硕士学位论文

基于图自监督学习的微生物-药物关联及药物性质预测

Prediction of Microbe-Drug Associations and Drug Properties Based on Graph Self-supervised Learning

专业学位类别	电子信息
专 业 领 域	软件工程
作 者 姓 名	黄毅彪
指 导 教 师	邓磊教授

中 南 大 学

2023 年 5 月

中图分类号 TP391　　学校代码 10533
UDC 004.41　　学位类别 专业学位

硕士学位论文

基于图自监督学习的微生物-药物关联及药物性质预测

Prediction of Microbe-Drug Associations and Drug Properties Based on Graph Self-supervised Learning

作 者 姓 名	黄毅彪
专业学位类别	电子信息
专 业 领 域	软件工程
研 究 方 向	生物信息学
二级培养单位	计算机学院
指 导 教 师	邓磊教授

论文答辩日期 2023年5月22日　　答辩委员会主席_____

中 南 大 学

2023 年 5 月

基于图自监督学习的微生物-药物关联及药物性质预测

摘要：基于人工智能的药物研发已经成为近年来广受关注的研究领域,其中筛选微生物-药物关联和预测药物性质对药物研发大有裨益。微生物-药物关联预测能够揭示微生物和药物间相互作用和机制，同时可以帮助寻找良好治疗效果和减少副作用的药物。药物性质预测可以在药物设计和开发过程中提供有用的信息，帮助优化药物结构，提高药物的药效和可靠性。但药理学数据集的质量问题以及微生物和药物的多样性和复杂性可能导致基于微生物-药物数据构建的异质网络存在难以处理的噪声。此外，由于药理学数据中被标记的分子性质数据有限，基于监督学习的分子表示算法只能搜索有限的化学空间，泛化性较差。

本文基于图自监督学习来研究微生物-药物关联及药物分子性质预测。主要研究内容如下：

1、针对微生物-药物的异质网络中难以处理的噪声，本研究提出了一种基于多模态变分图嵌入的微生物-药物关联预测模型Graph2MDA。首先构建两个微生物与药物的关联网络，同时结合药物的分子结构、微生物的基因序列和功能注释以及相似性网络拓扑特征等多种特征构建了多模态属性图。以多模态属性图为输入，训练图变分自编码器学习每个节点和整个图的有信息和可解释的潜在表示，然后使用深度神经网络分类器进行微生物与药物关联预测。超参数分析和模型消融研究显示了本研究模型的敏感性和稳健性，同时在三个独立数据集上的实验评估结果表明本研究提出的方法优于现有的六种最先进的方法。此外，本研究还探讨了药物在训练中潜在表征的意义，发现药物表现出明显的聚类模式，且与药物 ATC 分类显著一致。最后，对两种微生物和两种药物在 PubMed 文献进行了案例研究，验证了模型的有效性。大量评估实验验证了本研究方法的有效性。

2、为了充分利用大规模无标记的药物数据，本文提出了一种自监督对比学习预训练，并用于分子表示学习和性质预测的模型ATMOL。本研究设计开发了一种新的分子图增强策略，称为注意力

权重图遮掩，并以此策略生成具有挑战性的负样本进行对比学习。使用图注意力网络作为分子图编码器，并利用所学习到的注意权值作为遮掩引导，生成分子增强图。该模型通过最小化原始图和增强图之间的对比损失，来捕捉到重要的分子结构和高阶语义信息。本研究通过大量的实验表明，这种注意力图遮掩对比学习能够在一些下游分子性质预测任务中表现出了最先进的性能。此外，还验证了在更大规模的未标记数据上预训练的模型能提高了学习到的分子表示的泛化能力。最后，通过注意力热图的可视化方法，表明原子和原子基团对特定的分子性质很重要。

图 11 幅，表 17 个，参考文献 124 篇

关键词：深度学习；图自监督学习；微生物与药物；关联预测；药物性质；对比学习；自注意力；化学亚结构
分类号：TP391

Prediction of Microbe-Drug Associations and Drug Properties Based on Graph Self-supervised Learning

Abstract: Artificial intelligence-based drug discovery and development has gained significant attention in recent years. Screening for microbe-drug associations and predicting drug properties are critical tasks in drug discovery. The prediction of microbe-drug associations can aid in the identification of drugs with optimal therapeutic effects and reduced side effects, while also providing insights into the interaction and mechanism between microbes and drugs. On the other hand, drug property prediction can offer valuable information in drug design and development, facilitating the optimization of drug structure, improving drug efficacy, and reliability. However, the quality of pharmacological datasets, as well as the diversity and complexity of microbes and drugs, may result in noise in heterogeneous networks constructed from microbe-drug data. Moreover, molecular representation algorithms based on supervised learning can only explore a limited chemical space and generalize poorly due to the limited availability of labeled molecular properties in pharmacological data.

This paper proposes the use of graph self-supervised learning for predicting associations between microbes and drugs and for molecular property prediction. The main research contents of this paper are as follows:

1. Aiming at the difficult noise in microbe-drug heterogeneous networks, a microbe-drug association prediction model (Graph2MDA) based on multimodal view embedding was proposed in this study. Firstly, two microbe and drug association networks were constructed, and multi-modal attribute maps were constructed based on drug molecular structure, microbial gene sequence and functional annotation, similarity network topology and other features. With the input of a multimodal attribute graph, the graph variational autoencoder is trained to learn the informed

and interpretable potential representation of each node and the entire graph, and then a deep neural network classifier is used for microbe and drug association prediction. Hyperparametric analysis and model ablation studies demonstrated the sensitivity and robustness of our model, while experimental evaluation results on three independent data sets showed that the proposed method outperformed six existing state-of-the-art methods. In addition, the significance of potential characterization of drugs in training was also explored, and it was found that drugs showed a clear clustering pattern that was significantly consistent with drug ATC classification. Finally, a case study was carried out in PubMed literature to verify the validity of the model A large number of evaluation experiments have verified the effectiveness of this method.

2. In order to make full use of large-scale unlabeled drug data, a self-supervised contrast learning pretraining model ATMOL is proposed for molecular representation learning and property prediction. This study developed a new molecular design enhance the strategy `and strategies to generate challenging negative contrast study samples. The graph attention network is used as the molecular graph encoder, and the learned attention weights are used as the masking guidance to generate the molecular enhanced graph. The proposed model captures important molecular structure and higher-order semantic information by minimizing the contrast loss between the original and enhanced graphs. In this study, a large number of experiments show that this attention-force masking contrast learning can show the most advanced performance in some downstream molecular property prediction tasks. In addition, it is verified that the model pre-trained on a larger scale of unlabeled data can improve the generalization ability of the learned molecular representation. Finally, the visualization method of attention heat map shows that atoms and atomic groups are important for specific molecular properties.

Keywords: Deep Learning; Graph self-supervised learning; Microbe and drug; Association prediction; Self-attention; Chemical substructure
Classification: TP391

目　录

第 1 章 绪论

1.1 研究背景和意义

微生物群落，包括细菌、古生菌、病毒、原生动物和真菌，与人类宿主密切相关[1, 2]。它们存在于人体内各种器官中，如皮肤、胃肠道、口腔和其他组织。大量的研究证实，微生物在维持人体内部环境的稳态方面起着基础性的作用，人体及其体内的微生物群落属于共生互利关系[3]。宿主为微生物提供保护和丰富的营养，而微生物又支持和配合在宿主的消化、免疫和神经系统发育等方面。它们的有益功能涉及相当多的方面，如改善新陈代谢，合成必需维生素，抵抗病原体，增强免疫力[4]。例如，肠道微生物可以将摄入的碳水化合物分解为单糖，而后被分解成二氧化碳和短链脂肪酸，为宿主提供能量[5]；肠道微生物通过 Treg 细胞调节在调节免疫反应或免疫耐受中发挥关键作用[6]。此外，微生物构建菌膜屏障或促进肠道上皮细胞增殖分化等方式形成保护屏障[7]。例如婴儿肠道的双歧杆菌和乳杆菌能够帮助稳定肠道的生态稳定，通过母乳喂养来帮助肠道有益菌的定植和肠道屏障功能的恢复，来抵抗和排斥外源致病菌入侵[8]。

另一方面，微生物对人体的影响并不只有益处，它们与人体疾病与药物治疗之间存在着复杂的相互作用和调节机制。微生物群落的失衡往往会导致各种疾病，如糖尿病[9]，肥胖[10]，心脑血管疾病[11]，肠易激综合症[11]，癌症[12]及更多疾病。例如肥胖的原因与人体内菌群的失衡存在直接的关联，Clarke 等人对小鼠进行饮食诱导后导致其肥胖，研究使用前后期肠道菌群组成的差异，其中消化球菌科含量明显升高，而理研菌科和紫单胞菌科含量大幅下降[13]。口腔中的微生物是消化道炎症的重要诱发因素，对消化道炎症患者口腔菌群测序分析对比消化道炎症患者唾液菌群相对健康人群口腔菌群含量，发现拟杆菌门与变形杆菌门含量发生明显变化[14]，此外越来越多的研究结果表明口腔微生物也与消化系统癌症的发生存在关联。患有牙周病或缺牙的人患消化系统癌症的风险增加，原因就来自口腔内的微生物具核梭杆菌，有研究表明食管癌患者体内癌组织存在的具核梭菌会大幅减少患者的生存概率[15]。

微生物与药物之间同样存在复杂多样的作用机制。微生物可以通过多种方式影响药物的作用，包括药物吸收、药物代谢、药物排泄、药物耐药性、药物毒性等诸多关联。多项研究表明，微生物参与药物吸收和代谢，从而调节药物疗效和药物毒性[16]。例如，肠道放线菌类中的埃格特氏菌负责心脏药物地高辛的失活[17]。在宿主细胞将替诺福韦转化为药物活性形式之前，阴道微生物群可有效降解替诺福韦[18]。中药何首乌可能导致微生物失衡从而产生毒性，引起肝

损害、腹泻等症状[19]，有研究采用高通量测序等生物技术手段研究何首乌诱导的大鼠肠道内的微生物，发现毛孢子菌、乳杆菌等微生物丰度发生显著变化[20]。此外长期复用何首乌会导致更多的肠道微生物和内毒素通过肝肠轴进入肝门静脉，从而导致肝损伤的加重和肠道疾病的发展[21]。

相对的，药物反过来也改变了生活在人体内的微生物群落的多样性和功能，药物也能够抑制或者杀死人体内的微生物。例如，广谱抗生素可以杀灭大量的细菌，不仅包括致病菌，还包括对人体有益的细菌，这种杀菌作用可能导致微生物群落的失衡，从而引发肠道疾病等问题。此外，一些药物如非甾体抗炎药(NSAIDs)、质子泵抑制剂等也可能通过抑制某些细菌的生长或代谢来影响微生物。目前一些针对微生物的新药已经被开发出来，以维持微生物群落结构的适应性[22]，如胆汁酸类药物可以影响肠道微生物的代谢和生长，从而影响微生物群落的平衡和功能。微生物与药物之间的相互作用导致耐药细菌的传播，这对人类健康构成了另一个严重威胁[22]。抗生素的长期广泛使用也造成了耐药性等问题，对应的细菌和病毒对药物的敏感性降低，导致药效下降甚至无效化，如鲍曼不动杆菌和铜绿假单胞菌对抗菌药物的耐药性显著增加，并出现多重耐药和泛耐药菌株[23]。基于上述研究，可以得出药物与微生物关联的重要性以及深入研究的必要性。为了更好地探究药物与微生物关联，需要结合现代生物技术手段例如微生物基因组学、微生物转录组学、药物代谢组学以及在研究药物与微生物互作机制研究中起着决定性作用的药物分子性质。

药物性质与微生物之间存在密切的关联。一方面，微生物是药物的生产者之一，许多药物都是从微生物中提取或合成出来的，例如青霉素和链霉素等。微生物还可以作为药物靶点或药物传递途径，参与药物的作用和传输过程。另一方面，微生物对药物的代谢和转化对药物的吸收、分布、代谢和排泄过程产生重要影响。此外，药物的疗效和副作用很大程度上取决于药物与微生物的相互作用，如药物的毒性和副作用也可能会导致微生物的生长受到抑制或死亡，从而影响人体微生物群落的稳定性和功能。因此，了解药物与微生物的相互作用对于药物研发、治疗方案的制定以及个体化药物治疗的实现都至关重要，同理，预测药物性质在微生物学研究中也发挥重要作用。例如通过预测药物对微生物毒性，可以避免药物对微生物群落的不利影响，从而提高药物的安全性和耐受性。其次，如水溶性、亲脂性等影响微生物抗性、药物代谢吸收及药物疗效的分子性质研究对于药物研发也是不可或缺的一部分。

然而，现有已知药物详细分子性质的数据相对现有已知的药物数量而言是十分稀少的。现有的药物数据库中所收纳的上亿药物分子数据中只有相当微小

的药物分子性质是已知验证的。尽管目前仍有实验室进行新药性质测验，但除了实验本身就费时费力、成本巨大，这海量的药物规模就足以让人望而却步。那么通过一种安全快速的药物性质预测模型，能够在现有的未标记药物数据中对候选药物的毒性、药效和代谢稳定性等进行前期预测，从而在后续研究中优先筛选出具有潜在疗效的药物候选物，降低研发时间和成本，提高研发效率，将具有至关重要的意义及价值。

综上所述，不仅需要对微生物与药物之间的相互作用进行研究预测，还要探寻药物性质对于机体及体内微生物的影响，高效可靠对药物性质预测是药物发现和再利用领域的一个理想目标。本研究期望通过深入这些问题本质来探索基于深度学习领域的关联预测技术来获得一些重要突破，期望能够对微生物与药物关联以及药物性质对微生物等的相互机制有进一步的探索解释。此外为了推动药物疗效机制及微生物与药物调节互作机制的探索，高效预测药物分子性质的技术也至关重要。可以看出，本研究对相关研究如药物重定位、疾病预防治疗等有着一定社会价值和学术意义。

1.2 国内外研究现状

微生物组不仅是健康和疾病的重要调节因子，也是临床上重要的药物靶点。大量的研究人员将目光放在人体微生物群对药物代谢、毒性和疗效等影响的研究上，微生物与药物之间的相互作用通常需要进行生物实验室通过实验验证。实验室通常会采用一些依赖培养技术，例如最小抑菌浓度测定 MIC[24]、最小杀菌浓度测定 MBC[25]、突变体筛选[26]、动物试验[27]、细胞培养实验[27]等方法来研究不同药物对微生物的抗菌作用及耐药性。但在实验室培养大量微生物类群即使不是不可能，也是有问题的，因为研究人员很难对特定微生物组的个体成员进行全面的分类，也很难理解微生物群落的功能和如何影响宿主与病原体的相互作用。随着测序技术和计算工具的进步使得越来越多的宏基因组研究得以进行，这些研究为人类微生物组和其他环境中的其他微生物群落提供了关键的见解。不依赖于培养的聚合酶链式反应(PCR)等技术方法通过对多种病原体的 DNA 或 RNA 片段进行体外扩增，用于微生物基因的定性和定量鉴定。基于 PCR 技术为微生物宏基因组学的建立铺平了道路，为宏基因组学进一步的研究揭示了微生物如何以意想不到的方式影响宿主的基因，提供适用于广泛学科的有用信息，包括病原体监测、生物技术、宿主相互作用，功能失调，进化生物学等[28]。

基于聚合酶链式反应(PCR)的运用，出现了一系列高效准确的测序技术，能

够可以对 DNA 或 RNA 分子进行序列测定。从最早的第一代 Sanger 测序技术，到目前最常见的能同时对多个 DNA 分子进行测序的高通量测序技术 NGS[29]，再到第三代单分子测序技术及当前最新的纳米孔测序技术[30]。测序技术的出现极大地提高了识别和描述微生物参与代谢和调节机制的能力，宿主和微生物通过这些机制相互作用，从而影响宿主生物体的健康或疾病状态。通过对微生物菌群的测序，可以更深入地了解微生物的组成、数量、功能以及微生物之间和宿主之间的相互作用，从而为药物研发提供重要信息和指导。例如，通过 16S rRNA 基因测序技术可以对微生物菌群的组成进行分析，从而发现与某种药物治疗相关的微生物群落的变化[31]。454 焦磷酸测序用于评估甲硝唑治疗和益生乳杆菌活菌制剂治疗细菌性阴道病前后阴道微生物群落菌群多样性和丰度差异，该测序技术帮助探索细菌性阴道病相关致病菌在发病及治疗期间中的微生物与药物互作和机制[32]。综上所述，测序技术可以对微生物的基因组和代谢组进行深入分析，发现微生物之间的多样性和功能差异，为药物研发提供更加精确的目标。但随着测序技术的进步导致了大量相关数据集的产生，这些数据集越来越难以分析[33]。随着更大数据集的产生，需要更复杂的计算资源和生物信息学工具。对为微生物宏基因组研究的解释和理解取决于可用于分析巨大数据集并挖掘有关被研究的微生物群落的有价值、有用和有效信息的计算工具。由于上述传统的药物与微生物相互作用研究方式都是由相关专业领域人物利用各样生物技术对微生物或药物在人体内进行研究，这些湿实验室的技术实验往往需要耗费大量的实验周期，操作专业度高，要求更为严苛，而且传统技术难以处理大规模的数据集和挖掘其中有价值的微生物与药物的潜在关联。人们强烈地认识到需要一种能更便捷高效、低成本的生物信息学技术来分析挖掘这些庞大数据中关于微生物多样性与药物互作间的有用信息。

随着人工智能技术在生物领域的不断发展，人们将目光放在能够预测性能高效且适用于大规模数据集的机器学习方法上，目前一些基于机器学习方法处理大规模的数据例如微生物基因组数据、表观基因组数据、药物代谢组数据等。这些数据通常难以用传统的统计方法进行处理和分析，而机器学习算法能够自动识别和提取数据中的特征，帮助研究人员快速预测微生物与药物间的关联，减少人工成本和进一步筛选所需的微生物与药物组并采取生物实验技术来验证。HMDAKATZ 模型[34]采用 KATZ 度量其基于微生物与药物相似性网络所构建的异质网络中各节点的关联性，并由此来预测微生物与药物间存在的潜在关联。而 NTSHMDA[35]在通过网络拓扑相似性优化后微生物与疾病异质网络上使用重启随机行走(RWR)来预测新的微生物与疾病关联。

近些年来，多个数据库发布其收集整理的大量经实验验证的微生物与药物的关联，如收集了微生物对抗生素和其他药物的抗性信息的 MDAD[36]、抗生物膜剂的资源及其在针对抗生素药物抗性的 Abiofilm 数据库[37]和药物与病毒相互作用的数据库 Drugvirus[38]，以及手工收集的大量药物与病毒相互作用条目的人类药物病毒数据库 HDVD[39]。这些能够免费获取的数据集的出现推动了基于深度学习的微生物与药物的关联预测方法的发展。例如，EGATMDA[40]利用微生物、疾病和药物的异质网络中存在的“微生物—疾病—药物” 虚拟元路径作为潜在的微生物-药物新关联网络，并利用分层的注意机制来构建了一个图注意网络来预测人类微生物-药物之间的关联。HNERMDA[41]将 metapath2vec 与二分网络推荐相结合，提出了微生物-药物关联预测的异质网络嵌入表示框架。GCNMDA[42]首先通过一个随机行走的预处理方法来捕获药物网络和微生物网络中权重高的、有价值的特征，然后设计了一个准确聚集网络邻域表征的关注机制并与条件随机场(CRF)相结合，最后嵌入到图卷积网络(GCN)框架中来预测微生物与药物的关联。此外在其他的一些关联预测领域，如药物与疾病，微生物与疾病的潜在关联预测任务也被人们所关注。例如，LAGCN[43]是另一种基于分层注意力映射卷积网络的计算模型，用于预测疾病与药物关联。GATMDA 结合了归纳矩阵补全(IMC)，采用带有会话头的图注意变种网络在微生物疾病的二分网络上进行微生物与疾病的关联预测。上述的研究都表明，与传统分类器相比，它们在预测性能方面取得了进展。但这些模型仍然存在严重限制其有效性的局限性。这些模型构建了具体的微生物与药物或者其他关联的集成网络来提取特征来预测新的关联，然而，他们未能在特征提取中处理离群节点，这经常导致不正确的预测。此外，这些模型考虑了网络结构信息或节点特征，但未能同时处理两者。虽然目前已经存在一些基于人工智能的微生物与药物关联预测算法，但这些算法仍然存在着一些难以突破的瓶颈。因此，如何避免数据噪声的影响并且能够准确有效地预测微生物与药物相互作用依旧是值得挑战的技术难题。

现有的研究大都只关注药物对于微生物作用某一个方面的性质，例如药物的毒性、耐药性、溶水性等。然而药物分子所有的理化性质，如水溶性、亲脂性、膜透性和解离度等，对药物重定位及药物开发中先导化合物的筛选至关重要。如果人们需要筛选出能够针对特定病原体的靶标药物时，首先需要知道靶标药物的所有分子性质，此时只能够对药物单一的性质逐步进行生物实验验证测定。然而传统的药物发现实验耗时耗力，不可能覆盖数亿个候选分子，对这些候选分子进行标签标注都需要昂贵、耗时的实验室测试，所以在药物性质识别领域中数据稀缺仍然是药物开发和药物重定位任务中需要克服的难题，如果

能够将规模庞大的无性质药物分子数据运用到分子性质预测任务中，也会是一个值得探讨的解决方案。基于上述的期望与挑战，人们提出了许多预测分子性质的计算机模拟方法，这些方法极大地提高了药物开发的效率和投资回报率[44]。

近年来，自监督学习因其在多个领域具有较好的泛化能力而受到广泛关注。一些基于预训练的自监督学习对分子性质预测上取得较好的成果[45]，这些方法采用规模庞大的预训练数据集，其中所包含的分子数量可达百万级[46]。自监督学习首先在大规模无标记数据集上进行预训练，得到潜在表示[47-49]，然后迁移到下游任务以获得更好的鲁棒性[50, 51]。对于分子性质预测任务，已经提出了一些自监督方法来学习分子表示[52-59]。这些方法大致分为两类:基于生成的方法和基于对比学习的方法。生成式方法通过建立特定的借口任务来学习嵌入，鼓励编码器提取高阶结构信息。例如，MG-BERT 通过将图神经网络(GNNs)的局部消息传递机制集成到强大的 BERT 模型[53]中来增强分子图的表示学习，从而学会了预测隐藏原子[52]。MolGPT[54]训练了一个 transformer-decoder 模型，用于下一个标记预测任务，使用隐藏的自我注意来生成新分子。对比学习鼓励相同分子的增强 (对比视图)，与不同分子生成的嵌入相比，具有更多相似的嵌入。例如，MolCLR[55]通过遮掩节点或边或子图提出了三种不同的图增强，然后最大化同一分子增强的一致性，同时最小化不同分子增强图的一致性。CSGNN[56]设计了一个深度混合跳跃的 GNN 来捕捉高阶依赖关系，并引入了一个自我监督的对比学习框架。MolGNet[57]同时使用配对子图识别和属性遮掩来实现节点级和图级预训练，这被证明可以提高从分子图中提取特征的能力。MoTSE[58, 59]同时考虑多个相似任务，通过迁移学习提高分子预测性能，方法是将单个任务投射到一个潜在空间，在该空间中计算任务相似度，并选择相似任务进行多任务学习。MV-GNN[59]采用多视图 GNN 进行分子性质预测，通过从节点(原子)和边(键)建立两个分子视图，然后进行相互依赖的消息传递，加强两个视图的信息通信。这些方法大都通过随机遮掩一些节点和边来生成对比性视图，然而随机遮掩的方式不能指导编码器检测最重要的子结构。

目前仍然存在诸多需要攻克的技术难点和挑战，主要体现在以下几点：

（1）受到数据噪声的影响难以精确预测。现有的微生物与药物关联预测算法大都通过构建微生物药物集成网络来提取潜在特征进行新关联预测，然而这些模型可能受到数据不平衡、数据噪声、数据缺失等问题的影响，仍然存在严重限制其有效性的局限性，导致出现不精确的预测结果。此外这些模型很少探究作用微生物的药物间可能存在潜在关联，无法充分利用微生物组、药物模型中潜在的特征信息，使其预测结果不可解释。

（2）微生物与药物互作机制缺乏可解释性。理解药物与微生物相互作用对于优化药效学和药代动力学具有重要意义。构建药物与微生物间的异构网络可以帮助人们挖掘可能存在的内在互作联系，但目前现有的关联预测算法仍然需要改进。此外在设计和使用这些药物时，需要考虑个体差异以及微生物因素，同时药物本身的分子性质存在着多样性与复杂性，难以解释二者互作机制。目前对微生物与药物关联预测模型中仍存在一些需要解释的问题，例如分子特征经过深度学习模型训练后会有什么变化；微生物与药物间是如何相互作用的；以及如何解释预测模型的可行性。

（3）未标记药物性质数据集规模庞大。随着对药物发现和药物设计的高度关注及研究，采用高通量筛选技术和用于临床的药物规模与日俱增。但关于药物性质数据如药物序列、二维结构、三维结构、生理化学、药物动力学等需要专业的生物化学实验验证，但被其昂贵的成本和冗长的周期限制，无法跟上新药发现的速度。因此能够有效地发现药物各项性质将帮助研究人员更详细地了解微生物与药物的互作机制。由于当前被标记的分子性质数据相较于海量未标记数据规模差距悬殊。需要一种高效的模型能够挖掘无标签的药物分子数据中共性特征，并迁移至有标签的药物数据进行验证是后续药物性质预测的主要研究方向。

1.3 研究目标及内容

在本研究中，希望通过构建微生物与药物关联网络，并结合网络特征挖掘方法和图变分自编码器提出一种新颖的关联预测模型，探寻训练后的药物特征在表征空间的表现。此外为了探索广阔的药物空间，计划结合自监督学习中对比学习和图自注意力网络作为预训练，并设计一个有效的增强策略提高模型的泛化性和鲁棒性。此外结合注意力权重来探索药物分子训练时特征变化的可解释性，探寻一种新的方法推进相关领域的研究。

本文具体的研究目标及内容如下：

（1）基于多模态变分图嵌入的微生物-药物关联预测

越来越多的临床证据表明， 存在人体中的微生物在药物开发，耐药性研究及药物分解等过程中起到重要的作用。因此确定潜在的与微生物相关的药物能够有效的帮助微生物药物互相作用机制的了解，也可以促进药物发现和药物用途的发展。随着微生物基因组和药理学数据集的增加，开发系统合理的计算方法以鉴定微生物药物互相作用至关重要。在本研究中，期望构建一种基于多模式属性图的体系结构，通过现有的数据集中丰富的生物学信息为微生物和药物

构建了多种模式的网络，考虑结合微生物与药物的相似性特征到网络中，利用图变分自编码器来提取网络中微生物药物潜在特征表示，最后通过深度神经网络(DNN)来预测新的微生物与药物关联。在不同的交叉验证设置下与目前最新关联预测方法进行对比。此外，将预测得分最高的几组微生物（或药物）在 PubMed 文献数据库中进行验证。最后研究与微生物相关的药物在潜在表征空间中的表现。

（2）基于注意力图遮掩对比学习的药物分子性质预测

由于药物分子性质标注的数据有限，基于监督学习的分子表征算法只能搜索有限的化学空间，并且泛化能力较差。本研究希望探索一种自适应、泛化性强的对比学习通用框架，同时提供一种可解释性的表征方式。在这项研究里，计划将药物分子建模成分子图，利用一种自监督的学习方法用于分子表征学习和属性预测。此外考虑采用图注意网络作为分子图的编码器，设计一种基于注意力权重的分子图增强策略，以产生具有挑战性的阳性样本进行对比学习。此外在不同规模的无标签数据上进行了预训练，探寻本研究的模型是否能够在大规模的数据集上性能有所提升，并在一些分子性质预测的下游任务中与目前先进的模型进行对比验证。最后分析对比学习中的网络来捕获重要的分子结构和高阶语义信息，并通过可视化的方式表现出来。本研究任务期望寻找到一种优异的、泛化性更强的对比学习遮掩策略。

1.4 本文组织结构

本文包含五个章节，其详细内容主要如下：

第一章：绪论。本章节主要介绍了微生物与药物相互作用及药物性质研究的相关背景意义，以及收集分析了相关领域的国内外研究进展，包括传统测序技术、微生物与药物关联预测、药物性质预测等，确定本研究的研究目标和内容。

第二章：相关知识和理论部分。本章节中对研究所需的相关理论知识进行了详细的阐述。描述了微生物-药物关联与药物性质数据集的相关知识，并介绍本研究所涉及到的深度学习模型以及对应性能评价指标。

第三章：基于多模态变分图嵌入的微生物-药物关联预测。此章节提出了一种基于多模态网络和图变分自编码器的深度学习模型，来预测微生物与药物的潜在关联。首先介绍了多模态网络的构建流程，包括微生物与药物关联的二分网络、异质网络的搭建，以及相似性特征和拓扑网络特征的构建融合。其次介绍了模型的总体架构，将微生物与药物的网络及特征嵌入作为图变分自编码器

的前馈输入，来挖掘潜在共同特征，最后使用 DNN 分类器进行预测。并在基准数据集上将该模型与目前最新的方法进行了性能比较。此外，对提出的模型进行超参数敏感度研究、消融实验探索模型的鲁棒性，并在独立数据集上将本研究的模型与竞争方法一起对比验证，还通过可视化技术解释本研究的模型所挖掘的与微生物存在关联的药物出现聚类现象。最后对两种流行的抗菌药物芦荟凝胶和氯唑西林，以及常见的微生物人类免疫缺陷病毒(HIV)和结核分枝杆菌中关联预测得分最高的相关微生物（和相关药物）在 PubMed 文献数据库中找到实验证实的协同作用。

第四章：基于注意力图遮掩对比学习的药物分子性质预测。在本章中提出了一个基于对比学习的图注意力网络(GAT)并通过权重遮掩的分子图增强策略的分子性质预测模型，构建了一个用于大型未标记分子数据集的自监督学习框架。首先使用药物的 SMILEs 建立一个分子图，然后通过 GAT 对分子图中点和边分配权重，然后通过利用分子图增强策略对每个药物生成对比学习样本，利用对比学习来最大限度地提高分子与其增强图的一致性。模型采用多种权重遮掩策略来对模型进行预训练，再将获得的药物分子特征学习网络迁移到分子性质预测任务。此外，将本研究的模型与其他先进的模型进行性能对比，并对模型做了可视化分析以验证预训练的效果，探索了模型中注意力网络的可解释性，揭示药物的重要化学亚结构与其药物性质的关联。

第五章：总结与展望。本文的最后一章对本文的研究内容进行了总结。主要包括本研究的创新点，并针对本研究的不足提出进一步改进的可行。

第 2 章 相关知识和理论

2.1 微生物与药物关联预测

2.1.1 微生物与药物数据集

MDAD[36]收集了微生物对抗生素和其他药物的抗性信息。该数据库向用户提供临床或实验证实的微生物和药物之间直接或全面的关联来促进微生物与药物研究。其中每条记录提供了详细的注释，包括药物的分子形式、来自 DrugBank 的超链接，来自 Uniprot 的微生物目标信息和原始参考链接。访问网址为 http://chengroup.cumt.edu.cn/MDAD。

Abiofilm[37]是抗生物膜剂的资源及其对抗生素药物抗性。其中包括多种生物，包括革兰氏阴性、阳性细菌等，主要是化学品、噬菌体、纳米粒子和提取物等。访问网址为 http://bioinfo.imtech.res.in/manojk/abiofilm/。

Drugvirus[38]是药物与病毒相互作用的数据库。该数据库被用于探索和分析广谱抗病毒药物（BSA，抑制几种人类病毒的化合物）和含有 BSA 的药物组。访问网址为 https://drugvirus.info/。

HDVD[39]手工收集了大量人类药物与病毒相互作用条目，其中包含与各种病毒相关的实验支持药物，为药物重新定位提供了基础，有助于筛选抗病毒药物。访问网址为 https://github.com/luckymengmeng/HDVD。

2.1.2 关联预测流程

随着计算能力和计算方法的研究与发展，人们将目光放在了生物学与计算机的交叉领域，期望通过一些不同于传统生物实验的技术方法来解决一些费时费力的推测诊断任务，例如使用关联预测、分子模拟、药物设计等任务。在生物学领域中的关联预测是一种利用已知的生物数据和信息来推断未知的生物现象或特征的方法，如蛋白质与非编码 RNA[60]、miRNA 与疾病[61]、ncRNA 与疾病[62]、疾病与药物[63]、药物与微生物[64]等各类关联预测任务。关联预测流程通常包括以下几个步骤：

（1）数据收集和整理。从不同的来源和平台获取相关的生物数据，如基因序列、表达谱、蛋白质相互作用、药物性质、疾病语意相似性等，对数据进行归一化、异构信息网络构建等处理，使数据可以进行特征挖掘和分析验证。

（2）数据挖掘和建模。根据研究目的和假设，选择合适的数据挖掘方法如聚类分析、分类分析、回归分析等，或适当的算法如机器学习、深度学习等，探索性地分析数据，建立数学模型、机器学习模型或神经网络模型，挖掘数据

中潜在特征以期望获得内在规律和联系。

（3）结果预测和评估。对挖掘的潜在特征进行新关联进行预测结果评分排名，并通过文献和临床数据验证预测结果的可信度和有效性是否高，对生物学问题或应用是否有意义。

（4）结果展示和应用。通过可视化的手段将预测结果以适当的形式展示，如图表、热力图、3D 图等，最后根据实验结果提出可改进提升的地方及新的研究方向，或将结果应用于生物学领域，如药物开发、疾病诊断、基因工程等。

2.2 药物性质预测

2.2.1 预训练药物数据集

在药物性质预测中，首先会使用大量未标记的药物数据集进行自监督模型的预训练，而后迁移到下游任务中来。其中所使用的未标记数据集是 ZINC 数据集。ZINC[65]是一个包含超过 1 亿个化合物结构的公共化合物数据库，其中，包括有机小分子化合物、药物分子、化学试剂、天然产物、多肽、核苷酸和配体等多种类型的化合物数据。这些化合物可用于药物设计和虚拟筛选等领域。ZINC 数据库以自动化的方式从多种来源收集化合物，并提供用户友好的搜索和过滤功能。访问地址为 https://zinc.docking.org。

2.2.2 药物性质数据集

在药物性质预测中，将预训练好的药物 embedding 编码器迁移到下游任务中的药物性质数据库 MoleculeNet[66]进行分子性质预测任务。MoleculeNet 是一个针对分子机器学习的公共数据库，由深度学习和化学专家联合开发。该数据库提供了超过 20 个基准数据集，涵盖了 800 多个各种分子属性的预测任务，如量子力学、物理化学性质、生物大分子的生物物理亲和性及活性、和对人体的宏观的生理影响等层次。这些数据集的目的是评估不同机器学习模型在不同任务上的性能。MoleculeNet 还提供了方便的数据加载框架、特征化方法、数据拆分方法和学习模型。

2.2.3 药物分子表征

分子表征对于分子建模和识别分子的各种理化性质至关重要[67-69]。但如何进行分子表征也是一个挑战，因为它是建立模型的主导因素之一。现有的部分模型那个从原始数据中学习与任务相关的特征，并将分子表征细化为一个标准。但在早期药物再利用阶段，只能依赖于由科学文献、专利和临床试验结果。后续随着人工智能的发展，为分子表征的构建及使用提供了新思路。在分子建模

中，有两种分子表征方式被广泛应用。其中一种是基于特征工程的化学指纹，每个比特代表某种生化特性或子结构的存在或缺失，将结构信息或特性转换为固定长度的向量[70]。例如，PubChem 指纹[71]和扩展连接指纹[72]是常用的分子表示。PubChem 指纹通过人工预先设置分子子结构，涵盖了广泛的不同子结构和特征，是 PubChem 用于相似性搜索和邻接的指纹。扩展连通性指纹(ECFPs)为通过迭代的方式不断地结合分子中的非氢原子初始特征与其邻居原子特征，在到达特定直径结束。然而，大多数化学指纹依赖于领域知识，只包含特定于任务的信息，这通常导致应用于下游任务时性能有限。另外一种分子表征方式是基于 SMILES 的分子图构建。SMILES 通过一条 ASCII 串来描述分子的组成和化学结构。为了使用分子二维或三维的内部结构信息，SMILES 的分子表征可以自然地建模为图，其中节点和边分别表示原子和化学键。

2.3 深度学习算法

2.3.1 图自监督学习

图自监督学习(Graph Self-supervised Learning, GSSL)是一种利用未标记的图数据来学习图结构和节点特征的深度学习方法。图自监督学习的目标是通过设计一些预测任务，允许模型自动从数据中提取有用的信息，而无需人工标注，并生成适合于各种下游任务的节点或图表示。图自监督学习可以分为两种类型：基于生成模型(Generative Model)的学习和基于对比学习(Contrastive Learning)的学习。

基于生成模型的方法是对生成问题的模型进行优化，使生成的数据尽可能接近真实数据。模型生成的常见任务包括图生成(Graph Generation)，即根据给定条件或无条件地生成新颖合理的图结构；属性生成(Attribute Generation)，即根据给定的条件或无条件的生成新颖和合理的节点属性；子图生成(Subgraph Generation)是指根据给定的条件或无条件地生成新颖合理的子图结构。基于生成模型的方法包括变分自编码器(Variational Autoencoder)和生成对抗网络(Generative Adversarial Network)。

基于对比学习的方法是通过构造正样本对和负样本对来优化二元分类问题中的模型，使正样本对之间的表示相似度高，而负样本对之间的表示相似度低。常见的对比学习任务包括图重构(Graph Reconstruction)，即基于部分图结构或受损图结构预测完整或原始图结构；属性遮掩(Attribute Masking)，是一种基于部分或随机的节点属性来预测完整或原始节点属性的方法；上下文预测(Context Prediction)，即基于节点及其邻居或路径的信息，对其他相关节点或子图进行预

测；聚类对齐(Cluster Alignment)表示基于同一图像在不同视角或域的跨视角或跨域表示匹配。目前，有许多基于对比学习的图自监督学习模型被提出，比如MoCo[48]、SimCLR[51]、MoCo V2[73]、BYOL[74]等。

图自监督学习具有以下优点：第一是可以充分利用大量未标注数据，减少人工标注的成本和时间；其次，提高了模型的泛化能力和鲁棒性，降低了过拟合风险和噪声干扰；第三，它能增强模型的表现力和解释性，能够数据揭示潜在的、多层次的语义信息。

2.3.2 图变分自编码器机制

近年来，变分图自编码器（Variational Graph Autoencoders，简称 VGAE 或 GVAE）[75]作为一种基于变分自编码器的无监督图学习框架出现。变分图自编码器是一种用于处理图数据的无监督学习框架。它可以从已知的图结构中学习节点的低维向量表示，并可以重构图的邻接矩阵或预测缺失边。

变分自编码器(Variational Auto-Encoders，VAE) [76]是一种融合变分贝叶斯推理和神经网络的深度生成模型，它提供了一种描述潜在空间变量的通用概率模型。VAE 是一种生成模型，它可以从高维数据中学习潜在变量的概率分布，并从分布中收集新数据。VAE 分为编码器和解码器两部分。编码器将输入数据映射到潜在空间中的隐式向量，该向量服从多元高斯分布，其均值和方差由编码器输出。解码器需要隐藏的向量作为输入并重建原始数据。

图变分自编码器的基本思想是将 VAE 迁移到图的领域。VGAE 扩展 VAE 来学习图结构数据的潜在表示。它使用一个 GCN 作为编码器的骨干，学习潜在的节点级和图级表示，解码器使用它来重建图。

VGAE 与 VAE 的模型构建的区别在于:（1）输入数据是图结构，而不是像素或特征向量;（2）编码器使用图像卷积网络(GCN)来捕获节点间的拓扑关系和特征信息;（3）解码器利用内积或其他函数计算节点对之间的相似度，并根据相似度阈值重构邻接矩阵或预测边是否存在。

图变分自动编码器可以有效地处理稀疏、不完整、异构和动态的图结构数据，VGAE 生成具有良好聚类特性和表达能力的节点向量表示，并利用潜在空间的随机性，增加模型的鲁棒性和泛化能力。此外它可以用作训练的方法来为其他基于提供初始化参数或特征输入神经网络的任务。VGAE 在 miRNA -疾病关联预测[77]、lncRNA -疾病关联预测[78]等生物信息学任务中取得了进展。然而，VGAE 还没有被应用于预测微生物与药物的关联。

2.3.3 图注意力机制

图注意机制的起源可以追溯到一个在 20 世纪 90 年代反复出现的视觉图形思想，但其真正意义上的使用是在 2014 年 Google Mind 团队将一种注意力机制引入到循环神经网络模型上来帮助模型对图像进行分类[79]。如今图注意机制广泛应用于许多领域，如社会网络分析、知识图推理、生物信息学、计算机视觉等。图注意机制是将注意机制引入图神经网络中，实现邻接节点加权聚集的一种方法，通过学习邻接节点之间的相关性，提高图卷积网络的表示性和鲁棒性。图注意机制的基本算法流程 是：在一个图中，对于给定的节点，它的特征向量可以通过其相邻节点特征向量的加权平均得到，其中权重通过注意函数计算。注意函数只需要满足非负、归一化等一些基本条件，并不拘泥于固定的形式。

图注意力机制根据对图结构数据学习方法的不同，可以分为两种类型：基于空间域的图注意力网络和基于谱域的图注意力网络。（1）基于空间域的图注意力网络是直接在图结构上进行操作，通过对邻居节点的特征进行加权求和来更新当前节点的特征，其中权重是由一个共享的参数矩阵和一个注意力函数来计算得到的。这种类型的网络能够学习所有形式的图结构，并通过多头注意力机制来增强模型表达能力，无需提前设定滤波器或者正则化项来计算节点特征。但正是因为如此，导致其计算复杂度较高，难以处理大规模或者动态变化的图数据，出现了一定的局限性。（2）基于谱域的图注意力网络是在图信号处理的框架下进行操作，通过对拉普拉斯矩阵或者其变换形式进行傅里叶变换或者切比雪夫多项式展开，然后利用注意力机制来学习不同频率分量之间的关系，从而得到节点特征的更新。基于谱域的图注意力网络可以有效地利用全局信息和频域特征，而且计算复杂度较低。然而基于谱域的图注意力网络也存在一些问题，比如需要人工提前固定输入输出尺寸，难以处理非欧几里得空间或者异构数据。两种方法各有优缺点，需要结合实际任务进行抉择。

图注意机制中最典型例子是图注意网络(Graph Attention Network, GAT)，图注意网络是一种基于空间域的图注意网络，GAT 通过多头自我注意来实现多个不同邻居的聚合更新自身节点特征。GAT 可以处理不同类型和大小的输入图，并可以自适应地学习节点之间的依赖关系。本研究选择使用图注意机制来提取分子图的特征的主要原因是：首先 GAT 能够有效捕获图结构中的复杂依赖关系，并动态调整相邻节点之间的权值，可以有效地处理分子图中存在的噪声和孤立的节点，因为它会赋予这些离散点更低的权重，使得模型在学习过程中忽略这些会导致性能下降的节点。此外图注意力网络通过使用了多头自我注意来融合不同方面的信息，可以实现对分子图结构数据进行多尺度和多角度的特征提取，

最后图注意力网络由于不会对输入图做太多的假设或限制，可以灵活地适应不同的任务和场景，具有更好的鲁棒性和泛化性。

2.3.4 对比学习算法

对比学习是一种自监督学习或无监督学习的机器学习方法。对比学习的基本思想是学习一个能够使得同一样本的变换或增强样本在特征空间中尽可能接近同时尽可能远离不同样本的编码器。对比学习的来源可以追溯到心理学和神经科学中的视觉感知理论，例如 Hebbian learning[80]和 slow feature analysis[81]。这些理论认为，视觉系统通过对比不同刺激之间的相似性和差异性来提取有意义的特征，并且这些特征应该具有稳定性和鲁棒性。同时受到了信息论和统计物理中最大熵原理和最小互信息原理等概念的启发，对比学习也引入了目标函数和损失函数等这些有效的概念。此外，对比学习还借鉴了机器学习领域中其他相关的方法，例如度量学习、多示例学习、多任务学习、半监督学习等。近年来，随着深度神经网络和大规模数据集的发展，对比学习在计算机视觉、自然语言处理等领域取得了显著的进展和应用。对比学习的过程主要包括以下三个步骤：

（1）数据增强。在对比学习中，数据增强手段是非常重要的，因为它可以提高样本的多样性和学习难度，从而促进模型的泛化能力。通常会对数据集中的每个样本选择执行两个或两个以上的随机变换或增强手段获取一组样本对。当我们需要对不同领域的数据进行处理时，通常需要使用不同的数据增强手段。例如，在处理图像数据时，我们通常会使用旋转、剪裁、遮盖、加噪声等手以改变图像的外观和视角，但不改变其图像信息。在处理自然语言数据时，我们会使用替换、插入、删除词语等手段，以改变句子的结构和表达方式，但不改变其语意内容。在处理图结构数据时，我们会使用遮掩点或边、特征扰动等手段，以改变图的拓扑结构和节点属性，但不改变其全局特征。在处理知识图谱数据时，可以使用 K-阶邻居作为负样本的方式进行数据增强，以利用知识图谱中的隐含关系信息。

（2）特征编码。当进行对比学习时，重要的是要设计一个合适的编码器，该编码器可以将原始数据映射到低维特征空间中。在这个低维特征空间中，同类数据之间的特征距离应该尽可能小，不同类数据之间的特征距离则应该尽可能大。常见有三种主要的特征编码方式：基于投影头的编码方式，在基础编码器的输出层后增加一个或多个全连接层，将高维度特征向量投影到较低维度子空间中；基于预测头的编码方式，在基础编码器和投影头之后增加一个预测头，用于从一个视图生成另一个视图的特征向量；基于协议网络的编码方式，在基

础编码器后增加一个协议网络，用于从每个类别中抽取代表性特征向量作为协议，并根据数据与协议之间的距离计算相似度或损失函数。这三种编码方式分别用于不同的对比学习任务和应用领域，如增强视图之间的互信息、利用类别信息指导对比学习、提高分类精度和鲁棒性等。例如，SimCLR[51]、MoCo[48]、SimCLRv2[50]、BYOL[74]、ProtoNCE[82]都采用了这些编码方式。

（3）相似度计算和损失函数优化。核心思想是通过对比不同样本之间的相似度来学习数据的特征表示。为了实现这一目标，对比学习需要定义一个合适的相似度计算方式和一个有效的损失函数优化方式。相似度计算方式是指如何衡量两个样本之间的相似程度。常见的相似度计算方式有欧氏距离、余弦相似度、向量内积等。而余弦相似度和向量内积通常被认为是更适合用于对比学习的相似度计算方式，因为它们可以消除特征向量长度对结果的影响，并且可以简化损失函数形式。损失函数优化方式是指如何根据相似度计算结果来更新模型参数。常见的损失函数优化方式有 InfoNCE[83]、NT-Xent[51]等。它们的优缺点不同，可能会导致不同的收敛速度和泛化能力。例如，InfoNCE 和 NT-Xent 都需要构造正负样本对，并且需要调节温度参数来控制分布平滑程度。而 InfoNCE 是在一个固定大小的字典中选择负样本，而 NT-Xent 是在一个动态变化的 batch 中选择负样本。这意味着 NT-Xent 可以利用更多的负样本信息，提高模型性能。

综上所述，对比学习适用于无标签或标签很少的数据集的学习方法，可以利用数据本身的信息来学习有意义的特征表示。为了实现这一目的，对比学习需要选择合适的数据增强方法和编码器模型，以确保正样本对之间有足够的相似性和差异性，负样本之间有足够的差异性。此外，为了平衡正、负样本之间的相似度分布和梯度更新效率，需要合理设置超参数，如批次大小、温度系数、投影头尺寸等。此外，对比学习也可以作为预训练模型被用于下游任务前的微调或特征提取，从而提高模型的泛化能力和准确性。因此，当数据集缺乏标签或标签很少，并且期望通过数据本身的信息来学习有意义的特征表示，那么对比学习是一个很好的选择。

2.4 模型性能评价指标

本研究使用了广泛应用于分类任务的性能度量指标来评估微生物与药物关联预测和药物性质预测任务的结果。这些指标包括 ROC 曲线下面积(ROC AUC)、PR 曲线下的面积(AUPR)、召回率(Recall)、F1 得分。

ROC 曲线下面积(ROC AUC)是一个常用的性能度量指标，它在分类问题中

被广泛应用。ROC 曲线是以假正例率(FPR)为横轴，真正例率(TPR)为纵轴的曲线。ROC 曲线反映了模型在不同阈值下对正负样本的分类能力。在 ROC 曲线下的面积即为 ROC AUC，ROC AUC 的值越接近 0 表示模型的性能越差。当 ROC AUC 的值为 0.5 时，表示模型的性能和随机猜测的效果相当；当 ROC AUC 的值为 1 时，表示模型的性能完美，即完全正确地分类了所有样本。

PR 曲线下面积(AUPR)也是一个常用的性能度量指标，它在分类问题中也被广泛应用。PR 曲线是以召回率(Recall)为横轴，精确率(Precision)为纵轴的曲线。PR 曲线反映了模型在不同阈值下对正样本的查准查全能力。在 PR 曲线下的面积即为 AUPR，AUPR 越大，说明分类器对正例的识别能力越强。与 ROC AUC 不同的是，PR 曲线和 AUPR 更适合评估样本不平衡的情况下的模型性能。

召回率(Recall)表示模型正确预测为正例的样本数占真实为正例的样本数的比例。召回率是一个重要的性能度量指标，因为它可以帮助评估模型的误报率和漏报率，对于不同应用场景下的模型性能评价都有着重要的作用。当召回率较高时，表示模型能够正确识别出更多的正例样本；而当召回率较低时，则表示模型漏识别了一些正例样本。

F1 得分是精确率(Precision)和召回率(Recall)的加权平均数，用于综合评估模型的预测能力。F1 得分可以同时考虑精确率和召回率，是一个更全面的评价指标。当 F1 得分较高时，表示模型具有更好的预测性能；而当 F1 得分较低时，则表示模型的预测能力有待提高。

2.5 本章总结

本章节从微生物与药物关联预测任务中数据数据集和常见的关联预测流程、药物性质预测任务中预训练和下游任务所使用的数据集以及药物分子表征、本研究使用到的深度学习算法和模型性能评价这四个方面介绍后续研究中使用到的相关理论知识和相关算法。例如在第三章对于微生物与药物关联预测的研究内容就是采用了图变分自编码器作为微生物和药物潜在特征提取的神经网络。而在第四章中采用先进的对比学习算法在收集到的 ZINC 数据库中进行预训练，并将训练好的编码器用于下游的药物性质预测，并结合注意力机制来解释分析了药物性质与化学亚结构的相关性，为药物研发和药物涉及提供了一种新的思路。本章所介绍的相关知识和算法为后续的实验提供了重要的理论支持。

第 3 章 基于多模态变分图嵌入的微生物-药物关联预测

随着生物领域和计算机领域的研究人员对微生物与药物复杂多样的互作机制的研究，出现了许多微生物-药物关联数据集，同时一些基于机器学习和深度学习的计算模型也被用于预测微生物-药物关联。但现有的计算模型大多数都通过整合多种类型的生物学数据（如药物结构相似性和微生物序列相似性信息）来构建关联网络，并对其提取有效的特征来预测新微生物-药物关联。但是由于关联网络的稀疏导致所提取的特征中存在噪声，从而出现不完美的预测结果。此外，这些模型很少探究这些作用在微生物上的药物间可能存在某些潜在的关联，并通过可视化的方法对其进行解释。

为了解决上述的问题，在本章中提出了一个新的模型 Graph2MDA，它是图变分自编码器(VGAE)和深度神经网络(DNN)的集成框架，用于预测微生物-药物关联。为了充分利用微生物和药物的多种生物信息属性，本研究构建了由微生物和药物及其关联组成的多模态属性图。对于药物，考虑了分子结构、药物相互作用谱和网络拓扑属性；对于微生物，考虑了基因组序列和功能注释。基于构建的多模态属性图，本研究训练 VGAE 学习整个图的信息和每个节点中存在的潜在表示。最后，使用 DNN 分类器学习潜在表示并输出代表微生物-药物关联存在的概率。本研究已经进行了广泛的性能评估，并证明 Graph2MDA 可以有效地利用多模态属性网络来提高性能。在其他独立数据集上，Graph2MDA 优于其他六种最先进的方法。特别是，本研究发现药物在潜在表征空间中表现出明显的聚类模式，且聚类与药物 ATC 分类显著一致。这验证了习得的潜在表征的可解释性。最后，本研究对两种微生物(即人类免疫缺陷病毒和结核分枝杆菌)和两种药物(即氯唑西林和芦荟凝胶)的案例研究也证明了本研究提出的模型的有效性和稳健性。

3.1 数据来源及预处理

本研究使用的微生物-药物关联来自 MDAD 数据库[36]。该数据库共收集了 2470 份临床报告，包括通过实验验证的 1373 种独特药物与 173 种微生物之间的关联。

药物与药物相互作用从 DrugBank 数据库检索[84]。本研究选择了与 MDAD 数据集中的药物相关的相互作用，得到了 5586 种药物与药物相互作用，涵盖 1228 种药物。

同样，本研究选择了 MDAD 数据集中的微生物相关的相互作用，从 MIND 数据库(http://www.microbialnet.org/mind_home.html)得到了 138 种微生物-微生物

相互作用，覆盖 123 种微生物。上述数据集的详细数量如表 3-1 所示。

表 3-1 微生物与药物网络的细节

网络	微生物	药物	关联
二分网络	173	1273	2470
药物与药物网络	—	1228	5586
微生物与微生物网络	123	—	138

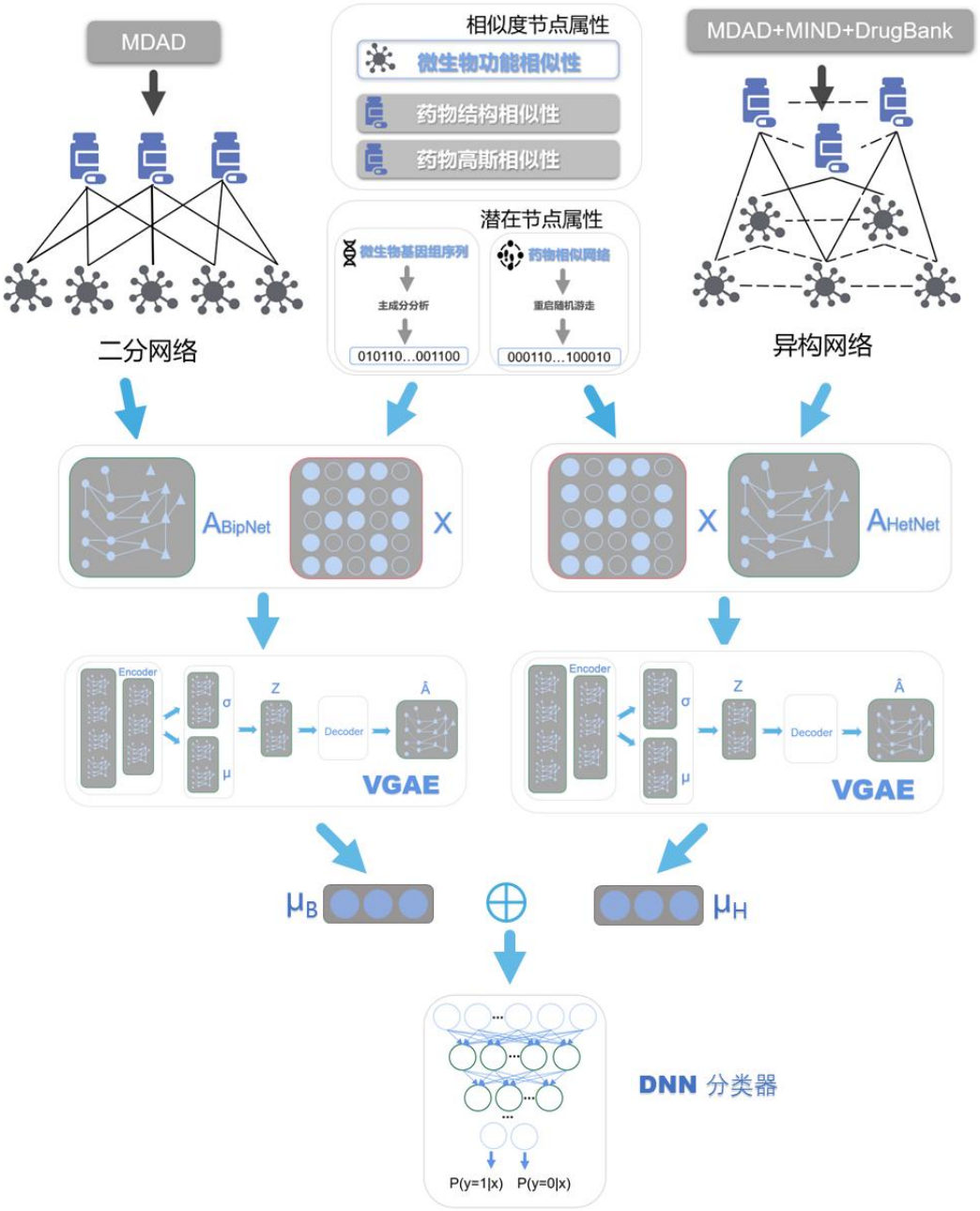

图 3-1 微生物–药物关联预测学习框架

3.2 多模态属性图的构建

考虑到药物和微生物的数据来源，本研究仅使用微生物-药物关联来构建 BipNet (BipNet)。具体而言，BipNet 网络的节点集只包括已知关联的微生物和药物，此外，本研究还构建了包含微生物-药物关联的异构网络(HetNet)，以及微生物的功能相似性和序列相似性、药物结构相似性和高斯核相似性、微生物与药物的潜在相似性，如图 3-1 所示。需要注意的是，HetNet 网络中包含 BipNet 网络中没有的附加节点。本研究用$A \in R^{nd\times nm}$表示网络的相邻矩阵，其中nd和nm分别代表微生物和药物的数量。如果节点i和j之间存在已知关联，则元素A_{ij}为 1，否则为 0。

3.2.1 药品相似属性构建

（1）药物分子结构相似。本研究使用 SIMCOMP2[85]工具计算药物结构相似性。SIMCOMP2 根据药物的化学结构信息来衡量药物之间的相似性。本研究计算成对药物结构相似性，然后构造相似矩阵$DS^{struct} \in R^{nd\times nd}$，其中$DS^{struct}(d_i, d_j)$表示药物和药物之间的分子结构相似性。

（2）药物高斯核相似度。在假设具有相似治疗效果的药物与相似微生物相互作用密切的前提下，本研究利用药物的高斯核相互作用来计算另一种相似度量。用$DIP(d_i)$表示药物d_i的药物-药物相互作用谱，即药物-药物相互作用矩阵的第i行表示药物与所有其他药物的相互作用，药物d_i与药物d_j之间的高斯核相似度$DS^{\text{gauss}}(d_i, d_j)$，公式如下:

$$DS^{\text{gauss}}(d_i, d_j) = \exp\left(-\mu\|DIP(d_i) - DIP(d_i)\|^2\right) \quad \text{公式(3-1)}$$

其中μ为标准化的内核带宽，定义如下:

$$\mu = \mu^{'}\left(\frac{1}{nd}\sum_{i=1}^{nd}\|DIP(d_i)\|^2\right)^{-1} \quad \text{公式(3-2)}$$

其中$\mu^{'}$为原始带宽，一般设为 1。

（3）综合药物相似性。本研究将分子结构相似度$DS^{struct}(d_i, d_j)$与药物高斯核相似度$DS^{struct}(d_i, d_j)$相结合，得到综合药物相似度$S_d(d_i, d_j)$，新药物相似度计算如下:

$$S_d(d_i, d_j) = \frac{DS^{struct}(d_i, d_j) + DS^{\text{gauss}}(d_i, d_j)}{2} \quad \text{公式(3-3)}$$

3.2.2 药物网络拓扑属性构建

由于几种药物相似性度量方面并不能包括所有的药物，本研究在药物-药物网络上运行 RWR 以获得另一种药物属性。RWR 可以有效地捕捉网络局部和全局拓扑信息的固有特征，被广泛应用于图像识别以降低噪声[86]。RWR 的正式定义为:

$$p_i^{(t+1)} = (1-\theta)p_i^{(t)}T + \theta p_i^{(0)} \quad \text{公式(3-4)}$$

其中θ为重启概率，T为转移概率矩阵，$p_i^{(0)} \in R^{n\times 1}$表示第$i$个节点的启动概率向量，$p_i^{(t)} \in R^{n\times 1}$表示节点$i$在$t$时刻移动到其他节点的概率。将 RWR 收敛后，得到每种药物的概率分布向量。因此，将概率分布向量作为网络拓扑属性矩阵$F_d \in R^{nd\times nd}$。通过对结构特征的整合，有效丰富了离群节点的属性，得到了信息量大、可解释的潜在表征。

3.2.3 微生物功能相似性构建

本研究使用 Kamneva 工具[87]来计算微生物的功能相似性。为了获得两种微生物之间的功能相似性，构建了微生物蛋白质-蛋白质功能关联网络，其中节点代表基因组编码的任何基因家族，链接代表基于最新 STRING 数据库的遗传邻居评分[88]。然后，用连接两个微生物的链接得分与两个微生物基因家族的所有链接得分之和的比值来计算微生物的功能相似性。最后用矩阵$S_m \in R^{nm\times nm}$表示微生物函数的相似度，其中$S_m(m_i, m_j)$表示微生物与微生物的相似度。

3.2.4 微生物序列属性构建

从 NCBI 数据库下载了微生物的 FASTA 基因组序列，在 173 个微生物中找到了 131 个。原始基因组序列采用 one-hot 编码，所有序列均用 0 填充，使所有序列长度相同。对于在 NCBI 中没有发现序列的微生物，本研究使用其他已知微生物的均值来代替。最后，本研究使用主成分分析(PCA)提取微生物的主要维度特征[89]。微生物序列属性用 k 维矩阵表示为$F_m \in R^{nm\times k}$。

3.2.5 多模态属性构建

为了使微生物(药物)的描述值在多种测量方法之间具有可比性，本研究分别对药物的基于相似性的属性矩阵S^d和基于拓扑的属性矩阵F^d以及微生物的S^m和F^m进行了归一化，然后构建了微生物和药物的多模态属性。事实上，本研究可以通过组合上述不同的属性来构造一些模态。为了简单起见，本研究只考虑两种不同的模态。第一种仅仅基于相似度度量，用$X_{\text{similarity}} \in R^{(\text{nd+nm})\times(\text{nd+nm})}$

表示。定义如下:

$$X_{similarity} = \begin{bmatrix} 0 & S_m \\ S_d & 0 \end{bmatrix} \quad \text{公式(3-5)}$$

另一种模式整合了从药物-药物相互作用网络中提取的次要特征和微生物序列的主要成分。因此，将其称为次要属性，用表示:

$$X_{secondary} = \begin{bmatrix} 0 & F_m \\ F_d & 0 \end{bmatrix} \quad \text{公式(3-6)}$$

药物和微生物的两种不同属性被单独或组合作为 VGAE 学习框架的输入，本研究在实验中评估了它们对预测性能的影响。

3.3 预测模型

本研究提出了一种基于 VGAE 和深度神经网络(DNN)的集成框架 Graph2MDA 模型，在多模式网络中预测微生物和药物之间的关联。如图 3-1 所示，VGAE 通过将从原始数据构建的图和节点属性作为输入来学习潜在表示(嵌入)。深度神经网络(DNN)分类器接收 VGAE 学习到的嵌入来预测微生物与药物的关联。需要注意的是，当 BipNet 和 HetNet 同时输入时，本研究使用了两个 VGAE，将输出的嵌入向量串联起来，形成 DNN 分类器的输入。

3.3.1 基于图变分自编码器的 Graph2MDA

Graph2MDA 模型中定义了一个无向图 G=(V, E)，其中$V = \{V^{(m)}, V^{(d)}\}$为微生物节点和药物节点集合，每条边表示一对不同类型节点(微生物和药物)或相同类型节点(两个微生物或两种药物)之间的关联。用 A 表示G的邻接矩阵，用$X_{similarity}$或$X_{secondary}$或两者的组合形式表示的属性矩阵。如图 3-1 所示，GCN 编码器通过聚合来自其局部邻域节点和属性本身的信息，将图中的每个节点v_i转换为低维潜在表示。接下来，解码器尝试重构节点对v_i和v_j对应的邻接关系。编码器和解码器通过在异质网络中微生物和药物之间的渐进式信息传播来优化潜在表征。最后，将学习到的潜在表征输入 DNN 分类器，以预测新的微生物-药物关联。

(1)编码器。编码器是一个两层 GCN[90]，它以网络邻接矩阵和节点属性矩阵作为输入和输出潜在变量。具体来说，本研究尝试建模条件概率分布$q(Z|X, a)$采用两层 GCN 正态分布，如下所示:

$$q(Z|X, A) = N(Z; \mu, \sigma^2 I) \quad \text{公式(3-7)}$$

其中μ和σ分别是高斯分布相对于潜在变量的均值和方差。是单位矩阵。两

层 GCN 定义如下:

$$\overline{\mathrm{X}} = \mathrm{GCN}(\mathrm{X}, \mathrm{A}) = \widetilde{A}ReLU(\widetilde{A}\mathrm{X}\mathrm{W}^0)\mathrm{W}^1 \quad \text{公式(3-8)}$$

其中

$$\widetilde{A} = D^{\frac{1}{2}}(A + I)D^{\frac{1}{2}} \quad \text{公式(3-9)}$$

W^{i}表示本研究需要训练的第 i 个 GCN 层的参数，$ReLU = max(0，\cdot)$为元素级激活函数，D为A的度矩阵，$\widetilde{A}$为对称归一化邻接矩阵。A首先通过对称归一化来保持特征向量的大小，使所有行的和为 1。通过 GCN 输出，本研究计算了高斯分布的均值和方差:

$$\mu = \mathrm{GCN}_{\mu}(\overline{\mathrm{X}}, \mathrm{A}) = \widetilde{A}\overline{\mathrm{X}}\mathrm{W}_{\mu} \quad \text{公式(3-10)}$$

$$\log\sigma = \mathrm{GCN}_{\sigma}(\overline{\mathrm{X}}, \mathrm{A}) = \widetilde{A}\overline{\mathrm{X}}\mathrm{W}_{\sigma} \quad \text{公式(3-11)}$$

其中W_{μ}和W_{σ}代表本研究需要训练的μ层和σ层的参数。一旦本研究获得μ和σ的值，就可以使用重新参数化技术[76]从分布$q(Z|X, A)$中采样，即可以通过以下公式计算:

$$z_i = \mu + \sigma * \varepsilon_i \quad \text{公式(3-12)}$$

其中$\varepsilon_i \sim N(0,1)$。

直观地，GCN 编码器逐步聚合来自邻居(包括微生物和药物)的信息，以更新每个节点的属性，以便每个节点(微生物或药物)的潜在变量产生集成结构和属性信息的信息表示。这对于注释很少的微生物(或药物)特别有用。

(2)解码器。考虑到编码器学习到的潜变量Z已经携带了足够的信息，解码器运行一个简单的内积来重构邻接矩阵A。形式上，设$p(A_{ij}|z_i, z_j)$是给定潜变量z_i和z_j在节点i和j之间有边的条件概率，则有

$$\mathrm{p}(\mathrm{A}|\mathrm{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(A_{ij}|z_i, z_j) \quad \text{公式(3-13)}$$

其中$p(A_{ij}|z_i, z_j)$定义为

$$p\left(A_{ij}\middle|z_i, z_j\right) = \varphi(z_i{}^T z_j) \quad \text{公式(3-14)}$$

其中 $\varphi(\cdot)$是逻辑函数。本研究使用的逻辑函数变换内积z_i和z_j,来表示微生物和药物间存在相互作用的概率。解码器的输出$\widehat{\mathrm{A}}$是邻接矩阵A的相似值，本研

究通过不断的优化模型使它们尽可能接近

(3)损失函数。损失函数包括两部分，第一部分是输入A和输出$\widehat{A}$之间的二元交叉熵，第二 部分是 q(Z|X, A)与 p(A|Z)之间的 KL 散度：

$$L = E_{q(Z|X,A)}[logp(A|Z)] - KL[q(Z|X,A)||p(Z)] \quad \text{公式(3-15)}$$

损失函数帮助本研究的模型去权衡重建输入网络的准确度和潜在变量匹配的紧密程度$p(Z)$。根据 VAE 中的设定，本研究假设$p(Z)$~$N(0,1)$。在训练过程中，本研究使用随机梯度下降训练 VGAE 以最小化损失函数。

3.3.2 DNN 分类器

本研究使用每个网络训练出来的潜在特征表示构建一个嵌入特征矩阵，并训练一个完全连接的 DNN 作为最终的监督分类器。DNN 分类器由三种类型的层组成：一个输入层、几个隐藏层和一个输出层。DNN 的输入由输入层节点接收，通过多个隐藏层以非线性方式转换，并成为输出层的最终输出。每个隐藏层都应该根据前一层的输出来提取越来越多的概括性特征，通过堆叠多个隐藏层，DNN 可以学习从输入到输出的极其复杂的非线性函数映射。因此 DNN 在分类任务中表现出了卓越的区分能力。

本研究将微生物与药物之间的关联预测设定为二元分类任务，因此采用多层全连接 DNN 作为分类模型。DNN 分类器将微生物和药物的潜在表征作为输入，通过多个非线性隐藏层进行信息转换，输出代表两个节点之间存在关联的概率。交叉熵被用作损失函数，首先使用自适应优化器 Adam 进行预训练，然后使用 SGD 优化器进行微调。由于超参数对 DNN 分类器性能的影响，本研究在后续的实验中对超参数对分类器性能的影响进行了详细的实证评估。根据 VGAE 中的设置，本研究假设 p(Z)~N(0,1))。在训练过程中，使用 SGD 算法对 VGAE 进行训练，使损失函数最小化。DNN 分类器的性能可以在没有许多复杂神经网络结构的情况下非常出色，因为嵌入已经包含足够的信息并且在学习的低维向量空间中具有很高的代表性。

3.4 实验结果

3.4.1 超参数敏感性分析

本模型中的神经网络主要包括 VGAE 和 DNN 分类器，其中有几个超参数。为了探索超参数的影响，本研究在 MDAD 数据集上执行了 10-flod CV，以观察性能的变化趋势，从而优化它们的值。本研究首先研究了隐藏层数对 VGAE 编码器和 DNN 分类器的影响。由于过多的隐藏层会导致过拟合和无意义的潜在

表示，本研究在 VGAE 编码器和 DNN 分类器中考虑了常见的四个隐藏层。本研究采用塔形 DNN 结构，何等人认为塔形 DNN 模型可以在更高的层次上提取更紧凑、更有鉴别性的特征[91]。

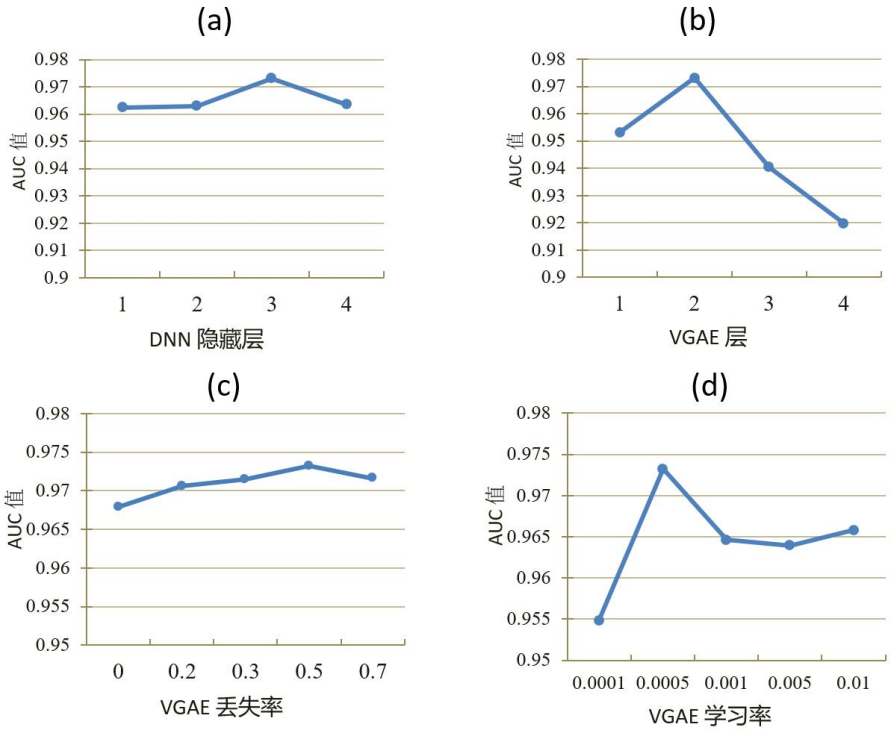

图 3-2 超参数对性能影响的实证分析。(a)-(d)子图分别表示了 DNN 和 VGAE 的隐层数对应的 AUC 值，以及 VGAE 的学习率和丢失率

本研究分别验证了 1 到 4 层的隐藏层数，隐藏层的大小分别设置为 1024、512、256、128。如图 3-2a 所示，随着 DNN 分类器隐藏层数的增加，AUC 值增加，在 3 层时性能最好，层数超过 3 层后 AUC 值下降。因此，本研究选择了三层塔式结构作为 DNN 分类器模型。此外如图 3-2b 所示，当 VGAE 隐藏层数为 2 时，性能最高，一旦层数超过 2，性能急剧下降。此外，本研究还探讨了 VGAE 编码器中丢失率(dropout)和学习率的影响。如图 2c 和 d 所示，当丢失率概率大于 0.5 时，性能略有下降，因为过高的丢失率会阻碍信息在隐藏层之间的传播。本研究还注意到学习率过低或过高都会导致性能下降，当学习率为 0.0005 时 AUC 最高。因此，本研究将 VGAE 编码器的丢失率设置为 0.5，学习率设置为 0.0005。

此外，还有几个重要的参数会影响模型的性能，例如 VGAE 训练的 epoch 和模型的时间复杂度。为了探索 epoch 的影响，本研究在 MDAD 数据集上执行了 10 倍 CV，以观察性能的变化趋势，从而优化它们的值。对于模型上的多模

态属性图，VGAE 上的 epoch 数在模型中起着重要的作用。本研究执行了两种不同的输入图，包括二分网络、异构网络。本研究评估了两种不同的输入图网络在不同的网络 epoch 组合模式下该模型在{100,200,300,400}范围内的性能，实际表现如表 3-2、表 3-3 所示，BipNet 在 epoch 为 300 时性能最佳，HetNet 在 epoch 为 200 时性能最佳。同时也评估了组合输入后最优的 epoch，从表 3-4 可以看出该模型在分别在 BipNet 训练 epoch 为 300、HetNet 在 epoch 为 200 时性能最佳。

表 3-2 二分网络在不同 epoch 组合模式下模型的性能

Epoch(BipNet)	AUC	AUPR	Recall	F1
100	0.9518	0.8880	0.8392	0.5964
200	0.9682	0.9051	0.9492	0.6204
300	0.9729	0.9280	0.9570	0.7091
400	0.9603	0.9006	0.9305	0.6535

表 3-3 异构网络在不同 epoch 组合模式下模型的性能

Epoch(HetNet)	AUC	AUPR	Recall	F1
100	0.9139	0.7788	0.8504	0.4289
200	0.9348	0.8311	0.8906	0.5521
300	0.9204	0.7762	0.8648	0.4245
400	0.9139	0.7809	0.8474	0.4629

表 3-4 二分网络与异构网络组合在不同 epoch 组合模式下模型的性能

Epoch(HetNet)	AUC	AUPR	Recall	F1
100+100	0.9128	0.9128	0.8357	0.6525
100+200	0.9556	0.8965	0.9198	0.6809
100+300	0.9495	0.8941	0.9071	0.6666
200+100	0.9571	0.9062	0.9226	0.6831
200+200	0.9602	0.9240	0.9213	0.7344
200+300	0.9570	0.8957	0.9224	0.6833
300+100	0.9585	0.9044	0.9258	0.6711
300+200	0.9721	0.9310	0.9531	0.7751
300+300	0.9592	0.8961	0.9268	0.7516

此外，为了评估本研究模型的时间复杂度，本研究在两个不同大小的数据集(即 aBiofilm 和 DrugVirus)上进行模型，以评估其运行时间。数据集的大小可以在表 3-7 中看到。对于每个数据集，本研究将完整的微生物-药物关联作为训练数据(例如，aBiofilm 的关联对为 2884 对，DrugVirus 的关联对为 933 对)。如图 3-3 所示，本研究观察到随着 epoch 的增加，运行时间也呈线性增加。例如，在 epoch 200，aBiofilm 的运行时为 2383 秒，而 DrugVirus 的运行时为 246 秒。

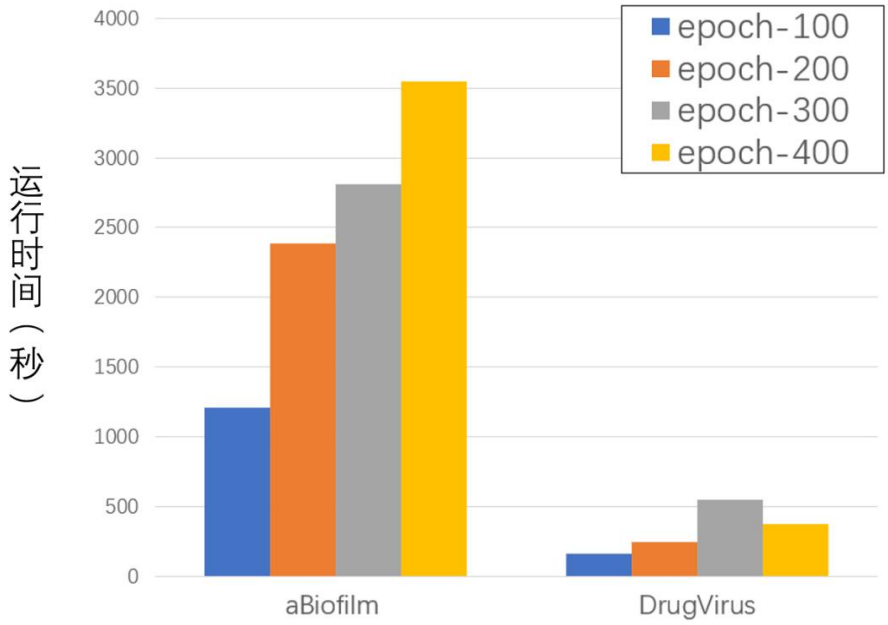

图 3-3 模型在不同数据大小下的运行时间

3.4.2 与 SOTA 方法的性能比较

为了对本研究提出的方法的性能进行基准测试，将本研究的方法与现有的微生物-药物关联预测方法以及生物信息学领域中链接预测问题的一些方法进行了比较。这些对 比方法简要总结如下:

（1）EGATMDA[40]通过构建一种新的集成图注意网络来预测人类微生物-药物关联。

（2）GCNMDA[42]是一种基于 GCN 和条件随随机场预测人体微生物-药物关联的方法。

（3）HNERMDA[41]提出了一种基于异构网络的嵌入表示，用于预测人类微生物-药物关联。

（4）HMDAKATZ[34]是专门针对基于 KATZ 指标的微生物药物预测而开发的。

（5）NTSHMDA[35]采用 RWR 模型预测微生物疾病相关性。

（6）LAGCN[43]是一种基于卷积神经网络的模型，具有预测药物-疾病关联的注意机制。

本研究在 MDAD 数据集上运行这些方法，并使用它们的默认模型特定参数，并调整这些竞争方法的深度学习定义超参数。本研究对所有方法进行了 10 次交叉验证。已知的微生物-药物关联作为阳性样本，训练集和测试集中的阴性样本随机生成。10 次折叠的平均预测精度被用作最终的性能度量。为消除随机抽样的偏差，该过程重复 10 次，最终 AUC 评分由 10 次重复的平均值计算。如图 3 所示，本研究的 Graph2MDA 模型的 AUC 值最高，为 0.9732，其次是 LAGCN，AUC 值为 0.9645，而 NTSHMDA 的 AUC 值最低，为 0.8893。为了综合比较，本研究在补充表 S6 中给出了这些方法的 AUPR 值。总体而言，基于深度学习的方法优于基于传统机器学习的方法。

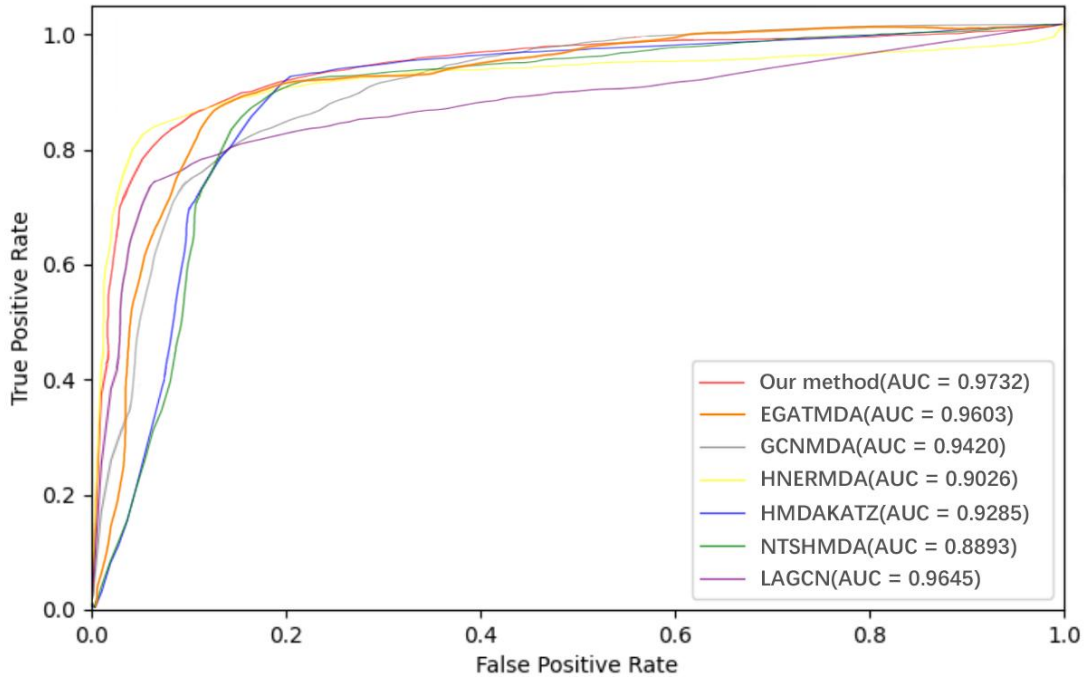

图 3-4 本方法与六种竞争方法在 MDAD 数据集上预测微生物-药物关联的 ROC 曲线和 AUC 值

本研究从两个方面总结了该方法的优点。首先，与其他基线方法相比，本研究基于微生物和药物及其已知关联的各种本体论和相似性信息构建了多模态属性图，使本研究的方法能够充分利用多种类型的属性和关联。特别是，本研究将拓扑结构集成到多模态属性图中，缓解了冷启动问题的影响，即某些微生物(或药物)只有一种或更少的相关药物(或微生物)导致结构非常稀疏。而其他一

些方法，如 KATZHMDA 和 NTSHMDA，则遇到了冷启动问题，因为它们只能利用过于稀疏的现有关联来识别新的关联。此外，由于相似度度量有一定的噪声，可能会影响基于图表示的算法，如 EGATMDA， GCNMDA，而本研究的方法受益于多模态属性图，可以减轻噪声相似度的影响。此外，本研究的方法得 益于骨干编码器 VGAE，它可以有效地解决冷启动问题。

3.4.3 模型消融研究

本研究进一步进行了消融实验，以评估多模态属性图对模型性能的影响。首先，本研究测试了三种不同的输入图，包括二分网络、异构网络和两者都有，对 MDAD 数据集进行消融研究。其中，二分网络仅包括微生物-药物组合，而异质网络包括微生物-药物、药物-药物和微生物-微生物组合。表 3-5 显示了本研究的模型在三个不同的输入图上实现的 AUC 值。为了进行综合对比，本研究同时进行了 5 倍和 10 倍交叉验证。本研究发现，当这两个网络都被输入到模型中时，可以获得最好的性能，这表明多源信息的组合有利于特征提取，以预测微生物和药物之间的关联。本研究的方法在异构网络上表现不如二分网络，本研究认为主要原因是异构网络中包含了大量与药物(微生物)关联较少的微生物(药物)节点，而这些节点不包含在二分网络中。在异构网络的图卷积过程中，这些类似离群点的节点导致信息传播不良。因此，异构网络的性能要低于二分网络。

表 3-5 不同输入图的性能比较

网络	5-flod CV	10-flod CV
二分网络	0.9477 ± 0.0015	0.9710 ± 0.0030
异构网络	0.9190 ± 0.0045	0.9358 ± 0.0056
二分网络+异构网络	0.9567 ± 0.0039	0.9732 ± 0.0037

表 3-6 不同节点属性性能比较

节点属性	5-flod CV	10-flod CV
相似属性	0.8940 ± 0.0204	0.9088 ± 0.0320
辅助属性	0.9512 ± 0.0022	0.9693 ± 0.0024
相似属性+辅助属性	0.9567 ± 0.0039	0.9732 ± 0.0037

本研究还评估了基于多模态节点属性模型的鲁棒性。在使用二分网络与异构网络的前提下，本研究对相似属性、辅助属性以及两者进行了独立测试。请

注意，当使用两个类型属性时，属性向量将连接到每个节点。5 折和 10 折交叉验证得到的 AUC 值如表 3-6 所示，当两个节点属性都输入模型时，性能最佳。同时，本研究发现，单独的辅助属性也获得了非常高的性能。本研究认为原因可能在于辅助属性包含高阶信息。

3.4.4 在两个独立数据集的性能评估

为了验证本研究模型的通用性，本研究将本研究的方法与其他两个独立数据集上的竞争方法进行了比较，这两个数据集是 aBiofilm[37]和 DrugVirus[38]。aBiofilm 数据库包含 1720 种独特的抗生物膜药物，针对 140 多种微生物。在过滤掉重复序列后，本研究得到了 2884 种微生物-药物关联，涵盖 1720 种药物和 140 种微生物。DrugVirus 数据库从药物数据库和相关出版物中手动收集了 175 种药物与 95 种人类病毒之间的关联，有 933 种临床或实验证实的药物病毒关联。这两个数据集的详细信息见表 3-7。

表 3-7 两个独立的微生物–药物联合数据集的细节

数据库	微生物	药物	关联
aBiofilm	140	1720	2884
DrugVirus	95	175	944

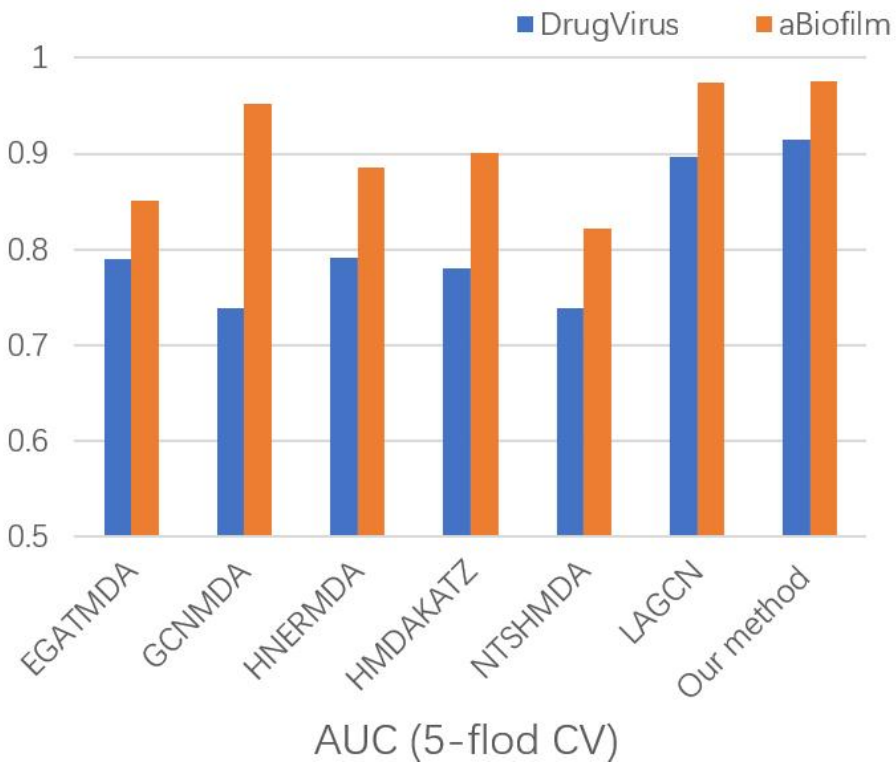

图 3-5 Graph2MDA 与六种竞争方法在 aBiofilm 和 DrugVirus 数据集上的性能比较

同样，本研究在 aBiofilm 和 DrugVirus 数据集上运行了所有竞争方法，并采用 AUC 值作为性能度量。由于 DrugVirus 数据集相对较小，本研究执行了 5 次交叉验证，以避免测试集太小。图 3-5 显示了 Graph2MDA 和六种竞争方法的性能。与其他方法相比，本研究的模型在两个独立的数据集上仍然表现出最好的性能。性能评价表明，Graph2MDA 是一种预测微生物与药物之间关系的有效且强大的计算模型。

3.5 潜在表征的解释研究

本研究的模型可以从原始输入图和节点属性中学习高级特征表示。为了挖掘可解释性，本研究使用 t-SNE[92]将学习到的药物潜在表征在 MDAD 数据集中可视化，t-SNE 是一种通过将高维特征嵌入到二维(2D)图像中实现高维数据可视化的工具。如图 3-6a 所示，散点图显示了使用原始属性的药物分布情况。可以看出，在训练前药物分布是混乱的，而通过本研究的模型学习到的潜在表征，药物呈现出清晰的聚类模式，如图 3-6b 所示。

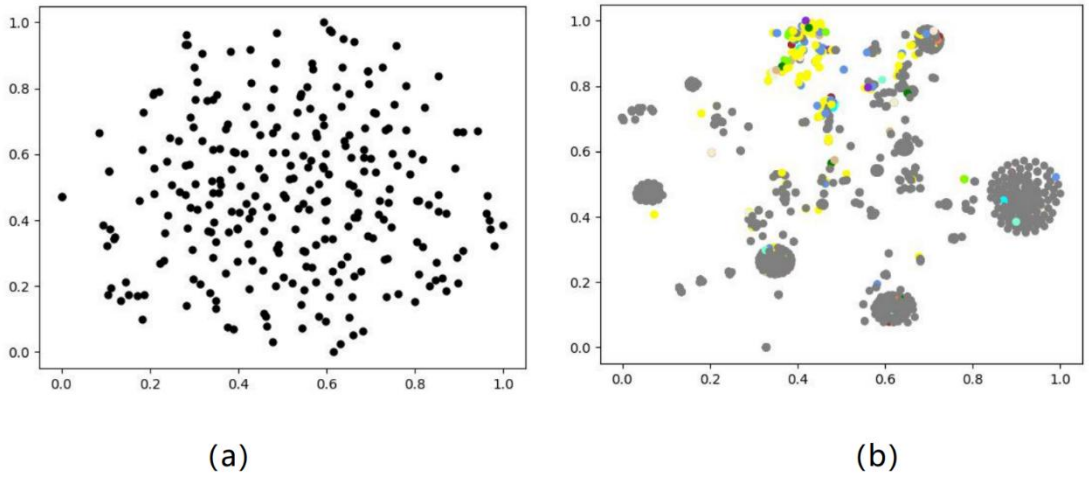

图 3-6 利用 t-SNE 工具将 MDAD 数据集中药物的原始属性和学习到的潜在表征可视化。(a) 药物分布在原始属性空间中可视化，(b)学习后表示空间中可视化药物分布，具有 ATC 编码映射的药物为彩色，其他药物为灰色

为了进一步探索潜在表征的意义，本研究使用药物的解剖学治疗化学(ATC)代码来验证聚类模式与 ATC 分类之间的一致性[93]。ATC 代码根据药物作用的器官或系统及其工作方式分配给药物。本研究手工整理了 MDAD 数据集中 1373 种药物的 ATC 代码，并成功映射了 269 种药物的 ATC 代码。根据 ATC 的分类系统，属于不同类别的药品用不同的颜色标注。为了进行比较，没有 ATC 代码的药物也用灰色表示，如图 3-6b 所示。本研究发现黄色散斑(黄色对应抗感染

药物系统使用)明显聚集在一起，并与其他簇分离。这一观察证实了学习到的药物特征是有信息的和可解释的，本研究大胆猜测，本研究的模型可以有效地学习到可以转化为药物药理功能的特征。

3.6 案例研究

为了进一步验证 Graph2MDA 的有效性，本研究将 Graph2MDA 应用于两种流行的抗菌药物芦荟凝胶和氯唑西林，以及两种微生物人类免疫缺陷病毒(HIV)和结核分枝杆菌作为本研究的案例研究。对于排名前 20 位的预测微生物或药物，本研究通过搜索 MeSH 和 DrugBank 来交叉核对它们的同义词，然后通过搜索 PubMed 文献来验证预测的微生物或药物关联是否已被报道。

表 3-8 前 20 位预测氯唑西林相关微生物及相关出版物

排名	微生物	证据
1	Bacillus cereusreus	PMID24876650
2	Bacillus subtilis	PMID25945113
3	Baker's yeast	PMID2089228
4	Burkholderia cepacia	NA
5	Candida albicans	PMID2713774
6	Candida dubliniensis	PMID16353125
7	Candida spp.	PMID21496537
8	Clostridium pasteurianum	NA
9	Enterobacter aerogenes	PMID22001269
10	Francisella novicida	NA
11	Helicobacter pylori	PMID10748053
12	Klebsiella planticola	NA
13	Klebsiella pneumoniae	PMID20597925
14	Micrococcus luteus	PMID7771695
15	Pantoea agglomerans	PMID33666040
16	Salmonella Typhi	PMID15490798
17	Schistosoma	PMID15490798
18	Staphylococcus aureus	PMID15490798
19	Streptomyces sp.	PMID6970744
20	Thermus thermophilus	NA

3.6.1 药物关联案例研究

氯唑西林是一种半合成青霉素类抗生素，广泛用于治疗 β-溶血性链球菌和肺炎球菌感染以及葡萄球菌感染。氯唑西林对大多数葡萄球菌的青霉素酶具有失活作用，并对许多产生青霉素酶的金黄色葡萄球菌和表皮葡萄球菌具有活性[94]。对于氯唑西林预测前 20 位的微生物，本研究也发现有 15 个(75%)微生物被研究，如表 3-8 所示。例如，Saengsai 等研究了氯唑西林对环烯酮内酯 Plummericin 对粪肠球菌和枯草芽孢杆菌的抗菌活性[95]。Orogade 和 Akuse 还报道了氯唑西林可以抑制金黄色葡萄球菌、血吸虫和伤寒沙门氏菌 50%的活性。

表 3-9 前 20 位预测芦荟凝胶相关微生物及相关出版物

排名	微生物	证据
1	Actinomyces oriss	PMID33354266
2	Aggregatibacter actinomycetemcomitans	PMID22466882
3	Bacillus cereus	PMID33466284
4	Bacteroides eggerthii	NA
5	Burkholderia pseudomallei	NA
6	Candida albicans	PMID 28781531
7	Citrobacter freundii	PMID30675330
8	Klebsiella variicola	NA
9	Listeria monocytogenes	PMID23477211
10	Proteus mirabilis	PMID19263248
11	Pseudomonas aeruginosa	PMID31391416
12	Pseudomonas putida	NA
13	Streptococcus gordonii	NA
14	Streptococcus mitis	PMID22087805
15	Streptococcus oralis	PMID22087805
16	Streptococcus parasanguinis	NA
17	Streptococcus pneumoniae	PMID25362808
18	Streptococcus pyogenes	PMID25362808
19	Streptococcus salivarius	NA
20	Streptococcus sanguinis	PMID22087805

长期以来，芦荟凝胶一直被用作促进伤口愈合的传统药物。它是一种天然产品，现在被用于化妆品行业。虽然有多种使用说明，芦荟的益处归因于叶子

凝胶中所含的多糖[96]。Kaithwas 等人[97]还研究了芦荟在便秘、炎症、癌症、溃疡和糖尿病中的使用。因此，在芦荟凝胶预测的 20 种靶向微生物中，已有 16 种微生物(80%)在之前的研究中被报道过。如表 3-9 所示，本研究列出了微生物的名称和报道微生物与芦荟凝胶之间关系的出版物的 PMID。例如，Fani M 等人研究了芦荟凝胶对从龋齿和牙周病患者中分离出的某些致龋细菌(变形链球菌)、牙周细菌(放线菌聚集杆菌、牙龈卟啉单胞菌)和机会性牙周细菌(脆弱拟杆菌)的抑制活性[98] 。

表 3-10 前 20 位预测结核分枝杆菌相关药物及相关出版物

排名	药物	证据
1	Amikacin	PMID29311078
2	Amprenavir	NA
3	Apramycin	PMID25136009
4	Azithromycin	PMID7849341
5	Calanolide A	PMID14980631
6	Cloxacillin	PMID25104892
7	Darunavir	PMID28193650
8	Dirithromycin	NA
9	Dolutegravir	PMID33315751
10	Efavirenz	PMID28874142
11	Gatifloxacin	PMID 16714850
12	Genistein	NA
13	Gentamicin	PMID22143521
14	Indinavir	PMID21442799
15	Lopinavir	PMID21442799
16	Minocycline	PMID30597040
17	Nevirapine	PMID2039216
18	Raltegravir	PMID30350998
19	Rilpivirine	NA
20	Ritonavir	PMID21442799

3.6.2 微生物关联案例研究

结核分枝杆菌是一种革兰氏阳性需氧菌。这种细菌可以侵入身体的所有器

官，但它是结核病的最常见原因。在预测针对结核分枝杆菌的前 20 种药物中，80%的药物得到了文献的支持。表 3-10 列出了出版物的药品名称和 PMID。例如，绝对浓度法显示阿米卡星对结核分枝杆菌[99]的活性高于卡那霉素和卷曲霉素。阿霉素是一种独特的氨基糖苷类药物，具有抗菌活性和耳毒性。在两个感染小鼠模型中，阿泊拉霉素对结核分枝杆菌和金黄色葡萄球菌显示出显著的抗菌效果[100]。

表 3-11 前 20 位预测人类免疫缺陷病毒(HIV)相关药物及相关出版物

排名	药物	证据
1	Amantadine	PMID30085647
2	Aminosalicylic Acid	PMID25114132
3	Amprenavir	PMID15748098
4	Bedaquiline	PMID30846058
5	Capreomycin	PMID25909847
6	Cloxacillin	PMID12092473
7	cyclophosphamide	PMID32430507
8	Daclatasvir	PMID28182611
9	Delavirdine	PMID11152019
10	Desipramine	PMID7649718
11	Efavirenz	PMID31095608
12	Erythromycin	PMID18719105
13	Fosamprenavir	PMID15748098
14	Fusidic Acid	PMID2443777
15	Gatifloxacin	PMID23843980
16	Nevirapine	PMID31017649
17	Pyrazinamide	PMID29095954
18	Rilpivirine	PMID29794818
19	Salicylic Acid	PMID21371895
20	Sinapic acid	NA

人类免疫缺陷病毒(HIV)是一种可以攻击人体免疫系统的病毒。HIV 会诱导关键 T 细胞失效和功能损伤，导致整个免疫系统功能障碍和缺陷，最终导致机会性感染和肿瘤[101]。本研究检查了排名前 20 的预测药物，其中 95%已经被确认，如表 3-11 所示。以红霉素为例，有报道称红霉素可通过调节 MAPK 活性诱

导 C/EBP β 的小亚型来抑制 HIV-1 在巨噬细胞中的复制[102]。

3.6.3 新型冠状病毒关联药物案例研究

2019 年新型冠状病毒病(SARS-CoV-2)大流行在全球范围内造成了大规模的健康危机，并颠覆了全球经济。为了测试本研究的方法预测对这种新型病毒具有治疗效果的药物的能力，本研究在 HDVD 数据集[39]上运行了本研究的模型，该数据集从与 COVID-19 相关的文献中收集了药物-病毒相互作用的词条，并建立了人类药物-病毒数据库。HDVD 数据集包括 34 种病毒、219 种药物和 455 种已确认的人类药物-病毒相互作用。本研究在表 3-12 中列出了治疗 SARS-CoV-2 的前 10 种预测药物。对于每种药物，本研究显示了预测评分、药物库登录号、规范名称和报告支持证据的 PubMed ID。在 10 种预测药物中，有 7 种药物(70%)之前已被报道过。

表 3-12 Graph2MDA 预测 SARS-CoV-2 的前 10 个预测药物

排名	注册编号	药物	分数	证据
1	DB00811	Ribavirin	0.93355	PMID34616188
2	DB01299	Sulfadoxine	0.90622	PMID32438446
3	DB00715	Paroxetine	0. 90405	PMID34416332
4	DB00608	Chloroquine	0. 84819	PMID34631362
5	DB13609	Umifenovir	0. 72018	PMID34539392
6	DB00756	Hexachlorophene	0. 66164	NA
7	DB11753	Rifamycin	0. 58912	NA
8	DB14761	Remdesivir	0. 53219	PMID34632942
9	DB01029	Irbesartan	0. 52509	PMID33735271
10	DB00644	Gonadorelin	0. 51181	NA

3.7 本章小结

在本章中，提出了一种基于图变分自编码器 VGAE 和深度神经网络(DNN)的集成框架 Graph2MDA，用于预测微生物与药物之间的相关性。实验结果表明，本研究提出的 Graph2MDA 模型优于现有的最先进的方法。本研究的主要贡献至少体现在三个方面：（1）构建了多模态属性图，有效整合了微生物和药物丰富的相似性和本体信息。（2）提出了一种基于 VGAE 和 DNN 的预测微生物与药物之间关系的新框架。与目前最先进的基于图神经网络和链接预测的微生物-

药物关联预测方法相比，本研究的方法在三个独立的数据集上获得了更好的性能。（3）本研究验证了学习得到的潜在表征在语义上与药物药理功能相关。具体而言，本研究发现药物在潜在表征空间中表现出明显的聚类模式，且聚类与药物 ATC 分类显著一致。

虽然本研究的方法已经取得了优异的性能，但本研究的方法至少可以从两个方面进行改进。首先，与整个药物-微生物关联空间相比，已知的关联数量非常有限，导致数据结构稀疏。结果，由于图卷积过程中节点之间的信息传播不足，从 VGAE 学习的嵌入表示仍然是次优的。为了解决这个问题，本研究考虑了遮掩机制来提高本研究的模型性能。更准确地说，对于没有监控信息的节点，在损失计算时使用遮掩，以限制噪声信息的传播。此外，由于本研究的方法可以通过多模态属性图集成各种类型的节点信息，本研究可以利用属性，如基于副作用的药物相似性，或新的网络信息，如微生物与疾病关联，疾病与药物关联，来提高性能。

第 4 章 基于注意力图遮掩对比学习的药物分子性质预测

在上一章的研究中提出了基于多模态变分图嵌入的微生物和药物关联预测模型 Graph2MDA。基于微生物与药物的生物特征与理化性质构建的多模态特征图来预测两者间的潜在关联，相较于其他几种主流的机器学习方法和深度学习都取得最优的预测性能表现。最后探索了药物潜在表示的含义，发现与微生物存在潜在关联的药物表现出明显的聚类模式。为了进一步探索药物作用于微生物的互作原理、耐药性等关联的底层机制，因此了解药物分子的生物理化性质也能够进一步促进微生物的研究，同时也推动着微生物相关的药物重定位、新药开发等相关领域的发展。

近年来，深度学习在自然语言处理[47]、计算机视觉[48]和图结构预测[90]等方面取得了显著的成功。许多研究将深度学习应用于化学建模[103-108]和药物发现[48, 109, 110]等。然而，全监督深度学习的性能依赖于大量手工标记的样本[104, 111]，例如本研究中将会使用的具有已知性质的分子，然而获取分子的性质只能通过耗时又昂贵的化学实验。同时，全监督模型应用于小型数据集时容易发生过拟合，训练后的模型的鲁棒性和泛用性较差。

对比学习通过参与最具挑战性的负样本的训练而更新后的网络能够更好地学习到更具有效的表示，能够很好地在下游任务泛化。目前主流的方法大都通过随机遮掩来生成对比视图。然而，随机遮掩的方法并不能引导编码器识别最重要的子结构。此外，在对比学习中负样本的复杂性和多样性对模型的学习性能有很大程度的影响。人们通常采取这两种方法来构建对比学习中所需负样本：一些方法会维护一个负样本队列同时采取 FIFO 的迭代方法更新[50]，其他方法中的负样本是选择训练批次中除去当前样本的其他样本[51]。因此，为解决上述问题。本研究希望通过结合对比学习和新颖的遮掩策略来推进分子性质预测工作。在这项工作中提出了一种注意力对比学习框架来预测药物分子性质。首先利用药物分子的 SMILES 表征来构建对应的分子图，将分子图输入到图注意网络 (GAT) 中提取药物分子的潜在表征。在此期间，根据图注意机制中对节点或边的注意力得分遮掩特定比例的节点或边来生成药物分子的增强视图。通过最小化原视图和增强图之间的对比损失以达到识别分子图中关键的子结构，进而得到一个能够高效地编码信息丰富的分子表示的编码器来迁移到下游任务中。本研究广泛的实验表明，通过本研究的方法学习的分子表示在各种下游分子性质预测任务中表现出最先进的性能。性能对比验证了本研究的方法优于其他竞争方法。此外，本研究还对其可解释性进行了探讨，发现注意权重揭示了重要子结构的重要性模式。

4.1 数据来源及预处理

为了满足预训练的大规模数据，从 ZINC 数据库[65]下载了相应的大规模数据。同时，为了证明经过在大规模数据集上预训练后模型的泛化能力，从 MoleculeNet 数据库[66]中获得了药物的生物物理学习和生理学等相关药物性质数据集。详细数据介绍如下。

4.1.1 用于预训练的药物分子大规模数据集

ZINC 是加州大学实验室由提供的一个免费的商业化化合物数据库，用于虚拟筛选。ZINC 收集整理了可购买的化合物与高价值的化合物连接起来，如药物、文献中注释的化合物等。本研究从 ZINC 数据库的两个药物分子数据集分别选取了两个不同规模的药物分子集作为预训练的数据。

第一个药物分子数据集是 in-vitro 数据集，其中包括 306 347 种已知或推断在直接结合试验中具有 10 μM 或更高生物活性的独特物质。in-vitro 数据集中的所有分子均用于本研究的评价实验。此外，为了对比不同数据规模对预训练的影响，本研究另一个预训练数据集是由 ZINC 中 now 数据集构建的，其中包括所有目前被库存及可立即交付的代理化合物。该数据集包含更多的独特的分子结构，共 9 814 569 个独特的分子，本研究随机选取该数据集中三分之一数量的分子，既 in-vitro 数据的 10 倍规模。

开源工具 RDkit[112]能够将药物分子数据集中的 SMILES 表征构建为节点既原子、边为化学键的分子图。同时通过 DeepChem 计算药物 SMILES 所包含的二值特征，包括原子自身信息、邻居原子信息以及分子结构信息等这些特征作为分子图中节点的特征。

4.1.2 用于下游任务的药物性质数据集

本研究选择了 MoleculeNet 中共包含 60 多个子任务的 7 个数据集作为下游任务。表 4-1 显示了每个数据集中分子的总数、任务的数量以及数据集的具体描述。这些数据集涵盖了生物物理学和生理学方面的分子特性。

根据 MoleculeNet 的建议，这七个任务类型均是分类任务，因此使用 AUC 均值作为下游任务的评价指标。下游任务的 AUC 指标越好，表明模型对于分子共同特征信息提取能力越强，泛化能力也越好。对于每个数据集，本研究使用从 DeepChem[113]提供的数据集分割方法 scaffold-split 来创建训练集、验证集和测试集。scaffold-split 分割方法以分子中存在的子结构作为划分依据，这种方法将贴近真实情况，但也会导致预测任务会更加复杂和挑战。

表 4-1 七个数据集代表不同的分子特性

数据集类型	数据集	任务数	分子数	描述
生物物理学	HIV	1	41913	抑制艾滋病毒复制的能力
	BACE	1	1522	人类 BACE-1 抑制剂的结合结果
	BBBP	1	2053	血脑屏障穿透
	Tox21	12	8014	毒性测量
生理学	SIDER	27	1427	对 27 个系统器官的药物不良反应
	ClinTox	2	1491	临床试验毒性和 FDA 批准状态
	MUV	17	93087	提炼的最近邻分析，用来验证虚拟筛选技术

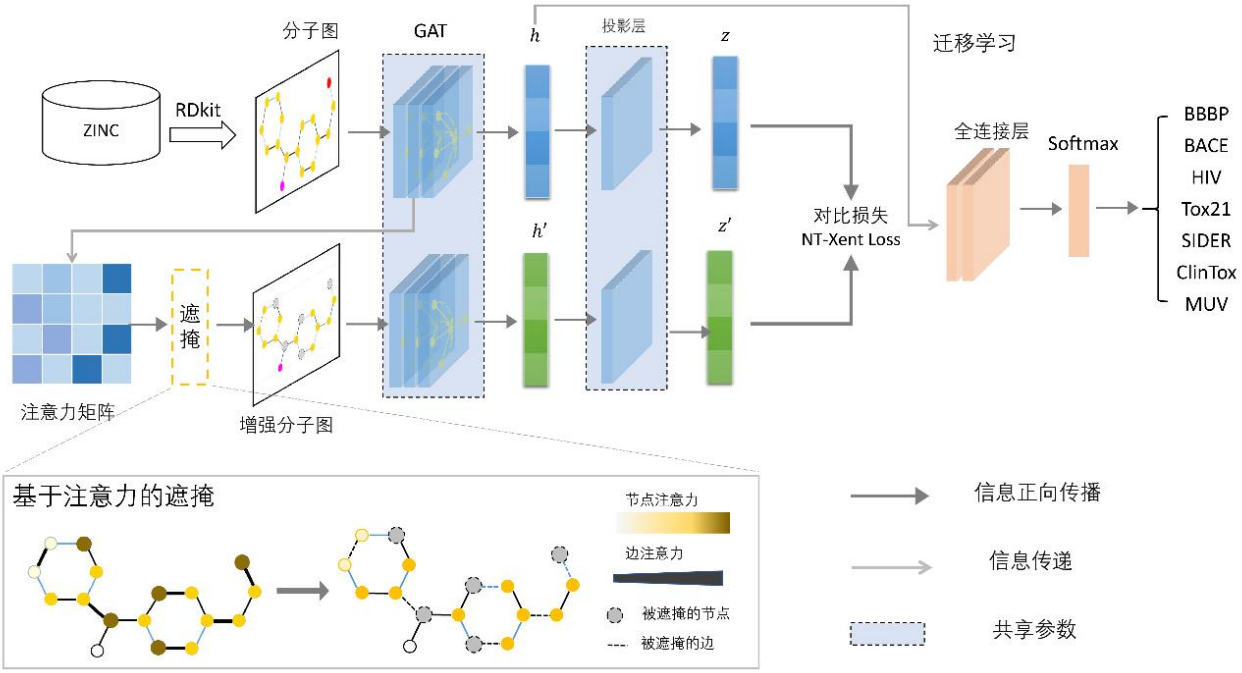

图 4-1 分子性质预测的 ATMOL 对比学习框架

4.2 预测模型

本研究的框架包括两个阶段：预训练和迁移学习。如图 4-1 所示，本研究首先在大规模未标记数据集上进行对比学习以获得分子表示，然后应用迁移学习来预测分子属性。将分子图作为输入，通过 GAT 编码器映射到潜在空间。与此同时，本研究设计了一个密切跟踪 GAT 编码器的注意遮掩模块，利用注意分数来遮掩一些节点或边产生一个增强视图。本研究特意设计了遮掩模块来产生增强图，这对 GAT 编码器区分正样本和负样本提出了挑战，使来自同一分子的图被赋予了类似的嵌入，但其他分子生成不同的嵌入。因此，对比学习模型被迫捕获重要的化学结构和高阶语义信息。本研究在几个分子性质预测任务上验证了通过对比学习获得的分子表示。

在对比学习的预训练过程中，本研究将预训练数据划分成训练集、验证集和测试集。共 10 个部分，其中 8 个部分用于对比模型的训练，其余 2 个部分分别作为验证集和测试集来评估训练模型的优缺点。模型的损失由对比学习中常见的 NT-Xent 损失函数计算，除了自己和对应的增强图外都看作负样本对，计算正负样本间距离，最终使得相似的分子相聚，异类相离。

在对比模型的预训练过程中，为了充分利用服务器的资源和提高训练的效率。本研究采用基于 Pytorch 的多 GPU 并行分布式训练，同时采用 Adam 作为模型的梯度下降，其通过一阶矩均值来计算自适应参数的学习率，得到当前最优的更新步长以防止过拟合或陷入局部最优解。模型中 GAT 的初始权值按照多 GPU 并行的方式进行初始化。在每次正向传播阶段后通过验证集上评估对比模型的收敛性，计算对应 NT-Xent 损失函数，并计算多 GPU 的梯度均值，然后反向传播自动更新梯度来调整模型中 GAT 网络的权重。

4.3 基于 GAT 的分子图特征表示

图注意力网络(GAT)是一种基于注意力的多头图嵌入学习体系结构。GAT 架构由多个图注意层构建，每个层对节点级表示应用线性转换来计算注意分数。图注意力网络通过自注意力机制来学习分子图中节点的隐藏表示，将分子图的节点集及其初始权值作为输入，通过权重矩阵对每个节点进行线性变换。设$\mathrm{h_i}$为节点i的嵌入，W为可学习的注意权重矩阵。计算节点i与其一阶邻居节点j之间的注意得分α_{ij}为

$$\alpha_{ij}=\frac{\exp\left(\mathrm{elu}\left(\mathrm{a^T}\left(\mathrm{Wh_i},\mathrm{Wh_j}\right)\right)\right)}{\sum_{k\in N(i)}\exp\left(\mathrm{elu}\left(\mathrm{a^T}\left(\mathrm{Wh_i},\mathrm{Wh_k}\right)\right)\right)} \qquad \text{公式(4-1)}$$

其中 a 是一个可学习向量，elu 是指数线性单位激活函数，N(i)表示节点i的一阶邻居。注意分数α_{ij}实际上是节点i与其邻居之间的 SoftMax 归一化信息。一旦计算出了注意力分数，节点i的输出特征就会被计算出来，方法是将节点i的相邻特征按相应的注意力分数加权进行聚合:

$$h'_i = \sigma(\alpha_{ii}Wh_i + \sum_{j\in N(i)} \alpha_{ij}Wh_j) \quad \text{公式(4-2)}$$

其中$\sigma(.)$为 ReLU 激活函数。在本研究的模型中，本研究使用了两个 GAT 层。第一层采用多头注意机制，头数设置为 10 个。考虑到节点级特征，使用全局最大池化层来获得图嵌入。隐藏特征的维数设置为 128。将第一层图中的多头注意权重矩阵平均到第二层中的注意矩阵，用于注意力遮掩模块生成增强图。

4.4 基于注意力遮掩的分子图增强

本研究设计了一种新颖的注意力权重遮掩作为分子图增强策略，该策略依据 GAT 所学习到的注意力权重值来遮掩分子图中一定百分比的节点(或边)。如果节点(或边)被遮掩的话，将编码器中对应的嵌入设为 0，阻塞该节点(或边)传递的消息。因此本研究提出了一个遮掩率 r 的概念，表示遮掩百分之 r 的分子图节点(或边)。在遮掩过程中，根据选择的遮掩策略和注意力权重来计算对应的遮掩率 r，而后使遮掩节点(或边)的百分比达到 r。为了探索遮掩不同边和节点的效果，本研究尝试了不同的遮掩策略:

（1）最大注意遮掩：为了生成与输入的分子图差异最大的增强图，遮掩百分之 r 的注意力权重最大的节点(或边) 。

（2）最小注意遮掩：为了生成与输入的分子图差异最小的增强图，遮掩百分之 r 的注意力权重最小的节点(或边) 。

（3）随机遮掩：将常用的随机遮掩纳入本研究中。随机遮掩分子图百分之 r 的节点(或边)，不使用学习到的注意力权重。

（4）轮盘遮掩：每个节点(或边)被遮掩的概率与其注意力权重成正比。通过 SoftMax 函数对注意权重矩阵 W 进行归一化得到遮掩概率分布。

由于 GAT 的注意权重矩阵在预训练阶段是变化的，因此根据注意权重矩阵得到的增强图也会动态变化。特别是，最大注意遮掩实际上类似于图对抗学习[114, 115]，它已被证明可以增强深度神经网络的鲁棒性和对扰动的泛化能力。

4.5 对比学习模块

GAT 编码器将输入的分子图及其增强图转化为嵌入的h_i和$h_i^{'}$，然后由非线

性投影仪映射到z_i和z_i'。然后，计算两个投影视图之间的相似度 $\mathrm{sim}(z_i, z_i')$。本研究采用归一化温度尺度交叉熵(NT-Xent)作为对比损失函数。

$$L_w = \log \frac{\exp(\mathrm{sim}(z_i, z_i')/\tau)}{\sum_{k=1}^{2N} 1_{|k \neq i|} \exp(\mathrm{sim}(z_i, z_k)/\tau)} \qquad \text{公式(4-3)}$$

其中$1_{|k \neq i|} \in \{0,1\}$是$k \neq i$时取值为 1 的指标函数，$\tau$表示温度参数; N 是小批量样品的数量。在本研究的研究中，余弦距离被用来评估来自一个分子的两个视图的相似性。

负样本的数量和多样性在自监督表示学习中起着至关重要的作用，以往的研究已经证实大量的负样本有助于提高表现。因此，除了每一批次的负样本外，本研究还对负样本池注意遮掩生成的增强分子图进行了补充，从而大大扩展了负样本的数量。更重要的是，增强的分子图丰富了负样本的多样性。注意遮掩用于图增强的优势体现在两个方面。首先，注意遮掩产生了具有挑战性的样本对，这增加了对比学习的难度，从而防止了学习潜在表征的崩溃。同时，它丰富了负样本的多样性，有助于本研究的模型学习具有良好泛化能力的分子表征。

4.6 实验结果

4.6.1 药物性质预测模型的超参数设置

本研究的实验环境如下：计算资源包含 3 张 24G 内存的 NVIDIA GeForce RTX 3090 显卡的工作台。基于对比学习预训练的参数设置如下： epoch 为 20、学习率为 1e-4、batch 为 128。

在分子性质预测的迁移学习中，本研究直接在 GAT 编码器后面添加了两个完全连接的层。本研究冻结了 GAT 的权重，在微调阶段只训练全连接层。下游的分类任务中损失函数选择交叉熵，选择 ROC-AUC 作为预测性能评价指标，选择 Adam 作为优化器并设置参数如下：学习率为 1e-7、batch 为 100。学习期间设定早期停止和退出策略防止过拟合。

每个用于分子性质预测的下游数据集以 8:1:1 的比例被分为训练、验证和测试数据集。传递模型在训练集上进行训练，并在验证集上进行验证。为了避免随机偏差，重复 5 次测试集评估流程，并选择平均 AUC 值作为最终性能报告。

在迁移学习阶段，迁移使用大规模的数据预训练的 GAT 网络加上进行下游任务微调的全连接层，并在全连接层后添加了随即抛弃层(dropout)，以防止过拟合情况及提高下游任务的鲁棒性。最后通过对下游任务集的交叉验证和损失值选择最优参数，然后使用下游任务最优的模型进行测试集中药物分子性质的

预测。

4.6.2 对比学习提升分子性质表现

本研究探寻对比学习作为预训练对于下游任务的性能提升情况。因此，将本研究的方法与直接进行不同分子性质预测任务的全监督方法进行了比较。对所有的下游任务，训练了相似的预测架构，包括处理分子图的 GAT 编码器和预测的两层全连接层。为了验证遮掩内容的有效性，本研究考虑遮掩节点、遮掩边或同时遮掩节点和边这三种不同分子增强策略(其中遮掩比 r 设定为 25%)。表 4-2 显示了这些模型和策略在下游任务上的 ROC-AUC 值。从表中不难发现，通过对比学习预训练后，模型在各种分子性质预测任务中的表现的显著提升。

同时，本研究发现在绝大部分下游任务中选择同时遮掩分子图的节点和边的效果相较于单一的遮掩节点或者边，结果有不同程度的提升。因此，可以认为基于对比学习的预训练能够使编码器学会如何分子图中信息丰富的表征，使其能够下游任务的性能。

表 4-2 预训练模型和全监督模型在分子性质预测任务中的 roc-auc(%)值

模型及策略	BBBP	BACE	HIV	ClinTox	Tox21	SIDER	MUV
全监督	85.5±1.4	76.2±0.1	72.6±0.1	92.6±0.2	76.2±0.1	80.5±0.2	69.8±0.1
ATMOL（遮掩节点）	91.1±0.4	84.3±0.6	75.1±0.1	97.1±0.1	79.1±0.2	81.4±0.3	**79.1±0.1**
ATMOL（遮掩边）	90.3±0.8	82.5±0.4	80.7±0.3	96.5±0.3	76.2±0.1	80.2±0.1	79.0±0.4
ATMOL（遮掩节点和边）	**92.1±0.1**	**87.3±0.3**	**81.2±0.5**	**97.5±0.3**	**77.1±0.3**	**81.9±0.2**	78.8±0.1

4.6.3 遮掩策略对特征提取的影响

本研究提出了基于注意力权重的四种遮掩策略，为了验证哪种遮掩策略与本研究的基准任务更匹配，在 7 个下游任务上进行基准测试。如图 4-2 所示，基于最大权重遮掩的策略在多数下游任务中取得最好的预测性能，而随机遮掩性能表现最差。在本研究所提出的基于对比学习的 GAT 编码器看来，对具有高注意力权重的节点和边进行遮掩会生成了一个与正面对应样本有很大不同的增强图。因此，本研究得出结论，最大权值遮掩策略对对比学习提出了一个挑战，即区分来自其他分子的一对正样本和一组负样本。这一挑战鼓励模型学习信息丰富的分子表示。

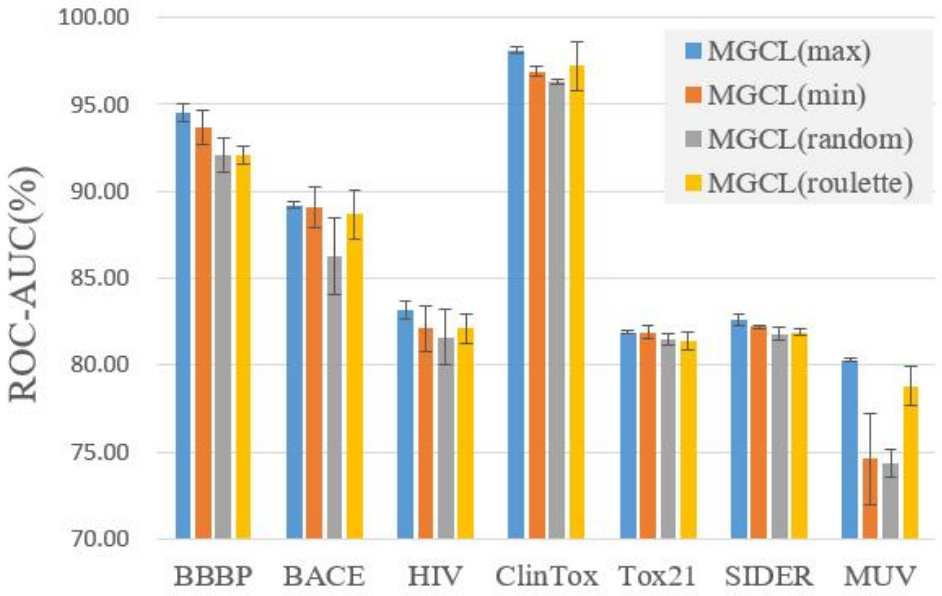

图 4-2 四种遮掩策略在 7 种分子性质预测任务中的图增强性能比较

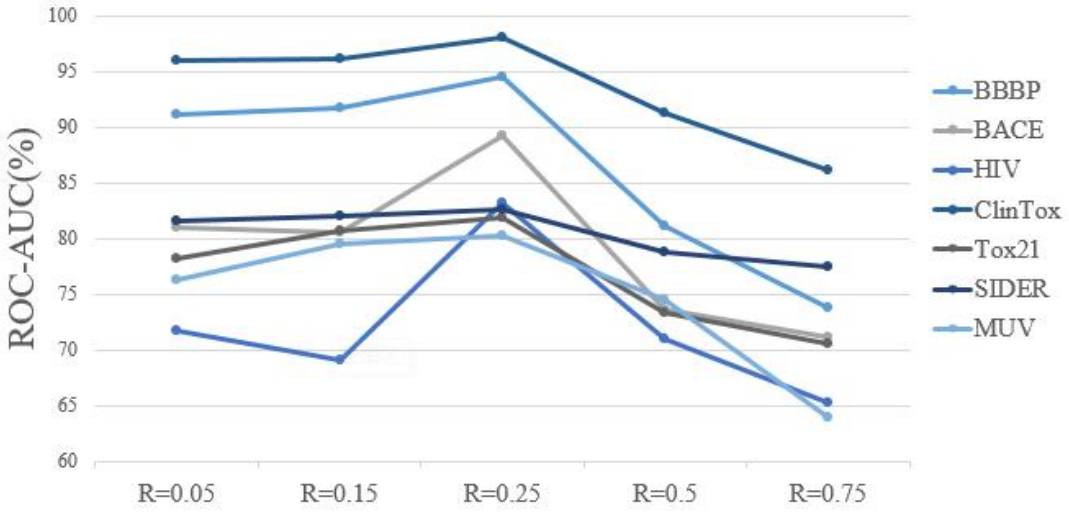

图 4-3 在 7 个下游任务上通过不同的遮掩率实现的性能

4.6.4 遮掩率对模型性能的影响

本研究进一步研究了不同遮掩率对于模型性能的影响，即遮掩分子图中不同比例的节点(或边)的预训练模型在不同下游任务的表现。在前面的实验中验证同时遮掩节点和边并且选择最大权重进行遮掩的策略能够获得目前最佳性能，因此选择这个遮掩策略作为基准的研究策略对不同遮掩率进行性能评估，其中

遮掩率的选值范围为{5%, 15%, 25%, 50%, 75%}。如图 4-3 所示，当遮掩率为 25%时，ROC-AUC 值最高。此后，性能迅速下降。在所有的下游任务中都可以观察到相似的性能表现趋势。这个现象说明当分子图的遮掩率过低时生成的增强图效果不是很理想，然而当选择过高的遮掩率时将会导致分子中大部分的子结构被破坏，编码器难以从中学习有效的内容。

4.6.5 不同规模数据集对模型的影响

本研究对于不同规模的未标记数据对预训练的学习是否有影响很感兴趣。因此，在已有的 *in vitro* 数据集的数据规模上，选择从 ZINC 的 *now* 数据集中选择了大约是 *in vitro* 数据集的 10 倍的分子作为对比数据集，其中共包含 300000 个分子。为了方便描述，本研究将它们分别称为小集和大集。本研究分别对这两个数据集进行了对比学习的预训练，而后验证在不同规模数据集预训练的模型在分子预测任务的性能情况。如表 4-3 所示，在规模更大的数据集预训练后，不同的遮掩策略在所有下游任务上都取得更优异的性能。因此本研究得出结论，大规模数据集上的自监督学习产生了具有更好泛化性的分子表示。

表 4-3 小型和大型未标记数据集在 7 个分子性质预测任务中的性能比较

数据集	遮掩策略	BBBP	BACE	HIV	ClinTox	Tox21	SIDER	MUV
in vitro（小）	最大权值	92.1±0.5	87.3±0.3	81.2±0.5	97.5±0.3	79.1±0.2	81.9±0.3	79.1±0.1
	最小权值	89.4±0.8	82.3±1.2	72.1±1.3	96.9±0.5	77.1±0.4	81.2±0.1	77.6±1.6
	随机	83.5±0.1	78.1±0.2	68.9±0.6	96.6±0.3	76.1±0.2	80.8±0.4	76.6±0.2
	轮盘	81.5±0.5	78.5±0.2	69.6±0.2	96.2±0.1	75.5±0.2	80.5±0.2	76.5±0.1
now（大）	最大权值	**94.5±0.5**	**89.2±0.2**	**83.2±0.5**	**98.1±0.5**	**82.5+0.4**	**82.6±0.3**	**80.3±0.1**
	最小权值	93.7±1.0	89.1±1.2	82.1±1.3	96.9±0.5	81.9±0.4	82.2±0.1	74.6±2.6
	随机	92.1±1.0	86.3±2.2	81.6±1.6	96.3±0.3	81.5±0.3	81.8±0.4	74.4±0.8
	轮盘	91.9±0.5	88.7±1.4	82.1±0.8	97.2±0.1	81.4±0.5	81.9±0.2	78.8±1.1

4.6.6 与其他方法的性能比较

为了验证本研究的方法的优越性能，与目前优异的方法进行对比验证。这些方法都采用自监督学习的方法进行分子特征提取。本研究简要介绍的方法如下:

（1）HU 等人[116]设计了一个节点层次和图层次 GNN 预训练架构，以便得到一个高表现力的模型来学习有用的局部和全局表示。

（2）N-Gram[117]运行节点嵌入，然后通过在图中的短步行中组装节点嵌入来构造图的紧凑表示。

（3）GROVER[118]通过结合 GNN 和 Transformer 来实现一个能够完成上下文预测任务以及主题功能预测任务。

（4）MolCLR[55] 提出了一个新颖的基于 GNN 和图随机去除的对比学习架构，该架构策略会随机去除节点、边或子图来生成增强视图。

（5）MGSSL[119]是一个基于 GNN 的自监督主题生成框架，同时考虑了原子层级和官能团层级。

为了与其他方法进行性能比较，本研究报告了在小型和大型数据集上预训练的模型的性能。对于其他具有训练和发布模型的方法，本研究在相同的下游数据集上运行它们。对于那些没有发布源代码的人，本研究使用他们在出版物中报告的结果。表 4-4 显示了竞争方法以及本研究的方法在小型和大型未标记数据集在 7 个分子性质预测任务中的性能比较。除了 MUV 外的所有下游任务，本研究的模型在小型数据集上的性能表现已经高于其他竞争方法。特别在 SIDER 上，本研究的方法对于其他方法取得接近 10%的提升。此外，预训练选择更大数据集上进行时，本研究的方法获得了更大的性能优势。

表 4-4 小型和大型未标记数据集在 7 个分子性质预测任务中的性能比较

方法	BBBP	BACE	HIV	ClinTox	Tox21	SIDER	MUV
HU	70.8±1.5	85.9±0.8	80.2±0.9	78.9±2.4	78.7±0.4	65.2±0.9	81.4±2.0
N-Gram	91.2±3.0	87.6±3.5	83.0±1.3	85.5±3.7	76.9±2.7	63.2±0.5	81.6±1.9
Grover	68.0±1.5	79.5±1.1	77.8±1.4	76.9±1.9	76.3±0.6	60.7±0.5	75.8±1.7
MolCLR	73.6±0.5	89.0±0.3	80.6±1.1	93.2±1.7	79.8±0.7	68.0±1.1	88.6±2.2
MGSSL	70.5±1.1	79.7±0.8	79.5±1.1	80.7±2.1	76.5±0.3	61.8±0.8	78.7±1.5
ATMOL (small)	92.1±0.5	87.3±0.3	81.2±0.5	97.5±0.3	79.1±0.2	81.9±0.3	79.1±0.1
ATMOL (large)	**94.5±0.5**	**89.2±0.2**	**83.2±0.5**	**98.1±0.5**	**82.5+0.4**	**82.6±0.3**	**80.3±0.1**

4.7 模型可解释性的探索

4.7.1 分子表征的空间定位研究

分子表征的空间分布有助于验证所提方法的有效性。本研究应用 UMAP 工具[120]分别对预训练前后的分子表示进行可视化，这是一种用于降维的学习算法，能够很好地保存数据全局结构。图 4-4 显示了在不同时期中 BBBP 和 SIDER 下游任务中 UMAP 降维后的分子嵌入。初始分子表征在空间上无序，经过预训练后，属于同一类的分子聚集在一起，与其他类明显分离。通过 UMAP 降维可视化的图可以看出，该方法可以有效地从化学结构上检测分子的物理化学性质，使物理化学性质相似的分子获得相似的潜在表征。

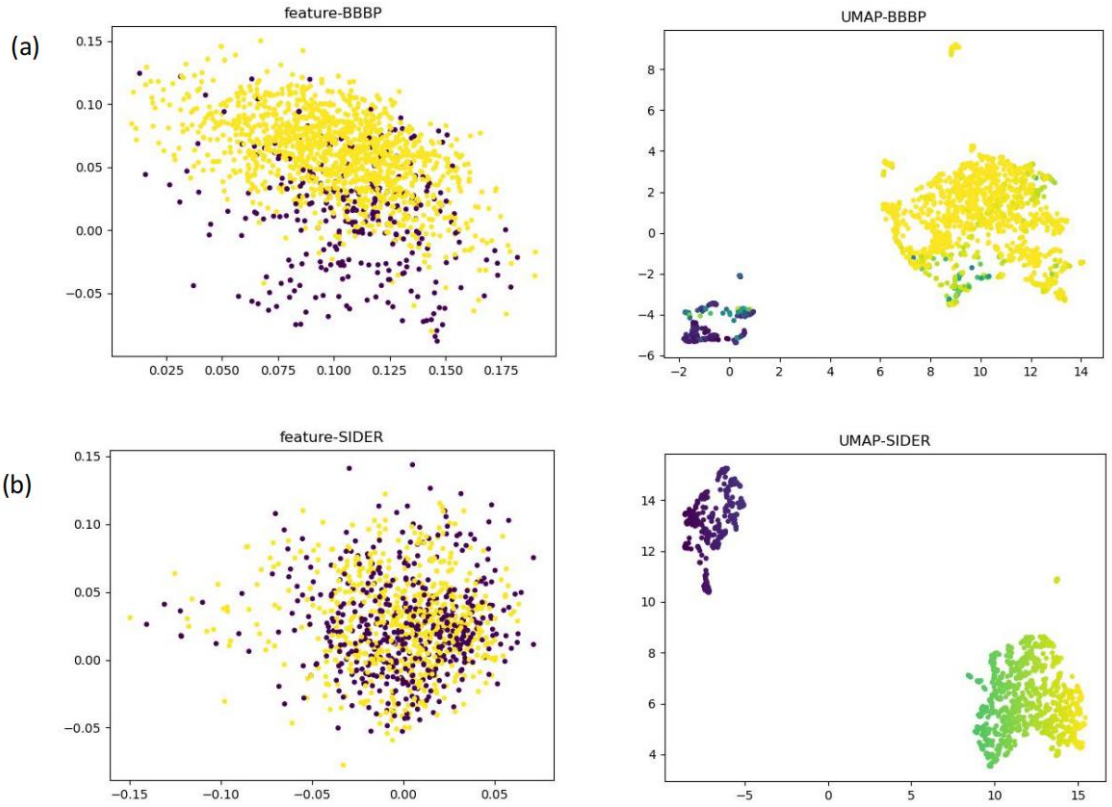

图 4-4 BBBP 和 SIDER 数据集中分子表征的空间定位可视化。左列显示初始嵌入，右列显示预训练嵌入

4.7.2 探究注意权值对重要的化学子结构的关联

在分子表示学习阶段，本研究尝试使用 GAT 来识别对特定预测任务重要的化学成分。本研究旨在探究注意力机制如何关注对比学习过程中分子表征的影响，在下游任务中对于 GAT 的参数进行了微调。为了能够方便理解研究结果，本研究在分子图中可视化了注意力权重。在 BBBP 膜透性数据集[121]中，本研究随机选取一个分子 SMILES 为 C[S](=O)(=O)c1ccc(cc1)[C@@h](O)[C@@h](CO)NC(=O)C(Cl)Cl 的作为样本。本研究通过计算每对原子之间的注意权重的 Pearson 相关系数，并以热图的形式展示计算结果。如图 4-5a 所示，热图显示了一些相关性密切的原子团的存在，这些原子团可能会影响分子的某些特定性质。在分子的二维结构上显示权重后，发现分子中的苯环可能在调节膜的通透性方面起关键作用。同样，本研究从 BACE 数据集[122]中随机选取一个 SMILES 为 FC1(F)COC(=NC1(C)c1cc(NC(=O)c2nn(cc2)C)ccc1F)N 的分子，该分子的热图中也存在一些关键的原子团，如图 4-5b 所示。

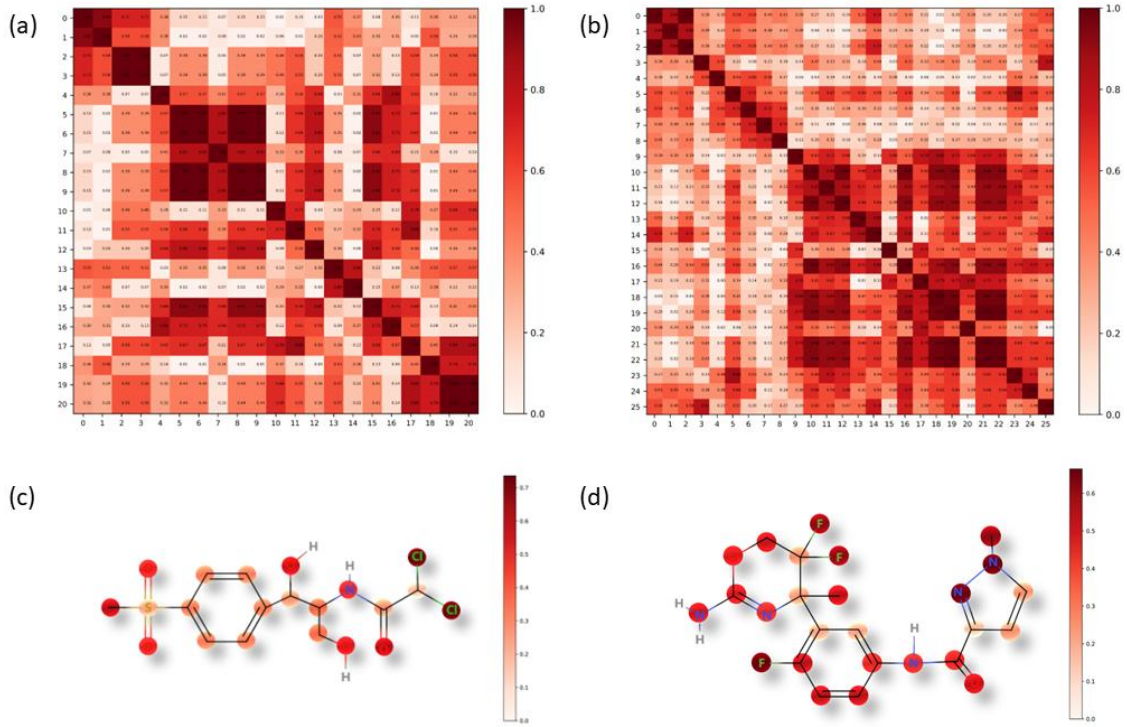

图 4-5 注意权重和分子结构的相关热图的可视化与原子的注意权重着色。(a)和(c)为从 BBBP 集合中选取的样本分子，(b)和(d)为从 BACE 集合中选取的样本分子

除了原子团层级的影响研究外，本研究同时探究了分子中特定原子对其分子特性的影响，因此在分子的二维结构上可视化了每个原子的注意力权重。由于 BBBP 数据集中的任务是关于膜的通透性，如图 4-5c 所示，在选取的样本分子中的两个 Cl 原子拥有着最高的注意力得分。根据现有的研究表明 Cl 原子具有对电子存在着很强的吸引力，本研究认为 Cl 原子能够改变分子的极性，进而影响了膜的通透性。此外，分子中存在促进亲水性的羟基，其也有着较高的注意力权重[123]。

同理，对 BACE 数据集中的人 β -分泌酶 1 (BACE-1)抑制剂样本做了原子级的可视化研究。Mureddu 等人[124]的研究表明异胞嘧啶芳香族会抑制 BACE-1 分泌酶。如图 4-5d 所示，该分子中的异胞嘧啶结构获得了最高的注意力权重。本研究通过可视化的手段解释了模型中注意力机制如何识别到与分子性质相关的分子子结构和原子。

4.7.3 探究分子图的网络属性

本研究进一步检查了不同数据集中分子的化学结构是否会影响预测性能。本研究计算了一些网络属性，包括聚类系数、中间度、距离和度。此外，本研

究还计算了平均分子式长度。如表 4-5 所示，本研究发现 MUV 和 HIV 数据集中的分子平均度相对较小。特别是 Tox21、MUV 和 HIV 数据集的分子式长度较短。相应地，本研究的方法在这些数据集上获得了相对较低的性能。这些数据集中的药物分子包含相对较少的原子和简单的化学结构，这导致小分子图。因此，本研究推测，在太小的图上，基于注意的遮掩图增强会产生折叠视图，其中被遮掩的节点特征不能再从其邻居中重建。因此，在这些数据集上，本研究得到了相对较低的性能。

表 4-5 7 个数据集的平均网络性质和分子式长度

网络属性	BBBP	BACE	HIV	ClinTox	Tox21	SIDER	MUV
平均聚集系数	0.001829	0.001501	0.006202	0.002184	0.002400	0.003914	0.000623
平均介数	0.178243	0.177894	0.166343	0.207503	0.184122	0.174443	0.190406
平均距离	0.247832	0.227349	0.174330	0.304597	0.241782	0.232310	0.209278
平均度	0.232785	0.207849	0.141425	0.334637	0.235802	0.226125	0.197246
平均分子式长度	48.9169	44.82665	63.91494	31.44367	58.11045	65.62198	40.5214

4.8 本章小结

本章研究了基于注意力遮掩策略的对比学习的药物分子性质预测，并研究了负样本对分子性质预测的影响。负样本的多样性已被证明极大地影响表征学习。在本研究中，除了将 minibatch 中的样本作为负样本外，本研究还将通过注意遮掩策略生成的增强分子图纳入到负样本队列中，使负样本得到了极大的扩展和多样化。

所提出的分子图注意力遮掩生成了不同的对比视图。通过对比，本研究发现，同时遮掩边和节点生成的对比视图在几乎所有下游任务中都获得了最佳性能。此外，本研究发现最大权重遮掩，即遮掩具有较大注意力权重的边或节点，获得了最佳性能。直观上，最大权重遮掩策略类似于对抗性学习的思想，从对比损失的角度来看，每次生成一个与正面对等视图差异最大的增强分子图。另外，注意权重的变化动态影响图增强过程，使得本研究的模型能够识别分子图的不同组成部分，最终达到稳态。因此，本研究得出结论，具有挑战性的对比观点有助于学习重要的语义结构。

综上所述，本研究对大规模未标记分子的自监督表示学习显著提高了各种分子性质预测任务的性能。这是一个任务不可知的预训练，从而产生了具有理想的表达性和泛化性的分子表示。

第 5 章 总结与展望

5.1 研究工作总结

越来越多的临床和实验证据表明，人类体内的微生物与人类健康有着密不可分的联系和复杂的互作关系。微生物群不仅是健康和疾病的重要调控因素，也是重要的临床药物靶点，然而药物的毒性和副作用等药物性质也可能会导致微生物的生长受到抑制或死亡，从而影响人体微生物群落的稳定性和功能。因此，识别微生物与药物之间的联系不仅仅有助于理解微生物与药物互作机制，更是对于促进药物和个性化药物的有效开发提供了微生物衍生法。同时，预测药物性质在微生物学研究中也发挥重要作用。在传统生物实验、临床验证和人工智能的不断积累下，药物与微生物关联数据及分子性质数据呈现出不断增长的趋势，因此人们也逐渐将目光放在能够处理大规模数据的深度学习上，如图注意力网络、图卷积网络等架构被广泛用于微生物与药物的潜在关联预测。但是由于数据的不平衡和多噪声的情况，这些模型还存在着一些可以改进的地方。

此外，预测得到微生物与药物新关联后，将面临着基于人工智能筛选的药物需要作用到人体内微生物组时是否有疗效和是否安全的新问题。由于体外实验的环境并不完全等同于真实的人体环境，而且个体间也存在着巨大差异，极有可能出现体内外实验效果差异明显的情况，甚至会出现药物不良反应现象，因此在体内实验前得到药物的相关性质具有重要的意义。但由于药物性质的多样性、复杂性以及常规的实验室方法仅能对药物进行特定性质验证，传统方法不但耗时费力、花费昂贵，而且效率低下。此外，目前有标记的药物性质数据规模量并不大，然而目前存在着庞大的无标签分子数据，因此，如何利用庞大的无标记药物集挖掘泛用和稳定的药物特征并用于下游药物性质预测任务，是当前亟待解决的问题之一。本文主要工作如下：

（1）提出了一个基于图变分自编码器(VGAE)和深度神经网络(DNN)的集成框架 Graph2MDA，用于预测微生物和药物之间的相关性。首先 Graph2MDA。构建了多模态属性图，有效整合了微生物和药物丰富的相似性和本体信息，通过 VGAE 对多模态属性图进行潜在特征提取，最后使用 DNN 分类器进行新关联预测。与目前最先进的基于图神经网络和关联预测的微生物-药物关联预测方法相比，本研究的方法在三个独立的数据集上获得了更好的性能。此外本研究验证了模型学习得到的潜在表征在语义上与药物药理功能相关，表现出与药物 ATC 分类相关的明显的聚类模式，且聚类分布显著一致。

（2）基于注意力机制的最大权重遮掩策略。本研究除将训练批次中样本作

为负样本外，还将注意力遮掩策略产生的增强分子图加入到负样本队列中，极大地扩展和多样化了负样本。本研究中所提出的最大注意权重遮掩策略即遮掩具有较大注意力权重的边或节点，获得了最佳性能。直观上，最大注意权重遮掩类似于对抗性学习的思想，从对比损失的角度来看，每次生成一个与正样本视图差异最大的增强分子图。另外，本研究通过动态跟踪图增强过程中注意力权重的变化，发现所提出的模型会检查分子图的不同组成部分，最终达到稳定状态。从实验结果可知，具有挑战性的对比观点有助于学习重要的语义结构。

（3）基于对比学习的预训练有助于下游的药物分子性质预测任务。本研究所提出的 ATMOL 是一种利用对比学习预训练模型预测分子性质的算法，首先从收录大量活性药物分子数据的 ZINC 数据库中整理了两种大小的数据集，分别训练图注意力编码器，同时尝试使用不同的遮掩策略，将该编码器迁移到 MoleculeNet 中的七个分子性质预测任务上。实验结果表明，ATMOL 在 7 个基准任务上均优于现有的方法，并且在数据规模更大的数据集上取得了更优异的性能表现。通过对比不同的分子图注意力遮掩策略，基于最大注意权重遮掩并同时遮掩边和节点生成的对比视图在大部分下游任务中都获得了最佳性能。此外，通过可视化手段分析展示了模型提取的药物分子表征以及权重与分子化学子结构的关联。最后计算了一些包括聚类系数、中间度等在内的网络属性进一步验证了不同数据集中分子的化学结构是否会影响预测性能。综上所述，通过对大规模未标记分子进行自监督表示学习的预训练，可以提高各种分子性质预测任务的性能。这种预训练方式不需要事先了解分子的具体任务，因此可以生成理想中的表达性和泛化性的分子表示。

5.2 研究工作展望

在本文的研究中采用了不同的图自监督学习模型对药物与微生物关联和药物性质预测两个任务进行了研究，得到一些创新性且有价值的结果。但是在数据处理和算法模型上都存在一定的改进空间：

（1）构建的药物与微生物初始关联网络稀疏。与整个药物与微生物关联空间相比，已知的关联数量非常有限，导致数据结构稀疏。结果，由于图卷积过程中节点之间的信息传播不足，从 VGAE 学习的嵌入表示仍然是次优的。为了解决这个问题，考虑了遮掩机制来提高模型性能。更准确地说，对于没有监控信息的节点，在损失计算时使用遮掩，以限制噪声信息的传播。此外可以通过多模态属性图集成更多类型的节点信息，例如可以利用属性，如基于副作用的药物相似性，或新的网络信息，如微生物与疾病关联，疾病与药物关联，以此

来提高输入网络的层次度和丰富度，提供更多的潜在特征，进而提升预测的性能。

（2）药物分子图初始特征维度不够丰富。在将药物 SMILES 转化为分子图的过程中，会构建药物分子图点边的邻接矩阵以及初始特征，本研究所构建的药物原子节点初始特征维度较低，仅包含由 DeepChem 计算得来的二值特征，包括原子信息、邻居原子信息以及分子子结构信息。这些信息相对较少，可能会对模型的性能产生负面影响。因此，可以考虑将更多的原子信息进行特征初始化，例如原子类型、电荷、杂化状态、相邻原子类型等。

（3）药物分子性质预测模型对下游任务中小分子的药物预测泛化性较差。在第四章中，计算了下游任务中各数据集的分子平均度和分子式平均长度。发现 MUV 和 HIV 数据集中的分子平均度相对较小，特别是 Tox21、MUV 和 HIV 数据集的分子式平均长度较短，相应地在这些数据集上获得了相对较低的性能。这些数据集中的药物分子包含相对较少的原子和简单的化学结构，这导致小分子图。因此，基于注意权重的遮掩策略的图增强方法会产生坍缩视图，其中被遮掩的节点特征不能再从其邻居中重建，在这些数据集上得到了相对较低的性能。可以考虑根据以上问题，设计一些更合理的分子遮掩策略，或者使用一些启发式规则来保证被遮掩的部分是有意义且可恢复的。

参考文献

[1] HUMAN MICROBIOME PROJECT C. Structure, function and diversity of the healthy human microbiome [J]. Nature, 2012, 486(7402): 207-214.

[2] SOMMER F, BAECKHED F. The gut microbiota - masters of host development and physiology [J]. Nature Reviews Microbiology, 2013, 11(4): 227-238.

[3] 王庆忠, 范云, 沈祁烨. 人类微生物组学与健康及其临床检验的需求 [J]. 分子诊断与治疗杂志, 2018, 10(1): 6.

[4] VENTURA M, O'FLAHERTY S, CLAESSON M J, et al. Genome-scale analyses of health-promoting bacteria: probiogenomics [J]. Nature Reviews Microbiology, 2009, 7(1): 61-U77.

[5] 林璋, 祖先鹏, 谢海胜, 等. 肠道菌群与人体疾病发病机制的研究进展 [J]. 药学学报, 2016, 51(6): 10.

[6] MALLA M A, DUBEY A, KUMAR A, et al. Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment [J]. Frontiers in Immunology, 2019, 9.

[7] 张萌萌,姜宁,张爱忠.肠道微生物对肠道屏障功能完整性的维护机制研究概况[J].微生物学通报,2020,47(03):933-940.

[8] 贺锐,张丽秀,叶萍等.低出生体重儿肠道菌群及肠道屏障功能的研究[J].中国微生态学杂志,2017,29(09):1022-1026+1036.

[9] WEN L, LEY R E, VOLCHKOV P Y, et al. Innate immunity and intestinal microbiota in the development of Type 1 diabetes [J]. Nature, 2008, 455(7216): 1109-U1110.

[10] ZHANG H, DIBAISE J K, ZUCCOLO A, et al. Human gut microbiota in obesity and after gastric bypass [J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(7): 2365-2370.

[11] 郭慧玲. 肠道菌群与疾病关系的研究进展 [J]. 微生物学通报, 2015, 42(2): 11.

[12] SCHWABE R F, JOBIN C. The microbiome and cancer [J]. Nature Reviews Cancer, 2013, 13(11): 800-812.

[13] CLARKE S F, MURPHY E F, ORLA O S, et al. Targeting the Microbiota to Address Diet-Induced Obesity: A Time Dependent Challenge [J]. PLoS ONE,

2013, 8(6): e65790-.

[14] 朱崟, 秦环龙. 口腔微生物与口腔疾病,肠道菌群,肠道疾病的关联性研究进展 [J]. 上海预防医学, 2020, 32(3): 6.

[15] 张志民, 程博群. 口腔微生物与消化系统癌症关系的研究进展 [J]. 口腔医学研究, 2020, (2): 93-97.

[16] ZIMMERMANN M, PATIL K R, TYPAS A, et al. Towards a mechanistic understanding of reciprocal drug-microbiome interactions [J]. Molecular Systems Biology, 2021, 17(3).

[17] HAISER H J, GOOTENBERG D B, CHATMAN K, et al. Predicting and Manipulating Cardiac Drug Inactivation by the Human Gut Bacterium Eggerthella lenta [J]. Science, 2013, 341(6143): 295-298.

[18] KLATT N R, CHEU R, BIRSE K, et al. Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women [J]. Science, 2017, 356(6341): 938-944.

[19] 蔡佩. 肠道微生物对药物毒性影响的研究进展 [J]. 中国处方药, 2022, 20(6): 3.

[20] 邓红, 吴纯启, 赵春雪, 等. 高通量测序和实时荧光定量 PCR 分析何首乌肝损伤与肠道微生物组的关系 [J]. 中文科技资料目录-中草药, 2017, (040-004).

[21] 邓红, 王全军. 何首乌肝损伤与肠道微生物的相关性研究, F, 2016 [C].

[22] KASHYAP P C, CHIA N, NELSON H, et al. Microbiome at the Frontier of Personalized Medicine [J]. Mayo Clinic Proceedings, 2017, 92(12): 1855-1864.

[23] 黄洁,孙景勇,毛恩强等.重症急性胰腺炎细菌感染的菌谱与耐药性分析[J].中国感染与化疗杂志,2009,9(05):372-376.

[24] 孙长贵, 张丽君, 曾贤铭, 等. 微量稀释法测定抗真菌剂对酵母菌 MIC 的评价 [J]. 临床检验杂志, 2000, 18(6): 3.

[25] 温旺荣, 王德春. 常用消毒剂对念珠菌最小杀菌浓度的测定 [J]. 中国消毒学杂志, 1994, 11(2): 4.

[26] 张文波, 吴移谋, 尹卫国, 等. 解脲脲原体耐氟喹诺酮类药物突变体的分离及其突变位点的研究 [J]. Chinese Medical Journal, 2002, 115(10): 1573-1575.

[27] MACPHERSON A J, MCCOY K D. Standardised animal models of host microbial mutualism [J]. Mucosal Immunology, 2014, 8(3): 476.

[28] ROSARIO K, BREITBART M. Rosario K, Breitbart M.. Exploring the viral world through metagenomics. Curr Opin Virol 1: 289-297 [J]. PubMed, 2011.

[29] 段墨, 肖炜, 王永霞, 等. 454 测序技术在微生物生态学研究中的应用 [J]. 微生物学杂志, 2011, 31(5): 6.

[30] 乌日拉嘎, 徐海燕, 冯淑贞, 等. 测序技术的研究进展及三代测序的应用 [J]. 中国乳品工业, 2016, 44(4): 5.

[31] 刘禧杰, 辛德莉, 李靖, 等. 肺炎支原体对大环内酯类抗生素的耐药机制 [J]. 实用儿科临床杂志, 2008, 23(22): 4.

[32] 凌宗欣. 女性生殖道微生物群落菌群多样性变化与生殖道感染的相关性研究 [D]; 浙江大学, 2012.

[33] LU H, GIORDANO F, NING Z. Oxford Nanopore MinION Sequencing and Genome Assembly [J]. 基因组蛋白质组与生物信息学报(英文版), 2016, (5).

[34] CHEN X, HUANG Y-A, YOU Z-H, et al. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases (vol 33, pg 733, 2017) [J]. Bioinformatics, 2018, 34(8): 1440-1440.

[35] LUO J, LONG Y. NTSHMDA: Prediction of Human Microbe-Disease Association Based on Random Walk by Integrating Network Topological Similarity [J]. Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2020, 17(4): 1341-1351.

[36] SUN Y-Z, ZHANG D-H, CAI S-B, et al. MDAD: A Special Resource for Microbe-Drug Associations [J]. Frontiers in Cellular and Infection Microbiology, 2018, 8: 424.

[37] RAJPUT A, THAKUR A, SHARMA S, et al. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance [J]. Nucleic Acids Research, 2018, 46(D1): D894-D900.

[38] ANDERSEN P I, IANEVSKI A, LYSVAND H, et al. Discovery and development of safe-in-man broad-spectrum antiviral agents [J]. International Journal of Infectious Diseases, 2020, 93: 268-276.

[39] MENG Y, JIN M, TANG X, et al. Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study [J]. Applied Soft Computing, 2021, 103: 107135.

[40] LONG Y, WU M, LIU Y, et al. Ensembling graph attention networks for human microbe-drug association prediction [J]. Bioinformatics, 2020, 36: I779-I786.

[41] LONG Y, LUO J. Association Mining to Identify Microbe Drug Interactions Based on Heterogeneous Network Embedding Representation [J]. Ieee Journal of Biomedical and Health Informatics, 2021, 25(1): 266-275.

[42] LONG Y, WU M, KWOH C K, et al. Predicting human microbe-drug

associations via graph convolutional network with conditional random field [J]. Bioinformatics, 2020, 36(19): 4918-4927.

[43] YU Z, HUANG F, ZHAO X, et al. Predicting drug-disease associations through layer attention graph convolutional network [J]. Briefings in Bioinformatics, 2021, 22(4): bbaa243.

[44] SONG C M, LIM S J, TONG J C. Recent advances in computer-aided drug design [J]. Brief Bioinform, 2009, 10(5): 579-591.

[45] HONDA S, SHI S, UEDA H R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery [J]. arXiv preprint arXiv:191104738, 2019.

[46] WANG S, GUO Y, WANG Y, et al. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction[C]; proceedings of the Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics, F, 2019: 429-436.

[47] PAN S J, TSANG I W, KWOK J T, et al. Domain Adaptation via Transfer Component Analysis [J]. Ieee Transactions on Neural Networks, 2011, 22(2): 199-210.

[48] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2020: 9729-9738.

[49] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks [J]. Advances in Neural Information Processing Systems 27 (Nips 2014), 2014, 27.

[50] CHEN T, KORNBLITH S, SWERSKY K, et al. Big self-supervised models are strong semi-supervised learners [J]. Advances in neural information processing systems, 2020, 33: 22243-22255.

[51] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations [J]. International Conference on Machine Learning, Vol 119, 2020, 119.

[52] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J/OL] 2018, arXiv:1810.04805[https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D].

[53] ZHANG X C, WU C K, YANG Z J, et al. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction [J]. Brief

Bioinform, 2021, 22(6): bbab152.

[54] BAGAL V, AGGARWAL R, VINOD P, et al. MolGPT: Molecular Generation Using a Transformer-Decoder Model [J]. Journal of Chemical Information and Modeling, 2021, 62(9): 2064-2076.

[55] WANG Y, WANG J, CAO Z, et al. MolCLR: molecular contrastive learning of representations via graph neural networks [J]. arXiv preprint arXiv:210210056, 2021.

[56] SONG Y, GU Y, LI X, et al. CSGNN: Improving Graph Neural Networks with Contrastive Semi-supervised Learning [Z]. 2022: 731-738.10.1007/978-3-031-00123-9_58

[57] LI P, WANG J, QIAO Y, et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery [J]. Brief Bioinform, 2021, 22(6).

[58] LI H, ZHAO X, LI S, et al. MoTSE: an interpretable task similarity estimator for small molecular property prediction tasks [M]. 2021.

[59] MA H, BIAN Y, RONG Y, et al. Multi-View Graph Neural Networks for Molecular Property Prediction [Z]. 2020

[60] 张慧. 长链非编码 RNA-蛋白质相互作用及疾病关联算法的研究 [D]; 吉林大学, 2019.

[61] JIANG L, ZHU J. Review of MiRNA-disease association prediction [J]. Current Protein and Peptide Science, 2020, 21(11): 1044-1053.

[62] ALAIMO S, GIUGNO R, PULVIRENTI A. ncPred: ncRNA-disease association prediction through tripartite network-based inference [J]. Frontiers in bioengineering and biotechnology, 2014, 2: 71.

[63] YU Z, HUANG F, ZHAO X, et al. Predicting drug–disease associations through layer attention graph convolutional network [J]. Briefings in Bioinformatics, 2021, 22(4): bbaa243.

[64] MA Q, TAN Y, WANG L. GACNNMDA: a computational model for predicting potential human microbe-drug associations based on graph attention network and CNN-based classifier [J]. Bmc Bioinformatics, 2023, 24(1): 35.

[65] STERLING T, IRWIN J J. ZINC 15 – Ligand Discovery for Everyone [J]. Journal of Chemical Information and Modeling, 2015: 2324.

[66] WU Z, RAMSUNDAR B, FEINBERG E N, et al. MoleculeNet: a benchmark for

molecular machine learning [J]. Chem Sci, 2018, 9(2): 513-530.

[67] BARTOK A P, KONDOR R, CSANYI G. On representing chemical environments [J]. Physical Review B, 2013, 87(18).

[68] GHIRINGHELLI L M, VYBIRAL J, LEVCHENKO S V, et al. Big data of materials science: critical role of the descriptor [J]. Phys Rev Lett, 2015, 114(10): 105503.

[69] DAVID L, THAKKAR A, MERCADO R, et al. Molecular representations in AI-driven drug discovery: a review and practical guide [J]. Journal of Cheminformatics, 2020, 12(1): 1-22.

[70] BADE R, CHAN H F, REYNISSON J. Characteristics of known drug space. Natural products, their derivatives and synthetic drugs [J]. European Journal of Medicinal Chemistry, 2010, 45(12): 5646-5652.

[71] CERETO-MASSAGUE A, OJEDA M J, VALLS C, et al. Molecular fingerprint similarity search in virtual screening [J]. Methods, 2015, 71: 58-63.

[72] ROGERS D, HAHN M. Extended-connectivity fingerprints [J]. J Chem Inf Model, 2010, 50(5): 742-754.

[73] CHEN X, FAN H, GIRSHICK R, et al. Improved baselines with momentum contrastive learning [J]. arXiv preprint arXiv:200304297, 2020.

[74] GRILL J-B, STRUB F, ALTCHé F, et al. Bootstrap your own latent-a new approach to self-supervised learning [J]. Advances in neural information processing systems, 2020, 33: 21271-21284.

[75] KIPF T N, WELLING M. Variational Graph Auto-Encoders [J]. 2016.

[76] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes; proceedings of the International Conference on Learning Representations, F, 2013 [C].

[77] DING Y, TIAN L-P, LEI X, et al. Variational graph auto-encoders for miRNA-disease association prediction [J]. Methods, 2021, 192: 25-34.

[78] SHI Z, ZHANG H, JIN C, et al. A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations [J]. Bmc Bioinformatics, 2021, 22(1): 1-20.

[79] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention [J]. Advances in neural information processing systems, 2014, 27.

[80] KEMPTER R, GERSTNER W, VAN HEMMEN J L. Hebbian learning and spiking neurons [J]. Physical Review E, 1999, 59(4): 4498.

[81] WISKOTT L, SEJNOWSKI T J. Slow feature analysis: Unsupervised learning of invariances [J]. Neural computation, 2002, 14(4): 715-770.

[82] MEDINA C, DEVOS A, GROSSGLAUSER M. Self-supervised prototypical transfer learning for few-shot classification [J]. arXiv preprint arXiv:200611325, 2020.

[83] OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding [J]. arXiv preprint arXiv:180703748, 2018.

[84] 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.

[85] HATTORI M, TANAKA N, KANEHISA M, et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses [J]. Nucleic Acids Research, 2010, 38: W652-W656.

[86] JAIN D K, ZHANG Z, HUANG K. Random walk-based feature learning for micro-expression recognition [J]. Pattern Recognition Letters, 2018, 115: 92-100.

[87] KAMNEVA O K. Genome composition and phylogeny of microbes predict their co-occurrence in the environment [J]. Plos Computational Biology, 2017, 13(2): e1005366.

[88] SZKLARCZYK D, GABLE A L, NASTOU K C, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets [J]. Nucleic Acids Research, 2021, 49(D1): D605-D612.

[89] WEBER K S, KNEBEL B, STRASSBURGER K, et al. Associations between explorative dietary patterns and serum lipid levels and their interactions with ApoA5 and ApoE haplotype in patients with recently diagnosed type 2 diabetes [J]. Cardiovascular Diabetology, 2016, 15: 1-13..

[90] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [J]. arXiv preprint arXiv:160902907, 2016.

[91] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [J]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[92] VAN DER MAATEN L, HINTON G. Visualizing Data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

[93] LIBRARY W P. Anatomical Therapeutic Chemical Classification System [J]. Dictionary of Pharmaceutical Medicine, 2009: 8-8.

[94] ENNA S, BYLUND D B. Astressin [J]. 2007: 1-2.

[95] SAENGSAI J, KONGTUNJANPHUK S, YOSWATTHANA N, et al. Antibacterial and Antiproliferative Activities of Plumericin, an Iridoid Isolated from Momordica charantia Vine [J]. Evidence-Based Complementary and Alternative Medicine, 2015, 2015.

[96] GUPTA V K, MALHOTRA S. Pharmacological attribute of Aloe vera: Revalidation through experimental and clinical studies [J]. Ayu, 2012, 33(2): 193-196.

[97] KAITHWAS G, SINGH P, BHATIA D. Evaluation of in vitro and in vivo antioxidant potential of polysaccharides from Aloe vera (Aloe barbadensis Miller) gel [J]. Drug and Chemical Toxicology, 2014, 37(2): 135-143.

[98] FANI M, KOHANTEB J. Inhibitory activity of Aloe vera gel on some clinically isolated cariogenic and periodontopathic bacteria [J]. Journal of Oral Science, 2012, 54(1): 15-21.

[99] DIJKSTRA J A, VAN DER LAAN T, AKKERMAN O W, et al. In Vitro Susceptibility of Mycobacterium tuberculosis to Amikacin, Kanamycin, and Capreomycin [J]. Antimicrobial Agents and Chemotherapy, 2018, 62(3): e01724-17..

[100] MEYER M, FREIHOFER P, SCHERMAN M, et al. In Vivo Efficacy of Apramycin in Murine Infection Models [J]. Antimicrobial Agents and Chemotherapy, 2014, 58(11): 6938-6941.

[101] STINGL G. HIV-1 infection: pathogenesis of immune suppression [J]. Wiener medizinische Wochenschrift (1946), 1988, 138(19-20): 487-492.

[102] KOMURO I, SUNAZUKA T, AKAGAWA K S, et al. Erythromycin derivatives inhibit HIV-1 replication in macrophages through modulation of MAPK activity to induce small isoforms of C/EBP beta [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(34): 12509-12514.

[103] DUVENAUDT D, MACLAURIN D, AGUILERA-IPARRAGUIRRE J, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints [J]. Advances in Neural Information Processing Systems 28 (Nips 2015), 2015, 28.

[104] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural Message Passing for

Quantum Chemistry [J]. International Conference on Machine Learning, Vol 70, 2017, 70: 1263-1272.

[105] KARAMAD M, MAGAR R, SHI Y T, et al. Orbital graph convolutional neural network for material property prediction [J]. Physical Review Materials, 2020, 4(9): 093801.

[106] CHMIELA S, SAUCEDA H E, MULLER K R, et al. Towards exact molecular dynamics simulations with machine-learned force fields [J]. Nature Communications, 2018, 9: 3887.

[107] DERINGER V L, BERNSTEIN N, BARTOK A P, et al. Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics [J]. Journal of Physical Chemistry Letters, 2018, 9(11): 2879-2885.

[108] WANG W J, GOMEZ-BOMBARELLI R. Coarse-graining auto-encoders for molecular dynamics [J]. Npj Computational Materials, 2019, 5(1): 125.

[109] ALTAE-TRAN H, RAMSUNDAR B, PAPPU A S, et al. Low Data Drug Discovery with One-Shot Learning [J]. ACS Cent Sci, 2017, 3(4): 283-293.

[110] CHEN H, ENGKVIST O, WANG Y, et al. The rise of deep learning in drug discovery [J]. Drug Discov Today, 2018, 23(6): 1241-1250.

[111] HAO Z K, LU C Q, HUANG Z Y, et al. ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction [J]. Kdd '20: Proceedings of the 26th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2020: 731-739.

[112] LANDRUM G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling [Z]. Academic Press Cambridge. 2013, 8.

[113] RAMSUNDAR B, EASTMAN P, WALTERS P, et al. Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more [M]. O'Reilly Media, 2019.

[114] HU Q, WANG X, HU W, et al. AdCo: Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries [C]; proceedings of the Computer Vision and Pattern Recognition, F, 2021: 1074-1083.

[115] SURESH S, LI P, HAO C, et al. Adversarial Graph Augmentation to Improve Graph Contrastive Learning [J]. Advances in Neural Information Processing Systems, 2021, 34: 15920-15933.

[116] HU W, LIU B, GOMES J, et al. Strategies for pre-training graph neural networks [J]. arXiv preprint arXiv:190512265, 2019.

[117] LIU S, DEMIREL M F, LIANG Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules [J]. Advances in neural information processing systems, 2019, 32.

[118] RONG Y, BIAN Y, XU T, et al. Self-supervised graph transformer on large-scale molecular data [J]. Advances in neural information processing systems, 2020, 33: 12559-12571.

[119] 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 : 15870-15882.

[120] MCINNES L, HEALY J, MELVILLE J. UMAP: uniform manifold approximation and projection for dimension reduction [J]. arXiv preprint arXiv:1802.03426, 2018..

[121] MARTINS I F, TEIXEIRA A L, PINHEIRO L, et al. A Bayesian approach to in silico blood-brain barrier penetration modeling [J]. J Chem Inf Model, 2012, 52(6): 1686-1697.

[122] SUBRAMANIAN G, RAMSUNDAR B, PANDE V, et al. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches [J]. Journal of Chemical Information and Modeling, 2016, 56(10): 1936-1949.

[123] FANG Y, ZHANG Q, YANG H, et al. Molecular Contrastive Learning with Chemical Element Knowledge Graph [J]. arXiv preprint arXiv:211200544, 2021.

[124] MUREDDU L G, VUISTER G W. Fragment-Based Drug Discovery by NMR. Where Are the Successes and Where can It Be Improved? [J]. Frontiers in molecular biosciences, 2022: 110.