C：51　　学　　号：20222451

大连交通大学

# 硕 士 学 位 论 文

# 基于深度学习的药物分子表征与性质预测研究

# Studies on drug molecule representation and property prediction based on deep learning

学 生 姓 名：张琦
导 师 及 职 称：刘立伟 教授
学 科 门 类：理学
专 业 名 称：数学
研 究 方 向：生物信息学
申请学位级别：硕士学位
论文答辩日期：2025 年 6 月 10 日
学位授予单位：大连交通大学

# 摘　要

本文聚焦于药物分子的特征表示及其在多性质预测中的应用，以深度学习理论与数据挖掘算法为工具，深入探讨了药物研发过程中的关键挑战。面对药物分子及其关联生物实体在结构和功能层面的高度异质性，如何建立精准的表征模型成为提升预测性能的核心挑战。针对这一难题，本研究从药物-靶标相互作用的双模态视角出发，通过协同优化单模态特征自学习与跨模态关联建模，深入解析药物-靶标互作的内在机制。

在研究方法上，本文首先系统梳理了药物分子表征和性质预测的主流方法，涵盖了文本、指纹、图形等多种表示方式，以及理化性质、毒性、药物间相互作用、药物-靶标结合亲和力及结合位点等关键性质。随后，本文提出了一种创新的基于对比学习的药物特征表示方法——DrugDL。该方法通过强化单模态内部特征间的联系，深入学习和理解跨模态特征间的交互模式，实现了对药物分子的高效表征。实验结果显示，DrugDL 在分子表征能力上相较于传统方法具有显著优势。

在模型构建与应用方面，本文将 DrugDL 所提取的特征与多样化的网络结构相结合，构建了一个全面的药物分子多性质预测模型。该模型能够精确预测药物研发中的一系列关键属性，包括药物的理化性质、毒性评估、药物-药物相互作用、药物-靶标结合亲和力以及结合位点等。通过与一系列先进的基线模型进行对比分析，结果显示 DrugDL 在所有对比任务上均展现出压倒性的优势，充分证明了其全面且卓越的预测能力。实际应用中，该模型已成功应用于抗 SARS-CoV-2 药物和代谢酶抑制剂的高通量筛选，并为 BRAF、ALK 等在内的多种癌症靶点药物研发提供理论支持。

总体而言，本论文通过应用深度学习方法，对药物分子表征和性质预测进行了深入研究。研究结果表明，本文提出的 DrugDL 方法凭借其强大的表示学习能力，能够解析复杂且多样的生物医药数据，揭示药物-靶标作用的内在规律，进而展现出卓越的药物分子性质预测能力。未来，该方法有望为药物研发领域带来更多的创新和突破。

**关键词：分子表征；分子性质预测；药物靶标预测；深度学习**

# Abstract

This study focuses on drug molecule feature representation and multi-property prediction, employing deep learning and data mining to address critical drug development challenges. Confronting structural and functional heterogeneity in drug molecules, we propose establishing precise representation models through dual-modal drug-target interaction analysis. By synergistically optimizing intra-modal self-learning and cross-modal correlation modeling, we unravel drug-target interaction mechanisms.

Methodologically, this work first systematically reviews mainstream approaches for drug molecular representation and property prediction, encompassing text-based, fingerprint-based, and graph-based representations, along with critical properties including physicochemical characteristics, toxicity, drug-drug interactions, drug-target binding affinity, and binding sites. Subsequently, we propose DrugDL, an innovative contrastive learning-based drug feature representation method. By reinforcing intra-modal feature relationships and comprehensively learning cross-modal interaction patterns, this approach achieves efficient characterization of drug molecules. Experimental results demonstrate DrugDL's significant advantages in molecular representation compared to conventional methods.

In model construction and application, we integrate features extracted by DrugDL with diverse network architectures to develop a comprehensive multi-property prediction model. This framework enables accurate prediction of crucial attributes in drug development, including physicochemical properties, toxicity assessment, drug-drug interactions, drug-target binding affinity, and binding sites. Comparative analyses with state-of-the-art baseline models reveal DrugDL's overwhelming superiority across all evaluated tasks, fully validating its comprehensive predictive capabilities. Practically, the model has been successfully applied to high-throughput screening of anti-SARS-CoV-2 compounds and metabolic enzyme inhibitors, while providing theoretical support for drug development targeting various cancer-related targets including BRAF and ALK.

In conclusion, this dissertation conducts in-depth research on drug molecular representation and property prediction through deep learning methodologies. The results demonstrate that the proposed DrugDL method, empowered by its robust representation learning capacity, effectively analyzes complex biomedical data to reveal intrinsic patterns of drug-target interactions, thereby exhibiting exceptional predictive performance. This approach holds potential to drive future innovations in pharmaceutical research and development.

**Key Words: Molecular Representation; Molecular Property Prediction; Drug Targets Prediction; Deep Learning**

# 目 录

# 第一章 绪　论

## 1.1 研究背景及意义

在当今快速发展的生物医学领域，药物研发是一个至关重要的环节，它不仅直接关系到人类健康水平的提升，也是推动医药产业进步和创新的关键动力[1]。药物研发的核心在于理解药物分子与目标生物大分子 (如蛋白质、DNA 等) 之间的相互作用机制[2]，以及药物分子本身的各种物理化学性质[3]。例如，药物分子间的相互作用、药物靶标的相互作用以及药物的毒性。这些性质直接决定了药物的生物活性、药代动力学特性等关键参数，是药物筛选和优化过程中不可或缺的信息。

药物间的相互作用，依据其效果的不同，可细分为增强、抑制及拮抗三种类型[4]。举例来说，当某些抗生素与 β-内酰胺酶抑制剂协同作用时，会显著增强抗生素的抗菌效能，这一过程即被称为增强作用[5]。相反，有些药物可能会抑制胃酸的分泌，进而妨碍那些依赖胃酸辅助吸收的药物的摄取，这则体现了抑制作用[6]。再者，当钙通道阻滞剂与β受体阻滞剂并用时，两者可能会产生拮抗效应，使得原本的降压效果有所削弱。因此，药物-药物相互作用的预测有助于优化药物组合，提高治疗效果[7]。通过预测不同药物之间的相互作用，研究人员可以筛选出最佳的药物组合，从而在治疗中达到更好的疗效，同时减少不必要的药物相互作用和副作用。

药物靶标，作为人体内与疾病紧密相关且能被药物特异性识别的生物大分子 (如蛋白质、核酸等)，扮演着促使药物发挥疗效的关键角色[8]。药物通过与这些针对特定疾病的靶标精准结合，从而达到治疗疾病的目的。药物靶标的发现对于深入剖析药物作用机制、阐释药物副作用及理解疾病病理具有重大意义[9]。在药物研发的复杂流程中，靶标的选择与确认构成了至关重要的步骤。这一环节不仅为后续研发奠定了坚实的基础，还直接关系到药物的疗效与安全性。借助对药物与靶标相互作用的精准预测，研究人员能够高效地筛选出具有潜力的药物候选分子，从而显著加速药物研发的进程[10]。此外，这一预测技术还为实现药物的精准治疗开辟了道路。通过深入分析药物与特定靶标的结合特性，研究人员能够设计出更加个性化的治疗方案，从而不仅提升了治疗效果，还大幅度降低了副作用的发生。更为深远的是，药物靶标相互作用的预测还为新药研发领域的创新注入了强大动力。通过对药物与靶标相互作用机制的深入探索，研究人员能够不断挖掘新的药物作用靶点和药物类型，为新药研发提供了更为广阔的思路和更为先进的方法[11]。这不仅丰富了药物研发的工具箱，更为人类战胜疾病提供了更多的可能性。

预测药物的理化性质和各种毒性在药物研发领域具有深远的意义[12]。它不仅能够显著提升研发效率，通过早期筛选潜力药物并减少因毒性或理化性质不佳导致的失败，从而节省研发成本[13]；还能有力保障药物的安全性[14]，降低毒性风险，优化药物设计，确保患者在使用过程中的安全[15]。此外，这一技术还为个性化医疗的发展提供了重要支撑，通过预测药物在个体内的代谢和毒性情况，实现精准治疗，提高治疗效果[16]。更重要的是，深入研究药物的理化性质和毒性有助于发现新的药物作用靶点和药物类型，推动新药研发领域的创新，促进跨学科合作，形成更加综合和系统的药物研发体系。

然而，传统的药物性质预测手段面临着诸多挑战。实验测定方法虽然准确，但耗时费力、成本高昂，且部分实验因涉及动物而引发伦理和道德上的争议[17]。另一方面，理论计算方法虽能提供一定程度的指导，但其精度和适用范围有限，难以应对大规模药物筛选的需求。据统计，从化合物的发现到最终药物的上市，往往需要耗费数年甚至更长时间，经济成本更是高达数十亿美元[18]。这一现状迫切要求科学家们探索更为高效[19]、精准的药物研发方法[20]。

近年来，人工智能技术的蓬勃发展，特别是深度学习技术的异军突起，为药物研发领域带来了前所未有的变革[21]。深度学习凭借其强大的数据处理能力和模式挖掘能力，在药物分子表征与性质预测方面展现出巨大的潜力[22]。通过多层神经网络结构，深度学习能够自动学习数据的高维特征表示，实现对药物分子结构与性质之间复杂关系的有效建模和预测[23]。这一技术的引入，为药物研发提供了一种全新的、基于大数据和机器学习的解决方案，极大地提高了药物发现的效率和准确性[24]。

基于深度学习的药物分子表征与性质预测研究不仅具有深远的科学意义，更在实践中展现出巨大的应用潜力。它加速了药物研发进程，降低了研发成本，提升了研发成功率，为药物研发领域带来了革命性的变化[25]。甚至深度学习技术所实现的准确药物性质预测，现已转化为对药物分子结构优化具有重要指导意义的信息[26]。依托深度学习所揭示的药物分子功能机理，研发能够治疗特殊疾病的安全有效药物也成为了可能，为患者的治疗带来了全新的希望[27]。此外，深度学习技术还有助于推动个性化医疗的发展，通过对药物分子与个体遗传信息的综合分析，实现药物的精准匹配，为个性化医疗提供有力支持。同时，这一领域的研究还促进了计算机科学、数学与生物医学的交叉融合，为其他领域如材料科学、环境科学等提供了方法学上的借鉴和启示。

## 1.2 国内外研究现状

鉴于传统药物性质验证实验存在的种种局限，计算机辅助预测方法以其成本低廉、快速高效且安全可靠的显著优势，为药物性质的精准预测开辟了一条全新的路径[28]。近年来，国内外学者在这一前沿领域取得了令人瞩目的显著进展，成功开发出众多性能卓越的药物性质预测算法[29]，这些创新成果不仅推动了药物研发领域的深刻变革[30]，更为新药发现带来了革命性的突破[31]。

深度学习模型在药物分子的特征表示、理化性质及毒性预测方面展现出了巨大的潜力和独特的优势。多个研究团队利用深度学习技术对药物分子的水溶性、穿透血脑屏障能力、亲脂性等理化性质，以及致癌性、肝损伤等毒性数据进行了深入的学习和分析，成功构建了一系列预测精度高的模型[32]。这些模型通过构建复杂精细的网络结构，从药物分子特征中精准提取关键信息[33]，从而实现了对药物毒性的高精度预测[34]。在早期研究阶段，药物毒性和理化性质的预测主要依赖于简单的特征表示方法和传统的分类器，如贝叶斯模型和支持向量机等[35]。然而，随着技术的持续进步，研究者们开始积极探索将深度学习模型与更为复杂多样的特征表示方法相结合的新途径。例如，他们尝试将分子的扩展连接指纹作为输入，通过深度神经网络来预测药物毒性[36]。除了传统的分子指纹特征外，研究者们还不断挖掘和探索其他类型的特征表示方法[37]。其中，一种创新性的方法是将分子序列转化为分子 RGB 图像，并利用深度残差神经网络对这些图像进行高效处理，这种分子图像的表征方式为药物性能的预测提供了全新的视角和思路[38]。目前，基于分子图的表示方法已成为应用最为广泛的药物特征表示手段之一。该方法以原子为节点，化学键为边缘，构建出二维的分子图[39]，并通过先进的图神经网络 (GNN) 进行特征提取和精准预测[40]。最近，该领域又取得了新的突破，一些基于分子图的药物子结构分割方法和药物基序 (motif) 图开始崭露头角[41]，它们为药物研发提供了更为详尽可靠的信息，进一步推动了药物研究的深入发展[42]。

药物-药物相互作用预测是药物研发领域的重要课题之一。目前，计算机辅助的药物间相互作用预测方法主要分为两大类。第一类是基于药物相似性的算法。这类方法从药物的特征出发，认为具有相似特征的药物可能具有相似的性质或相互作用类型[43]。例如，早期的研究中，大多数模型通过结合多个特征计算的药物相似性来预测未知的药物间相互作用[44]。后期，有些模型整合了已知药物间相互作用、药物-靶标相互作用和药物化学结构信息，对药物间的相互作用进行了更为精准的预测[45]。近年来，很多方法借助药物亚结构、靶点、副作用等多元数据，利用神经网络构建预测模型，取得了显著的效果。此外，还有一部分模型创新性地提出了利用医学文本中包含的药物相

似性信息进行药物间相互作用预测的方法，通过双向长短期记忆 (BiLSTM) 模型和注意力机制，生成新的特征向量，显著提高了药物-药物相互作用预测的准确性[46]。第二类方法是基于网络的算法[47]。这些算法利用了已知的药物相互作用网络[48]，将药物间相互作用预测任务视为链路预测任务[49]。例如，很多模型利用元路径来计算和衡量药物之间的关系[50]，并综合考虑药物和其他生物医学实体之间的丰富语义信息，通过随机森林和神经网络等先进算法预测潜在的药物对相互作用[51]。同时，有些模型设计了元路径级别的注意力机制，通过组合不同长度的元路径获得的语义特征，进一步提升了药物特征的表示能力和预测精度[52]。随着技术的不断发展，药物-药物相互作用预测方法正在不断演进和完善，为药物研发提供了强有力的支持和保障[53]。

药物-靶标相互作用及亲和力预测构成了药物发现流程中的核心环节，对于加速新药研发具有至关重要的意义。众多深度学习模型将该类预测视为二进制分类任务或回归任务 (亲和力的预测)，利用深度神经网络、GNN 或 Transformer 架构等深度编码与解码模块，从大规模药物靶标相互作用数据中自动学习药物和蛋白质的数据驱动表示，从而实现了高效准确的预测[54]。这些模型不仅充分利用了药物和蛋白质的线性或二维结构信息[55]，还从化学基因组学角度构建了统一的端到端预测框架[56]，将化学空间[57]、基因组空间[58]和相互作用信息有机融合[59]。同时，越来越多的深度网络被应用于药物-靶标结合位点预测的研究中。诸多模型利用先进的注意力模块，在实验结果中成功可视化了药物-靶标结合位点预测结果，为结合位点预测提供了新的思路和方法[60]。此外，通过为蛋白质子序列分配不同权重、应用双侧注意力机制、使用反卷积揭示重要相互作用区域等方法的相继推出和不断优化，药物-靶标相互作用结合位点预测的准确性得到了显著提升[61]。随着技术的不断进步和数据的持续积累，相信未来会有更多高效准确的预测模型不断涌现[62]，为药物研发注入新的活力，推动新药发现领域迈向更高的台阶[63]。

现有的基于深度学习的方法在药物分子表征和性质预测领域尽管已取得一定进展，但仍存在诸多不足与缺陷。在特征提取层面，这些方法显得较为单一，缺乏足够的丰富性。传统的分子指纹方法虽然能从分子结构中提取信息，但可能会遗漏重要的化学或生物信息，进而削弱预测的准确性。同时，分子图在揭示分子内部局部结构和关系方面表现出色，但在表达全局特性或整体结构时却存在局限性，这在一定程度上限制了其在某些任务中的表示能力。此外，以往的预测模型往往缺乏对模型决策过程的深入理解，导致在药物设计和开发过程中难以发挥显著作用。这不仅影响了模型在实际应用中的效果，也增加了对模型预测结果解释的难度。

在药物分子表征方面，药物分子由化学子结构组成，这些子结构决定了其所有的药代动力学和药效学特性。然而，以往的研究大多依赖整个 SMILES 序列来进行学习和预测，而很少考虑分子子结构之间更加细致的相互作用。即使有部分方法提取了子结构，但其固定个数和范围的提取方式往往导致模型过拟合，仅在数据充足、模型已部分训练完成的热启动场景中表现良好，而在数据稀缺、模型需从头训练的冷启动验证中效果并不理想。同时，在药物-药物和药物-靶标相互作用的预测方面，现有方法也存在诸多不足。大多数模型独立地提取每个药物的特征，缺乏药物之间的充分特征交互，这使得模型难以深入理解更深层次的相互作用机理，进而影响模型的预测精度和可解释能力。此外，在药物-靶标亲和力和药物-靶标结合位点的预测中，也存在一些不足，如特征整合方式过于简单、蛋白质三维结构表示和预测方法存在误差等问题。这些问题都亟待解决，以推动药物分子表征和性质预测领域的进一步发展。

## 1.3 本文工作

本文针对以上问题提出了解决方案。本文借助于对比学习的思想提出了一种新的药物分子表征方法，其首先通过跨模态交互学习模块将存在相互作用的药物分子与靶标分子进行特征整合学习，借助药物-靶标的关联性和靶标的特征信息深度挖掘药物分子的潜在信息。然后采用矩阵奇异值分解 (SVD) 和高斯核函数来计算药物分子以及靶标之间的相似性，并通过引入单模态相似性损失来增强药物分子自身的特征属性，有效防止特征信息的丢失。最后，模型联合两个模块的损失目标进行联合优化，旨在获得强表达力的药物分子特征。与其它药物分子特征表示方法相比，本文提出的分子表征方法的主要优点在于：(1) 增强的特征表达能力。通过跨模态交互学习以及单模态特征强化，该方法能够整合药物分子与靶标分子的特征信息，捕获它们之间的复杂关系，并保持药物和靶标各自模态的内在结构，保持各自特征的独立性和完整性。(2) 广泛的应用场景。该方法可以支持多种下游任务，如药物-靶标结合亲和力预测、药物重定位、新靶标发现等。(3) 提供药物先导优化建议。通过药物-靶标相互作用关系以及细致的结合位点预测分析，该方法对药物分子潜在特性进行深入理解，为后续的药物结构优化工作提供相关的指导。

此外，我们依托深度学习网络框架，在多个应用场景下对所获取的药物分子特征进行了详尽的分析与验证。实验结果表明，无论是在药物理化性质的预测、毒性的评估、药物间相互作用的解析，还是在药物-靶标相互作用的识别、亲和力的估算以及结合位点的定位等方面，该方法均展现出了卓越且令人满意的性能。

本文余下的主要内容安排为：第二章将深入探讨药物分子的表征方法，详细阐述如何从不同维度对药物性质进行相关的预测。这一章将为我们理解药物分子的基本特性和其性质预测提供坚实的理论基础。第三章将提出一种基于对比学习的药物特征表示方法。该方法将展示其在药物分子表征方面的强大能力，并通过与当前常用的表征方法进行对比分析，凸显其优越性和独特性。第四章将基于深度学习技术，利用我们提取的药物分子特征进行一系列药物分子性质的预测。这些预测任务涵盖了药物理化性质、毒性评估、药物-药物相互作用、药物-靶标相互作用识别、结合亲和力估算以及相互作用位点定位等多个方面。通过这一章的研究，我们将进一步验证所提特征表示方法的有效性和实用性。第五章将对全文进行总结，并展望未来的研究方向。我们将回顾本文的主要研究成果，指出当前研究的局限性，并提出可能的改进方案和未来的探索方向，为药物分子表征和性质预测领域的发展贡献新的思路。

# 第二章 药物分子表征和性质研究概述

## 2.1 药物分子表征方法

### 2.1.1 文本表示

药物分子表征是药物研发过程中的重要环节，其中文本表示是药物分子表征的一种常用方法。文本表示是指将药物分子的结构信息以文本的形式进行编码和表示。这种表示方法能够捕获药物分子的基本结构特征，并为其在后续的药物研发过程中提供重要依据。药物分子的文本表示方法多种多样，其中应用最为广泛的为 SMILES 表示法、InChI 表示法等。

SMILES (Simplified Molecular-Input Line-Entry System) 是一种描述分子结构的规范语言，它使用简短的文本字符串来表示化学结构[64]。该表示法具有简洁性、唯一性、可逆性等特点。在特定条件下 (如使用相同的生成规则和算法)，每个分子都可以生成一个唯一的 SMILES 字符串。尽管存在多个合理的 SMILES 表示同一个分子的情况，但可以通过标准化算法 (如规范 SMILES) 来生成唯一的表示。同时，SMILES 字符串可以方便地转换回分子结构，这使得它在分子设计和药物研发中具有重要价值。

SMILES 表示法不仅包含分子的骨架结构，还可以表示支链、环结构、双键、三键、芳香键以及手性、电荷等复杂信息。对于常见的原子 (如 B、C、N、O、P、S、F、Cl、Br 等) 直接用其元素符号表示。其他原子 (如金属原子) 需要加方括号表示，如[Na]表示钠原子。芳香体系的原子用小写字母表示，如“c”表示苯环中的碳原子。氢原子通常不显示，但在需要时可以隐含地添加到分子结构中。单键可以省略或用“-”表示。双键用“=”表示。三键用“#”表示。芳香键在芳香体系中省略，SMILES 自动识别。支链用小括号表示，如异丁酸的 SMILES 表示为 CC(C)C(=O)O。环结构断开一根键后以数字表示成环位置，环编号顺序任意。例如，环己烷的 SMILES 可以表示为 C1CCCCC1。手性中心的绝对构象可以通过符号“@”和“@@”来分别定义取代基的逆时针或顺时针取向。电荷可以在方括号中定义，如[OH-]表示带负电荷的羟基。

如图 2.1 所示，我们以阿司匹林为例进行 SMILES 表示。阿司匹林的分子结构式为 C9H8O4，其 SMILES 表示可以为：CC(=O)Oc1ccccc1C(=O)O。这个 SMILES 字符串遵循了上述 SMILES 表示法的基本规则，其中：“C”和“O”分别表示碳原子和氧原子。“c1ccccc1”表示苯环结构，其中小写字母“c”表示苯环中的碳原子。“C(=O)O”表示羧基结构，其中“=”表示双键，“O”表示氧原子。整个字符串没有显示氢原子，但根据价键饱和原则，氢原子可以隐含地添加到分子结构中。

### 2.1.2 指纹表示

药物分子指纹表示的基本原理是将药物分子的化学结构信息转化为一种易于计算机处理和比较的格式。这种格式通常是一维的二进制序列 (向量)，其中每一位 (或位点) 代表了药物分子中某一特定的结构特征或属性的有无。例如，某一位可能代表药物分子中是否含有苯环，或者某个特定的官能团是否存在。现在最为常用的分子指纹特征有 MACCS 指纹 (Molecular ACCess System Fingerprints)、PubChem 指纹 (PubChem Fingerprint) [65]以及 ECFP 指纹 (Extended-Connectivity Fingerprints) 等[66]。

MACCS 指纹是一种用于表示分子结构的二进制指纹，它基于分子中是否含有特定的亚结构来定义。MACCS 指纹共包含 166 个不同的分子特征，每个特征都对应于一个特定的化学子结构，如羟基、苯环或氮原子等。如果分子中存在某个特征，则相应位置上的值为 1，否则为 0。PubChem 指纹由 PubChem 数据库开发并用于分子相似性搜索和化学信息学分析。PubChem 指纹基于分子中是否存在预定义的子结构进行编码，生成一个由 881 个位组成的二进制字符串。每个位表示特定子结构的存在与否，1 表示该子结构存在，0 表示不存在。这些子结构是根据 PubChem 数据库中广泛的化合物进行挑选和定义的，旨在为分子相似性比较提供一致性。ECFP 指纹是一种圆形拓扑指纹，可用于分子表示、相似性搜索、构效关系建模等。ECFP 指纹通过圆形原子邻域来表示分子结构。它首先对给定分子的每一个非氢原子分配一个初始整数标识符，该标识符包含相应原子的化学信息。然后，通过一系列迭代操作，将某一原子的初始标识符与邻近原子的标识符合并，直到到达设置的半径为止。每一轮迭代都会捕捉距中心原子越来越远的原子信息，最终经哈希运算编码成为一个整数值，这些整数值合并形成一个整数标识列。ECFP 指纹还可以通过“折叠”操作形成定长比特串，简化比对和相似性计算。

此外，还有基于字典的、圆形的 (circular)、拓扑的[67]、药效团的 (pharmacophore)、蛋白质-配体相互作用的、基于形状的、强化的等[68]多种类型的药物分子指纹表示方法。

### 2.1.3 图形表示

药物分子的图像表示主要分为两大类：分子图和分子图像 (RGB 图像)。这些表示方法通常依赖于特定的软件或库，如 RDkit [69]，它们能够解析 SMILES 字符串，识别其中的原子和化学键信息，并据此构建分子图和生成分子图像。

分子图是一种基于节点和边缘的抽象表示方法，专注于展示药物分子的结构骨架和原子间的连接关系。分子图 $G_a=(V_a,E_a)$ 是一种直接反映药物分子内部原子排列和化学键连接的图模型。在 $G_a=(V_a,E_a)$ 中，$V_a$ 代表原子节点集合，而原子之间的化学键边

集合则被表示为$E_a$。药物分子中包含$n$个原子，每个原子$i\,(i=1,2,\cdots,n)$在图中对应一个节点$v_i \in V_a$。若原子$i$与原子$j$之间存在化学键，则在图中对应一条从$v_i$到$v_j$的边$e_{ij} \in E_a$。这种图模型能够直观地展示药物分子的空间构型和原子间的相互作用。如图 2.1 所示，我们以阿司匹林为例进行分子图表示。随着药物分子子结构和官能团研究的深入，药物 motif 图也应运而生。与药物原子图不同，药物 motif 图$G_m=(V_m,E_m)$是一种更高级别的图模型，它关注的是药物分子中的结构片段 (motif) 以及它们之间的相互作用。在$G_m=(V_m,E_m)$中，$V_m$表示由 motif 组成的节点集合，而节点之间的边的集合则被表示为$E_m$。药物分子中包含$m$个 motif，每个 motif $k\,(k=1,2,\ldots,m)$在图中对应一个节点$w_k \in V_m$。如图 2.1 所示，我们同样以阿司匹林为例进行子结构分割，并构建相应的 motif 图。

分子图像是一种二维图形表示方法，通常基于红、绿、蓝三个颜色通道来构建，以直观且简洁的方式展示分子的结构特征。在这种图像中，分子中的原子被特定的符号 (一般为元素符号) 所表示，并通过线条 (常见为键线) 相连，精确地描绘了原子间的化学键连接。分子图像能够清晰地展现分子的形状、大小以及空间构型，为科学家提供了直观理解分子间相互作用和反应机制的重要工具。在分子图像中，为了增强视觉效果并优化信息传达，分子的不同部分会采用不同的颜色或阴影进行区分。此外，分子图像还具备展示药物分子特定性质的能力，例如通过颜色编码的方式呈现分子的电荷分布、电子云密度以及分子轨道等关键性质。这些性质对于深入理解药物分子的反应活性和生物活性至关重要。对于具有复杂形状或特殊构象的化合物，传统的基于分子指纹和分子图的特征表示方法可能难以准确捕捉其结构细节。相比之下，图像处理技术能够通过像素级别的分析来精确捕捉分子的细微结构变化，因此，对分子的图像进行表征在药物研发领域具有重要意义。如图 2.1 所示，我们通过 RDKit 算法库中的 Draw.MolToImage 函数将阿司匹林的 SMILES 序列转化为分子图像。

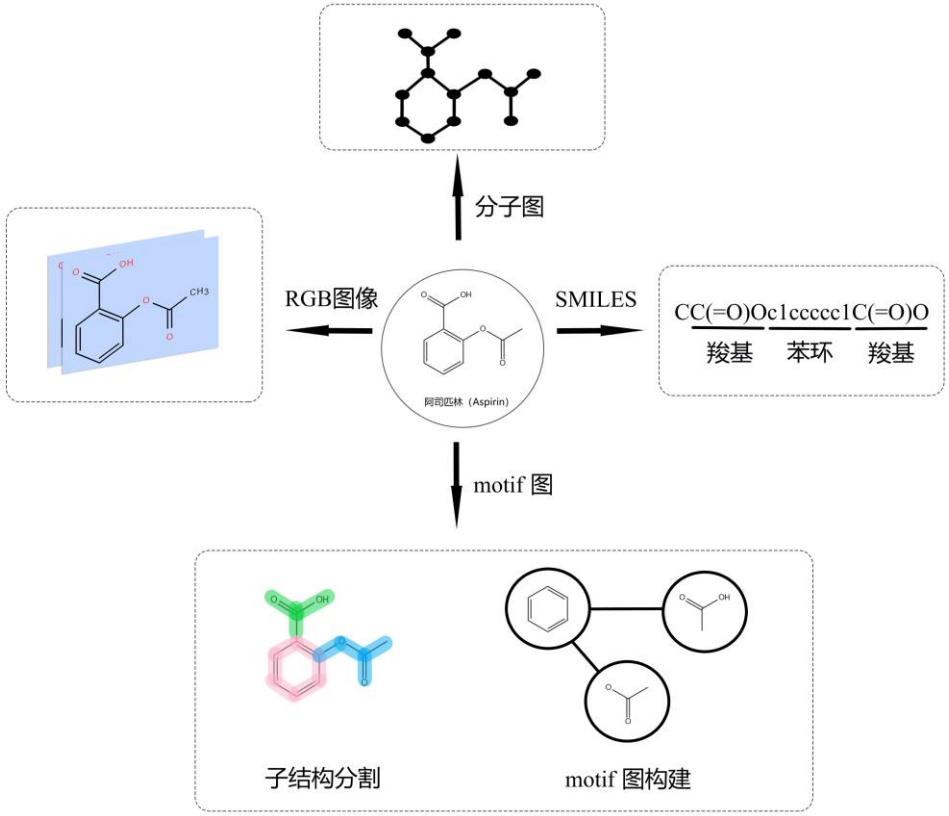



图 2.1 药物分子表征方法示意图

Fig. 2.1 Schematic diagram of the representation method of drug molecules

## 2.2 药物多性质研究

### 2.2.1 药物理化性质研究

在药物研发过程中，药物理化性质的预测是至关重要的一环。这些性质不仅影响着药物的生物利用度和药代动力学，还直接关系到药物的疗效和安全性。当前，药物理化性质的预测涵盖了多个关键方面，包括但不限于药物对 β-分泌酶 1 的抑制活性 (BACE)、药物血脑屏障穿透性 (BBBP)、药物水溶性 (ESOL)、脂溶性 (Lipophilicity) 以及自由溶解度 (FreeSolv) 等性质。

BACE 作为阿尔茨海默病 (AD) 病理机制中的核心酶类，其抑制活性的预测对于开发新型 AD 治疗药物至关重要[70]。当前，基于结构生物信息学的预测方法，如分子对接 (Docking) 技术，已成为预测化合物对 BACE 抑制活性的主流手段。该方法通过模拟化合物与 BACE 活性位点的相互作用，评估其结合亲和力，进而预测抑制活性。此外，

机器学习算法，特别是深度学习技术，正逐步应用于 BACE 抑制剂的筛选与活性预测中，通过挖掘大量已知抑制剂数据中的潜在规律，实现更为精准的预测。

血脑屏障 (BBB) 作为保护中枢神经系统免受外界干扰的天然屏障，对药物的 BBB 穿透性提出了严格要求。当前，基于物理化学描述符的定量构效关系 (QSAR) 模型已成为预测药物 BBB 穿透性的主流方法。这些模型通过整合药物的分子量、极性表面积、氢键供受体数量等关键描述符，结合机器学习算法，构建出能够准确预测药物 BBB 穿透性的预测模型。此外，基于分子动力学的模拟方法也被用于研究药物分子与 BBB 中转运蛋白的相互作用，进一步提升了预测的精确度。

水溶性作为影响药物生物利用度的关键因素，其预测方法经历了从传统实验测定到现代计算预测的演变。当前，基于不同描述符的统计学方法，如多元线性回归、偏最小二乘法等，以及量子力学 (QM)、分子力学 (MM) 计算方法，如密度泛函理论 (DFT)、蒙特卡洛模拟等，共同构成了预测药物水溶性的强大工具箱。这些方法通过模拟药物分子在水溶液中的行为，计算其溶解自由能等关键参数，进而预测水溶性。此外，基于大数据的机器学习方法，通过挖掘大量药物水溶性数据中的潜在规律，实现了更为快速、准确的预测。

脂溶性作为药物发现与设计中的关键参数，其预测对于优化药物的膜渗透性、活性强度及选择性具有重要意义。当前，基于子结构的加和模型法，如 logP 值预测，已成为预测药物脂溶性的主流方法。这些方法通过计算药物分子中各原子或片段的疏水贡献，累加得到整体的 logP 值，进而评估药物的脂溶性。此外，基于性质的方法，如分子表面积、分子体积、偶极矩等描述符，也被用于构建更为精细的脂溶性预测模型。

药物的自由溶解度是指在一定温度和压力下，药物在纯溶剂 (通常是水) 中达到溶解平衡时所溶解的最大量。FreeSolv 数据集为药物自由溶解度的预测提供了宝贵的实验数据资源，为构建精准的溶解度预测模型提供了坚实基础。当前，基于机器学习算法的预测模型，如随机森林、支持向量机等，已成为预测药物自由溶解度的主流方法。这些模型通过挖掘 FreeSolv 数据集中的溶解度数据与药物描述符之间的关联，实现了对药物在不同溶剂中溶解度的准确预测。

### 2.2.2 药物毒性研究

药物毒性涵盖多个方面，包括致癌性、致突变性、药物性肝损伤 (DILI) 以及由人类醚-à-去甲肾上腺素能受体基因相关基因 (hERG) 介导的心脏毒性等。这些毒性不仅直接影响患者的生命健康，也是药物研发过程中必须严格监控的要素。以下将对这些毒性及其预测方法进行深入解析。

药物致癌性是指药物或其代谢产物在人体内具有潜在的诱发癌症的能力[71]。致癌性的预测通常依赖于结构警报分析、遗传学毒性测试以及基于机器学习的预测模型。结构警报分析通过识别药物分子中潜在的致癌性结构特征，如芳香胺、烷基化剂等，来初步评估药物的致癌风险。遗传学毒性测试，如 Ames 试验，能够直接评估药物分子对 DNA 的损伤能力，从而间接反映其致癌潜力。此外，基于机器学习的预测模型通过挖掘大量已知致癌性和非致癌性化合物的数据，构建出能够准确预测新化合物致癌性的模型，为药物研发提供有力支持。

药物致突变性是指药物能够引起生物体遗传物质发生可遗传的变异，这种变异可能导致基因突变、染色体畸变等，进而引发疾病[72]。致突变性的预测主要依赖于遗传毒性测试、体外和体内基因突变试验以及基于计算的方法。遗传毒性测试能够直接评估药物分子对 DNA 的损伤能力，是预测致突变性的重要手段。体外和体内基因突变试验则通过模拟药物在生物体内的代谢过程，观察其对基因突变的诱导作用。此外，基于计算的方法，如 QSAR 模型和机器学习算法，通过挖掘药物分子结构与致突变性之间的关联，实现对药物致突变性的快速预测。

DILI 是指药物或其代谢产物在肝脏内代谢过程中引起的肝脏损伤，是药物研发中常见的安全性问题[73]。DILI 的预测通常依赖于基于机器学习的预测模型以及代谢组学和蛋白质组学研究方法。基于机器学习的预测模型通过整合药物的物理化学性质、代谢途径、肝毒性机制以及患者遗传信息等数据，构建出能够准确预测药物肝毒性的模型。代谢组学和蛋白质组学研究方法则通过监测药物代谢产物的变化和肝脏蛋白质表达谱的改变，来评估药物的肝毒性风险，为 DILI 的预测提供更为全面的信息。

hERG 心脏毒性是指药物阻断 hERG 钾离子通道所引起的心脏毒性反应，是药物研发中需要重点关注的安全性问题[74]。hERG 钾离子通道在心脏动作电位的复极化过程中起着关键作用，其被抑制可能导致心律失常、QT 间期延长等严重心脏问题。hERG 心脏毒性的预测主要依赖于计算机模拟预测以及基于机器学习的预测模型。计算机模拟预测通过构建 hERG 钾离子通道的三维结构模型，模拟药物分子与通道的相互作用，评估药物对通道的抑制强度。基于机器学习的预测模型则利用大量已知 hERG 抑制剂的数据，构建出能够准确预测新化合物 hERG 抑制活性的模型，为药物研发提供有力的安全保障。

### 2.2.3 药物-药物相互作用研究

药物-药物相互作用指的是两种或多种药物同时使用时，它们之间可能产生的药效学或药动学上的相互影响，这种影响可能导致药物疗效的改变、毒性增加或不良反应的发生。

药物-药物相互作用通常分为药效学相互作用和药动学相互作用两大类。药效学相互作用是指药物之间在作用机制上的相互影响，如一种药物可能增强或减弱另一种药物的生理效应。这种相互作用可能发生在受体、酶、离子通道等生物靶点层面。药动学相互作用则涉及药物在体内的吸收、分布、代谢和排泄等过程，如一种药物可能改变另一种药物的代谢速率或排泄途径，从而影响其在体内的浓度和持续时间。

药物-药物相互作用的研究方法和技术多种多样，包括临床试验、体外实验、计算模拟和数据分析等。临床试验是评估药物间相互作用最直接、最可靠的方法，但成本高昂且耗时较长。体外实验则通过模拟药物在生物体内的代谢过程，如使用肝细胞、微生物或酶系统，来评估药物之间的相互作用。计算模拟方法，如基于结构的药物设计、QSAR 模型和机器学习算法，通过预测药物分子与生物靶点 (如受体、酶、离子通道) 的结合能力和相互作用模式，来评估潜在的药物-药物相互作用风险。此外，随着大数据和人工智能技术的发展，数据挖掘和机器学习技术也被广泛应用于药物间相互作用的预测和研究中，通过挖掘和分析大量药物使用数据和患者信息，发现潜在的药物-药物相互作用模式和规律。

药物-药物相互作用研究是药物研发与临床治疗中不可或缺的一部分。通过综合运用多种研究方法和技术手段，深入研究药物-药物相互作用的分类、机制、影响因素和预测方法，可以为药物的安全使用、疗效优化和药物组合方案的制定提供科学依据和技术支持。

### 2.2.4 药物-靶标结合亲和力和结合位点研究

药物-靶标结合亲和力，即药物分子与生物体内特定靶标 (如蛋白质、酶、受体等) 结合的紧密程度，是评估药物疗效的关键因素。高亲和力意味着药物与靶标结合稳定，能够在体内维持有效浓度，从而发挥更强的药效。而结合位点，则是靶标上与药物分子结合的具体位置，其结构和性质决定了药物与靶标的结合模式、选择性和亲和力。

在研究方法上，实验技术如 X 射线晶体学、核磁共振 (NMR) 等，能够直接观察或测量药物与靶标的结合状态和结构变化，为理解药物-靶标相互作用提供直观证据。同时，计算方法如分子对接、分子动力学模拟等，也能预测药物分子与靶标结合位点的相互作用模式、亲和力以及可能的构象变化，为药物设计提供理论指导。此外，数据

挖掘与机器学习技术的应用，能够从大量药物-靶标相互作用数据中挖掘潜在规律，预测新的药物-靶标关系。

然而，药物-靶标结合亲和力和结合位点的研究也面临诸多挑战。靶标的复杂性和多样性、药物分子的多样性以及药物-靶标相互作用的动态性，都增加了研究的难度。因此，需要不断开发新的实验技术和计算方法，提高研究的准确性和效率，以推动药物研发的创新和发展。

## 本章小结

本章详细探讨了药物分子表征和性质研究的多个关键方面。在药物分子表征方法上，我们介绍了文本表示、指纹表示和图形表示三种主要方法。文本表示，特别是 SMILES 和 InChI 等表示法，以其简洁性、唯一性和可逆性等特点，在药物分子结构编码和表示中占据重要地位。指纹表示则将复杂的化学结构信息转化为易于计算机处理的二进制序列，为药物分子相似性比较和数据库搜索提供了便利。而图形表示，包括分子图和分子图像，直观地展示了药物分子的结构骨架和原子间的连接关系，为药物分子的可视化和结构分析提供了有力工具。

在药物多性质研究方面，我们深入分析了药物的理化性质、毒性、药物-药物相互作用以及药物-靶标结合亲和力和结合位点。理化性质的预测涵盖了药物的抑制活性、血脑屏障穿透性、水溶性、脂溶性和自由溶解度等关键方面，这些性质对于药物的生物利用度、药代动力学以及疗效和安全性具有重要影响。药物毒性研究则关注致癌性、致突变性、DILI 以及 hERG 心脏毒性等，这些毒性不仅影响患者的生命健康，也是药物研发过程中必须严格监控的要素。药物-药物相互作用研究揭示了两种或多种药物同时使用时可能产生的药效学或药动学上的相互影响，为药物的安全使用、疗效优化和药物组合方案的制定提供了科学依据。而药物-靶标结合亲和力和结合位点的研究，则深入探讨了药物分子与生物体内特定靶标的结合紧密程度和结合位置，为理解药物作用机制、优化药物设计以及提高药物疗效和降低副作用提供了关键信息。

# 第三章 基于对比学习的药物特征表示方法

## 3.1 引言

近年来，对比学习作为一种前沿的深度学习方法，在图像识别、自然语言处理等领域取得了显著成效[75]。受此启发，为了更透彻地揭示药物与靶标之间的相互作用机制，并加速新药研发的进程，我们提出了一种基于对比学习的药物特征表示方法——DrugDL (图 3.1 所示)。该方法的核心在于精准捕捉药物与靶标之间潜在的、复杂的关系，以期增强药物与靶标特征的表征能力，从而高效地提取药物分子的关键特征。本章将全面而深入地介绍 DrugDL 方法的核心组件，即跨模态交互学习模块和单模态特征强化模块。这两个模块各司其职，协同工作，共同构成了 DrugDL 方法的坚实基石。此外，我们还将通过实验分析，充分验证 DrugDL 方法的有效性和优越性，为新药研发领域带来新的突破和可能。

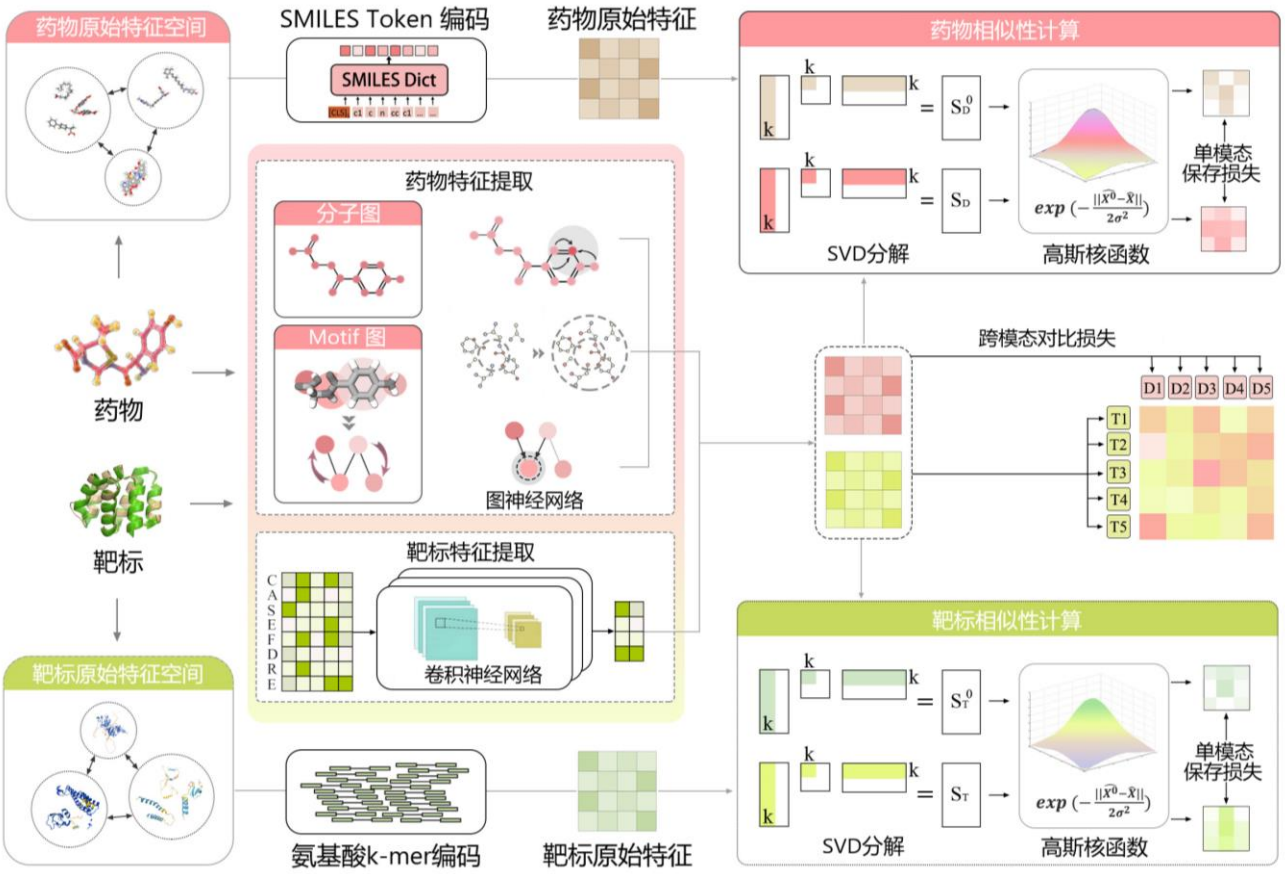



图 3.1 DrugDL 框架图示

Fig. 3.1 Diagram of the DrugDL framework

## 3.2 模型概述

### 3.2.1 跨模态交互学习模块

在本节中，我们将详细介绍药物和靶标的特征提取过程以及基于对比损失的跨模态交互学习过程。对于每个药物分子我们构建了两种类型的图：一种是以原子为节点、化学键为边的药物原子图，另一种是以 motif 为节点、通过多种原则在节点之间构建边的药物 motif 图。这两种图分别从不同角度揭示了药物分子的结构特征和相互作用关系。如 2.1.3 节所述，我们把药物分别表示为原子图$G_a=(V_a,E_a)$，其中，$V_a$代表原子节点集合，$E_a$表示原子之间的化学键组成的边集合。药物分子中包含$n$个原子，每个原子$i(i=1,2,\cdots,n)$在图中对应一个节点$v_i\in V_a$。对于$G_a=(V_a,E_a)$中的每个节点 (即原子)，我们采用一个 78 维的向量来作为其初始特征的表示。这些特征是基于原子的直接属性进行分类的，每个独特的属性均对应于一个特定的特征维度。例如，“是否属于芳香族原子”便是一个重要的特征维度，任何具备这一特征的节点，在特征向量中都会有相应的特殊标识。因此，每个节点的初始特征不仅能够详细反映分子中原子自身的类型，还能揭示其周围环境的特性，包括原子间的连接关系等。这样的设计使得图中的每个节点都能精准地体现原子的局部结构特征，进而在药物原子图中实现对原子节点的精确表示。同时，这种节点初始特征的表示方式具有固定长度，计算效率高，为药物原子图的构建与分析提供了坚实的支撑。

此外，我们又构建了药物 motif 图$G_m=(V_m,E_m)$，其中，$V_m$表示由 motif 组成的节点集合，$E_m$表示节点之间的边的集合。构建药物 motif 图的关键之处在于药物分子的 motif 提取和节点之间的边的构建。首先，我们把每个药物分子拆分为多个分子片段。具体来说，子结构片段大致分为三种类型：环结构，非环状部分以及碳-碳单键。通过查询预训练后的子结构列表，其中的子结构覆盖了我们基准数据集中所有药物，并使每个药物的 motif 拆分具有唯一性和确定性。然后，我们在领域知识的指导下，通过设定多个原则来构建连接节点 (motif) 的边。第一点，当两个 motif 具有共享的原子时，我们将设定这两个motif节点之间有边相连接。第二点，若两个motif没有共享的原子时，则考虑其中一个motif的所有原子中是否存在原子与另一个motif中的原子是相邻原子，若存在则设定这两个 motif 节点之间有边相连接，否则不相连。通过这种方式，我们成功地构建了药物 motif 图$G_m=(V_m,E_m)$，为揭示药物分子之间的复杂相互作用和功能关系提供了有力的工具。而针对于$G_m=(V_m,E_m)$中每个 motif 节点的初始特征我们设置为其包含的所有原子特征的串联。

而对于每一条蛋白质序列$T=a_1,a_2,\cdots,a_n$，其中$a_i$表示第$i$个氨基酸，$n$为序列长度。为充分提取序列中氨基酸相邻信息，我们把每$k$个相邻的氨基酸表示成一个独热编码。例如$k$为 3 且$n$为 5 时，则蛋白质序列$T$表示为$(a_1,a_2,a_3),(a_2,a_3,a_4),(a_3,a_4,a_5)$。本文中每一个蛋白质原始嵌入均以$20^3=8000$种独热编码组合而成。

在获得药物和靶标的嵌入特征后，我们将采取不同的深度学习方法以缩短序列长度并降低计算复杂度和时间，同时保留序列中的原始相关性和全局信息。对于药物分子来说，为了提取药物分子中更深层次的信息，我们使用多通道 GNN 来学习药物的表示。具体来说，除了使用两个并行的图注意力网络 (GAT) 来对药物原子图和 motif 图进行特征提取外，还使用具有共享权重的图卷积网络 (GCN) 来提取两个图的一致性特征。首先，给定原子图$G_a=(V_a,E_a)$作为输入，对于每个原子节点$v_i$，我们计算其与邻居节点$v_j$之间的注意力系数 $\alpha_{ij}$，该计算过程可以公式化为：

$$\alpha_{ij}^{k}=\frac{\exp(LR(e_{ij}))}{\sum_{t\neq i}\exp(LR(e_{it}))}, \tag{3.1}$$

$$e_{ij}=a_k^T[W_k h_i \| W_k h_j], \tag{3.2}$$

其中，$e_{ij}$表示相邻节点间的注意力权重，$a_k^T$，$W_k$是第$k$个注意力机制的可训练的参数，$\|$是串联操作。$LR(\cdot)$表示 LeakyReLU 激活函数。$h_i$，$h_j$分别表示药物原子图中节点$v_i$和节点$v_j$的特征表示，$k$表示 GAT 中注意力机制的头数。在获得节点$v_i$和节点$v_j$之间的第$k$个注意力权重$\alpha_{ij}^k$后，我们便使用以下等式来聚合邻居节点的特征并完成节点更新：

$$h_i^{new}=\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\neq i}\alpha_{ij}^{k}W_k h_j\right), \tag{3.3}$$

其中，$h_i^{new}$表示药物原子图中节点$v_i$更新后的特征表示。$\sigma(\cdot)$表示激活函数。最后我们得到每个药物原子图的特征$H_a$：

$$H_a=BN\left(\sum_{i=1}^{n}h_i^{new}\right), \tag{3.4}$$

其中，$BN(\cdot)$表示批处理归一化层。同理，我们以药物 motif 图$G_m=(V_m,E_m)$作为输入，经过多层 GAT 特征提取后我们得到每个药物 motif 图的特征$H_m$：

$$H_m = BN\left(\sum_{i=1}^{m} \bar{h}_i^{new}\right). \tag{3.5}$$

因此，我们通过两个并行的 GAT 对药物原子图和 motif 图进行特征提取后，分别获得相应的特征表示矩阵 $H_a$ 和 $H_m$。为了利用不同图结构中学习到的知识来改进对其他图结构的理解，并且确保从原子图和 motif 图中提取的特征在语义上是相关的，我们又建立了共享权重的 GCN 模型来提取两种图结构上的一致性的特征表示。

对于给定药物原子图特征表示矩阵 $H_a$ 和 motif 图特征表示矩阵 $H_m$ 作为输入，我们使用以下等式来聚合邻居节点的特征并完成节点更新：

$$h_i^{cnew} = \sigma\left(\sum_{j\in N_i}\left(\frac{1}{c_i} W_c h_j^{new} + b\right)\right), \tag{3.6}$$

$$\bar{h}_i^{cnew} = \sigma\left(\sum_{j\in \bar{N}_i}\left(\frac{1}{\bar{c}_i} W_{\bar{c}} \bar{h}_j^{new} + \bar{b}\right)\right), \tag{3.7}$$

其中，$h_i^{cnew}$ 和 $\bar{h}_i^{cnew}$ 分别表示药物原子图和 motif 图中节点 $v_i$ 和 $\bar{v}_i$ 更新后的特征表示，$W_c$，$W_{\bar{c}}$，$b$，$\bar{b}$ 分别表示 GCN 中可训练的参数矩阵和偏置，$c_i$，$\bar{c}_i$ 是归一化常数。由此，我们得到每个药物原子图和 motif 图在公共空间的特征表示 $H_a^c$ 和 $H_m^c$：

$$H_a^c = BN\left(\sum_{i=1}^{n} h_i^{cnew}\right), \tag{3.8}$$

$$H_m^c = BN\left(\sum_{i=1}^{m} \bar{h}_i^{cnew}\right). \tag{3.9}$$

最后，我们将两种图在公共空间的特征表示的平均值作为药物分子的公共特征 $H_{am}$：

$$H_{am} = \frac{H_a^c + H_m^c}{2}. \tag{3.10}$$

而对于每个蛋白质，为了充分提取序列中的相邻信息，我们使用卷积神经网络 (CNN) 来进行蛋白质序列的特征提取。我们在每个卷积层中使用多个卷积核来学习该区域的嵌入。对于每个序列，使用卷积核进行卷积计算，每个卷积核负责提取序列中特定段的信息。计算表达式如下：

$$X_i' = W_i X + b_i, X' = [X_1' \| X_2' \| \cdots \| X_N'], \tag{3.11}$$

其中，$X$ 表示给定序列的原始嵌入表示；$W_i \in \mathbb{R}^{3\times L}$，$b_i \in \mathbb{R}^1$ 分别表示第 $i$ 个卷积核的权重和偏置；$X' \in \mathbb{R}^{L-2\times N}$ 表示经过 $N$ 个卷积核的卷积处理后序列的表示；‖表示特征拼接。在之后我们使用两层相同的卷积层，充分提取序列区域嵌入与其相邻嵌入间的信息。使用 $N$ 个的卷积核进行卷积计算，提取序列中特定的区域嵌入与其左右邻居间的交互信息，计算表达式如下：

$$X_i'' = \sum_{k=1}^{N}\left(W_{k,i}\sigma\left(X_i'\right)+b_{k,i}\right), X'' = \left[X_1''\| X_2''\|\cdots\| X_N''\right], \tag{3.12}$$

其中，$X_i'$ 表示 $X'$ 的第 $i$ 个通道，$W_{k,i} \in \mathbb{R}^{3\times 1}$，$b_{k,i} \in \mathbb{R}^1$ 分别表示第 $i$ 个卷积核中第 $k$ 个通道的权重和偏置；$\sigma(\cdot)$ 表示 RELU 非线性激活函数；$X'' \in \mathbb{R}^{l-2\times N}$ 表示经过 1 个卷积层的卷积处理后序列的表示。

受残差网络的启发，我们连接了一个由池化层和两个卷积层组成的特征聚合模块。我们将池化层设置为“最大池化”的非线性池化函数，这样每次池化后序列特征向量的序列长度都会减半。之后，连接的卷积层相当于对非线性函数作用结果的线性加权，这强化了池化层在减少信息冗余方面的作用，同时也减少了池化造成的信息损失。最后，我们将池化结果与卷积结果进行拼接，计算表达式如下：

$$X^{(t+1)} = P(X^t) + \theta(P(X^t)), \tag{3.13}$$

其中，$X^0 = X'''$，$X''' \in \mathbb{R}^{(l-2)\times N}$ 表示经过 2 个卷积层的卷积处理后序列的表示; $P(\cdot)$ 为最大池化函数；$\theta(\cdot)$ 为两层卷积层卷积计算。

接下来我们将利用对比学习损失来对齐药物和靶标两种模态的特征。我们假设有 $N$ 对药物-靶标数据，其特征表示为 $(D_i, T_i)$，其中 $D_i = \text{Concat}(H_a, H_m, H_{am})$，当药物 $i$ 和靶标 $i$ 存在相互作用时，我们视其为正样本，否则视为负样本。对比损失函数可以公式化为：

$$L_{\text{cot}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(t\cdot\cos(D_i, T_i^+))}{\sum_{j\in P(i)}\exp(t\cdot\cos(D_i, T_j))}, \tag{3.14}$$

其中，$D_i$，$T_i^+$ 分别表示第 $i$ 个正样本对中的药物特征向量和靶标特征向量，$\cos(D_i, T_i^+)$ 表示药物特征向量 $D_i$ 和靶标特征向量 $T_i^+$ 之间的余弦相似性。$t$ 是温度参数，用于调整损失函数的敏感度，$P(i)$ 表示与第 $i$ 个药物特征向量 $D_i$ 配对的所有靶标特征向量的集合，包括正样本 $T_i^+$ 和所有负样本 $T_i^-(j \neq i)$。

### 3.2.2 单模态特征强化模块

在单模态特征强化模块中，我们采用独特的相似性计算方法和损失函数来保存每个模态中存在的内在关系。对于特征提取后的药物特征矩阵和靶标矩阵，我们使用矩阵奇异值 (SVD) 分解提取出特征矩阵中的主成分，有效剔除冗余信息，同时精准保留关键信息。然后我们引入高斯核函数测量并计算在嵌入特征空间中不同药物之间以及不同靶标之间的相似性。同时，在药物和靶标原始特征空间中，我们进行归一化后，同样使用 SVD 分解进行信息提取，使用高斯核函数计算各模态数据的内在相似性。对于每个模态数据在原始特征空间的相似性和嵌入特征空间的相似性，我们采用相似性矩阵的一致性损失进行测量，并加和两种模态数据的损失值得到最终的单模态保持损失。我们希望通过单模态特征强化过程，有效避免在数据资源相对有限的情况下，因过度追求模态间的对齐而忽略了各模态内部特征分布的复杂性。因此，我们按如下方式设计单模态的保存损失：

$$L_{\text{pre}} = L_{\cos}(S_D^0, S_D) + L_{\cos}(S_T^0, S_T), \tag{3.15}$$

其中，$S^0$，$S$ 分别表示单模态数据在原始特征空间和嵌入特征空间中的相似性矩阵，$S = \exp\left(-\frac{\| \hat{X}_i - \hat{X}_j \|^2}{2\sigma^2}\right)$，$\hat{X}_i$ 和 $\hat{X}_j$ 分别代表不同模态数据在特征空间经 SVD 分解后，重构矩阵中第 $i$ 个与第 $j$ 个数据所对应的特征。$L_{\cos}(\cdot)$ 表示相似性矩阵的一致性损失：

$$L_{\cos}(S_1, S_2) = \frac{1}{N}\sum_{i=1}^{N}\left|1 - \sum_{j=1}^{N} S_1(i,j)\cdot S_2(i,j)\right|. \tag{3.16}$$

### 3.2.3 联合优化与输出模块

如前文所述，为了高效地捕获药物与靶标的特征表征，我们提出的框架融合了两大核心组件：跨模态交互学习模块与单模态特征强化模块。这两个模块协同工作，旨在深化对药物与靶标相互作用的理解。接下来，我们通过整合两个模块的损失目标来实现联合优化，并据此预测输出结果。具体而言，我们的模型损失函数完整形式如下：

$$L = \alpha L_{\cos}(S_D^0, S_D) + \beta L_{\cos}(S_T^0, S_T) + \gamma L_{\cot}, \tag{3.17}$$

其中，系数 $\alpha$，$\beta$ 和 $\gamma$ 作为权重，用于平衡不同损失项的重要性，$\alpha + \beta + \gamma = 1$。同时，我们将输出药物和靶标在特征空间中的嵌入表示 $D_i$, $T_i$，并用于下游各任务的学习。

## 3.3 实验分析

### 3.3.1 数据集预处理及实验流程设计

为了训练和评估我们的方法，我们精心收集和整理了多个公开的药物-靶标相互作用数据集，涵盖了超过 20000 个药物和 7000 多个蛋白质，累积了近 100000 个药物靶标相互作用数据。这些数据主要源自三个知名的公开数据集：BindingDB[76]、BioSNAP[59]和 Human[77]。BindingDB 是一个可在线访问的数据库，专注于小分子药物样分子与蛋白质之间的相互作用，并提供经过实验验证的结合亲和力数据。其数据源可访问于 https://www.bindingdb.org/bind/index.jsp。BioSNAP 数据集则是从 DrugBank 数据库中精心提取并处理得到的，它包含了丰富的药物和蛋白质相互作用信息。该数据集的资源可在 https://github.com/kexinhuang12345/MolTrans/tree/master/dataset/BIOSNAP/full_data 找到。Human 数据集则是一个专门针对人类药物-靶标相互作用研究的数据集。其数据源位于 https://github.com/lifanchen-simm/transformerCPI/blob/master/Human%2CC.elegans/dataset/human_data.txt。在实验设计中，为了提升模型的泛化能力，我们采取了特定的样本处理策略：对正样本实施了上采样，同时对负样本进行了下采样，以确保经过上采样后的正样本数量达到一致。随后，我们遵循严格的标准对数据集进行了划分，并据此开展实验。具体来说，我们抽取 80%的正负样本对作为训练集，而将剩余的 20%样本对保留作为测试集，用以评估模型的预测性能。并且在数据集划分时，我们采用了随机拆分和冷启动拆分两种方式。随机拆分意味着数据集依据预设比例被随机地分割为训练集与测试集；而冷启动拆分则在维持拆分比例的同时，确保测试集中的药物与蛋白质在训练集中未曾出现，从而避免模型对已知特征产生过度依赖。这种做法使得模型在面对测试数据时，能够展现出更为真实的预测能力，而不仅仅是基于已学习到的药物、蛋白质特征进行推断。此举旨在确保实验结果的坚实可靠性，并对模型性能进行切实有效的验证。

此外，为了反映出本文提出的特征表示方法的有效性，我们通过将获取的药物、靶标特征进行串联拼接然后输入到全连接神经网络和机器学习分类器中进行药物-靶标是否相互作用进行预测，并将测试结果在四个评估指标进行展示分析，它们分别是准确率 (Accuracy)、受试者工作特征曲线下的面积 (AUC)、特异度 (Specificity) 和召回率 (Recall)。首先，Accuracy 是一个基础且直观的指标，它表示模型正确预测药物与靶标相互作用的样本数与总样本数的比例。Accuracy 的提升意味着模型能够更准确地识别出潜在的药物-靶标相互作用，减少误判的可能性，这对于优化药物设计和筛选具有重要意义。AUC 是衡量二分类模型性能的重要参数。在药物-靶标相互作用预测中，由于

相互作用的发生往往与多种因素相关，且不同药物与靶标的结合特性可能存在差异，因此模型需要在不同分类阈值下保持稳定的预测性能。AUC 值越高，说明模型在不同情况下对药物-靶标相互作用的预测能力越强，这对于药物研发中的决策制定具有关键意义。Specificity 反映了模型对无药物-靶标相互作用样本的准确识别能力，即真实不存在相互作用的样本中被正确判定为“无相互作用”的比例。在药物-靶标相互作用预测中，由于真实相互作用样本可能高度稀缺，提高 Specificity 意味着模型能够更可靠地排除非相互作用对，从而避免将无生物学关联的分子误判为潜在候选 (减少假阳性)，显著降低后续实验验证的无效投入与资源消耗。最后，Recall 衡量了模型在所有真正存在药物-靶标相互作用的样本中成功预测出的比例。高 Recall 意味着模型能够覆盖更多的真实相互作用，减少漏检的可能性，这对于确保药物研发的准确性和完整性至关重要。各指标的公式如下所示：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{3.18}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{3.19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.20}$$

其中， TP (True Positive) 表示真正例，模型正确预测的正样本数，在药物靶标相互作用中即正确预测为存在相互作用的样本数。TN (True Negative) 表示真负例，模型正确预测的负样本数，在药物靶标相互作用中即正确预测为不存在相互作用的样本数。FP (False Positive) 表示假正例，模型错误地将负样本预测为正的数量，在药物靶标相互作用中即错误预测为存在相互作用的负样本数。FN (False Negative) 表示假负例，模型错误地将正样本预测为负的数量，在药物靶标相互作用中即错误预测为不存在相互作用的正样本数。

### 3.3.2 基线方法与结果对比分析

首先，本节将采用 3.2 节的特征表示办法对 3.3.1 节所述数据集中的药物、蛋白质进行特征表示，并通过连接药物-靶标特征到不同的分类器中进行药物-靶标相互作用的预测。我们采用的分类器共计四种，分别是：(1) 全连接神经网络 (Fully connected neural network，FCNN)；(2) 支持向量机 (Support vector machine, SVM)；(3) 逻辑回归 (Logistic regression，LR)；(4) 随机森林 (Random forest，RF)。所有这些算法的重要参数调优细节已详尽地记录在支撑材料中。各分类器的预测效果已在表 3.1 中详细列出。具体而言，FCNN、SVM、LR 和 RF 的平均 AUC 值分别为 0.8315、0.6835、0.7701 和

0.9585。从表中数据可以看出，RF 在 AUC、Accuracy、Recall 和 Specificity 等多个评估指标上均取得了最优表现。因此，从整体性能来看，RF 的表现明显优于其他分类模型。这一结果表明，随机森林分类器与我们的特征表示框架更为契合，能够更好地发挥特征的优势，提高预测的准确性。

接下来，我们深入对比 DrugDL 提取的药物特征与其他五种常见的分子指纹特征，具体包括：(1) Morgan 指纹；(2) ECFP 指纹；(3) PubChem 指纹；(4) MACCS 指纹；(5) Pharmacophore ErG 指纹。这些指纹特征的详细描述请参见 2.1.2 节。在与这些经典指纹特征的对比中，我们的 DrugDL 药物表示方法在药物-靶标相互作用数据集上展现出了卓越的适应性，并在其应用中取得了显著的预测性能 (见图 3.2)。在采用 RF 分类器的情况下，DrugDL 特征表示方法在 AUC 和 Accuracy 指标上均达到了最佳表现。此外，与其他指纹特征相比，DrugDL 模型在 Recall 和 Specificity 指标上也展现出了明显的优势。我们认为，这一优势主要源于药物-靶标相互作用数据集中阳性和阴性样本之间的不平衡。在这种情况下，所有指纹特征 (特别是 MACCS 和 Pharmacophore ErG 等) 在 Specificity 指标上的表现均显著优于 Recall 指标。这表明，在面临类别不平衡时，大部分指纹特征可能会错误地将许多真实存在相互作用的药物靶标对判定为负样本，从而增加了漏检的风险。然而，DrugDL 模型在 Recall 和 Specificity 指标上的良好表现表明，在药物-靶标相互作用预测任务中，DrugDL 所提取的潜在表示能够灵活地适应数据集大小的变化，并有效解决样本不平衡问题，从而显示出强大的稳健性和可靠性。

表 3.1 各分类器在药物-靶标相互作用基准数据集上的性能对比

Table 3.1 Performance comparison of various classifiers on the drug-target interaction benchmark dataset

| Classifiers | AUC | Accuracy | Recall | Specificity |
|---|---|---|---|---|
| FCNN | 0.8315 | 0.8316 | 0.8468 | 0.8161 |
| SVM | 0.6835 | 0.6324 | 0.6435 | 0.6215 |
| Logistics regression | 0.7701 | 0.7022 | 0.7194 | 0.6850 |
| Random forest | **0.9585** | **0.9536** | **0.9902** | **0.9171** |

此外，我们运用了 t 分布随机邻居嵌入 (t-SNE) 降维方法，对测试集上的预测结果进行了可视化分析 (见图 3.3)，旨在直观对比不同特征表示方法在药物-靶标相互作用预测任务中的性能差异。该可视化结果揭示了各类方法在药物-靶标相互作用数据分类上的表现特征。具体而言，五种传统的分子指纹特征，对于正负样本的区分能力并不显著，尤其是 Pharmacophore ErG 指纹，其对应的正负样本在特征空间中几乎完全重叠，这一可视化现象直观反映了该特征表示方法在预测性能上的局限性。相比之下，

DrugDL 的特征表示方法展现出了非凡的区分效能，成功地在特征空间中将正负样本清晰地分割为两个截然不同的聚类区域。这一显著的样本分离现象不仅有力证明了 DrugDL 在特征提取层面的卓越能力，还进一步印证了其能够引导学习模型生成更为精确的药物嵌入表示，从而极大提升了药物-靶标相互作用预测的准确性。因此，我们有理由认为 DrugDL 作为一种多功能、可靠且值得信赖的特征表示方法，在药物-靶标相互作用预测领域具有显著的优势和应用前景。

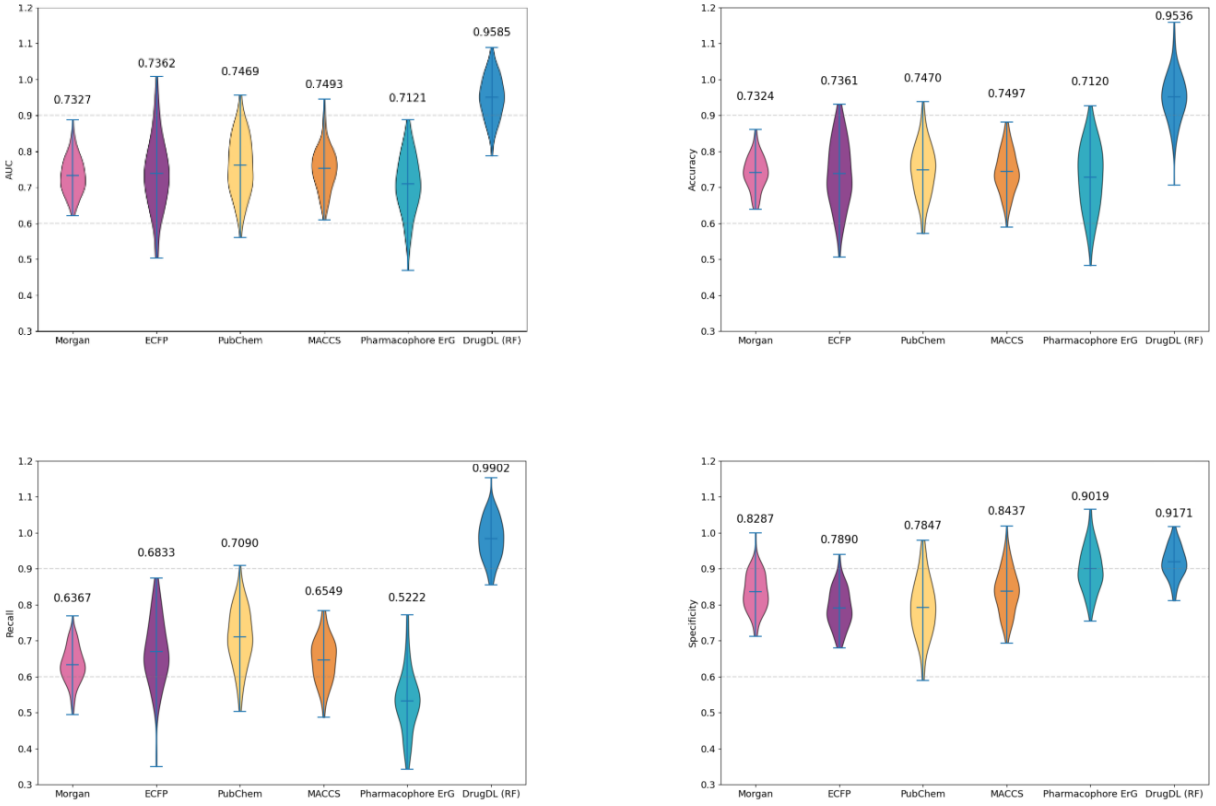



图 3.2 各分子指纹特征在药物-靶标相互作用数据集上的性能对比

Fig. 3.2 Performance comparison of various molecular fingerprint features on the drug-target interaction dataset

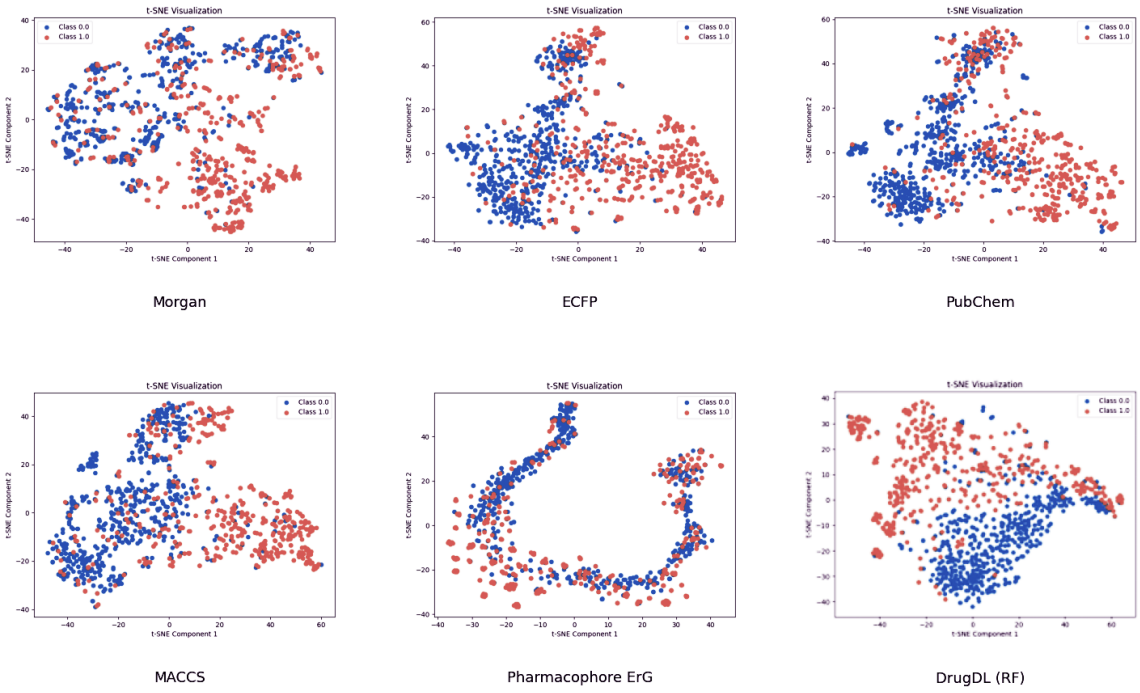



图 3.3 各分子指纹特征的 t-SNE 可视化结果对比

Fig. 3.3 Comparison of t-SNE visualization results of various molecular fingerprint features

最后，我们将 DrugDL 与最先进的药物-靶标相互作用预测模型进行比较分析。这些模型是目前可用于解决药物-靶标相互作用问题或任务的最先进和最前沿的技术的缩影。我们考虑的方法如下：(1) DrugBAN[78]；(2) ZeroBind[79]；(3) PSICHIC[80]。各模型在随机拆分数据和冷启动拆分数据上的详细评估结果如表 3.2 所示。在针对药物-靶标相互作用的数据集上，DrugDL 展现出了卓越的性能，不仅在随机拆分数据上，而且在更具挑战性的冷启动拆分数据上，均显著优于所有基线模型。在随机拆分数据的评估中，DrugDL 的 AUC 值高达 0.9585，相较于其他基线模型如 DrugBAN (0.9123)、ZeroBind (0.9348) 和 PSICHIC (0.9396)，展现出了明显的优势。同时，DrugDL 在 Accuracy、Recall 和 Specificity 等关键评价指标上也均取得了最优成绩，分别为 0.9536、0.9902 和 0.9171，进一步验证了其出色的预测能力和稳定性。在冷启动拆分数据的评估中，尽管所有模型的性能均有所下降，但 DrugDL 依然保持了相对领先的优势。其 AUC 值为 0.8891，远高于 DrugBAN (0.8296)、ZeroBind (0.7845) 和 PSICHIC (0.8034)。此外，DrugDL 在 Accuracy、Recall 和 Specificity 等指标上也均表现优异，分别为 0.8506、0.8787 和 0.8225，充分证明了其在处理冷启动问题上的强大能力。

表 3.2 DrugDL 与基线模型在药物-靶标相互作用数据上的性能对比

Table 3.2 Performance comparison between DrugDL and baseline models on drug-target interaction dataset

| 数据类型 | 模型 | AUC | Accuracy | Recall | Specificity |
|---|---|---|---|---|---|
| 随机拆分 | DrugBAN | 0.9123 | 0.8757 | 0.8462 | 0.9052 |
| | ZeroBind | 0.9348 | 0.9294 | 0.9511 | 0.9077 |
| | PSICHIC | 0.9396 | 0.9341 | 0.9563 | 0.9019 |
| | DrugDL (RF) | **0.9585** | **0.9536** | **0.9902** | **0.9171** |
| 冷启动拆分 | DrugBAN | 0.8296 | 0.8054 | 0.8101 | 0.8007 |
| | ZeroBind | 0.7845 | 0.7730 | 0.7912 | 0.7548 |
| | PSICHIC | 0.8034 | 0.8120 | 0.8312 | 0.7928 |
| | DrugDL (RF) | **0.8891** | **0.8506** | **0.8787** | **0.8225** |

### 3.3.3 方法性能优势解析

(1) 消融实验

本节中，我们进行了消融实验，以研究不同的特征提取方法 (GNN、CNN)、跨模态交互学习模块和单模态特征强化模块对 DrugDL 性能的影响。结果如图 3.4 所示。为了验证特征提取方法的有效性，我们分别尝试了包含 GAT、GCN 等和 CNN、MLP、FCNN 等在内的 12 种特征提取方法。如图 3.4 (a)、(b)、(c) 所示，结果表明在多通道 GNN 中 GAT 和 GCN 的联合使用是提取药物特征最佳的组合方式，而 CNN 是提取蛋白质特征最有效的方法。为了验证跨模态交互学习模块和单模态特征强化模块的有效性，我们分别剔除对比学习损失$L_{cot}$和单模态保存损失$L_{pre}$来生成两种变体：只保留跨模态交互学习模块的 DrugDL-C 以及只保留单模态特征强化模块的 DrugDL-P。图 3.4 (d) 显示，从 DrugDL 中删除任何一个模块均会导致其在药物-靶标相互作用中的预测能力显著降低。这些发现表明，DrugDL 中的各种模块对预测性能的贡献不同，只有将所有的模块并行使用时模型才能够在所有评估指标上获得最佳的预测性能。

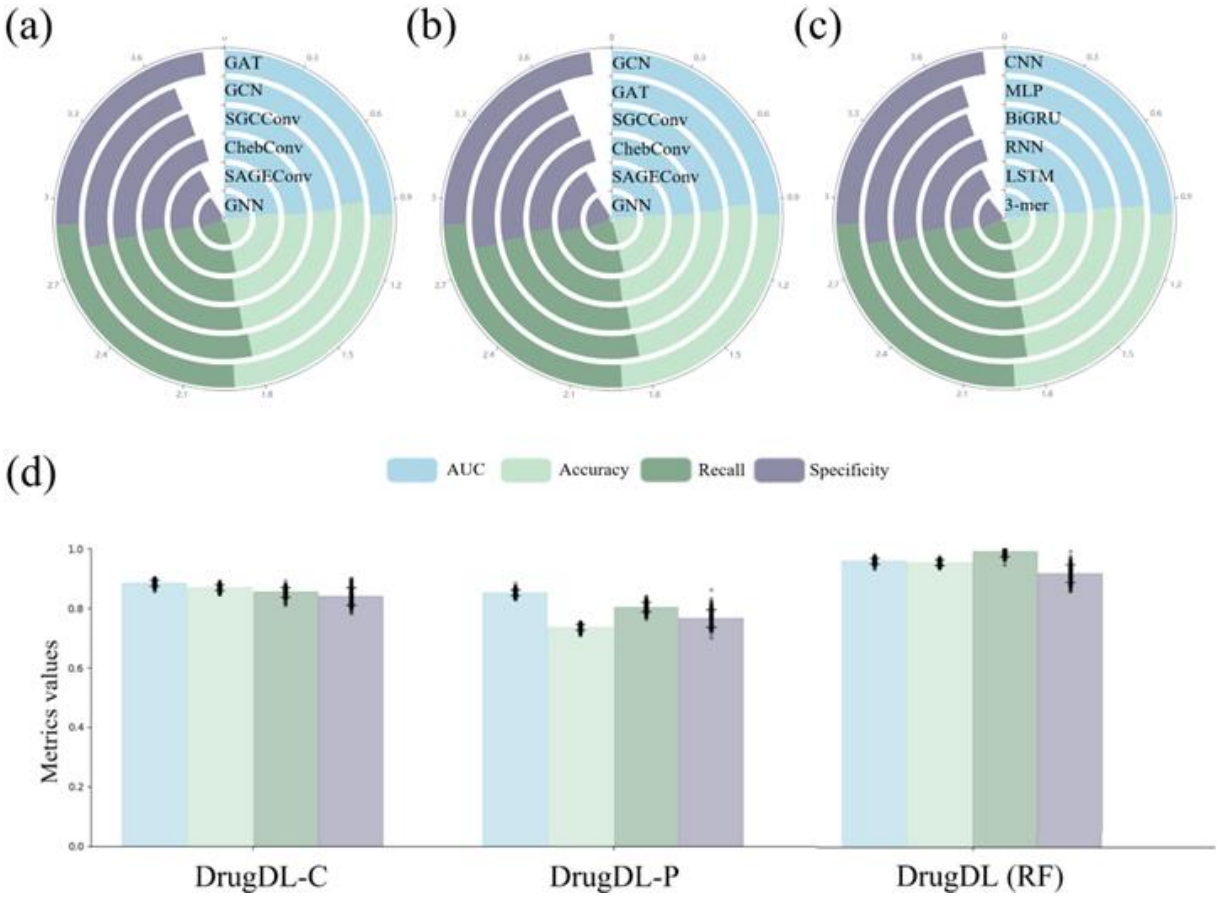



图 3.4 消融实验结果。(a) 各种 GNN 在药物图上的性能；(b) 各种 GNN 在共享权重药物特征上的性能；(c) 各种深度学习方法在靶标特征提取上的性能；(d) DrugDL 各模块的性能分析

Fig. 3.4 Results of the ablation experiments. (a) Performance of various GNNs on drug graphs; (b) Performance of various GNNs on drug features with shared weights; (c) Performance of various deep learning methods in target feature extraction; (d) Performance analysis of each module of DrugDL

(2) 个案研究

为了全面评估 DrugDL 模型在实际应用场景中的可靠性，我们精心挑选了多个与肿瘤密切相关的靶点进行深入案例研究，这些靶点包括 BRAF[81] (B-Raf 原癌基因，一种丝氨酸/苏氨酸激酶)、ALK[82] (间变性淋巴瘤激酶)、EGFR[83] (表皮生长因子受体，属于受体酪氨酸激酶家族)、CD47[84] (一种关键的跨膜蛋白) 以及 PLK1[85] (细胞周期调控的重要蛋白)。针对每一个靶点，我们首先利用基准数据集对 DrugDL 进行训练，随后运用随机森林分类器来预测可能与这些靶点发生相互作用的药物，并根据预测出的相互作用概率对候选药物进行排序。

在图 3.5 中，我们特别展示了针对每个靶点预测出具有较高相互作用概率的几种药物。值得注意的是，在预测与 BRAF 靶点可能相互作用的药物列表中，已经有两种药

物——Vemurafenib 和 Dabrafenib，被批准用于治疗相关癌症，如黑色素瘤[86]，这初步验证了我们的预测方法的有效性[87]。此外，尽管尚未有直接的临床试验证据，但我们的预测指出 Erlotinib、Osimertinib 可能与 BRAF 靶点存在显著的相互作用。值得一提的是，这两种药物在多项研究中已展现出潜在的抗癌活性，并得到了临床试验的支持。这些发现提示，DrugDL 预测出的与特定靶点相互作用的新药物，有望成为进一步探索癌症治疗的有前景的候选药物。

此外，在针对 PLK1 靶点的预测中，我们发现 Fenretinide 正处于 T 细胞非霍奇金淋巴瘤临床试验的 I 期阶段[88]，而 Cromolyn 和 Icosapent 则处于乳腺癌[89]和前列腺癌临床试验的 II 期[90]。对于 CD47 靶点，我们预测出的药物 Fluvastain 已进入治疗恶性黑色素瘤临床试验的早期 I 期阶段[91]。这些个案研究结果充分展示了 DrugDL 在识别肿瘤相关靶点的潜在治疗候选药物方面的强大能力。

这些发现不仅为抗癌药物的持续发现和开发提供了有力支持，还为肿瘤学领域的研究和实验开辟了新的道路。通过提供对药物-靶标相互作用的深入见解，DrugDL 有望推动癌症患者精准医疗和个体化治疗方法的革新与进步。

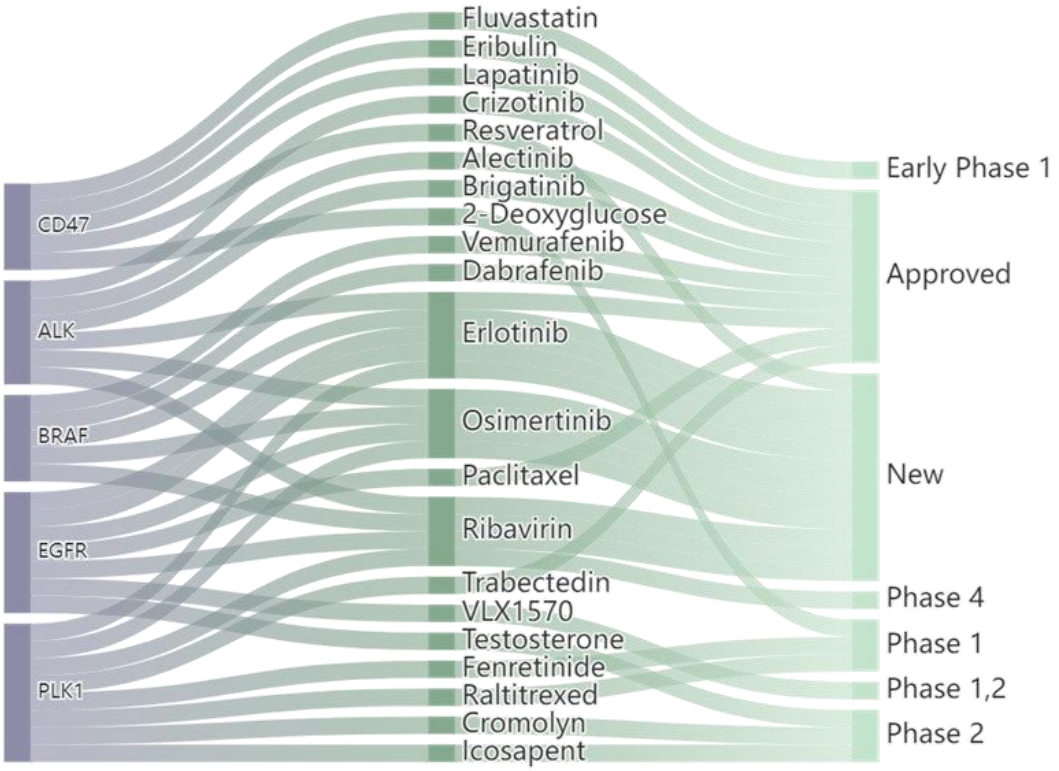



图 3.5 多肿瘤相关靶点个案研究结果

Fig. 3.5 Results of the case studies on multiple tumor-related targets

(3) 可解释性分析

为了深入阐释 DrugDL 模型的可解释性能力，我们挑选了药物-靶标相互作用数据集中与 PLK1 靶点存在相互作用的五个分子 (Fenretinide、Cromolyn、Trabectedin、Icosapent 和 Raltitrexed) 进行细致研究。在数据读出阶段，我们运用自我注意力池化技术，以直观展现原子和 motif 的注意力权重分布。图 3.6 描绘了这些分子的三维图形结构，以及与之对应的原子和 motif 注意力权重大小。其中，我们特别以绿色高亮标记了那些注意力权重较大的亚结构。这些高亮部分有助于我们深入分析和理解药物与靶标之间复杂而微妙的相互作用实例，促进了我们对药物分子中关键官能团的认识与把握。通过观察这些可视化结果，我们能够迅速锁定那些对药物-靶标相互作用起决定性作用的重要官能团。这一过程进一步强调了药物的 motif 信息在高性能药物-靶标相互作用预测中的不可或缺性，从而有力验证了我们所提出的分子表征方法的科学性和有效性。

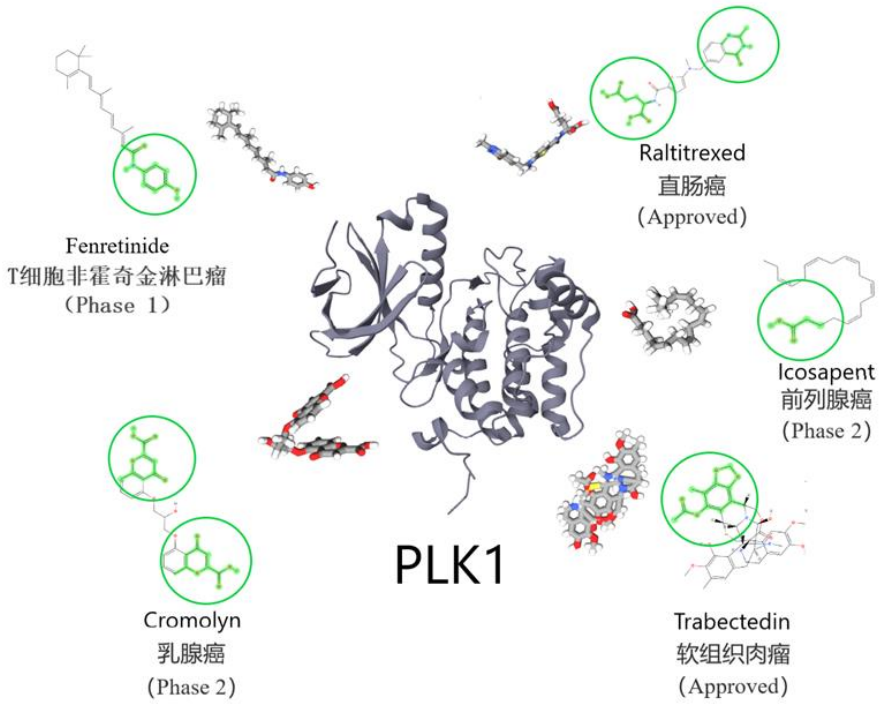



图 3.6 与 PLK1 靶点相互作用的药物分子的可视化结果

Fig. 3.6 Visualization results of drug molecules interacting with the PLK1 target

虽然，上述结果表明我们的模型能够为研究人员提供宝贵的机理见解，助力他们揭示与药物作用相关的更多信息。然而，我们也必须指出，尽管我们的模型在可解释性方面展现出了巨大潜力，但其识别的特殊官能团仅对模型认定药物-靶标相互作用的

相互作用起重要的作用，对于真实的作用关系和更多的细节仍需继续探索和湿实验的证实。

## 本章小结

本章提出了一种基于对比学习的药物特征表示方法 DrugDL，该方法基于对比学习的思想，融合跨模态交互学习模块和单模态特征强化模块，以全面捕捉药物与靶标之间的相互作用关系。通过实验分析，我们验证了 DrugDL 在药物-靶标相互作用预测任务中的卓越性能。实验结果显示，DrugDL 在多个关键评估指标上均显著超越了当前最先进的药物-靶标相互作用预测基线模型，并且在分子表征能力方面，也明显优于传统的分子指纹特征方法。在个案研究中 DrugDL 成功预测了与特定靶点具有相互作用潜力的新药物候选分子，这些发现不仅为抗癌药物的持续发现和开发提供了有力支持，还为肿瘤学领域的研究和实验开辟了新的道路。此外，消融实验和可解释性分析深入探讨了不同组件对模型性能的影响，并提取与特定靶点相互作用的药物分子亚结构。这些分析为我们进一步理解药物与靶标之间的相互作用机制提供了重要线索。

# 第四章 基于深度学习的药物分子性质预测方法

## 4.1 引言

在药物研发领域，药物特征表示对性质预测与设计至关重要。第三章中，我们提出的基于对比学习的 DrugDL 方法，为药物分子特征提取提供了新视角，并展现出卓越性能。为进一步挖掘其潜力，本章将深入拓展 DrugDL 的应用。如图 4.1 所示，我们将结合 DrugDL 提取的特征与不同网络组成，构建多性质预测模型，全面预测药物理化性质、毒性、药物-药物相互作用、药物-靶标结合亲和力及结合位点等药物研发中的关键任务。为验证其优越性，我们将与先进基线模型进行对比分析。本章旨在通过 DrugDL 的深入应用和多性质模型构建，探索其在药物性质预测方面的潜力和优势，为药物研发提供准确高效的工具。


图 4.1 DrugDL 应用工作流程示意图

Fig. 4.1 Schematic diagram of the application workflow of DrugDL

## 4.2 模型概述

### 4.2.1 网络组成

在药物分子性质的研究中，预测任务的多样性要求我们构建的网络模型必须具备高度的灵活性和准确性。本问研究的药物分子性质包含多个方面 (药物理化性质，毒性，药物间的相互作用以及药物靶标的结合亲和力和结合位点等)，这些性质不仅复杂多变，而且涉及了分类、回归等多个任务类型。因此，对于不同的预测任务我们将采取不同的网络组成。首先在分类任务上，我们主要关注药物分子是否具有某种特定的性质或属于某个特定的类别。为了应对这一挑战，我们延续了 3.3.2 节中提到的基分类器，这些基分类器包括全连接神经网络、支持向量机、逻辑回归和随机森林。这些分类器各具特色，能够捕捉药物分子性质中的不同特征。

与分类任务相比，回归任务的预测更为复杂，因为回归任务需要预测连续值的输出。为了应对这一挑战，在全连接神经网络进行多输入数据处理时，我们引入了交叉注意力机制 (Cross-Attention Mechanism)。交叉注意力机制的核心思想在于，通过计算不同输入特征之间的相关性，实现对关键特征的加权处理。具体而言，对于给定的多输入数据，我们首先通过嵌入层将其转换为高维特征向量。然后，利用注意力权重矩阵计算不同特征向量之间的相关性得分。该得分反映了特征之间的交互强度，得分越高的特征对预测结果的贡献越大。在得到注意力权重后，我们将其应用于特征向量的加权求和，从而得到融合了多输入特征信息的输出向量。该输出向量随后被传递给全连接层，进行进一步的特征提取和回归预测。交叉注意力机制的引入，不仅提高了模型对多输入数据的处理能力，还有效捕捉了特征之间的交互信息，提升了回归预测的准确性和鲁棒性。交叉注意力机制的公式化描述如下所示：

$$Q = f_Q(B_n(D)); K = f_K(B_n(T)); V = f_V(B_n(T)), \tag{4.1}$$

$$Attention\_energy = Softmax\left(\frac{QK^T}{\sqrt{C/d}}\right), \tag{4.2}$$

$$X_{D,T} = Attention\_energy * V, \tag{4.3}$$

其中，$X_{D,T}$ 是经过交叉注意力机制整合后的特征，$f = W^T x + b$ 是投影函数 (其中 $W$ 和 $b$ 分别是权重和偏差)，$BN(\cdot)$ 表示批量归一化操作，$C$ 和 $d$ 分别是头部的嵌入尺寸和数量。

### 4.2.2 训练损失函数的选择

在分类任务中，全连接神经网络使用的损失函数为交叉熵损失函数 (Cross-Entropy Loss)。交叉熵损失函数是多分类问题中常用的损失函数。它衡量的是两个概率分布之间的差异，其中一个概率分布是真实标签的分布，另一个概率分布是模型预测的输出分布。交叉熵损失函数越小，表示预测分布与真实分布越接近，模型的性能越好。对于二分类问题，其公式为：

$$\text{loss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right], \tag{4.4}$$

其中，$n$ 是样本量，$y_i$ 是第 $i$ 个样本的真实标签，$p_i$ 是第 $i$ 个样本的预测概率。对于多分类问题，交叉熵损失函数的公式可以表示为：

$$\text{loss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{m}\left[y_{i,c} \log(p_{i,c})\right], \tag{4.5}$$

其中，$m$ 是类别数，$y_{i,c}$ 是第 $i$ 个样本的真实标签，$p_{i,c}$ 是第 $i$ 个样本属于类别 $c$ 的预测概率。在多分类问题中，交叉熵损失函数通过计算每个样本在每个类别上的预测概率与真实标签之间的差异，并求和平均，来评估模型的性能。损失值越小，表示预测分布与真实分布越接近，模型的性能越好。

在全连接神经网络和随机森林 (分裂准则) 进行回归任务时，我们选择的损失函数为均方误差 (MSE):

$$\text{loss} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2, \tag{4.6}$$

其中，$p_i$，$y_i$ 分别表示第 $i$ 条数据的预测值和真实值。

## 4.3 实验分析

### 4.3.1 数据集预处理及实验流程设计

为了训练和精准评估我们的方法在药物分子性质预测领域的效能，我们精心策划并整合了多个权威且公开的数据集。在药物分子的理化性质探究方面，我们从 MoleculeNet 数据集[92]中精心筛选了近 10000 种药物分子的理化信息，这些信息涵盖了 BACE、BBBP、ESOL、Lipophilicity 以及 FreeSolv 等关键性质，其中 BACE、BBBP 为二分类任务数据，ESOL、Lipophilicity 和 FreeSolv 为回归任务数据。MoleculeNet 数据集的详细资源可访问于 https://paperswithcode.com/dataset/moleculenet。

针对药物分子的毒性评估，我们首先分别搜集了近10000种药物分子的致癌性和致突变性数据集，这些数据主要源自三个备受信赖的公开数据库：CPDB[93]、CCRIS 和 ISSCAN[94]。CPDB 是一个广受认可的分子致癌性毒理数据库，它记录了多种药物分子在小鼠、仓鼠等动物实验中的致癌性数据，其数据源可通过 https://www.nlm.nih.gov/databases/download/cpdb.html 获取。CCRIS 数据库则是由国家癌症研究所精心打造的，包含了超过 4500 种化学物质的致癌性研究数据，其资源可访问于 https://www.nlm.nih.gov/databases/download/ccris.html。ISSCAN 数据库的数据源则位于 http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7。其次，为了评估模型对药物分子 hERG 心脏毒性的预测效能，我们又构建了一个包含近 20000 种药物分子的数据集，这些数据来源于 ChEMBL[95] (https://www.ebi.ac.uk/chembl/)、PubChem (https://pubchem.ncbi.nlm.nih.gov/docs/bioassays) 以及之前提及的 BindingDB 这三个独立的公共数据库。我们特别选取了生物化学半最大抑制浓度 (IC50) 作为关键指标，该指标在生物学和药物学领域中被广泛采用，用于量化化合物或药物对生物系统的抑制或活性程度。具体而言，IC50 是指在特定条件下，能够抑制生物过程或活性达到 50%的化合物或药物的浓度。在此基础上，我们设定了一个明确的阈值，即 10 $\mu M$ 。根据这一标准，我们将 IC50 值低于 10 $\mu M$ 的化合物归类为 hERG 阻断剂(该药物具有 hERG 心脏毒性)，而 IC50 值等于或高于 10 $\mu M$ 的化合物则被视为 hERG 非阻断剂 (该药物不具有 hERG 心脏毒性)。最后，我们又整理了一个关于 DILI 的数据集。我们分别从美国食品药品监督管理局国家毒理学研究中心 (NCTR) 的肝脏毒性知识库[96] (LTKB) (https://www.fda.gov/science-research/bioinformatics-tools/liver-toxicity-knowledge-base-ltkb)、美国国家糖尿病、消化和肾病研究院 (NIDDK) 的 LiverTox 数据库[97] (https://www.ncbi.nlm.nih.gov/books/NBK547852/) 以及 Hepatox 数据库[98] (http://www.hepatox.org/) 收集了近 10000 种药物分子，并明确了它们是否可能导致肝脏不良反应。

在药物-药物相互作用的研究领域，我们利用 DrugBank 5.1.8 公开数据库[99]中的近 200000 条药物-药物相互作用数据，构建了一个包含四分类的分类数据集，以及一个采用 AUC FC[100] (药时曲线下面积的变化倍数) 来评估肇事者药物对受害者药代动力学 (PK) 影响的回归数据集。我们遵循 MeTDDI[101]的数据预处理标准，对 DrugBank 数据库中标签为 1 和 2 的药物对进行了顺序颠倒，从而获得了标签为 3 和 4 的药物对信息。具体而言，标签 1 代表药物 1 与药物 2 联合使用时能减少药物 1 的代谢，标签 2 则代表增加；标签 3 和 4 则分别代表药物 1 与药物 2 联合使用时对药物 2 代谢的减少和增加。在数据集拆分的过程中，我们不仅采用了随机拆分方法，还引入了单药不可见和双药不可见的数据拆分策略，使模型在完全未见过的单个药物以及药物对上进行预测，帮助

我们更加全面地评估模型的泛化能力。DrugBank 5.1.8 数据库的详细资源可访问于 https://go.drugbank.com/releases/5-1-8，而药物对的 AUC FC 值则可通过 https://github.com/harryscpt/pk-ddip 获取。

在药物-靶标结合亲和力预测方面，我们选择了 Davis[102]和 KIBA[103]这两个数据集进行验证。Davis 数据集由激酶蛋白质及其相关药物分子的解离常数 ($K_d$) 值组成，其数据可访问于 https://tdcommons.ai/multi_pred_tasks/dti/#davis。KIBA 数据集则涵盖了近 500 个靶点和 50000 多种药物，其数据来源可通过 https://github.com/TangSoftwareLab/DrugRepo 获取。此外，我们还使用了 sc-PDB 数据集[104]来探究药物-蛋白质的结合位点，该数据集包含了近 5000 个蛋白质和 6000 多种药物的 16034 个真实结合位点，相关数据可访问于 https://ngdc.cncb.ac.cn/databasecommons/database/id/57。

为深入探究模型在现实世界药物研发与临床应用中的实际效能，我们选取了来自抗 SARS-CoV-2 (严重急性呼吸综合征冠状病毒) 药物研发以及细胞色素 P450 2C9 (CYP2C9)酶抑制剂识别这两大关键领域的数据展开验证。在抗 SARS-CoV-2 相关数据方面，我们着重选用了 3CL 蛋白酶抑制剂数据集。该数据集详细内容可通过访问链接 https://opendata.ncats.nih.gov/covid19/assays 获取 。而 CYP2C9 酶抑制剂数据集获取链接为 https://drive.google.com/file/d/1mBsgGWXYqej5McsLwy1_fs_-VGGQnCro/view?usp=sharing。

在实验设计环节，为了进一步提升模型的泛化能力，我们针对分类任务采取了与 3.3.1 节相似的样本处理策略，即对样本进行了上采样和下采样处理，以确保不同类别的样本数量在采样后保持一致。同时，无论是分类任务还是回归任务，我们都随机抽取了 80%的正负样本对作为训练集，而将剩余的 20%样本对作为测试集，用以全面评估模型的预测性能。

在模型评估指标的选择上，针对分类任务，我们沿用了 3.3 节中的评估体系，具体包括 Accuracy、AUC、Specificity 以及 Recall，这些指标共同用于全面评估模型的预测性能。对于多分类场景，我们进一步采用了各类别的平均 Specificity 和平均 Recall，以提供更细致的性能展示。而针对回归任务，我们挑选了四个关键评估指标：均方根误差 (RMSE)、决定系数 ($R^2$)、平均绝对误差 (MAE) 以及一致性指数 (CI)。其中，RMSE 是衡量模型预测值与真实值之间偏差的直观指标，其值越小，预示着模型的预测结果越贴近真实情况，预测性能自然越优。$R^2$ 则揭示了模型预测值与真实值之间的拟合紧密程度，$R^2$ 值越趋近于 1，表明模型的拟合效果越佳，预测能力越强。此外，在部分任务中，为了与基线模型进行公平比较，我们采用平方相关系数值的变体 $r_m^2$ 进行评估分析。MAE 作为衡量模型预测误差平均水平的指标，与 RMSE 相比，它更侧重于反映

预测误差的整体趋势，能够有效评估模型在整个数据集上的预测稳定性。至于 CI，它专注于评估模型预测值排序与实际值排序的一致性，CI 值越高，意味着模型在捕捉数据间排序关系上的能力越强。各指标的公式如下所示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}, \tag{4.7}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y_i})^2}, \tag{4.8}$$

$$r_m^2=r^2\times\left(1-\sqrt{r^2-r_0^2}\right), \tag{4.9}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right|, \tag{4.10}$$

$$\mathrm{CI}=\frac{1}{Z}\sum_{y_i>y_j}h(p_i-p_j), \tag{4.11}$$

$$h(x)=\begin{cases}1, & x>0\\ 0.5, & x=0\\ 0, & x<0\end{cases}, \tag{4.12}$$

其中，$n$是样本数量，$\hat{y}_i$表示第$i$个样本的预测值，$y_i$表示第$i$个样本的真实值，$r^2$，$r_0^2$分别代表为有截距和无截距的平方相关系数，$Z$是归一化常数，$h(x)$是阶跃函数。

### 4.3.2 基线方法与结果对比分析

(1) 药物理化性质预测结果对比分析

我们采用 4.2 节的网络组成方式将 DrugDL 模型生成的药物分子特征应用于药物理化性质的分子预测中。我们采用了药物理化性质预测模型中最为先进的三种基线模型，分别是：(1) MoleculeNet[92]；(2) HiMol[105]；(3) HimGNN[106]。各模型的预测效果已在表 4.1 中详细列出。总体而言，我们的 DrugDL 模型在所有评估指标上均展现出了优于其他基线模型的性能。

在分类任务的 BACE 和 BBBP 数据集上，DrugDL 在 AUC、Accuracy、Recall 和 Specificity 四个关键指标上均取得了最优成绩。具体而言，在 BACE 数据集上，DrugDL 的 AUC、Accuracy、Recall 和 Specificity 分别为 0.9237、0.9249、0.9034 和 0.9465，均高于其他模型。同样，在 BBBP 数据集上，DrugDL 的这四个指标也均达到了最高值，

分别为0.9474、0.9475、0.9283和0.9665。尽管HiMol和MoleculeNet的预测效果相近，HimGNN在AUC和Accuracy上略优于前两者，但它们的表现均不及DrugDL，这进一步证明了DrugDL在分类任务中的优越性。

表4.1 DrugDL与基线模型在药物分子理化性质数据上的性能对比

Table 4.1 Performance comparison between DrugDL and baseline models on physicochemical property data of drug molecules

| 数据集 | 指标 | MoleculeNet | HiMol | HimGNN | DrugDL |
|---|---|---|---|---|---|
| BACE | AUC | 0.8028 | 0.8402 | 0.8564 | **0.9237** |
| | Accuracy | 0.8078 | 0.8444 | 0.8617 | **0.9249** |
| | Recall | 0.7812 | 0.8123 | 0.8345 | **0.9034** |
| | Specificity | 0.8345 | 0.8765 | 0.8890 | **0.9465** |
| BBBP | AUC | 0.8777 | 0.8654 | 0.9283 | **0.9474** |
| | Accuracy | 0.8678 | 0.8678 | 0.9211 | **0.9475** |
| | Recall | 0.8123 | 0.8234 | 0.8967 | **0.9283** |
| | Specificity | 0.9234 | 0.9123 | 0.9456 | **0.9665** |
| ESOL | RMSE | 1.0687 | 0.9861 | 0.8704 | **0.8025** |
| | $R^2$ | 0.4987 | 0.5876 | 0.7234 | **0.8554** |
| | MAE | 0.8345 | 0.7654 | 0.6789 | **0.5967** |
| | CI | 0.6987 | 0.7456 | 0.8123 | **0.8940** |
| Lipophilicity | RMSE | 0.7122 | 0.7163 | 0.6326 | **0.5927** |
| | $R^2$ | 0.5123 | 0.5032 | 0.5896 | **0.6298** |
| | MAE | 0.5589 | 0.5623 | 0.4987 | **0.5351** |
| | CI | 0.7089 | 0.7012 | 0.7654 | **0.8310** |
| FreeSolv | RMSE | 2.3989 | 2.7310 | 1.9214 | **1.6648** |
| | $R^2$ | 0.4567 | 0.3456 | 0.5678 | **0.7958** |
| | MAE | 1.6789 | 1.9876 | 1.3456 | **1.1094** |
| | CI | 0.7012 | 0.6543 | 0.7654 | **0.8882** |

在回归任务的ESOL、Lipophilicity和FreeSolv数据集上，DrugDL同样表现出色。在ESOL数据集上，DrugDL在RMSE、$R^2$、MAE和CI四个指标上均优于其他模型。这说明DrugDL在预测药物的溶解度方面具有高精度和可靠性。在Lipophilicity数据集上，DrugDL在RMSE、$R^2$、MAE和CI四个指标上均达到了最优水平，其中RMSE和MAE最低，$R^2$和CI最高，这表明DrugDL在预测药物的亲脂性方面也具有显著优势。最后，DrugDL在FreeSolv数据集上取得的最优成绩也验证了DrugDL在预测药物自由能方面的有效性。

(2) 药物毒性预测结果对比分析

在药物毒性预测的任务中，我们引入了三种前沿的基线模型——BAN[107]、MolCLR[108]和 NYAN[109]，并将它们与我们提出的 DrugDL 模型在四个关键的分子毒性数据集上进行了全面且深入的对比。对比结果如表 4.2 所示，清晰地展现了 DrugDL 模型的卓越性能。

表 4.2 DrugDL 与基线模型在药物分子毒性数据上的性能对比

Table 4.2 Performance comparison between DrugDL and baseline models on drug molecular toxicity data

| 数据集 | 模型 | AUC | Accuracy | Recall | Specificity |
|---|---|---|---|---|---|
| 致癌性 | BAN | 0.8732 | 0.7976 | 0.7358 | 0.8475 |
| | MolCLR | 0.8941 | 0.8243 | 0.7898 | 0.8544 |
| | NYAN | 0.9125 | 0.8482 | 0.8188 | 0.8738 |
| | DrugDL | **0.9550** | **0.8929** | **0.9126** | **0.8742** |
| 致突变性 | BAN | 0.8160 | 0.7529 | 0.7608 | 0.7433 |
| | MolCLR | 0.8349 | 0.7760 | 0.7692 | 0.7833 |
| | NYAN | 0.8414 | 0.8080 | 0.8359 | 0.7786 |
| | DrugDL | **0.9344** | **0.8456** | **0.8394** | **0.8525** |
| DILI | BAN | 0.8328 | 0.7360 | 0.7008 | 0.7669 |
| | MolCLR | 0.8955 | 0.8320 | 0.8966 | 0.7324 |
| | NYAN | 0.9116 | 0.8545 | **0.9027** | 0.7786 |
| | DrugDL | **0.9535** | **0.8805** | 0.8506 | **0.9112** |
| hERG 心脏毒性 | BAN | 0.8357 | 0.7600 | 0.7835 | 0.7327 |
| | MolCLR | 0.8739 | 0.8060 | 0.7867 | 0.8224 |
| | NYAN | 0.9198 | 0.8479 | 0.8850 | 0.8037 |
| | DrugDL | **0.9657** | **0.9111** | **0.8886** | **0.9343** |

具体而言，在 AUC、Accuracy 和 Specificity 这三个核心评估指标上，DrugDL 模型在四个数据集上均显著超越了所有基线模型，充分展示了其强大的预测能力和稳定性。这一结果表明，DrugDL 模型在捕捉药物分子毒性特征方面具有显著优势。在 DILI 数据集上，DrugDL 在 Recall 指标上的结果略逊色于部分基线模型 (NYAN、MolCLR)，但显而易见的是这些基线模型是通过将多数数据判定为阳性 (即预测为肝损伤) 来提高 Recall 的。这样的策略虽然能够提高 Recall，但也不可避免地导致了大量的假阳性结果，从而降低了模型的 Specificity 和整体的 Accuracy。相比之下，DrugDL 在平衡 Recall 与 Specificity 方面表现更为稳健，尽管其 Recall 略低，但在减少误报 (假阳性) 方面表现突出，从而提升了模型在实际应用中的可靠性和实用性。因此，DrugDL 在综合性能上具备更高的临床应用价值，尤其在减少不必要的诊断和误判方面，表现更为优异。在医

疗领域，尤其是涉及到 DILI 等药物毒性的诊断时，减少误诊 (假阳性) 通常比提高 Recall 更加重要。

此外，依据 3.3.2 节所阐述的亚结构提取方法，我们使用 DrugDL 模型对药物分子中在毒性判断方面起决定性作用的部分亚结构序列信息进行了精准提取，并将这些信息详细展示于表 4.3 中。表中不仅包含了亚结构的直观图像，还明确标注了它们各自关联的毒性类型以及对应的 SMILES 序列表示。具体而言，模型成功识别出氮原子连接的碳链结构 (SMILES 表示为 CCN(CC)CC(=O)O)，该结构与药物的致癌性呈现出高度的关联性。同样，胍基结构 (SMILES 表示为 CNC(=O)C(C)N) 被模型判定为与 DILI 紧密相关，从药物作用机制角度分析，该结构可能干扰肝细胞内的正常代谢途径、破坏肝细胞膜稳定性等，进而导致肝脏损伤。除此之外，表中还包含其他多种亚结构，如具有特定羟基结构 (SMILES: CC@HO) 也显示出致癌性；CCCCN、CC(N)CC(N)=O 等结构则与 hERG 心脏毒性相关，这些亚结构的识别与分析，为药物毒性评估提供了重要依据。

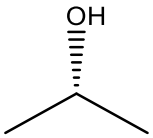
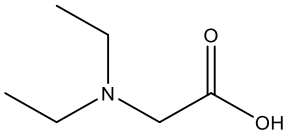
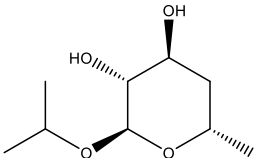
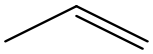
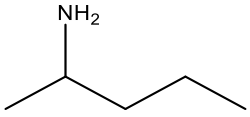
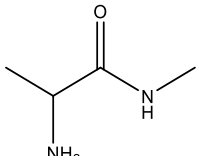
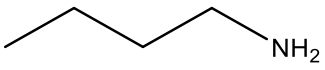
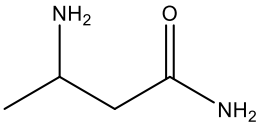
为验证这些识别出的亚结构对药物毒性的影响，我们进行了进一步的实验分析。在去除模型所识别的亚结构后，对药物分子进行了毒性检测，结果如图 4.2 (a) 所示。实验数据清晰地表明，在去除这些特定的亚结构后，包括 Chloroquine、Methylphenidate、Propranolol、Heptaminol、Cyclophosphamide 和 Allopurinol 在内的所有化合物分子的毒性均呈现出显著的降低趋势。这一实验结果不仅为这些亚结构在药物毒性机制中的核心作用提供了直接确凿的证据，也进一步加深了我们对药物毒性产生机制的理解。此外，图 4.2 (b) 直观详细的展示了药物分子中对于 DILI 和 hERG 心脏毒性最为敏感的 12 个亚结构。其中，Pentan-3-ol、Methylhydrazine、N-pentylbutan-1-amine、(S)-2-butanol、N-(2-aminoethyl)-2-aminopropanamide 和 prop-1-en-2-yl 等亚结构被模型精准识别为对 DILI 最为敏感的结构，它们的存在与 DILI 的发生密切相关。而 pentan-1-amine、ethyl butyrate、ethyl ether、2-propanol、2-(dimethylamino) ethyl chloride 和 N-propylamine 等亚结构则被模型识别为对心脏毒性最为敏感的部分，这些结构在药物分子中的存在可能显著增加心脏毒性的风险。

这些发现为药物设计与优化流程提供了关键的指导方向。在药物研发的复杂进程中，药物分子中对致癌性、致突变性、DILI 以 hERG 心脏毒性高度敏感的亚结构，需重点关注。这些亚结构可视为药物分子的“毒性开关”，其存在、缺失或特定状态，直接影响药物的毒性表现。为有效降低药物的潜在毒性风险，可通过合理的分子设计策略与结构优化手段，减少这些敏感亚结构在药物分子中的占比，从源头上降低毒性产

生的可能性；或通过巧妙的化学修饰，改变这些亚结构的电子云分布、空间构象等化学性质，使其失去原有的毒性活性。

表 4.3 DrugDL 提取的药物分子毒性相关亚结构信息

Table 4.3 information of substructures related to drug molecular toxicity extracted by DrugDL

| 亚结构 | 毒性 | SMILES |
|---|---|---|
|  | 致癌性 | C[C@H](C)O |
|  | 致癌性 | CCN(CC)CC(=O)O |
|  | 致突变性 | CC(C)O[C@@H]1O[C@@H](C)C[C@H](O)[C@H]1O |
|  | 致突变性 | C=Cc |
|  | DILI | CCCC(C)N |
|  | DILI | CNC(=O)C(C)N |
|  | hERG 心脏毒性 | CCCCN |
|  | hERG 心脏毒性 | CC(N)CC(N)=O |

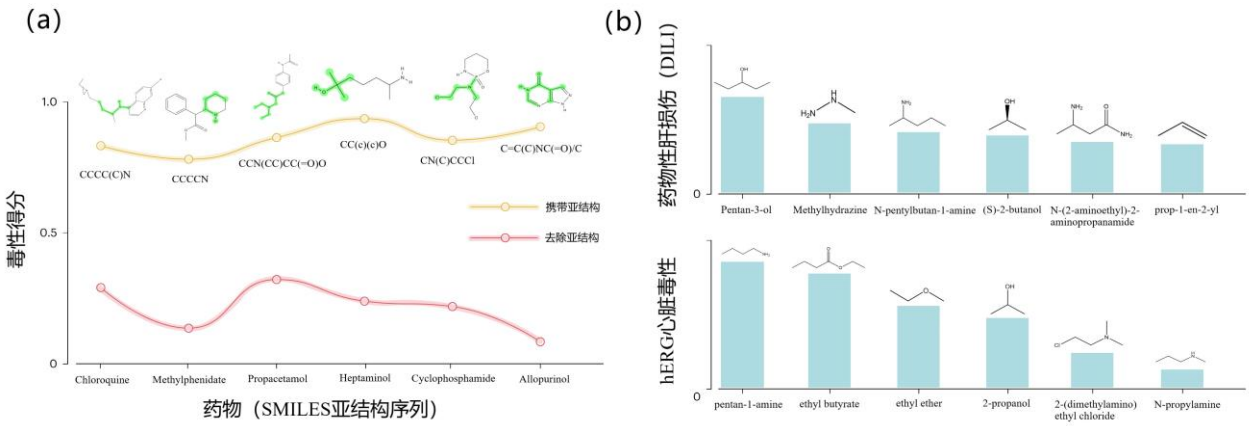



图 4.2 DrugDL 识别亚结构去除前后的毒性对比与敏感区域分析

Fig. 4.2 Toxicity comparison and sensitive region analysis before and after the removal of substructures identified by DrugDL

(3) 药物-药物相互作用预测结果对比分析

在药物-药物相互作用预测的任务中，我们引入了三种先进的基线模型——DeepDDI[110]、Molormer[111]和 MeTDDI[101]，与我们提出的 DrugDL 模型进行了深入对比。通过将药物间相互作用的预测转化为四分类问题 (详见 4.3.1 部分)，我们全面评估了各模型在不同场景下的预测性能。

DrugDL 与基线模型在药物-药物相互作用数据上的性能结果如表 4.4 所示。首先，在随机拆分的数据集中，DrugDL 模型展现出了卓越的性能。在所有评价指标上，包括 AUC、Accuracy、Recall 和 Specificity，DrugDL 均显著优于所有基线模型。具体而言，DrugDL 的 AUC 值高达 0.9969，Accuracy 为 0.9666，Recall 为 0.9595，Specificity 为 0.9708，这些指标均达到了最高水平。这一结果表明，在药物对随机分布的情况下，DrugDL 模型能够更准确地识别药物间的相互作用关系，显示出强大的预测能力和泛化性能。然而，当我们将数据集分为单药不可见和双药不可见时，所有模型的预测性能均出现显著下降。这主要是因为这两类数据集增加了模型的预测难度，要求模型在未见过的药物或药物对上进行预测，即所谓的冷启动问题。尽管如此，DrugDL 模型在这两类数据集上仍然保持了出色的预测性能。在单药不可见数据集中，DrugDL 的 AUC、Accuracy、Recall 和 Specificity 分别为 0.9006、0.7537、0.7140 和 0.7897，均优于基线模型。在双药不可见数据集中，尽管 DrugDL 的预测性能有所下降，但其 AUC、Accuracy、Recall 和 Specificity 仍然保持在较高水平，分别为 0.8051、0.5978、0.5218 和 0.6647。这进一步证明了 DrugDL 模型在处理冷启动问题上的优势。

此外，为了更全面地评估各模型在药物-药物相互作用预测中的性能，我们还考虑了药物-药物 AUC FC 值数据上的预测效果。AUC FC 值是衡量药物间相互作用强度变化的重要指标。如表 4.5 所示，DrugDL 模型在药物-药物 AUC FC 值数据上的预测性能同样优于基线模型。具体而言，DrugDL 的 RMSE 最低，为 0.6838，表明其预测值与真实值之间的偏差最小；$R^2$ 最高，为 0.5233，说明其模型对数据的拟合程度最好；MAE 也最低，为 0.4170，进一步验证了其预测的准确性。同时，DrugDL 较高的 CI 值也表明了其预测结果的可靠性。

表 4.4 DrugDL 与基线模型在药物-药物相互作用数据上的性能对比

Table 4.4 Performance comparison between DrugDL and baseline models on drug-drug interaction data

| 数据类型 | 模型 | AUC | Accuracy | Recall | Specificity |
|---|---|---|---|---|---|
| 随机拆分 | DeepDDI | 0.9711 | 0.8412 | 0.8206 | 0.8624 |
| | Molormer | 0.7362 | 0.8971 | 0.8851 | 0.9092 |
| | MeTDDI | 0.9920 | 0.9231 | 0.9350 | 0.9112 |
| | DrugDL | **0.9969** | **0.9666** | **0.9595** | **0.9708** |
| 单药不可见 | DeepDDI | 0.8811 | 0.6641 | 0.6301 | 0.6982 |
| | Molormer | 0.8867 | 0.6678 | 0.6609 | 0.6756 |
| | MeTDDI | 0.8852 | 0.6807 | 0.6502 | 0.6950 |
| | DrugDL | **0.9006** | **0.7537** | **0.7140** | **0.7897** |
| 双药不可见 | DeepDDI | 0.7811 | 0.5149 | 0.4983 | 0.5398 |
| | Molormer | 0.7775 | 0.4999 | 0.4815 | 0.5207 |
| | MeTDDI | 0.7895 | 0.5362 | 0.5013 | 0.5724 |
| | DrugDL | **0.8051** | **0.5978** | **0.5218** | **0.6647** |

表 4.5 DrugDL 与基线模型在药物-药物 AUC FC 值数据上的性能对比

Table 4.5 performance comparison between DrugDL and baseline models on drug-drug AUC FC value data

| 模型 | RMSE | $R^2$ | MAE | CI |
|---|---|---|---|---|
| DeepDDI | 1.1205 | 0.3210 | 0.8012 | 0.5567 |
| Molormer | 1.0137 | 0.4023 | 0.7234 | 0.6012 |
| MeTDDI | 0.9124 | 0.4876 | 0.6456 | 0.6543 |
| DrugDL | **0.6838** | **0.5233** | **0.4170** | **0.6611** |

为了直观展示 DrugDL 模型在药物-药物相互作用相关数据上的预测性能，我们将预测可视化结果展现在了图 4.3 中。具体来说，我们分别展示了 DrugDL 模型在三个分类任务上的 t-SNE 可视化结果以及在回归任务上的预测结果。在分类任务中，通过观察

t-SNE 可视化结果，我们可以发现 DrugDL 模型在随机拆分的数据集中成功地将 4 类数据分割开来，各类数据之间的界限相对清晰。这表明我们的模型在分类任务上具有较好的性能。同时，我们也关注了不可见数据集上的 t-SNE 可视化结果。尽管模型在这部分数据上的预测结果没有随机拆分数据上的表现出色，但这些数据点也在某种程度上呈现出类似的聚类趋势。在预测 AUC FC 值的回归任务上，我们展示了实际值与预测值的对比结果。通过观察散点图，我们可以明显看出预测值的分布与真实值的分布大致相同，真实值和预测值组成的二维点基本分布在对角线上。这进一步证明了 DrugDL 模型在回归任务上的准确性。此外，图中还展示了 95%预测区间，这为我们提供了预测结果的不确定性范围，有助于更深入地了解模型的预测可靠性。此外，三种分类数据集上的预测分类结果如图 4.4 所示，其混淆矩阵展现出来的分类结果与图 4.3 中的结果相对应。

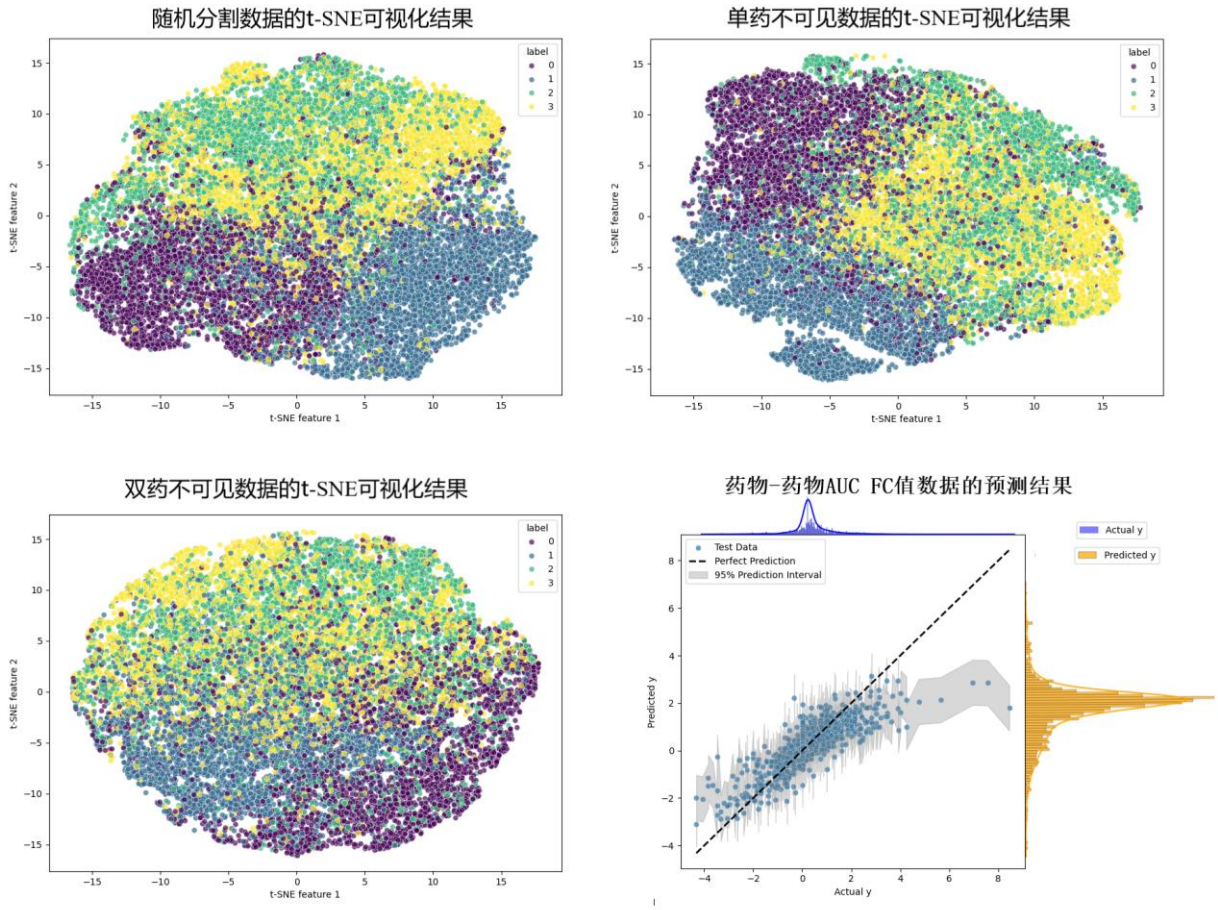



图 4.3 DrugDL 在药物-药物相互作用数据集上的预测可视化结果

Fig. 4.3 Prediction visualization results of DrugDL on the drug-drug interaction dataset

综上所述，无论是在药物-药物相互作用的分类预测上，还是在 AUC FC 值的回归预测上，DrugDL 模型均展现出了优于基线模型的预测性能。这一结果不仅证明了 DrugDL 模型的有效性和准确性，也为其在药物研发、药物相互作用评估等领域的实际应用提供了有力支持。

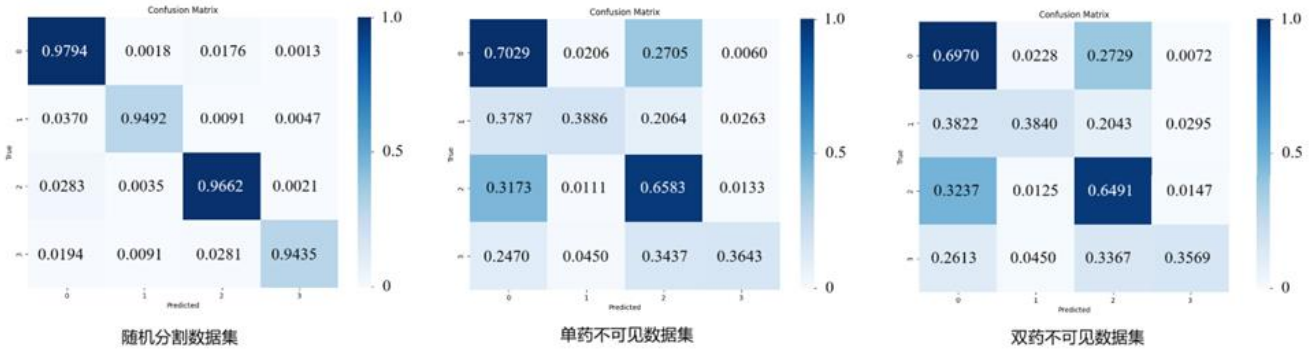



图 4.4 DrugDL 在药物-药物相互作用的三种分类数据集上的预测结果

Fig. 4.4 Prediction results of DrugDL on three classification datasets of drug-drug interactions

(4) 药物-靶标结合亲和力和结合位点预测结果对比分析

在药物-靶标结合亲和力预测的任务中，我们对 DrugDL 模型与 MFR-DTA[62]、KDBNet[112]和 MMD-DTA[113]这三种专为相关任务设计的基线模型进行了深入对比分析。通过在 Davis 和 KIBA 这两个权威数据集上的全面评估，我们详细考察了各模型的预测性能。

表 4.6 清晰地展示了所有模型在两个数据集上的性能结果。值得注意的是，我们提出的 DrugDL 模型在所有的评估指标上均显著超越了基线模型。其 $r_m^2$ 值远高于其他模型，充分证明了预测值与真实值之间的高度相关性。同时，DrugDL 模型的 RMSE 和 MAE 值也达到了所有模型中的最低水平，这再次有力地印证了其在该数据集上的卓越预测精度。此外，DrugDL 的一致性指数 CI 也相对较高，进一步增强了其预测结果可靠性的说服力。

图 4.5 直观地展现了 DrugDL 在两个数据集上的预测可视化结果。在 Davis 数据集中，由于存在不平衡的标签分布，即药物-靶标结合亲和力的实际值分布不均，这在一定程度上给模型在平方误差方面的表现带来了挑战。然而，DrugDL 模型依然凭借其出色的性能，相较于其他基线模型取得了最佳的预测结果。其散点图上的点紧密围绕对角线分布，清晰地表明了预测值与实际值之间较小的偏差。而在 KIBA 数据集中，数据

的标签分布相对正态，即药物-靶标结合亲和力的实际值分布较为均匀，这为模型学习更准确的映射关系提供了有利。

表 4.6 DrugDL 与基线模型在药物-靶标结合亲和力数据上的性能对比

Table 4.6 performance comparison between DrugDL and baseline models on drug-target binding affinity data

| 数据集 | 模型 | RMSE | $r_m^2$ | MAE | CI |
|---|---|---|---|---|---|
| Davis | MFR-DTA | 0.2211 | 0.7053 | 0.1532 | 0.9051 |
| | KDBNet | 0.5656 | 0.3681 | 0.3248 | 0.8471 |
| | MMD-DTA | 0.2201 | 0.7088 | 0.1469 | 0.9051 |
| | DrugDL | **0.2106** | **0.7217** | **0.1351** | **0.9104** |
| KIBA | MFR-DTA | 0.1361 | 0.7892 | 0.0997 | 0.8982 |
| | KDBNet | 0.4550 | 0.3906 | 0.2885 | 0.8379 |
| | MMD-DTA | 0.1341 | 0.7788 | 0.0935 | 0.9001 |
| | DrugDL | **0.1304** | **0.7974** | **0.0892** | **0.9104** |

此外，我们还对 DrugDL 模型在更加精细的药物-靶标结合位点预测任务中的性能进行了评估，并与 TransformerCPI[58]、MFR-DTA 和 MMD-DTA 等基线模型进行了对比分析。为了科学、客观地衡量不同模型的预测能力，我们遵循了标准的评估方法，将实际结合位点落入预测结合区域的概率作为核心评估指标。在具体操作中，我们设定了不同长度的预测结合区域，即 $S$ 个氨基酸组成的序列片段，其中 $S$ 的取值分别为 5、10、15 和 20，以全面评估各模型在不同长度测量方法下的预测性能。

结果如表 4.7 所示，该表详细记录了各模型在不同 $S$ 值下的预测精度。从表中数据可以清晰地看出，随着 $S$ 值的增加，所有方法的预测结合区域准确性均呈现上升趋势。当 $S$ 等于 5 和 10 时，尽管所有基线模型的预测精度相对较低，但 DrugDL 模型依然展现出了其在处理较短序列片段时的显著优势。而当 $S$ 取值为 15 和 20 时，DrugDL 模型的预测性能更是得到了进一步提升，Accuracy 分别高达 0.5396 和 0.5935。这一结果表明，随着预测结合区域长度的增加，DrugDL 模型能够更准确地捕捉药物与靶标之间的相互作用信息，从而显著提高了预测的准确性。与其他基线模型相比，DrugDL 在 $S=15$ 和 $S=20$ 时的预测性能均遥遥领先，这再次有力地验证了其在药物-靶标结合位点预测任务中的卓越有效性和高度可靠性。

为了更直观地展示 DrugDL 预测的药物-靶标结合区域与真实结合位点之间的关系，我们挑选了四个复合物，并对其预测结果进行了可视化处理，如图 4.6 所示。图中，绿色区域代表 DrugDL 预测的药物靶标结合区域 ($S=20$)，而红色则标记了通过实验测量

的真实结合位点。观察发现，对于蛋白质嘌呤核苷磷酸化酶、牛嘌呤核苷磷酸化酶以及二氢乳清酸脱氢酶，其结合位点精确地位于 DrugDL 预测的结合区域之内。尽管对于蛋白质细胞分裂素氧化酶/脱氢酶，预测的结合位点稍有偏移，未完全位于预测区域的正中央，但它仍然落在该结合区域之内。此外，我们还根据药物的 motif 特征对药物分子进行了可视化展示，其中红色区域突显了模型认定的、在药物与蛋白质结合过程中起关键作用的亚结构。

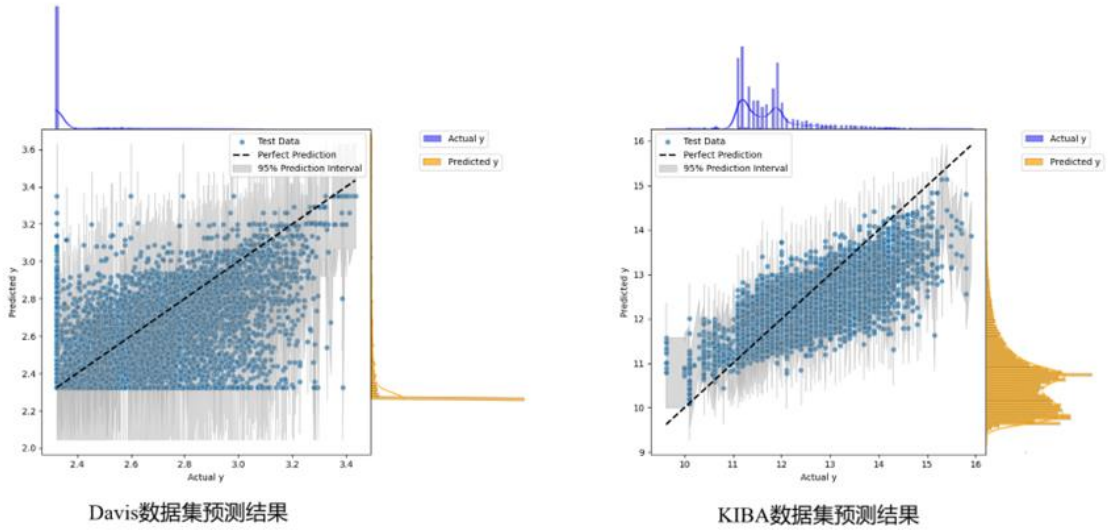



图 4.5 DrugDL 在药物-靶标相互作用数据集上的预测可视化结果

Fig. 4.5 Prediction visualization results of DrugDL on the drug-target interaction dataset

表 4.7 DrugDL 与基线模型在药物-靶标结合位点数据上的性能对比

Table 4.7 performance comparison between DrugDL and baseline models on drug-target binding site data

| 模型 | S=5 | S=10 | S=15 | S=20 |
|---|---|---|---|---|
| TransformerCPI | 0.0146 | 0.1031 | 0.1964 | 0.2008 |
| MFR-DTA | 0.3351 | 0.4221 | 0.5135 | 0.5333 |
| MMD-DTA | 0.3029 | 0.3606 | 0.3927 | 0.4768 |
| DrugDL | **0.3729** | **0.4497** | **0.5396** | **0.5935** |

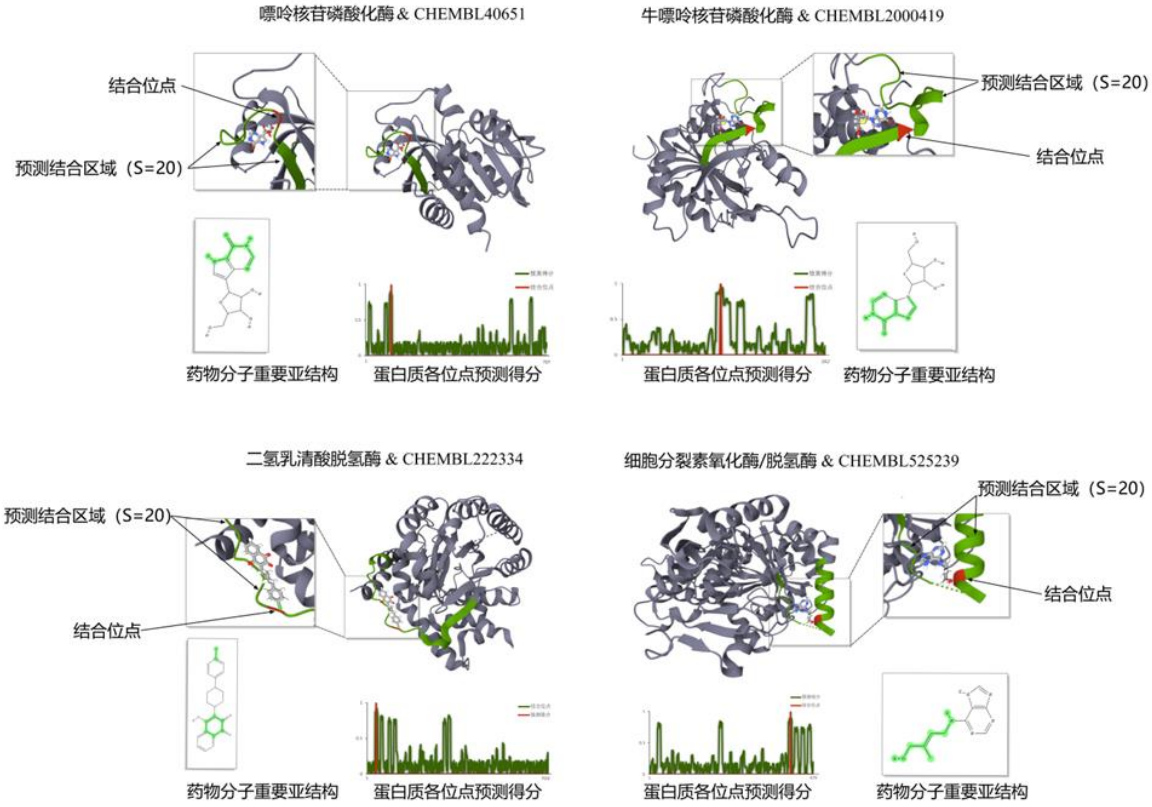



图 4.6 DrugDL 在药物-靶标结合位点预测上的可视化结果

Fig. 4.6 Visualization results of DrugDL on drug-target binding site prediction

### 4.3.3 模型真实应用效能测试

为深入探究 DrugDL 模型在现实世界药物研发与临床应用中的实际效能，我们分别围绕抗 SARS-CoV-2 药物研发以及 CYP2C9 酶抑制剂识别两大关键领域，系统评估 DrugDL 与其他基线模型的表现，以验证 DrugDL 模型在真实应用环境下的潜力。我们采用的三种基线模型分别是：(1) MolCLR；(2) NYAN；(3) ImageMol [38]。

SARS-CoV-2 作为导致 COVID-19 (2019 冠状病毒病) 的病原体，其引发的全球大流行给人类健康与社会经济带来了前所未有的沉重打击。在 SARS-CoV-2 的生命周期进程里，3CL 蛋白酶发挥着极为关键的作用，是极为重要且极具潜力的药物靶点。抑制 3CL 蛋白酶的活性，就好似切断了病毒繁殖的“生产线”，已然成为研发抗 SARS-CoV-2 药物的核心策略之一。因此，我们针对药物分子是否能够与 3CL 蛋白酶结合，以抑制其活性，进而展现出抗 SARS-CoV-2 活性展开了预测分析。相关结果呈现于图

4.7。首先，在 AUC 和 AUPR 这两项指标上，DrugDL 分别达到了 93.24 与 93.23 的分值，与三个基线模型相比优势十分显著。此外，我们运用 t-SNE 方法对 DrugDL 提取的特征进行了可视化分析。从图 4.7 中能够清晰地看到，DrugDL 所提取的药物分子特征实现了对分子是否属于抗 SARS-CoV-2 药物的良好聚类。图中还展示了部分具有抗 SARS-CoV-2 活性药物的结构。这充分表明，DrugDL 所提取的特征当中，蕴含着能够用于判别抗病毒活性的关键特征。

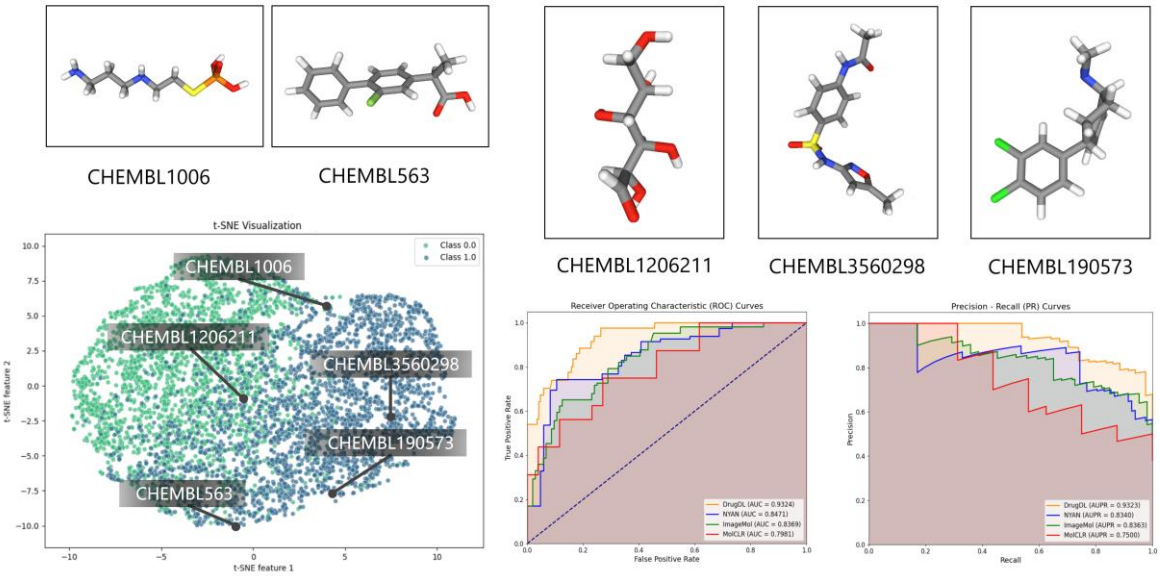



图 4.7 DrugDL 在抗 SARS-CoV-2 抑制剂数据上的评估结果

Fig. 4.7 Evaluation results of DrugDL on anti-SARS-CoV-2 inhibitor data

CYP2C9 参与众多临床常用药物的代谢过程，准确识别 CYP2C9 的抑制剂对于避免药物相互作用、保障用药安全至关重要。下面我们将对药物是否为 CYP2C9 的抑制剂或非抑制剂进行预测分析。结果如图 4.8 所示，在 ROC 曲线评估中，DrugDL 模型的 AUC 值达到了 0.9687，相比之下，其他基线模型的 AUC 值明显低于 DrugDL，这表明 DrugDL 模型在综合区分正样本和负样本方面具有更强的能力，能够更有效地将 CYP2C9 抑制剂和非抑制剂区分开来。在 PR 曲线评估中，DrugDL 的 AUPR 值为 0.9695，同样显著优于各基线模型，意味着在面对样本不均衡的情况下，DrugDL 模型在识别 CYP2C9 抑制剂时，能够保持较高的 Specificity 和 Recall 的平衡。

从上述结果我们可以相信，DrugDL 在现实世界药物研发中具备极大的应用潜力。该模型凭借其卓越的性能，能够深度挖掘并精准捕捉药物分子特征与性质间的潜在关

联。这一优势不仅为临床合理用药提供了坚实的保障，助力医生更科学地选择药物，降低药物相互作用风险；同时，也在新药研发环节中发挥着关键作用，能够有效帮助科研人员规避潜在的药物风险，加速研发进程，为推动创新药物的问世贡献力量。

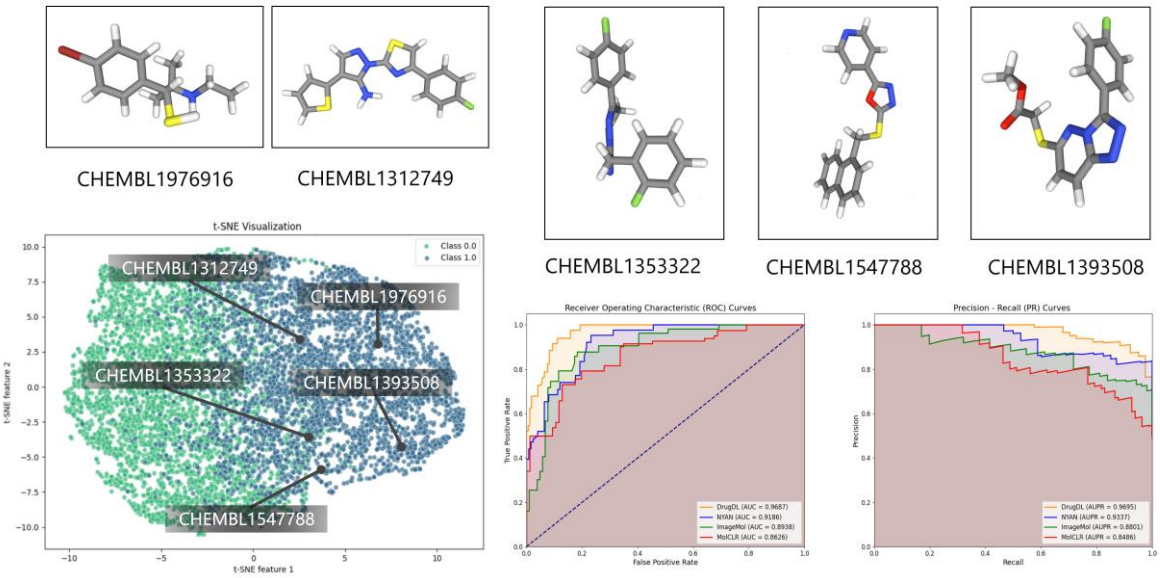



图 4.8 DrugDL 在 CYP2C9 抑制剂数据上的评估结果

Fig. 4.8 Evaluation results of DrugDL on CYP2C9 inhibitor data

## 本章小结

在本章中，我们运用了第三章所提出的基于对比学习的药物特征表示方法 DrugDL，并采用了多样化的网络组成策略，以执行一系列药物分子性质预测任务。通过系统的实验分析，我们充分验证了 DrugDL 在多个关键预测任务中的卓越性能，这些任务涵盖了药物的理化性质、毒性评估、药物-药物相互作用预测、药物-靶标结合亲和力预测以及药物-靶标结合位点的预测。实验结果表明，DrugDL 所提取的药物分子特征在多种与药物开发紧密相关的预测任务中均展现出了出色的表现，其性能超越了当前先进的基线模型。特别是在药物-靶标结合位点的预测中，DrugDL 所提取的药物和靶标特征能够精确地预测出药物与靶标结合的具体区域。这一能力不仅有助于我们更深入地探索药物分子的功能区域，还极大地增强了模型在生物医学领域的可解释性，为我们揭示药物分子与蛋白质结合区域的奥秘提供了有力的工具。

# 第五章 总结与展望

本文以药物分子的特征表示为研究起点，深入探讨了深度学习理论与数据挖掘算法在药物分子多性质预测中的应用。在药物研发的复杂过程中，药物分子性质的准确预测占据着举足轻重的地位。然而，由于药物分子及其相关生物实体 (如蛋白质等) 的高度复杂性和多样性，其表征成为了提升预测准确性的一个重大挑战。

本文深入剖析了药物与靶标两种模态间的相互作用关系，聚焦于单模态内部特征关系的挖掘以及跨模态间交互学习的两大核心要点，以此为基础探索复杂的药物-靶标相互作用机制，进而实现药物分子的精细特征表示。在第二章中，我们系统介绍了药物分子表征和性质预测的主流方法及其核心内容，涵盖了药物分子的文本、指纹和图形等多种表示方法，以及药物理化性质、毒性、药物间相互作用、药物-靶标结合亲和力及结合位点等关键性质。

在第三章中，我们提出了一种基于对比学习的药物特征表示方法——DrugDL。该方法通过强化单模态内部特征之间的联系，学习并理解跨模态特征间的交互模式，对药物分子进行高效表征。经过一系列对比实验分析，我们有力地证明了 DrugDL 在分子表征能力上的卓越表现，其性能显著优于传统的分子特征表示方法。

在本文的第四章，我们融合了 DrugDL 所提取的特征与多样化的网络结构，构建了一个全方位的药物分子多性质预测模型。该模型能够精准预测药物研发中的一系列关键属性，涵盖药物的理化性质、毒性评估、药物-药物相互作用、药物-靶标结合亲和力以及结合位点等。为了充分验证其卓越性能，本文在 4.3 节中，将 DrugDL 与一系列先进的基线模型进行了详尽的对比分析。结果显示，DrugDL 在所有对比任务上均展现出了绝对性的优势，其表现不仅超越了其他方法，甚至也优于当前已知的最佳方法，充分彰显了其全面且出色的预测能力。

尽管本文提出的 DrugDL 方法已经取得了较好的性能表现，但仍存在提升的空间。随着对药物研发的深入研究，越来越多的与药物相关的生物实体以及相互作用细节信息被不断挖掘出来。本文仅利用了蛋白质类靶标对药物分子进行跨模态的交互学习，如果能够合理探索并融合更多与药物存在关系的生物实体信息，将有望进一步提升药物的特征表示能力。此外，尽管 DrugDL 对药物和靶标间的相互作用显示出良好的可解释性潜力，但多数实验结果仍需通过实验室的湿实验进行进一步验证。

展望未来，我们将致力于收集更高质量的与药物分子相关的信息，以提高药物分子的表征能力。同时，我们将充分利用 DrugDL 的可解释能力，帮助药物研发人员筛选一些尚未发现的局部相互作用位点，学习药物分子的隐藏知识，为进一步的药物结构

优化提供重要信息。我们相信，通过持续的研究和探索，我们将能够为药物研发领域带来更多的创新和突破。

# 参 考 文 献


[1] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development[J]. Nature reviews Drug discovery, 2019, 18(6): 463-477.

[2] Ekins S, Puhl A C, Zorn K M, et al. Exploiting machine learning for end-to-end drug discovery and development[J]. Nature materials, 2019, 18(5): 435-441.

[3] Kim J, Park S, Min D, et al. Comprehensive survey of recent drug discovery using deep learning[J]. International Journal of Molecular Sciences, 2021, 22(18): 9983.

[4] König J, Müller F, Fromm M F, et al. Transporters and drug-drug interactions: important determinants of drug disposition and effects[J]. Pharmacological reviews, 2013, 65(3): 944-966.

[5] Wishart D S, Knox C, Guo A C, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets[J]. Nucleic acids research, 2008, 36(1): 901-906.

[6] Shitara Y, Sato H, Sugiyama Y. Evaluation of drug-drug interaction in the hepatobiliary and renal transport of drugs[J]. Annu. Rev. Pharmacol. Toxicol., 2005, 45(1): 689-723.

[7] Lv Q, Zhou J, Yang Z, et al. 3D graph neural network with few-shot learning for predicting drug–drug interactions in scaffold-based cold start scenario[J]. Neural Networks, 2023, 165(1): 94-105.

[8] Yıldırım M A, Goh K I, Cusick M E, et al. Drug—target network[J]. Nature biotechnology, 2007, 25(10): 1119-1126.

[9] Campillos M, Kuhn M, Gavin A C, et al. Drug target identification using side-effect similarity[J]. Science, 2008, 321(5886): 263-266.

[10] Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets[J]. Nature reviews Drug discovery, 2017, 16(1): 19-34.

[11] Marton M J, DeRisi J L, Bennett H A, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays[J]. Nature medicine, 1998, 4(11): 1293-1301.

[12] Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modelling[J]. Advanced drug delivery reviews, 2003, 55(9): 1163-1183.

[13] Hörter D, Dressman J B. Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract[J]. Advanced drug delivery reviews, 2001, 46(1): 75-87.

[14] O'Shea R, Moser H E. Physicochemical properties of antibacterial compounds: implications for drug discovery[J]. Journal of medicinal chemistry, 2008, 51(10): 2871-2878.

[15] Tran T T V, Surya Wibowo A, Tayara H, et al. Artificial intelligence in drug toxicity prediction: recent advances, challenges, and future perspectives[J]. Journal of chemical information and modeling, 2023, 63(9): 2628-2643.

[16] Basile A O, Yahi A, Tatonetti N P. Artificial intelligence for drug toxicity and safety[J]. Trends in pharmacological sciences, 2019, 40(9): 624-635.

[17] Raies A B, Bajic V B. In silico toxicology: computational methods for the prediction of chemical toxicity[J]. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2016, 6(2): 147-172.

[18] Mullard A. New drugs cost US $2.6 billion to develop[J]. Nature reviews drug discovery, 2014, 13(12): 1.

[19] Ashburn T T, Thor K B. Drug repositioning: identifying and developing new uses for existing drugs[J]. Nature reviews Drug discovery, 2004, 3(8): 673-683.
[20] Roses A D. Pharmacogenetics in drug discovery and development: a translational perspective[J]. Nature reviews Drug discovery, 2008, 7(10): 807-817.
[21] Boulougouri M, Vandergheynst P, Probst D. Molecular set representation learning[J]. Nature Machine Intelligence, 2024, 6(7): 754-763.
[22] Wigh D S, Goodman J M, Lapkin A A. A review of molecular representation in the age of machine learning[J]. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2022, 12(5): 1603.
[23] Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations[J]. Nature Machine Intelligence, 2021, 3(12): 1023-1032.
[24] Fang X, Liu L, Lei J, et al. Geometry-enhanced molecular representation learning for property prediction[J]. Nature Machine Intelligence, 2022, 4(2): 127-134.
[25] Zhu H, Zhou R, Cao D, et al. A pharmacophore-guided deep learning approach for bioactive molecular generation[J]. Nature Communications, 2023, 14(1): 6234.
[26] Swanson K, Liu G, Catacutan D B, et al. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics[J]. Nature Machine Intelligence, 2024, 6(3): 338-353.
[27] Munson B P, Chen M, Bogosian A, et al. De novo generation of multi-target compounds using deep generative chemistry[J]. Nature Communications, 2024, 15(1): 3636.
[28] Åqvist J, Medina C, Samuelsson J E. A new method for predicting binding affinity in computer-aided drug design[J]. Protein Engineering, Design and Selection, 1994, 7(3): 385-391.
[29] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, et al. A review on machine learning approaches and trends in drug discovery[J]. Computational and structural biotechnology journal, 2021, 19(1): 4538-4558.
[30] Gaudelet T, Day B, Jamasb A R, et al. Utilizing graph machine learning within drug discovery and development[J]. Briefings in bioinformatics, 2021, 22(6): bbab159.
[31] Stokes J M, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery[J]. Cell, 2020, 180(4): 688-702.
[32] Liu H, Huang Y, Liu X, et al. Attention-wise masked graph contrastive learning for predicting molecular property[J]. Briefings in bioinformatics, 2022, 23(5): 1-10.
[33] Xie A, Zhang Z, Guan J, et al. Self-supervised learning with chemistry-aware fragmentation for effective molecular property prediction[J]. Briefings in Bioinformatics, 2023, 24(5): 1-13.
[34] Zheng Z, Tan Y, Wang H, et al. CasANGCL: pre-training and fine-tuning model based on cascaded attention network and graph contrastive learning for molecular property prediction[J]. Briefings in Bioinformatics, 2023, 24(1): 1-13.
[35] Kim H, Park M, Lee I, et al. BayeshERG: a robust, reliable and interpretable deep learning model for predicting hERG channel blockers[J]. Briefings in Bioinformatics, 2022, 23(4): 1-15.
[36] Du W, Yang X, Wu D, et al. Fusing 2D and 3D molecular graphs as unambiguous molecular descriptors for conformational and chiral stereoisomers[J]. Briefings in Bioinformatics, 2023, 24(1): 1-12.
[37] Tian Y, Wang X, Yao X, et al. Predicting molecular properties based on the interpretable graph neural network with multistep focus mechanism[J]. Briefings in bioinformatics, 2023, 24(1): 1-9.

[38] Zeng X, Xiang H, Yu L, et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework[J]. Nature Machine Intelligence, 2022, 4(11): 1004-1016.
[39] Rao J, Xie J, Yuan Q, et al. A variational expectation-maximization framework for balanced multi-scale learning of protein and drug interactions[J]. Nature Communications, 2024, 15(1): 4476.
[40] Axelrod S, Gomez-Bombarelli R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation[J]. Scientific Data, 2022, 9(1): 185.
[41] Yu Z, Gao H. Molecular representation learning via heterogeneous motif graph neural networks[C]//International Conference on Machine Learning. PMLR, 2022: 25581-25594.
[42] Zhang Z, Liu Q, Wang H, et al. Motif-based graph self-supervised learning for molecular property prediction[J]. Advances in Neural Information Processing Systems, 2021, 34(1): 15870-15882.
[43] Chen X, Zhou C, Wang C C, et al. Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization[J]. Briefings in Bioinformatics, 2021, 22(6): 1-14.
[44] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties[J]. Journal of the American Medical Informatics Association, 2014, 21(2): 278-286.
[45] Chen X, Ren B, Chen M, et al. NLLSS: predicting synergistic drug combinations based on semi-supervised learning[J]. PLoS computational biology, 2016, 12(7): 1004975.
[46] Yan C, Duan G, Zhang Y, et al. Predicting drug-drug interactions based on integrated similarity and semi-supervised learning[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2020, 19(1): 168-179.
[47] 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): 1-10.
[48] 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.
[49] Yu H, Zhao S Y, Shi J Y. Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions[J]. Briefings in Bioinformatics, 2022, 23(4): 1-12.
[50] 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): 1-12.
[51] Tanvir F, Islam M I K, Akbas E. Predicting drug-drug interactions using meta-path based similarities[C]//2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, 2021: 1-8.
[52] Yuan X, Zhao W, Shen X, et al. Prediction of Drug-Drug Interactions Based on Meta-path-based Fusion Mechanism in Heterogeneous Information Network[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2022: 647-652.
[53] Yu H, Dong W M, Shi J Y. RANEDDI: Relation-aware network embedding for drug-drug interaction prediction[J]. Information Sciences, 2022, 582(1): 167-180.
[54] Gao K Y, Fokoue A, Luo H, et al. Interpretable drug target prediction using deep neural representation[C]//IJCAI. 2018, 2018: 3371-3377.

[55] 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): 1007129.
[56] Hinnerichs T, Hoehndorf R. DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug–target interactions[J]. Bioinformatics, 2021, 37(24): 4835-4843.
[57] Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. Bioinformatics, 2019, 35(2): 309-318.
[58] 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.
[59] 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.
[60] Hua Y, Song X, Feng Z, et al. CPInformer for efficient and robust compound-protein interaction prediction[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2022, 20(1): 285-296.
[61] He H, Chen G, Chen C Y C. NHGNN-DTA: a node-adaptive hybrid graph neural network for interpretable drug–target binding affinity prediction[J]. Bioinformatics, 2023, 39(6): 1-9.
[62] Hua Y, Song X, Feng Z, et al. MFR-DTA: a multi-functional and robust model for predicting drug–target binding affinity and region[J]. Bioinformatics, 2023, 39(2): 1-9.
[63] Lee I, Nam H. Sequence-based prediction of protein binding regions and drug–target interactions[J]. Journal of cheminformatics, 2022, 14(1): 5.
[64] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. Journal of chemical information and computer sciences, 1988, 28(1): 31-36.
[65] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening[J]. Expert opinion on drug discovery, 2016, 11(2): 137-148.
[66] Duan J, Dixon S L, Lowrie J F, et al. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods[J]. Journal of Molecular Graphics and Modelling, 2010, 29(2): 157-170.
[67] Duvenaud D K, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints[J]. Advances in neural information processing systems, 2015, 28(1): 1-9.
[68] Yang J, Cai Y, Zhao K, et al. Concepts and applications of chemical fingerprint for hit and lead screening[J]. Drug Discovery Today, 2022, 27(11): 103356.
[69] Lovrić M, Molero J M, Kern R. PySpark and RDKit: moving towards big data in cheminformatics[J]. Molecular informatics, 2019, 38(6): 1800082.
[70] Vassar R. Bace 1: The β-secretase enzyme in alzheimer's disease[J]. Journal of Molecular Neuroscience, 2004, 23(1): 105-113.
[71] Wan Y, Wu J, Hou T, et al. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation[J]. Nature Communications, 2025, 16(1): 413.
[72] Walmsley R M, Billinton N. How accurate is in vitro prediction of carcinogenicity?[J]. British journal of pharmacology, 2011, 162(6): 1250-1258.

[73] Andrade R J, Chalasani N, Björnsson E S, et al. Drug-induced liver injury[J]. Nature Reviews Disease Primers, 2019, 5(1): 58.
[74] Priest B, Bell I M, Garcia M. Role of hERG potassium channel assays in drug development[J]. Channels, 2008, 2(2): 87-93.
[75] Gao Y, Dong K, Gao Y, et al. Unified cross-modality integration and analysis of T cell receptors and T cell transcriptomes by low-resource-aware representation learning[J]. Cell Genomics, 2024, 4(5): 100553.
[76] Bai P, Miljković F, Ge Y, et al. Hierarchical clustering split for low-bias evaluation of drug-target interaction prediction[C]//2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021: 641-644.
[77] Liu H, Sun J, Guan J, et al. Improving compound–protein interaction prediction by building up highly credible negative samples[J]. Bioinformatics, 2015, 31(12): 221-229.
[78] 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.
[79] Wang Y, Xia Y, Yan J, et al. ZeroBind: a protein-specific zero-shot predictor with subgraph matching for drug-target interactions[J]. Nature Communications, 2023, 14(1): 7861.
[80] Koh H Y, Nguyen A T N, Pan S, et al. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data[J]. Nature Machine Intelligence, 2024, 6(6): 673-687.
[81] Mercer K E, Pritchard C A. Raf proteins and cancer: B-Raf is identified as a mutational target[J]. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2003, 1653(1): 25-40.
[82] Shaw A T, Engelman J A. ALK in lung cancer: past, present, and future[J]. Journal of clinical oncology, 2013, 31(8): 1105-1111.
[83] Normanno N, De Luca A, Bianco C, et al. Epidermal growth factor receptor (EGFR) signaling in cancer[J]. Gene, 2006, 366(1): 2-16.
[84] Jiang Z, Sun H, Yu J, et al. Targeting CD47 for cancer immunotherapy[J]. Journal of Hematology & Oncology, 2021, 14(1): 180.
[85] Liu Z, Sun Q, Wang X. PLK1, a potential target for cancer therapy[J]. Translational oncology, 2017, 10(1): 22-32.
[86] Chapman P B, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation[J]. New England Journal of Medicine, 2011, 364(26): 2507-2516.
[87] Robert C, Karaszewska B, Schachter J, et al. Improved overall survival in melanoma with combined dabrafenib and trametinib[J]. New England Journal of Medicine, 2015, 372(1): 30-39.
[88] Makena M R, Nguyen T H, Koneru B, et al. Vorinostat and fenretinide synergize in preclinical models of T-cell lymphoid malignancies[J]. Anti-Cancer Drugs, 2021, 32(1): 34-43.
[89] Samoszuk M, Corwin M A. Mast cell inhibitor cromolyn increases blood clotting and hypoxia in murine breast cancer[J]. International journal of cancer, 2003, 107(1): 159-163.
[90] Peterson B E, Bhatt D L, Steg P G, et al. Reduction in revascularization with icosapent ethyl: insights from REDUCE-IT revascularization analyses[J]. Circulation, 2021, 143(1): 33-44.
[91] Zhang S, Doudican N A, Quay E, et al. Fluvastatin enhances sorafenib cytotoxicity in melanoma cells via modulation of AKT and JNK signaling pathways[J]. Anticancer research, 2011, 31(10): 3259-3265.

[92] Wu Z, Ramsundar B, Feinberg E N, et al. MoleculeNet: a benchmark for molecular machine learning[J]. Chemical science, 2018, 9(2): 513-530.
[93] Gold L S, Manley N B, Slone T H, et al. Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997–1998[J]. Toxicological Sciences, 2005, 85(2): 747-808.
[94] Benigni R, Bossa C, Richard A M, et al. A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity[J]. Annali dell'Istituto superiore di sanità, 2008, 44(1): 48-56.
[95] Davies M, Nowotka M, Papadatos G, et al. ChEMBL web services: streamlining access to drug discovery data and utilities[J]. Nucleic acids research, 2015, 43(1): 612-620.
[96] Chen M, Vijay V, Shi Q, et al. FDA-approved drug labeling for the study of drug-induced liver injury[J]. Drug discovery today, 2011, 16(15): 697-703.
[97] Hoofnagle JH, Serrano J, Knoben JE, et al. LiverTox: A Website on Drug-Induced Liver Injury[J]. Wiley Online Library, 2013, 57(3): 873–874.
[98] Quinton A, Latry P, Biour M. Hepatox: database on hepatotoxic drugs[J]. Gastroenterol Clin Biol, 1993, 17(5): 116–120.
[99] 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(1): 1074-1082.
[100] Jang H Y, Song J, Kim J H, et al. Machine learning-based quantitative prediction of drug exposure in drug-drug interactions using drug label information[J]. NPJ Digital Medicine, 2022, 5(1): 88.
[101] Zhong Y, Li G, Yang J, et al. Learning motif-based graphs for drug–drug interaction prediction via local–global self-attention[J]. Nature Machine Intelligence, 2024, 6(9): 1094-1105.
[102] Davis M I, Hunt J P, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity[J]. Nature biotechnology, 2011, 29(11): 1046-1051.
[103] Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis[J]. Journal of Chemical Information and Modeling, 2014, 54(3): 735-743.
[104] Gaber Y, Rashad B, Fathy E. Biological 3D structural databases[J]. Essentials of Bioinformatics, Volume I: Understanding Bioinformatics: Genes to Proteins, 2019, 1(1): 47-73.
[105] Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction[J]. Communications Chemistry, 2023, 6(1): 34.
[106] Han S, Fu H, Wu Y, et al. HimGNN: a novel hierarchical molecular graph representation learning framework for property prediction[J]. Briefings in Bioinformatics, 2023, 24(5): 1-11.
[107] Wu C K, Zhang X C, Yang Z J, et al. Learning to SMILES: BAN-based strategies to improve latent representation learning from molecules[J]. Briefings in bioinformatics, 2021, 22(6): 1-9.
[108] Wang Y, Wang J, Cao Z, et al. Molecular contrastive learning of representations via graph neural networks[J]. Nature Machine Intelligence, 2022, 4(3): 279-287.
[109] Li R, Lu J, Liu Z, et al. Reusability report: exploring the utility of variational graph encoders for predicting molecular toxicity in drug design[J]. Nature Machine Intelligence, 2024, 6(1): 1457-1466.
[110] 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): 4304-4311.

[111] Zhang X, Wang G, Meng X, et al. Molormer: a lightweight self-attention-based method focused on spatial structure of molecular graph for drug–drug interactions prediction[J]. Briefings in Bioinformatics, 2022, 23(5):1-11.

[112] Luo Y, Liu Y, Peng J. Calibrated geometric deep learning improves kinase–drug binding predictions[J]. Nature machine intelligence, 2023, 5(12): 1390-1401.

[113] 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.