華中科技大學

结题报告

药物重定位研究

小组成员 鲍弈慎 冯子益 陈锦身 曹炜杰

学科专业 生物医学工程

指导教师 刘雪明

答辩日期 2024 年 12 月 06 日

摘要

当下新药的研发面临研发困难，上市周期长，安全性无法保障等诸多困境，随着人工智能的兴起，基于计算的药物重定位成为目前的研究热点。本文全面探讨了药物重定位研究的现状与进展，特别关注于如何利用基于计算的手段来提高药物重定位的效率与准确度。首先，本文总结了药物重定位模型的总体框架，并对异构网络的构建进行了系统介绍。接着，文章进一步整理归纳了基于图表示学习的药物重定位策略，介绍了图神经网络、图卷积网络和图池化网络等技术，并讨论了这些技术在图表征学习中的应用。同时，对近期研究者提出的一些新模型进行了原理阐释与总结评估，分别指出其各自优势以及面临挑战。此外，本文还探讨了基于机器学习的策略，如极限学习机和集成学习在药物重定位中的应用。最后，文章强调了模型可解释性的重要性，并提出了提升模型可解释性的多种方案，包括偏随机游走和基于语义多层关联推断的知识图谱学习模型。通过分析具体的案例，本文展示了如何利用这些方法提高药物重定位的准确性和可信度，为未来的药物重定位提供了有价值的参考。

关键词：药物重定位；图表示学习；机器学习；模型可解释性

Abstract

At present, the research and development of new drugs is faced with many difficulties such as research and development difficulties, long market cycle, and safety cannot be guaranteed. With the rise of artificial intelligence, computing-based drug repositioning has become a current research hotspot. In this paper, the current status and progress of drug relocation research are comprehensively discussed, with a special focus on how to improve the efficiency and accuracy of drug relocation using computational methods. Firstly, this paper summarizes the general framework of the drug reposition model and systematically introduces the construction of heterogeneous networks. Then, the drug repositioning strategies based on graph representation learning were further summarized, and the technologies such as graph neural network, graph convolutional network and graph pooling network were introduced, and the applications of these technologies in graph representation learning were discussed. At the same time, the principles of some new models proposed by recent researchers are explained and summarized, and their respective advantages and challenges are pointed out. In addition, this paper explores the application of machine learning-based strategies such as extreme learning machine and ensemble learning in drug repositioning. Finally, this paper emphasized the importance of model interpretability, and proposed various schemes to improve the interpretability of the model, including partial random walk and knowledge graph learning model based on semantic multi-layer association inference. By analyzing specific cases, this paper shows how these methods can be used to improve the accuracy and credibility of drug relocation, which provides a valuable reference for future drug reposition.

Key words： Drug reposition; Graph representation learning;machine learning ; model interpretability.

目录

1 绪论

1.1 研究背景和意义

药物发现是指研究人员识别出对治疗疾病有益的新药目标，通过进一步实验室研究寻找能够影响这些目标的化合物的过程。一种新药的发现与开发有着漫长的过程，涉及靶点的发现，先导化合物的发现与优化，临床前研究和临床研究，以及最后的药物许可申请和市场检测。虽然当下许多新兴的生物科学技术广泛应用于新药研发之中，但一种全新药物的开发仍然面临研发周期长，投入成本高，临床成功率低等问题。目前，成功研发一种药物的投入时间在 10 年左右，投入资金高达数亿美元。此外，新药由于崭新的结构特征，可能会产生不可预测的副作用，即使现在对药物的投入占比不断上升，新药审批的通过率不增反减。FDA 的评估实验中，只有 10%左右的药物可以通过评测，历年通过的新药屈指可数。

因此，提高药物研发的产投比，缩短研发周期，降低研发风险，已经成为生物医学研究中的一个重要课题。药物重定位便是一个合理有效的研究方向。药物重定位又称老药新用，即通过深入探索已批准上市的药物，寻找其原本功能之外的，对于其他疾病的治疗作用，开发并完善其新功能。由于是对已有药物进行再开发，药物风险相对全新药物而言更可控，且可以绕过许多批准前测试，有效降低成本与缩短周期，大大加快药物的研发效率。近几年已得到政府和相关科研机构的重视。

药物重定位最初的实现多源于偶然发现。比如用于治疗转移性结肠癌和非小细胞肺癌的血管生成抑制药物贝伐单抗，被发现对湿性黄斑变性视网膜血管异常的减缓甚至逆转有显著作用。此外，对于一些由于副作用显著而被撤市或禁止的药物，药物重定位可以扩展其使用范围，延拓其功能，使其获得“第二春”。比如沙利度胺最早上市时是作为孕妇妊娠时的止痛剂，但因为其会使婴儿致畸的恶性副作用而禁止使用。随后药学家发现沙利度胺和抗麻风药物有着良好的协同作用，且能用于治疗盘状红斑狼疮等多种皮肤病及其并发症，于 1998 年重获 FDA 的批准。在之后，这种药物又被开发出治疗血管炎，风湿性关节炎等众多疾病的功能。[1]

然而，上面几例药物重定位，其新功能的实现源于回顾性临床分析，虽然针对性强，成功率高，但多出于偶然，而非来自理性设计。由于疾病与药物种类繁多，想要通过实验一一检测药物的新功能，所需要的时间和成本都得不偿失。所以，基于人工智能等技术的

发展，各类数据库的充盈，药物重定位的研究重心逐渐转移至计算分析手段上。即利用生物信息学，系统生物学等知识，对药物重定位方案进行理性分析，使用计算预测的方法，加快大规模实验筛选的速度，缩小药物搜索空间，并通过整合多源数据，提高药物重定位的精度与效率，进一步降低成本。这种将理性设计和实验筛选相结合的办法，已成为药物重定位的主要研究策略。

1.2 研究现状

通过计算方法的药物重定位重点在于对药物开发体系中的各种元素进行关联预测，以更加精准地预测出药物可能具有的新功能。对于预测关系而言，存在三个基准任务：药物-疾病关联预测，药物-药物关联预测和药物-靶点关联预测。其中，药物-疾病关联预测和药物-药物关联预测可以较为直观地展示药物的适应症及药物间的互作关系。但是，由于大部分药物的作用机理是和生物体内的分子靶点（包括酶，离子通道，载体蛋白等蛋白质）相结合，抑制或提高相应蛋白的活性以发挥作用，比如抗炎药物布洛芬通过抑制 COX-2 的活性，影响可诱导炎症细胞的前列腺素的生成，从而起到解热镇毒的作用。因此，药物-靶点关联预测可以更好的揭示药物潜在的药理学机制，对药物功效及其副作用的研究有着重要意义。

然而，在当下研究之中，药物数据库中已知的药物，靶点数量相当之多，而能成功建立药物-靶标关联的只是少数。如何从庞大的数据资源中精准地发现药物与靶标的相互关联，是药物重定位的关键课题。若采取传统的生化试验方法，依然会面对成本高，周期长，风险高等经典问题。因此，以计算方法为核心的预测方式仍为研究的主流方向，有利于缩小候选药物的范围，提高筛选的效率与准确度。

基于配体的方法，基于模拟结构的方法，基于化学基因组的方法是药物重定位的三大算法。其中，基于配体的方法和基于结构的方法相对传统。基于配体的方法利用相似蛋白质往往会和相似的分子结合的思想，通过计算新老配体之间的相似度来预测可能的相互关联。基于结构的方法根据蛋白质和药物的三维结构在计算机上进行模拟实验，以预测其可能存在的相互关系。这两种方法都存在着不小的局限性。基于配体的方法在配体信息缺失或不足的情况下，预测的精准程度会大大下降。基于结构的方法在处理结构极其复杂或者结构尚未完全探清的蛋白质时，模拟计算的效率和准确度都会下降。而新兴的化学基因组方法可以充分利用广泛的生物医学相关数据，将已知的化学信息和靶点的基因组信息嵌入一个统一的药理学空间内，进而提高预测性能。该方法的优势在于将药物和蛋白质的辅助

信息和已知的药物-靶点信息紧密结合，而其也有明显的不足，如药物-蛋白质的关系稀疏，缺乏实验可以表明的阴性相互作用样本。

综上所述，基于化学基因组方法的药物-靶标预测是药物重定位中的一大关键，对于推动现代医药发展意义重大。但是，为了解决药物-蛋白质关系稀疏等问题，需要借鉴人工智能领域的网络技术，采取图学习等算法，将稀疏的关系嵌入低维空间中，从而提高有效信息的密度。该方面的探索是急需解决的问题。

2 生物异构网络的构建

2.1 概述

药物与靶点的相关特征,可以通过其在整个生物资源库中所代表的元素来表达与描述。随着大数据相关技术不断发展，生物医药的数据库也在不断更新与扩充。为了使越来越多的数据可以得到充分处理，以最大限度发挥数据库提供的多维度多视角的优势，我们可以引入异构信息网络。引入异构信息网络有以下优势，其可以一定程度上延伸药物或靶点间的长距离关系，有效地使相对稀疏的信息产生联系，从而增加预测的多样性；通过点与点之间丰富复杂的边关系，可以表达出大量语义信息，涵盖对应的语义相关性，能够辅助潜在链接的探索，提高预测的准确度；在网络中，待预测的药物-靶点关系与已知的药物-靶点关系的相关性更加可视化，能够为预测实例提供理性支持与新的见解，提高预测结果的可信度。但是，由于生物医学数据信息量巨大，且存在大量误差信息和残缺信息，噪声性强，对预测的准确度和效率都会产生消极影响。因此，基于生物异构网络的药物-靶点预测仍面临不小的风险与挑战。

在本章中，会先介绍异构网络的有关概念，明确网络构建对数据处理的意义。并介绍一些常用的生物数据库，从生物医药的角度给出构建网络的丰富数据资源，为生物异构网络的构建提供基础。接着通过生物数据库中的信息，模拟构建简单的网络，并提供生物异构网络中药物-靶点预测的具体解读方式。为后文有关信息特征提取的内容奠定基础。

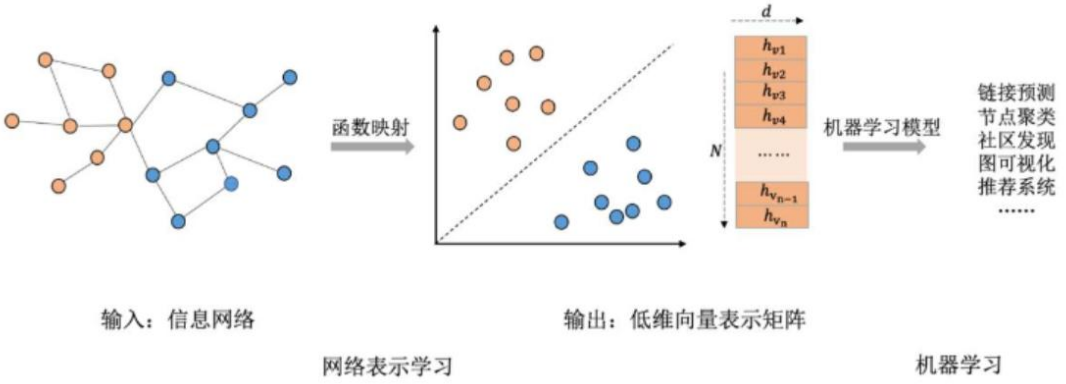
2.2 异构网络的相关概念

2.2.1 异构网络的背景

当下，网络数据呈现爆炸式增长，但其中未知的，有意义的信息虽然含量极大，但密度稀疏。因此，对网络数据进行挖掘与提取已成为研究的热点。传统的网络学习多采取建立同构网络的方式，以基于邻接矩阵或者邻接表的结构表示图数据。在同构网络中，所有的节点和边全部属于单一类型，仅仅表示了节点之间的连接关系，且雷同性较强。这种网络构建是对现实中的网络关系的简化，忽略了节点自身的属性特征，无法体现现实的高阶结构关系。随着数据量的快速膨胀，这种传统方式中的数据形式过于稀疏，很大程度上损害了网络的表示效果，导致计算大规模网络分析任务时，成本高昂且可行性不高。

随着深度学习技术的快速发展，广泛应用，在复杂图挖掘上优越性较强的网络表示学

习逐渐得到关注。网络表示学习又称为网络嵌入，主要功能是将原本稀疏的高维节点通过某些特定的数学模型或算法，嵌入到一个低维的向量空间中进行表示。这种方法可以有效地应对网络数据稀疏的问题，节省了不必要空间浪费。如图所示[3]，采取这种方法可以让信息网络中存在边联系，但是在高维空间中距离较远的两个节点投影进低维嵌入空间中，使二者距离大大缩短。由此得到的矩阵可以直接投入于机器学习中，应用于特征提取与数据分析预测。网络表示技术的实际应用相当广泛，比如将生物医学相关数据投入网络之中，则对药物-靶点关联预测起到重要帮助。



传统的同构网络的节点稀少，边关系简单，即使应用于网络表示学习技术，也无法得到充足的有效信息用于提取和预测。而异构网络可以对多类型的实体和边关系进行更加完整真实地表述，更接近于现实中的网络结构。而其中种类繁多的节点和连接关系，组成不同的网络结构，各自蕴含着丰富的语义信息。可以为后续的特征提取等步骤提供充足的资源。这也是异构网络的重要意义。

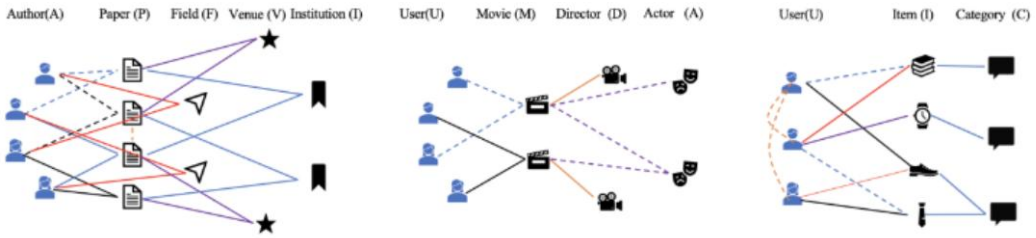
2.2.2 异构网络的定义

异构网络与同构网络最大的区别在于异构网络包含了更为丰富的节点类型和连接类型，构成了结构更为复杂的网络。

异构网络一般被定义为图 $= (\mathcal{V}, \mathscr{E}, \phi, \psi)$，其中$\mathcal{V}$表示节点集合，$\mathscr{E}$ 表示连接关系集合。ϕ：$\mathcal{V} \rightarrow \mathcal{A}$和$\psi$：$\mathscr{E} \rightarrow \mathscr{R}$ 分别表示节点类型映射函数和连接类型映射函数。其中$\mathcal{A}$和 $\mathscr{R}$ 分别表示节点类型集合和关系类型集合，每个节点$v \in \mathcal{V}$属于节点类型集合$\mathcal{A}$中的一个特定节点类型$\phi(v) \in \mathcal{A}$，每个连接$e \in \mathscr{E}$ 属于关系类型集合 $\mathscr{R}$ 中的一个特定关系类型$\phi(e) \in \mathscr{R}$，如果两个连接属于同一关系类型，则这两个连接共享相同的起始节点类型和结束节点类型。满足条件$|\mathcal{A}| + |\mathscr{R}| > 2$ 的网络的网络是异构，否则，被视为同构网络。

下图则介绍了三个典型的异构网络模式[2]。在学术合作网络中，存在五个节点，包括

作者，论文，机构，场所，领域。仅仅这五个节点中即存在复杂的连接关系。如关于论文与机构，场所相关；关于作者与论文，领域相关等等。在电影推荐网络中，由用户，电影，导演，演员四个节点构成的网络，包含了用户观看导演指导的电影，用户观看演员出演的电影，演员出演导演制作的电影等等。在电商网络中，只有用户，物品，种类三个节点，仍存在用户间的联系，用户对某种类的商品的关系，而同两个节点存在的关系也可能不同，比如用户对商品的关系可能是购买，也可能是收藏，加入购物车，好评等等。综上所述，异构网络通过复杂多样的联系，即使只有少数节点，也能产生各种各样的语义信息。



(a)学术合作网络　(b)电影推荐网络　(c)电商网络

2.3 生物异构网络的构建

2.3.1 构建生物异构网络的数据库

上文主要介绍了异构网络在各领域的一般情况。而为了实现对药物-靶点关联的预测，则需要构建出以生物医药相关信息作为节点的异构网络，为后续特征提取等一系列步骤奠定基础。在生物异构网络中，药物，靶点，疾病等元素作为网络中的节点，药物与靶点的作用关系，药物与疾病间的关系等等作为边关系。为了得到功能完整，语义丰富的生物异构网络，则需要大量的数据以填充异构网络中错综复杂的边-点关系。而随着分子生物学的发展，大量的基因组，药理学，化学等数据得以不断积累，并储存于人们建立的数据库之中，进行管理，维护，以及不断更新。其中相当多的数据库是完全公开的，供世界各地的研究人员使用。这些数据库不仅可以提供已获批药物，生物分子的信息，还可以提供一些正在实验中的实体信息，具有潜力的细胞靶点等。对于构建网络所需要的异构信息，这些数据库也能提供类型丰富的数据源。总而言之，这些数据库是构建生物异构网络不可或缺的基础。

下面介绍一些常用的数据库。

DrugBank 被誉为药物的百科全书，是一个包含海量生物信息学，化学信息学资源的庞大数据库。DrugBank 包含超过 17,000 个药物条目，其中包括 2800 个批准的小分子药物、1672 个批准的生物制品（蛋白质、肽、疫苗和过敏原）、135 种营养品以及超过 6726 种实

验性（发现阶段）药物。DrugBank 还链接了 5429 个非冗余蛋白序列（即药物靶标/酶/转运体/载体）。此外，DrugBank 还将详细的药物数据（包括药理学，制药等）与综合药物靶标信息（包括序列，途径等）相结合，并支持全面而复杂的搜索，便于使用者快速提取出有用的药物-靶点，药物-疾病相互作用信息等。DrugBank 中的所有信息都通过了生化实验的验证，准确性和权威性均较高。

HPRD 是专门储存人类蛋白质相互作用信息的数据库。该库对人类蛋白质功能的描述信息包括翻译后修饰，蛋白质-蛋白质相互作用，蛋白质-疾病关联，酶-底物关联等。除此之外，HPRD 还提供了蛋白质的表达谱、分类、结构域、亚细胞定位、转录后修饰、通路等其他信息。准确度方面，HPRD 的信息是由生物学家经过文献分析和实验考证得到的，数据质量上有明显优势。

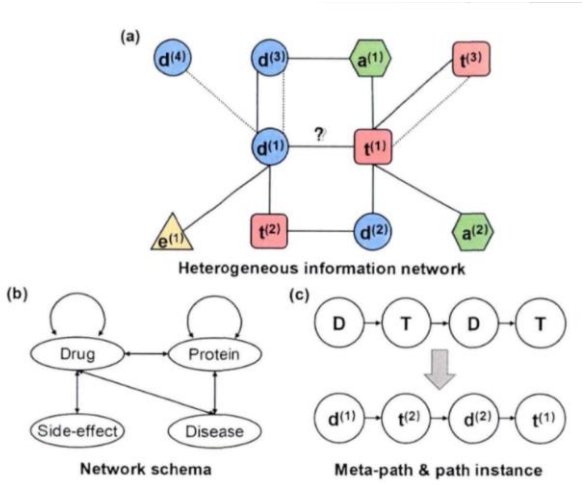
TTD 是由浙江大学和新加坡国立大学生物信息相关机构共同创建的药物和医疗知识库。提供了关于已知的和待探索的治疗蛋白、核酸靶标、通路信息以及靶标相关药物的信息。其 中靶标相关数据以及对应的用于临床中使用和治疗的药物数据，对于加速药物重定位有着 重要作用。该数据库中目前包含了 3419 种靶标和 37316 种药物，分别涵盖了 583 个蛋白生 化类别和 958 个药物治疗类别。世界卫生组织发布的最新版《国际疾病分类》（ICD-11）编码已被纳入 TTD 中，以便更明确地定义疾病的类别。

下图包括了更多常见的数据库。

数据库	Web 服务地址
DrugBank[40]	http://www.drugbank.ca
HPRD[41]	http://www.hprd.org/
CTD[42]	http://ctdbase.org/
SIDER[43]	http://sideeffects.embl.de/
PubChem[44]	https://pubchem.ncbi.nlm.nih.gov/
TTD[45]	http://db.idrblab.net/ttd/
SuperTarget[46]	http://bioinf-apache.charite.de/supertarget
STITCH[47]	http://stitch.embl.de/
TDR targets[48]	http://tdrtargets.org/
BRENDA[49]	http://www.brenda-enzymes.org/
ChEMBL[50]	https://www.ebi.ac.uk/chembldb
BindingDB[51]	http://www.bindingdb.org/bind
ChemBank[52]	http://chembank.broad.harvard.edu/
DCDB[53]	http://www.cls.zju.edu.cn/dcdb/
CancerDR[54]	http://crdd.osdd.net/raghava/cancerdr/
ASDCD[55]	http://asdcd.amss.ac.cn/
SuperPred[56]	http://prediction.charite.de/

2.3.2 模拟构建生物网络及其解读

在基于网络的药物-靶点关联预测研究中，为了构建语义丰富的生物异构网络，需要设立多种节点和边关系，并从数据库中检索取得相应的资源。在模拟构建的简单网络中，节点包括药物，蛋白质，副作用，疾病。边类型包括药物-蛋白质关系，药物-药物关系，蛋白质-蛋白质关系，药物-疾病关联，药物-副作用关联，蛋白质-疾病关联等。其中，涉及药物之间关联的数据资源可以从 DrugBank 中获取，涉及蛋白质关系的数据可以从 HPRD 中获取。涉及疾病的数据可以从 TTD 中获取，涉及副作用相关的数据可以在 SIDER（市售药物及其记录的副作用数据库）中提取。



上图[4]是基于这些节点和边关联模拟构建的一个简单的生物异构网络示意图。接下来对该示意图做出一些解释，使概念的了解更加直观。

该示意图上有着四个节点。d 代表药物，t 代表靶点蛋白，a 代表疾病，e 代表副作用。节点间的边代表节点间的关联。不同类型的边也代表着不同的语义关系。比如靶蛋白 t1 的功能与疾病 a2 的产生有关，药物 d1 和药物 d3 存在相互作用。如果两个节点不直接相连，而是通过其他节点间接联系在一起，其所表示出的交互语义将更加复杂。比如探究药物 d1 和靶蛋白 t1 间的联系。可以由两条不同路径将二者联系在一起。如 d1-d3-a1-t1 路径，传达的语义是“和药物 d1 相互作用的药物 d3 可以治疗疾病 a1，而疾病 a1 的发生与靶蛋白 t1 有关”，d1-t2-d2-t1 路径传达的语义是“药物 d1 和 d2 可以共同作用于靶蛋白 t2，且药物 d2 还可以作用于靶蛋白 t1。

由此可见，即使是相当简单的异构网络，经解读后也蕴含着极为丰富的信息。

2.3.3 生物异构网络的意义

在庞大的药理空间中，各种生物学，化学信息错综复杂，而存在价值的信息密度较低。不同元素在空间中的距离较远，关系稀疏，不易直接提取并利用。而通过构建异构网络的方式，可以有效缩短元素间的语义距离，使稀疏信息的可利用性变强。此外，由于大部分的深度学习方法无法直接利用稀疏的矩阵，通过构建网络，可以融合多个网络进行特征降维，为后续的一系列算法的实现做基础准备。

3 基于图表示学习的药物重定位策略-研究现状

3.1 引言

近年来，由于新药研发困难、上市周期长等因素，围绕着药物重定位的研究愈发热门，其中基于计算的药物重定位以其优秀的效率与准确度引得众多研究者在该领域进行研究。Xuan 等研究者[5]开发了一种融合卷积神经网络和双向长短时记忆网络的药物再定位方法。这种方法通过结合两种网络架构来学习药物与疾病之间的相互作用，并且采用注意力机制来评估不同路径对药物-疾病关联预测的影响。另一方面，Moridi M 等研究者[6]运用深度学习技术从不同来源的数据中提取药物和疾病之间的非线性特征，以此来提高药物-疾病关联预测的准确性。然而，这些传统的机器学习与深度学习方法对于非欧氏空间数据地处理能力有限，不能很好地捕捉药物与药物、药物与靶点、药物与疾病之间复杂地拓扑关系，处理 DTI 与 DDSI 预测时所需的多源数据环节相对乏力，预测精度不尽人意。而基于图表示学习的药物重定位策略能够较好地完成任务。图神经网络（GNNs）作为一种基于图结构的深度学习技术，发展势头迅猛，众多研究开始将药物与疾病的关系网络以图的形式表现，以探索两者之间潜在的联系。例如，Wang 等研究者[7]采用二部图卷积机制，以蛋白质节点作为信息传递的枢纽，构建了药物与疾病间的宏观和微观信息模型，充分利用了药物与疾病之间的相互作用信息，以预测药物可能治疗的疾病。在寻找 COVID-19 的抗炎药物方面，Wang 等研究者[8]构建了一个包含药物、疾病、蛋白和副作用的异构图，并运用深度图表征学习技术来识别潜在的新药物。Cai 等研究者[9]考虑到不同网络拓扑结构中的域内和域间信息，使用图神经网络方法来学习这些嵌入之间的关系，以此增强药物和疾病的表示，他们提出的 DRHGCN 方法在药物重定位任务的预测精度上取得了新的突破。Yu 等研究者[10]指出不同图卷积层对特征学习的贡献不同，因此他们提出了一种基于层次注意力机制的图卷积网络方法，用于药物重定位，通过捕获不同层级间的关系，提升了模型预测的准确性。

近年来，陆续有研究者提出新的基于图表示学习的方法用于药物重定位，相较传统的方法有诸多优势。Sun 等人[11]注意到之前大多数研究提出的方法中，药物表示是静态的，忽略了不同疾病环境对药物的影响导致预测精度降低，因此提出了基于图神经网络的合作伙伴图药物重定位方法（PSGCN）以解决上述问题。而 Jiao 等人[12]注意到生成节点嵌入时目前的研究采用的主要是基于元路径的方法，对于节点语义特征的捕捉能力较强，但是对于节点所处邻域拓扑特征的捕捉能力就相对较差。Jiao 等人针对现有模型存在的问

题进行了一定的改进，提出了一种基于子图聚合生成节点嵌入的方法。该方法可以更有效地捕捉节点的拓扑特征，以生成质量更好的节点嵌入，用于后续的预测任务。此外，相较于单独对各个数据源预学习而后集成结果的方式，该模型直接融合多源数据构建异构网络学习，能够尽可能地避免模型集成之前的信息损失。本章将对以上几种方法做详细介绍与评估。

3.2 图表示学习概述

处理高维且结构复杂的网络数据颇具挑战，因此网络分析往往依赖于有效的网络表征技术。这些技术旨在将非欧几里得空间中的原始网络数据转换为向量形式，以保留网络中的内在联系。在理想情况下，图中彼此接近的节点在经过嵌入后，在向量空间中的距离也应相近。这种转换过程，即图表示学习或图嵌入学习，其核心目标是通过学习图的拓扑结构来获得节点的低维、密集表示。图表示学习不仅能压缩数据、实现降维，还能使得到的向量更易于计算处理。

尽管图表示学习在处理图数据方面表现出色，但其有效性也依赖于几个关键因素。首先，生成的嵌入必须能够准确描述图的属性信息。其次，嵌入方法应具有良好的可扩展性。最后，需要在嵌入质量和维度之间找到平衡点。较高的维度虽然能保留更多原始信息、提升嵌入质量，但会增加时间和空间复杂度；而较低的维度虽能提升模型运行效率，却可能丢失部分原始信息。

在生物医学领域，药物-疾病关联预测是一个关键的链路预测问题，其中特征提取尤为关键，即如何将图表示学习应用于生成高质量的节点表征。节点表征的挑战在于将非数值型输入转换为数字化特征向量。对于图像数据，由于其具有固定的二维结构，可以通过二维矩阵来表示其特征。对于语音数据，可以基于频谱定义滑动窗口来进行特征表示。然而，对于图这种非欧几里得结构的数据，由于其结构的灵活性和缺乏平移不变性，特征提取变得更加困难。目前，图结构特征提取的常见方法主要有两种：线性化转换方法和图神经网络方法。线性化转换通过在网络中进行随机游走来生成节点序列，从而学习节点特征。而图神经网络则利用神经网络来学习图结构数据，提取和挖掘节点特征。

3.3 图表示学习相关技术

3.3.1 图神经网络

图神经网络（GNN）是一种专门处理图结构数据的深度学习模型，它通过模拟消息在图中的传播来学习节点或整个图的表示。GNN 的核心在于节点特征的聚合和更新，其中每

个节点会收集来自其邻居的信息，并结合自身特征进行表示的更新。这一过程可以跨越多个层次，使得模型能够捕捉到从局部到全局的不同尺度特征。GNN 的灵活性允许它处理包括节点分类、图分类和链接预测在内的多种任务，并且能够适应不同类型的图数据，包括同构图和异构图。随着深度学习技术的发展，GNN 已经成为图数据分析中一个强大的工具，其在社交网络分析、生物信息学、推荐系统等领域展现出广泛的应用潜力。

3.3.2 图卷积网络

图卷积网络（GCN）的概念最早由阿姆斯特丹大学的 Thomas N.Kipf 与 Max Welling 于 2017 年提出[13]，他们称之为“a scalable approach for semi-supervised learning on graph-structured data that is based on an efficient variant of convolutional neural networks which operate directly on graphs.”

Thomas 等人将传统卷积神经网络与新出现的图神经网络相结合，直接在图上进行卷积等操作。研究者通过局部一阶近似谱图卷积来激发所提出的卷积架构的选择。GCN 模型可以线性地扩展到图的边数，并且能够学习隐藏层表示，这些表示编码了节点的局部图结构和特征。

GCN 的核心是图卷积操作，它将归一化的邻接矩阵与节点特征矩阵相乘，并引入可学习的权重矩阵。公式如下：

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) .$$

通过这种方式，GCN 能够有效的学习到节点在图中的嵌入表示，这些表示可以用于下游任务。

3.3.3 图池化网络

图池化网络（Graph Pooling Networks）是一种用于图结构数据的神经网络技术，最早由 Mathias Niepert, Mohamed Ahmed, Konstantin Kutzkov 等人提出。它旨在对图进行下采样，减少节点数量，同时保留图的主要结构和特征信息。这种技术对于处理图数据的尺寸变化性、提高计算效率以及在图分类等任务中降低过拟合风险至关重要。

在图池化网络中，关键步骤包括节点聚合和节点选择。节点聚合操作将一组节点的特征合并为一个单一的表示，而节点选择操作决定哪些节点将被保留在下采样后的图中。这两个步骤通常在图的每一层中重复执行，直到达到所需的图大小或深度。

图池化网络的设计需要考虑如何有效地捕捉局部和全局的图结构信息。一些常见的图

池化策略包括基于聚类的方法，其中节点被分组到聚类中，并且每个聚类的中心或代表节点被选为下采样图中的节点；以及基于图卷积的方法，其中节点特征通过图卷积网络进行更新，然后基于更新后的特征进行节点选择和聚合。

图池化网络的应用包括但不限于图分类、图生成和图匹配任务。它们使得图神经网络能够处理更大规模的图，同时保持或甚至提高模型的性能。随着图神经网络领域的不断发展，图池化网络已成为构建高效图神经网络架构的重要组成部分。

3.4 基于图神经网络合作伙伴图的药物重定位模型（PSGCN）

由 Sun 等人提出的 PSGCN 方法通过提取与目标药物-疾病对相关的上下文信息，构建特定的合作伙伴图，然后利用图神经网络学习这些合作伙伴图的表示。该方法设计了层自注意机制来获取多尺度层信息，并使用图排序池化策略获取每个合作伙伴图的最终表示。最后，将药物-疾病关联预测问题转化为图分类问题，使用多层感知机对药物-疾病结果进行预测。

3.4.1 模型实现

如下图所示，PSGCN 模型可分为三个部分：

合作伙伴图的提取：以特定的药物-疾病对作为中心节点，围绕这些中心节点，提取了其周围 h 跳（即 h 层邻居）作为相关的上下文信息，以此来提取所谓的合作伙伴图。

合作伙伴图的表示学习部分：利用图卷积网络（GCN）对这些合作伙伴图进行处理，以学习图中节点的表示。通过图卷积操作有效地提取节点的局部连接模式和特征。采用图池化技术来整合图中的信息，从而得到一个综合的合作伙伴图表示。这一步骤的目的是将图结构数据转换为可以用于机器学习模型的向量形式。

关联预测部分：将药物-疾病关联预测（DDSI）问题转化为图分类问题，并利用多层感知机预测潜在的药物-疾病关联。

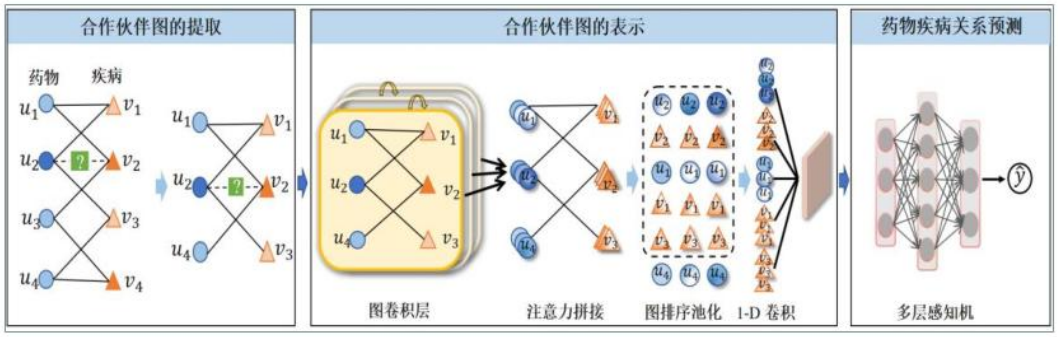


图 3-1 PSGCN 模型架构图

3.4.2 模型性能评估

为了全面评估 PSGCN 模型的性能，研究者采用了三个标准的药物重定位数据集：Gdataset、Cdataset 和 LRSSL，并执行了十次十折交叉验证实验。实验结果表明，PSGCN 在所有数据集上均展现出优越的性能，其 AUROC 和 AUPRC 值均高于其他对比方法，包括 SCMFDD、iDrug、GRMF、NRLMF、NIMCGCN 和 DRWBNCF。

表 3-2 PSGCN 和其它五种算法 AUROC 和 AUPRC 的比较

Dataset	SCMFDD	iDrug	GRMF	NRLMF	NIMCGCN	DRWBNCF	PSGCN
				AUROC			
Gdataset	0.7731	0.9078	0.7476	0.9097	0.8234	0.9061	**0.9485**
Cdataset	0.7896	0.9294	0.7469	0.9257	0.8393	0.9277	**0.9566**
LRSSL	0.7698	0.8993	0.6924	0.8854	0.7581	0.9232	**0.9395**
				AUPRC			
Gdataset	0.7749	0.9265	0.7978	0.9302	0.8590	0.9307	**0.9558**
Cdataset	0.7878	0.9454	0.8007	0.9441	0.8728	0.9476	**0.9627**
LRSSL	0.7860	0.9212	0.7689	0.9102	0.7962	0.9330	**0.9462**

在评估方法和指标方面，研究者采用了 ROC 曲线、PR 曲线、AUROC 和 AUPRC 等常用评价指标。这些指标能够全面反映模型在药物重定位任务中的预测性能，尤其是在数据不平衡的情况下。实验结果通过十次十折交叉验证的均值来确保结果的稳定性和可靠性。

与其他模型的对比分析显示，PSGCN 在 Gdataset、Cdataset 和 LRSSL 数据集上的 AUROC 值分别为 0.9485、0.9566 和 0.9395，均高于其他方法。在 AUPRC 值方面，PSGCN 同样表现优异，分别为 0.9558、0.9627 和 0.9462。这些结果不仅证明了 PSGCN 在药物重定位任务上的有效性，也显示了其在不同数据集上的泛化能力。

研究者还进行了消融实验来分析合作伙伴图提取时不同跳数深度对模型预测结果的影响，以及模型在不同稀疏数据程度下的鲁棒性。这些分析进一步验证了 PSGCN 模型的鲁棒性和对不同数据条件的适应性。综合实验结果，PSGCN 作为一种新的基于图神经网络的

药物重定位方法，其在预测药物-疾病潜在关联方面具有显著的优势和潜力。

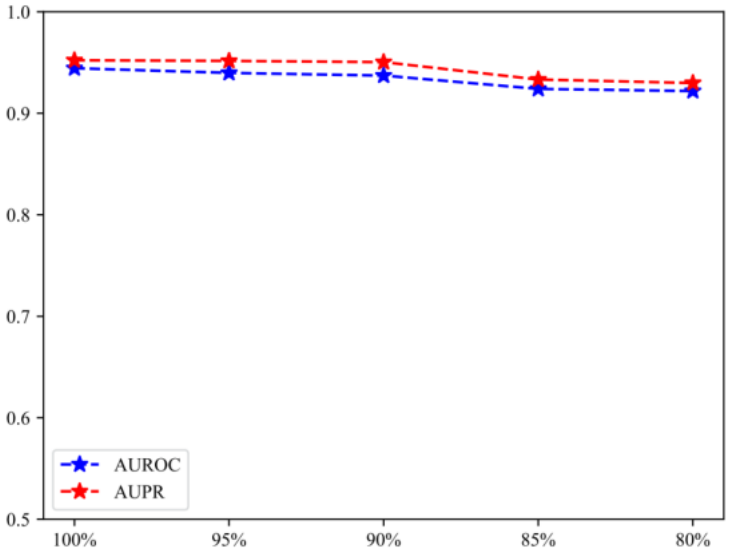


图 3-8 在 Gdataset 数据集上稀疏度对 PSGCN 模型的 AUROC 和 AUPRC 的影响

此外，研究者通过案例研究进一步验证了 PSGCN 模型的实际应用能力，选取了小细胞肺癌和乳腺癌作为研究对象。利用 Gdataset 数据集，研究者将已知的药物-疾病关联作为训练集，并对这两种疾病缺失的药物-疾病关联进行预测。预测结果经文献验证，显示 PSGCN 模型在小细胞肺癌的预测药物中命中率达到 80%，在乳腺癌中达到 100%，证明了 PSGCN 在发现特定疾病潜在治疗药物方面具有较高的准确性和实用价值。

3.4.3 挑战与展望

本研究提出的基于图神经网络的合作伙伴图药物重定位方法（PSGCN）虽然在实验中表现出色，但仍存在一些问题和限制。例如，PSGCN 在处理药物-疾病关联时，依赖于已知的药物-疾病关联数据的质量和数量。在数据稀疏的情况下，模型的性能可能会受到影响，因为有限的关联信息可能不足以提供丰富的上下文环境。也就是说，当数据集稀疏或存在空缺，以及面对新情境时，该模型可能遭遇冷启动问题。此外，尽管 PSGCN 通过合作伙伴图策略捕捉了药物在不同疾病环境下的差异性，但这种方法在处理大规模和复杂网络时可能会遇到计算效率和可扩展性的挑战。随着网络规模的增长，模型的计算复杂度也会增加，这可能影响模型在实际应用中的响应速度和处理能力。PSGCN 模型的可解释性也是一个挑战。虽然模型能够预测潜在的药物-疾病关联，但对于预测结果背后的生物学机制和逻辑推理的解释能力有限。这对于药物重定位领域尤为重要，因为研究人员和临床医生需要理解模型预测的依据，以确保药物的安全性和有效性。针对以上挑战，仍有待未来

研究取得进一步突破。

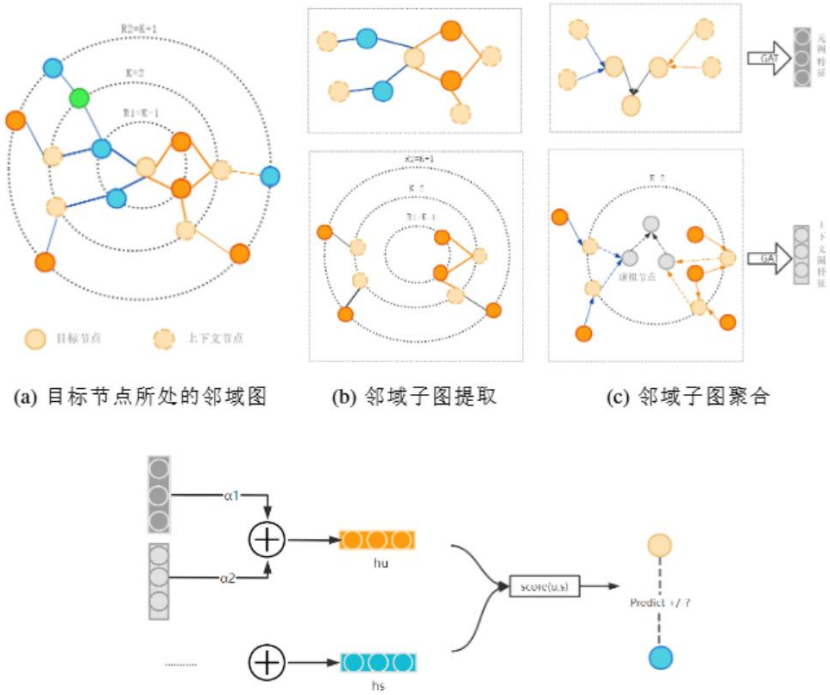
3.5 基于子图聚合的药物重定位模型

3.5.1 方法优势

在探讨节点嵌入生成的研究中，传统的基于元路径的方法虽然在捕捉节点的语义特征方面表现出色，但在识别节点的邻域拓扑特征方面则稍显不足。为了解决这一问题，Jiao[8]提出了一种创新的方法，即通过子图聚合来生成节点表征，这种方法能够有效地提升对节点拓扑特征的捕捉能力，进而生成更高质量的节点嵌入，以支持后续的预测任务。此外，Jiao 的方法相较于传统的多数据源预学习后集成结果的方式，通过直接融合多源数据构建异构网络学习模型，能够在模型集成前最大限度地减少信息损失。简而言之，Jiao 的研究通过子图聚合的方式，不仅提升了节点嵌入的质量，还优化了多源数据的融合过程。

3.5.2 模型实现

该模型 NSAP（Neighborhood subgraph aggregation method for durg-disease association prediction）基于子图聚合，由 4 个主要部分组成：邻域子图提取、邻域子图聚合、特征融合、药物-疾病关联预测。下图为整体框架：



(a) 目标节点所处的邻域图　(b) 邻域子图提取　(c) 邻域子图聚合

(d) 特征融合　(e) 药物疾病关联预测

K：目标节点对应的元路径长度　[R1，R2]：上下文节点的邻域范围

图 3-2 NSAP 框架图

这幅图展示了基于邻域子图聚合的模型 nsap 的整体框架。(a)目标节点所处的邻域图，图的中心位置是目标节点。图中不同颜色代表了不同种类的节点。目标节点能够通过元路径游走抵达的节点称为上下文节点，在图中用虚线圈表示，元路径长度记作 K，本图中 K＝2。R1 和 R2 分别是邻域的下界和上界，我们可以看到本图中 R1 是一，R2 是三。在邻域图中，我们提取元图和上下文图。元图是指从目标节点出发，基于特定元路径游走生成的同构图,它可以用来捕捉与目标节点相似的节点信息以及目标节点周围的局部拓扑结构。上下文图是指目标节点的同构邻居周边的一阶邻域图，它用于捕捉目标节点周边的全局拓扑结构及处于其中的节点信息。接下来，我们分别对元图和上下文图结合注意力机制 GAT 进行聚合。为了方便聚合，我们可以引入虚拟节点（c）来传递信息。最终，我们得到元图特征和上下文图特征。下一步，我们要对这其进行特征融合。此处 α1 和 α2 分别代表元图特征和上下文图特征所占的权重,经过计算之后得到 hu 也就是药物节点的最终特征。同理，我们得到 hs 也就是疾病节点的最终特征。基于 hu 和 hs 的数值，可以计算出一个分数，这个分数用于评估药物 u 和疾病 s 潜在的关联程度，从而用于药物重定位的指导。

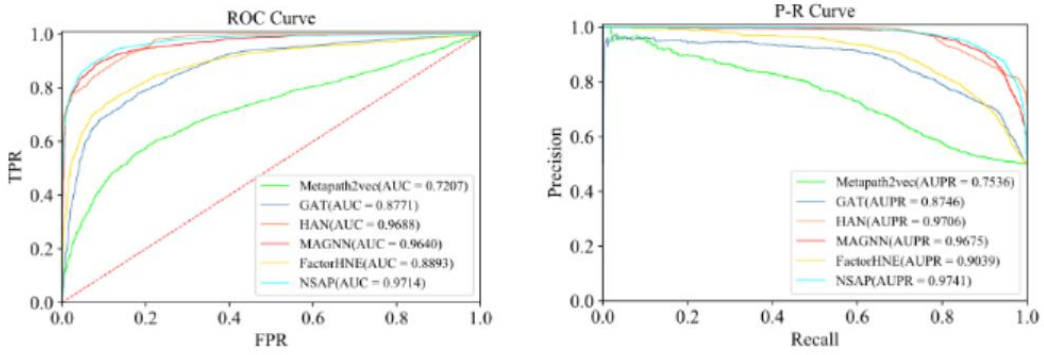
3.5.3 模型性能评估

对于 NASP 模型，研究者采用了 AUC、AUPR、F1-score、Accuracy、Precision 和 Recall 等指标来衡量模型性能。通过与多种基线模型的对比实验，包括 Metapath2vec、GAT、HAN、MAGNN 和 FactorHNE，验证了所提 NSAP 模型的有效性。实验结果表明，NASP 在 AUC，AUPR 和 AAccuracy 三个指标上相较目前最好的模型 MAGNN 提升了 0.09%，0.07%和 0.18%，而在 Precision 和 Recall 两个指标上略差于 MAGNN。

表 3-2 随机负采样策略下模型的预测效果

方法	AUC	AUPR	F1-score	Accuracy	Precision	Recall
Metapath2vec	0.7213	0.7543	0.6749	0.6746	0.6746	0.6752
GAT	0.8691	0.8647	0.7819	0.7919	0.8220	0.7465
HAN	0.9584	0.9609	0.8695	0.8695	0.8695	**0.8695**
MAGNN	0.9600	0.9643	0.8812	0.8892	**0.9438**	0.8289
FactorHNE	0.8893	0.9039	0.8108	0.8146	0.8276	0.7946
NSAP	**0.9609**	**0.9650**	**0.8848**	**0.8908**	0.9351	0.8397

所有模型的 ROC 和 PR 曲线如下图，由图可知，NASP 不仅最佳性能优于其他基线模型，而且运行平稳，平均性能也优于其他基线模型。而采取了全部邻居节点生成节点特征的 HAN 模型，虽然最佳性能与 NASP 平分秋色，但平均性能明显略逊一筹，可见选取邻域子图的重要性。



(a) ROC 曲线　　(b) PR 曲线

图 3-5 模型的 ROC 曲线和 PR 曲线对比图

综上所述，NASP 模型整体性能表现较好。它考虑到了不同节点的差异，对邻域节点进行了采样。同时，利用子图聚合的方式充分挖掘了目标节点的局部和全局拓扑特征，生成了高质量的节点嵌入。最后，它利用子图结构自动生成了节点间边的权重，这可以很好地评估不同节点之间的关联程度，使得下游的预测任务从中受益。

3.5.4 挑战与展望

从各项性能来看，NSAP 模型表现已经十分优秀。但该模型仍面临一些问题。注意力头的选取数量和邻居节点的采样数量对模型性能均有显著影响，需要谨慎决定。实际运用中，选取邻居节点数量较多会导致计算负担加重，性能下滑，较少又会导致准确度下降。因此，如何在选取足够邻居节点的情况下保证计算效率是亟需解决的问题，有待进一步探索。

3.6 本章总结

本章全面介绍了基于图表示学习的药物重定位研究现状，涵盖相关技术解释以及模型框架说明，着重强调表征学习步骤，并详细介绍了两个较新的性能优秀的基于图表示学习的药物重定位模型 PSGCN 与 NSAP。本章分析了图表示学习的优势所在，并对比现有模型的利弊，指出了有待进一步研究的方向。

4 基于机器学习的药物重定位策略-研究现状

4.1 引言

识别药物-靶标相互作用将大大缩小候选药物的搜索范围，因此可以作为药物发现的至关重要的第一步。考虑到体外实验极其昂贵且耗时，高效的计算预测方法可以作为药物-靶标相互作用（DTI）预测的有希望的策略。在这篇综述中，厦门大学陈若蓝团队的目标是专注于机器学习方法，并提供一个全面的概述。首先，该团队总结了药物发现中经常使用的几个数据库的简短列表。接下来，他们采用分层分类方案，介绍了每个类别中的几种代表性方法，特别是最新的最先进方法。此外，陈若蓝比较了每个类别中方法的优缺点。最后，陈若蓝团队讨论了机器学习在 DTI 预测中剩余的挑战和未来的展望。他们可能为未来的研究人员提供基于机器学习的 DTI 预测的参考和教程见解。

4.2 概述

在预测药物靶标关系的方法中，化学基因组学方法可以充分利用丰富的生物医学数据，有利于预测。在这样一个 DTI 预测问题中，主要挑战是已知药物-蛋白相互作用的稀缺性和未验证的负药物-靶标相互作用样本。这些化学基因组学方法可以分为不同的类别，如基于机器学习的方法、基于图的方法和基于网络的方法。在所有化学基因组学方法中，基于机器学习的方法因其可靠的预测结果而受到最多关注。这些方法通常利用药物和靶标的化学和生物学特征，并采用各种机器学习技术来预测药物和靶标之间的相互作用。

机器学习有以下两个特点：1，在研究环境计算资源有限的情况下，选择更轻量级的算法有助于合理利用实验室资源。2，传统机器学习算法通常具有较强的解释性，这为研究者理解模型预测背后的生物学意义提供了有力帮助。理解这两点有助于使药物重定位的效率获得大幅度提高。

本章专注于介绍应用于 DTI 预测的机器学习方法。具体来说，我们的目标是提供一个针对利用机器学习框架的化学基因组学方法子类的全面概述。与那些也使用机器学习策略的配体基方法相比，本文讨论的方法适用于已知配体不足的目标蛋白。首先，本章会涉及在药物发现中经常使用的数据库列表。接下来会采用分层分类方案。特别是，我们将机器学习方法分为两大类，更多的子类暂且不提及。我们试图分别介绍每类代表性方法。此外，

我们还呈现了每类方法的优点和缺点。最后，本章将讨论当前机器学习方法在 DTI 预测领域面临的挑战和进一步展望。

4.3 机器学习相关技术

4.3.1 机器学习

机器学习是一种人工智能 （AI），它通过学习和扩展以前的经验来训练计算机像人类一样思考。它采用最少的人工干预来分析数据和发现趋势。机器学习对社会有广泛的影响，包括生产线、医疗保健、教育、交通和食品。机器学习正在改变我们的生活以及住房和应用程序、汽车、零售、食品行业等行业。

机器学习是人工智能的一个分支，它使计算机系统能够利用数据来改进性能，无需进行明确的编程。通过从数据中学习并识别模式，机器学习算法能够预测结果或做出决策。这种方法广泛应用于各种领域，包括图像识别、自然语言处理、医疗诊断和药物发现等，它通过模拟人类的学习能力，使计算机能够从经验中学习和适应新数据。

4.3.2 极限学习机

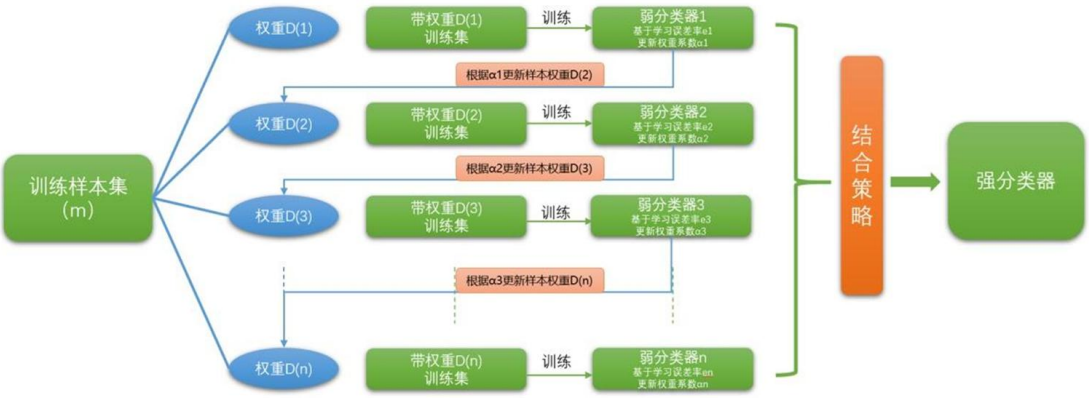
极限学习机（Extreme Learning Machine, ELM）是一种前馈神经网络的学习算法，由黄广斌教授于 2004 年提出。它的核心思想是在训练过程中随机生成输入层到隐藏层的权重和偏置，然后通过解析方法一次性确定隐藏层到输出层的权重。这种方法显著减少了传统神经网络训练中的参数数量，加快了学习速度，并有助于避免过拟合问题。

极限学习机的特点包括快速训练、良好的泛化能力以及对噪声数据的鲁棒性。ELM 不需要进行权重的微调，这意味着它可以在大数据集上快速地训练，同时保持较高的预测准确率。此外，ELM 的随机权重初始化允许模型在不同的初始化状态下进行多次训练，从而提高模型的稳定性和性能。

4.3.3 集成学习

集成学习（Ensemble Learning）是机器学习中的一种策略，它通过构建并结合多个学习器（可以是同种或不同种类的算法）的预测结果来提高整体模型的性能。这种方法的基本思想是，通过集合多个模型的优势，可以减少单一模型可能存在的偏差和方差，从而提升预测的准确性和稳定性。集成学习可以应用于分类、回归以及特征选择等多种机器学习

任务。



在集成学习中，各个学习器被称为“基学习器”或“弱学习器”，它们可以是简单的决策树、神经网络或者任何其他机器学习算法的实例。集成学习的关键步骤包括训练这些基学习器，然后将它们的预测结果进行整合。整合的方法有多种，如投票机制（majority voting）、平均（averaging）、堆叠（stacking）和提升（boosting）等。投票机制可以是让基学习器对测试样本进行预测，然后选择得票最多的类别作为最终预测；平均法则是取基学习器预测结果的平均值；堆叠方法通常涉及使用一个初始模型对数据进行预测，然后将预测结果作为新特征与其他特征一起输入到下一层模型中，如此迭代；提升方法则通过关注被前一轮基学习器错误分类的样本，给予这些样本更高的权重，以此来训练下一轮的基学习器。[14]

集成学习的优势在于它能够从基学习器的多样性中获益，减少过拟合的风险，并提高模型对未知数据的泛化能力。然而，集成学习也可能带来额外的计算成本和时间消耗，因为需要训练多个模型。此外，集成学习的成功也依赖于基学习器的选择和整合策略的设计。在实际应用中，集成学习已经被证明在许多机器学习竞赛和实际问题中都非常有效。

4.4 基于球形演化极限学习机的药物重定位方法

4.4.1 模型实现

算法构建包括种群初始化、参数优化、球形搜索和靶标预测四个基本步骤。参数优化涉及尺度因子 SF 和尺度选择因子 DSF，SF 控制搜索半径，DSF 决定搜索维度。ELM 通过随机生成输入权值和隐层偏置，并使用非线性激活函数将输入数据映射到新的特征空间。模型表达式为 $t_j = \sum_{i=1}^{N} \beta_i g(w_i x_j + b_i)$，其中 $(j = 1, \cdots, N)$。通过计算隐层输出矩阵的广义逆来确

定输出层权值 ($\beta = H^{\dagger}T$) 。为平衡数据集，筛选出高质量的负样本。将已知的药物-靶标对标记为 1 作为正样本，其他未知相互作用的药物-靶标对标记为 0 作为负样本候选集。并通过整合靶标和药物间的相似性，计算加权分数并排序，选择分数低的样本构成负样本候选集。最后，将正、负样本合并得到训练和测试数据集 。通过调整角度和半径实现球形搜索，寻找最佳的网络参数 (w) 和 (b)。

4.4.2 模型性能评估

为评估该模型，陈若蓝团队将 SEELM-DTI 算法在 YAMANISHI 等编制的金标准数据集上的性能与其他 5 个先进算法进行比较。数据集涵盖酶（E）、离子通道（IC）、G-蛋白偶联受体（GPCR）、核受体（NR）等重要蛋白靶标的药物-靶标对，包含 932 种药物、989 个靶蛋白和 5127 种相互关系 采用 ROC 曲线下的面积（AUC）和精确-召回曲线下的面积（AUPR）作为评价指标。AUPR 更能惩罚假阳性存在，适合评估算法性能。[15]

结果显示，SEELM-DTI 的性能相比其他先进算法有明显改善，AUC 和 AUPR 性能显著提高。在 GPCR 和 NR 中的 AUC 以及 E、GPCR、IC 和 NR 中的 AUPR 均优于其他算法。在 E 和 IC 中的表现略逊于 BLM-NII，因为 BLM-NII 更多利用药物与蛋白质邻居信息。在 E 和 IC 数据集中已知 DTIs 数量多，CPCR 和 NR 含更多未知药物-靶标对且结合特异性强，本文算法在这两个数据集上表现出色。

尽管为 4 个数据集设计，但 SEELM-DTI 也可作其他领域强大分类器。将未识别样本输入 SEELM-DTI 预测药物-靶标相互作用，按分数排序预测的高分相互作用，在医学生物数据库和科学文献中人工排序验证。例如，盐酸普萘洛尔和 5-羟色胺受体 1A 的相互关系已在 DRUGBANK 数据库中得到验证 。

4.5 基于 Adaboost 算法的药物重定位方法

4.5.1 模型实现

在本文中，团队基于 AdaBoost 算法构建药物—靶向蛋白作用预测模型。AdaBoost 算法通过集成多个弱分类器生成强分类器，将其应用于药物—靶向蛋白作用预测，把矩阵填补问题转化为分类问题进行处理。**然后**引入阈值(ϕ)，将矩阵分解模型的评分预测结果映射为二值分类问题，判断预测结果是否在($[\hat{r} * dt - \phi, \hat{r} * dt + \phi]$)区间内，以此确定模型的

准确性。最终通过 AdaBoost 算法训练集成(T)个基础训练模型，采用加权平均方法融合各个模型，以此构建最终模型。

4.5.2 模型性能评估

模型错误率评估：通过预测结果($\hat{r} * dt *$)*和真实结果*($* r * dt$)的差距评估错误率，当($\hat{r} * dt \notin [r * dt - \phi, r * dt + \phi]$ *)时计算错误率*($* \varepsilon * i$)，公式为($\varepsilon_i = \sum_{d=1}^{|d|}\sum_{t=1}^{|t|} w_{dt}$)。此处定义损失函数($argmin\sum_{d=1}^{|d|}\sum_{t=1}^{|t|} w_{dt}(r_{dt} - \hat{r}_{dt})^2$)，使用随机梯度下降法实现预测问题，根据错误率更新权重，权重更新公式为($\alpha_i = \mu ln(\frac{1-\varepsilon_i}{\varepsilon_i})$)，其中($* \mu *$)为调整参数。此外，为评估此算法，陈若蓝团队使用两个公测数据集 MATADOR 和 STITCH，MATADOR 包含 784 种药物、2431 种靶向蛋白和 13064 个药物—靶向蛋白相互作用；STITCH 包含 598 种药物、671 种靶向蛋白和 3296 个有效药物—靶向蛋白作用对。将数据集中作用分数均匀映射到[0,5]之间，采用 5 折交叉验证，1 份作测试集，4 份作训练集。评测指标：采用均方根误差（RMSE）、预测准确率（Accu）和召回率（Rec）作为评测指标。RMSE 定义为(RMSE=$\frac{\sqrt{\sum d=1^{|d|}\sum t=1|t|(r_{dt}-\hat{r}_{dt})2}}{|D|}$)；Accu 计算公式为(Accu=(TP + TN)/(TP + TN + FN + FP))；Rec 计算公式为(Rec = TP/(TP + FN))，其中 TP、FN、FP、TN 分别表示真阳性、假阴性、假阳性、真阴性的统计结果总数

表 1　药物—靶向蛋白作用预测的统计结果

Tab.1　Statistical results of drug-target protein interaction prediction

	预测有作用	预测无作用
实际有作用	*TP*	*FN*
实际无作用	*FP*	*TN*

表 2　AdaBoost 算法实验效果

Tab.2　Experimental results of AdaBoost algorithm

算法	RMSE			
	MATADOR		STITCH	
	未使用	AdaBoost	未使用	AdaBoost
SVD	0.923 4	**0.892 2**	0.832 1	**0.807 6**
PMF	0.943 8	**0.912 3**	0.859 8	**0.832 1**

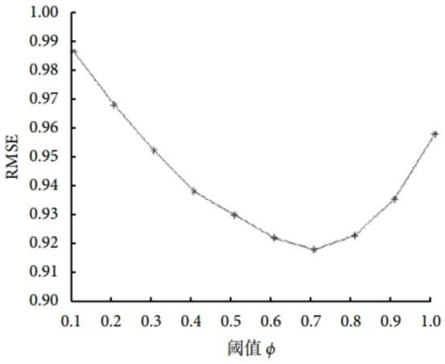


图 1　阈值 ϕ 对预测结果的影响

Fig.1　The influence of threshold ϕ

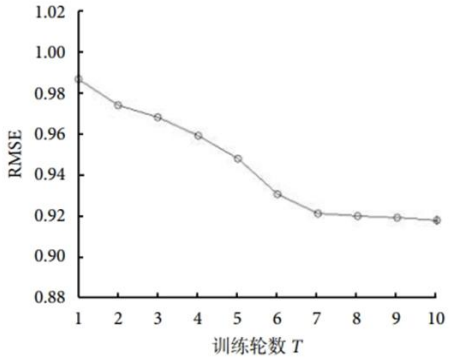


图 2　训练轮数 T 对预测结果的影响

Fig.2　The influence of training iterations T

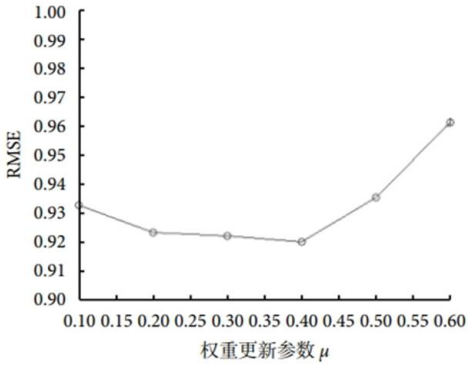


图 3　权重更新参数 μ 对预测结果的影响

Fig.3　The influence of weight updating parameter μ

实验结果与分析：与单个模型（如 SVD、PMF）相比，使用 AdaBoost 算法后误差减小，在 MATADOR 和 STITCH 数据集上，RMSE 值均降低，表明融合基础方法能提升预测准确度。算法框架参数对 RMSE 影响先大后小再变大，在[0.1,1.0]取值中，(ϕ= 0.7)时 RMSE 最小，需调整合适阈值。而训练轮数增长时精度先快速提高，7 次以上提高变缓。权重更新参数则在(* μ *= 0.4)时准确度最高。与经典和新近方法（如基于近邻用户/物品推荐算法、朴素贝叶斯、支持向量机、MTOI、基于二分图的核回归方法、基于化学相似度的方法）相比，本文算法（使用 AdaBoost 算法融合 SVD 方法）在 Accu 和 Rec 指标上表现较好，RMSE 值较低，获得较高预测准确度。 - 冷关联问题实验：在冷关联数据集（MATADOR - C 和 STITCH - C）上，各种方法实验结果普遍不佳，但本文方法除在 STITCH - C 数据集的 Rec 指标略差于 SVM 方法外，其余指标表现较好，表明本文方法在冷关联数据集下有一定有效性。[16]

4.6 总结

药物-靶标相互作用(DTIs)对于选择潜在药物并有效缩小生化实验的研究范围至关重要。此外，它们可以为药物的副作用和作用机制提供深刻的见解。因此，DTI 预测是药物发现的重要前提。实际上，已经建立了许多公开可用的数据库，并促进了创新 DTI 预测策略的发展。在这篇综述中，我们专注于整合化学空间和基因组空间的机器学习方法。我们总结了在 DTI 预测中常用的数据库和机器学习方法。特别是，我们关注近几年出现的一些最先进的预测模型。我们采用分层分类方案，将机器学习方法分为两大类，并提供更多的子类别。机器学习在未来几年的 DTI 预测中将是有希望的。然而，仍有很大的改进空间。因此，我们以一些建议作为未来研究者的参考。集成方法将多个独立分类器组合成一个模型，通常能获得更好的预测结果。其次，极限学习机是解决不平衡数据集问题的有力工具。

然而，这种方法解释性仍有不足。因此，需要更多关注极限学习机的研究。此外，考虑到药物-靶标对涉及结合亲和力和剂量依赖性，将 DTI 预测问题作为回归问题来研究更为实际和有意义。使用定量的生物活性数据将导致更准确、更可靠的预测结果。最后，随着高通量生物技术的发展，可用数据最近快速增长。依次可进一步利用更多不同类型的异构数据的机器学习技术。

5 基于图嵌入结合机器学习的药物重定位方法的补充

5.1 概述

相较于图表示学习，图嵌入方法更加注重于学习节点的低维向量表示，而这些低维表示往往能够被后续的机器学习算法更有效地处理。相较于端到端的图表示学习，图嵌入方法在大规模的图数据处理中呈现出了较高的计算速率与处理速度。[17,18]图嵌入方法不依赖特定的下游任务，因此生成的嵌入向量可以用于多种不同方向，具备较高的泛化能力，而且基于图嵌入的药物重定位对多源信息的整合能力较强，能够将药物与蛋白质以及相关疾病的相互作用进行更为全面的整合，有助于对药物-靶标的作用机制及潜在的适应症的分析呈现。[18,19]图嵌入方法不仅是在大规模数据处理中存在较为显著的优势，而且由于图嵌入方法通过在向量空间中保持节点的拓扑结构和属性信息，对噪声和异常值具有一定的鲁棒性，因此在数据不完整或存在噪声的情况下，图嵌入方法仍然能够提供可靠的药物重定位预测。[18,20]由于图嵌入方法通过生成低维向量，因此能够作为特征与其他的机器学习模型相结合，从而为构建起更为复杂的集成模型提供便利，有助于提升药物重定位过程的效率以及精确度。[18]本章节将着重介绍一种结合图嵌入技术和机器学习方法的药物重定位策略，旨在通过计算模型揭示药物与靶点间的复杂关系。

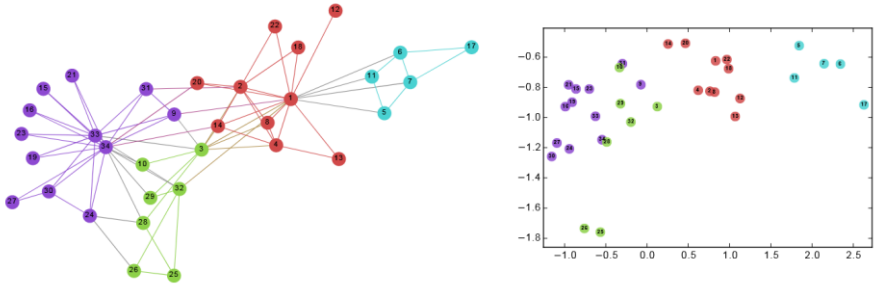
5.2 基于图嵌入结合机器学习的药物重定位方法的设计

本节所述的药物重定位方法分为两个主要部分：1.利用已经构建好的构建生物医学网络，并应用深度游走算法进行图嵌入，生成药物和疾病的低维特征表示；2. 利用随机森林算法对特征向量进行特征选择，并构建预测模型。在此之后，可通过模型评估和案例分析，验证所提方法的有效性和实用性。

5.2.1 深度游走算法的图嵌入过程

在生物医学网络中，节点通常代表基因、蛋白质、化合物等实体，而边则表示实体间的相互作用或关联。由于生物医学网络中的多模态数据无法直接导入统计机器学习进行分类或者其他下游任务，必须要把他们表示成向量的形式。利用图嵌入方法 encoding 出来的向量用少的数据就可以达到之前模型的效果，并且在一些常规模型不好处理的场景下它具

备优势。（图 5-1）深度游走算法作为图嵌入技术的一种，能够捕捉网络中的拓扑结构和节点间的复杂关系。该算法通过模拟随机游走过程，将网络中的节点映射到低维向量空间，从而实现图嵌入。这一过程不仅保留了节点的局部和全局网络结构信息，而且生成的特征向量可以用于后续的机器学习任务。



(a) Input: Karate Graph　　(b) Output: Representation

（图 5-1）图嵌入表示

5.2.1.1 采用深度游走算法的理论基础

DeepWalk 算法是一种图嵌入技术，其核心原理是通过随机游走和语言模型的框架将网络图的节点映射到低维向量空间。该算法借鉴了自然语言处理领域的 Word2Vec 模型，将图中节点序列视为文本中的词序列，利用 Skip-Gram 模型学习节点的上下文关系。在 DeepWalk 中，随机游走作为生成节点序列的手段，捕捉图中节点的邻接关系，从而模拟语言模型中的“句子”结构。每个节点序列被用作 Skip-Gram 模型的输入，通过当前节点的向量表示来预测其上下文节点，进而优化节点的嵌入表示。（图 5⁺-2）此外，为了降低多分类任务的计算复杂度，DeepWalk 采用了分层 Softmax 方法，将节点分类问题转化为二叉树结构的递归二分类问题。算法的训练过程通过随机梯度下降进行优化，支持并行化处理，有效提高了处理大规模图数据的效率。最终，每个节点被表示为一个 D 维向量，这些向量能够保留节点在原始图结构中的拓扑信息，所以 DeepWalk 算法能够为后续的机器学习任务提供较为丰富的特征表示。[21]

Algorithm 1 DeepWalk(G, w, d, γ, t)

Input: graph $G(V, E)$
window size w
embedding size d
walks per vertex γ
walk length t
Output: matrix of vertex representations $\Phi \in \mathbb{R}^{|V| \times d}$
Initialization: Sample Φ from $\mathcal{U}^{|V| \times d}$
Build a binary Tree T from V
for $i = 0$ to γ **do**
  $\mathcal{O} = \text{Shuffle}(V)$
  for each $v_i \in \mathcal{O}$ **do**
    $\mathcal{W}_{v_i} = RandomWalk(G, v_i, t)$
    SkipGram(Φ, $\mathcal{W}_{v_i}$, w)
  end for
end for

（图 5-2）DeepWalk 算法伪码

5.2.1.2 深度游走算法的优势

由于 DeepWalk 算法是在 RandomWalk 的基础上，结合了例如 Word2Vec 等用于自然语言处理的模型。[21]，因此其并非是用节点本身去预测，而是需要用一个 Embedding 为向量的输入模型进行预测，这个过程将离散的网络节点转化为连续的向量空间中的点，即图嵌入。作为一种无监督学习方法，DeepWalk 算法不需要标记数据，而是通过随机游走在图结构数据中生成节点序列[21]，因此能够关注到节点的多模态属性信息，兼顾到的信息较多，提升了数据挖掘的深度。故 DeepWalk 在药物重定位中是一种较为常用的方案，例如 DeepWalk 算法常被被用来构建药物和靶标的嵌入表示，并用于预测药物-靶标的结合强度和结合位点。[22]

DeepWalk 算法通过在整合了药物和靶标相关的信息后所构建的同质网络上执行随机游走来捕捉节点的邻域信息[23]，从而获得与特征节点局部相关的训练数据。DeepWalk 使用 SkipGram 模型来学习随机游走产生的序列节点并向量表示。这是个图嵌入表示过程，通过这些嵌入能够捕捉节点的拓扑结构和生物学属性。[23,24]在构建药物-靶标对网络（DTP-NET）以后，将随机游走过程中得到的相关数据嵌入表示后直接拼接，形成 DTP-NET 中节点的初始特征，并训练分类模型来实现药物-靶标相互作用（DTI）的预测。[25]在此步骤之后，通过预测所得的未知的药物-靶标相互作用，DeepWalk 方法可以同时评估多个潜在的药物候选者，从而提高了药物靶标识别的精确度，加快发现药物的新适应症的速率，提升药物重定位的效率。[24,26]

此外，由于 DeepWalk 生成的嵌入向量能够捕捉网络结构的复杂性，这些向量可以作为特征输入到其他机器学习模型中，如随机森林、支持向量机等，以提高药物重定位模型的泛化能力。[21]

5.2.1.3 深度游走算法存在的一些不足

由于 DeepWalk 算法采用完全随机的游走策略，因此有可能导致算法无法充分探索图的全局结构，特别是在图结构复杂、稀疏或具有高度异质性的情况下。这种随机性可能忽略了图中的全局信息和深层次的拓扑结构，从而限制了算法在某些应用中的有效性 。此外，虽然深度游走可以捕捉到节点附近的一些拓扑结构，但是随机游走生成的图本身没有意义，因此 DeepWalk 在解释性上仍然欠佳。[21]

5.2.2 结合图嵌入方法的随机森林算法的特征选择预测过程

在获取了节点的特征向量后，随机森林算法作为一种集成学习方法，被用于特征选择和预测模型的构建。随机森林通过构建多个决策树并集成它们的预测结果，能够提高模型的稳定性和准确性。在特征选择阶段，随机森林能够有效识别出对药物重定位任务最有影响的特征，从而简化模型并提高预测性能。[27]

5.2.2.1 采用随机森林算法的理论基础

随机森林 Random Forest 算法是一种集成学习方法，通过构建多个决策树模型并整合它们的预测结果来提高分类和回归任务的准确性和鲁棒性。在随机森林中，每个决策树是在数据集的一个随机子样本上构建的，这个子集包括了随机选择的样本和特征，该过程称为自助采样。这种方法引入了随机性，旨在减少模型的方差，从而提高泛化能力。每棵树在训练时使用不同的特征子集，这增加了模型的多样性。在预测阶段，对于分类问题，随机森林通过多数投票的方式确定最终的类别标签；对于回归问题，则通常取多棵树预测结果的平均值作为最终预测。（图 5-3）随机森林的核心在于这样集成策略，它通过综合多棵树的预测结果来作出最终决策，这样可以减少单棵决策树可能存在的过拟合问题。因此，在需要同时评估药物特征和靶标之间的复杂交互作用的药物重定位过程中，随机森林算法是一种预测的准确性和鲁棒性较高的方法。

5.2.2.2 随机森林算法的优势

随机森林算法通常是利用多个决策树的预测结果来提高整体模型的性能，决策树通过一系列通常是二元选择的问题来对数据进行分割，直到达到某个终止条件（如达到最大深度、节点中的样本数量小于某个设定值，或者节点中的样本都属于同一类别等情况）。通常情况下，随机森林是通过多数投票机制来确定最终的预测结果，决策树分别独立投票后，

得票最多的类别被选为最终预测结果；又或者取平均值作为最终预测。[27]在构建随机森林时，每个决策树是通过随机抽样整个训练集 bootstrap sampling 得到的，因此，随机森林的决策树对于特征选择的范围并非全体数据集，每个决策树在分裂节点时，会随机选择一部分特征，然后从中选择最佳分裂特征。[27] 由于在随机森林的决策树训练过中，每个特征对于模型预测的贡献都会被记录下来，所以随机森林算法能够根据特征对模型的预测能力贡献来判断特征重要性，[27,28] 因此随机森林在基于药物-靶标的药物重定位中具备较为显著优势。

Algorithm 1 Random Forest Algorithm for Drug-Target Interaction Prediction

```
procedure RF-DTI-PREDICTION(trainingdata, numberoftrees, numberoffeatures)
    forest ← ∅
    for i ← 1 to numberoftrees do
        subsample ← bootstrapsamplefromtrainingdata
        tree ← growadecisiontreefromsubsample
        tree.maxfeatures ← numberoffeatures
        forest.add(tree)
    end for
    return forest
end procedure
procedure GROW A DECISION TREE(subsample)
    while subsample is not empty do
        bestfeature, bestthreshold ← findthebestsplitonthecurrentnode
        splitthenodeintoleftandrightchildrenbasedonbestfeatureandthreshold
        if stopping criterion is met then
            returnthenodeasaleafnode
        else
            recursivelygrowtheleftandrightsubtrees
        end if
    end while
end procedure
procedure PREDICT(querydrug, forest)
    predictions ← ∅
    for each tree in forest do
        prediction ← getpredictionfromthetreefor thequerydrug
        predictions.add(prediction)
    end for
    finalprediction ← majorityvoteofpredictions
    return finalprediction
end procedure
```

（图 5-3）随机森林算法伪码

随机森林作为一种常用于特征选择的算法，能够有效地从大量数据中提取和选择对药物-靶标相互作用（DTI）预测有用的特征。例如在 Ahn、Kim 等人的研究中，他们通过计算药物的三维分子指纹（E3FP）和基于这些指纹的相似性矩阵，可以识别药物与靶标之间的独特性。[29]不仅如此，在存在多靶点特征的情况下，随机森林算法能够评估药物与多个

靶标之间的相互作用。这种多目标分类能力使得随机森林在预测药物的多靶点效应方面表现出色，有助于理解药物的复杂作用机制。

即使是面对处理异质性数据的情况下，随机森林算法也具备一定优势。随机森林算法能够处理和整合来自不同来源的异质性数据，如药物的化学结构、蛋白质序列相似性信息、药物的副作用信息等，这使得随机森林算法能够从多个角度分析药物-靶标的相关作用机制，从而提供更全面的预测结果。[27,30]

在随机森林的预测过程中，由于算法集成了多个决策树来提高预测的准确性和鲁棒性，因此可以更准确地预测药物的新适应症，对药物-靶标的结合关系判别也较为精准，从而提高药物重定位的成功率与可靠性。[23]

此外，随机森林算法通过整合药物诱导的基因表达谱和迭代特征选择，以预测药物可能产生的副作用，而这也有助于理解药物的作用机制和评估其安全性[30]。

5.2.2.3 结合深度游走图嵌入的随机森林算法

在药物-靶标的药物重定位实践中，相对于深度学习方法，随机森林算法作为一种传统的机器学习方法可能对于较为抽象的信息处理的效果略有欠佳。不过，结合了图嵌入的随机森林算法通过其强大的特征选择和集成学习能力，仍然在药物重定位领域展现出竞争力。[27,31]

结合图嵌入方法的随机森林算法除了图嵌入算法本身具备的减少过拟合的机制，随机森林本身使用的多个在不同的数据子集上训练的相互独立的决策树也具备一定的减少过拟合风险的能力。虽然单个决策树有可能会对训练数据过度拟合，但随机森林通过平均多个树的预测结果的机制能够降低整体过拟合程度，从而提高算法的泛化能力。[27] 随机森林对于数据集中的异常值和噪声具有一定的鲁棒性，但可能会消耗较多的内存和计算资源，尤其是在处理大规模数据集时，但是，在结合了图嵌入算法后，图嵌入方法生成的低维向量恰好弥补了随机森林算法的这一缺陷，二者刚好在处理大规模数据集这一问题上形成了互补，为基于大规模生物医学网络的药物重定位预测提供了一个较为完备的解决方案。[27,32]

5.3 对于深度游走结合随机森林的药物重定位方法的评估

深度游走结合随机森林的药物-靶标重定位方法是一种创新的计算模型，它通过结合图嵌入技术和集成学习方法来预测药物与靶标之间的潜在相互作用。该方法首先利用深度游

走算法，如 DeepWalk，对生物医学网络进行节点捕捉并生成图嵌入，这一过程通过模拟随机游走来捕捉网络中的拓扑结构和节点间的复杂关系，将节点映射到低维向量空间，从而保留节点的局部和全局网络结构信息。随后，该方法采用随机森林算法对采集到的特征向量进行特征选择和预测模型的构建。随机森林通过构建多个决策树并集成它们的预测结果，提高了模型的稳定性和准确性。在 Kim，Lee 等人的研究中，这种方法成功应用于药物-靶标相互作用的预测，展现了较高的预测准确率和 AUC 值。[29]此外，该方法的优势在于其能够处理大规模图数据，并且不需要对整个图结构进行完整的分析，而是通过随机游走来捕捉局部结构信息，这使得模型在处理具有复杂拓扑结构的生物网络时具有较高的效率和可扩展性。总体而言，深度游走结合随机森林的药物-靶标重定位方法为药物发现和重定位提供了一种高效、准确的预测工具，值得后续尝试。

5.4 本章小结

本章较系统地阐述了一种融合图嵌入技术与机器学习算法的药物重定位方法，强调了图嵌入技术在处理大规模生物医学网络数据时的显著优势，包括其在数据不完整性和噪声存在情况下的鲁棒性，以及在不同下游任务中的泛化能力，并且提出了一种药物重定位策略，该策略涵盖两个关键步骤：一是通过深度游走算法实现图嵌入，生成药物和疾病的低维特征表示；二是利用随机森林算法进行特征选择和预测模型的构建。

在图嵌入阶段，本章介绍了 DeepWalk 算法，该算法通过模拟随机游走过程，将网络节点映射至低维向量空间，以此保留节点的拓扑结构和网络关系。DeepWalk 算法的优势在于其能够无监督地学习节点的多模态属性信息，从而增强数据挖掘的深度。然而，该算法在探索图的全局结构方面存在局限性，可能忽略了图中的全局信息和深层次的拓扑结构。

进一步地，本章探讨了结合图嵌入技术的随机森林算法在特征选择和预测模型构建中的应用。随机森林算法作为一种集成学习方法，通过构建多个决策树并整合其预测结果，显著提升了模型的稳定性和准确性。该算法能够有效地识别对药物重定位任务具有决定性影响的特征，从而简化模型结构并提高预测性能。

在最后，本章对深度游走结合随机森林的药物重定位方法进行了综合评估。该方法成功地预测了药物与靶标之间的潜在相互作用，展现了较高的预测准确率和 AUC 值。发现其的优势在于其能够高效处理大规模图数据，且无需对整个图结构进行完整分析，而是通过随机游走捕捉局部结构信息，这使得模型在处理具有复杂拓扑结构的生物网络时展现出

较高的效率和可扩展性。

本章所述的深度游走结合随机森林的药物-靶标重定位方法是药物发现和重定位领域的一种较为高效、准确的预测工具。未来的研究可在此基础上进一步探索和完善该方法，以期在实际应用中实现更广泛的药物重定位预测。

6 对于药物重定位方法可解释性的探讨

6.1 概述

根据前面几章的内容可知，药物重定位领域已经发展出较为丰富多元且高效的方法体系。但是，随着药物重定位领域的快速增长，研究者们对于人工智能预测模型这个只能看到输入和输出、却无法观察其内部的逻辑和运作过程的“黑盒”的不透明性感到担忧。因此，为了使重定位的新药的安全性、可解释性和可信度得以保障，例如 Jürgen Bajorath 等人已经开始尝试开发解释复杂模型预测的方法并投入实践[33]。

在这一背景下，可解释性成为了药物重定位方法的关键因素。可解释性不仅能够提高模型预测的透明度，增强决策者对模型预测结果的信任度，而且对于监管审批、药物研发流程优化、后续研究指导以及新生物学知识的发现都具有重要的意义。监管机构在药物研发过程中需要对药物的安全性和有效性进行严格的审查，可解释的模型能够提供详细的预测依据和推理过程，有助于监管机构更好地评估药物的安全性和有效性，从而加快药物的审批流程。

此外，可解释性还有助于优化药物研发流程。通过揭示药物与靶点之间的相互作用机制，为药物研发人员提供有价值的见解，这些见解有助于指导药物的优化和改进，从而提高药物研发的成功率。在基础科学研究和药物开发中，可解释性通过解释药物如何通过特定的生物学途径或靶点对疾病产生影响，有助于揭示新的生物学知识和机制，促进对疾病本质的深入理解。

随着系统生物学、计算生物学、网络药理学等相关学科的快速发展，药物重定位已成为世界范围内关注的热点，在药物研发领域占据重要地位。因此，探索药物重定位方法的可解释性不仅是提高药物重定位成功率的关键，也是推动整个医药行业进步的重要驱动力。

6.2 现存的综合提升模型可解释性的方案分析

目前，研究者们提出了多种方案来综合提升药物重定位模型的可解释性，这些方案涵盖了从数据预处理到模型构建的各个阶段。

为了增强模型的可解释性，部分研究者们采用了多源数据融合技术。通过整合药物、疾病和靶点的多源信息，可以更全面地理解药物的作用机制和疾病之间的关联。例如，

DrugReAlign 框架利用大语言模型（LLMs）和多源提示技术，从海量人类知识库中学习靶点和药物的一般知识，克服了传统方法中数据可用性的限制[34]。

为了提升嵌入表示的准确性，研究者们提出了基于语义多层关联推断的知识图谱学习模型，如 DREAMwalk 模型。该模型通过语义引导的随机游走策略，提升了药物与疾病关系的预测准确度。此外，该模型还借助多层语义关联推断（GBA）提升了药物-疾病关联（DDA）预测能力，增强了药物-疾病关联的可解释性[35]。

在模型构建方面，也有研究者采用了基于异构图神经网络的药物重定位方法。这些方法通过引入多种生物学关系构建生物医学网络，并使用拓扑子网嵌入模块、图注意力模块和层注意力模块等方法学习药物和疾病表示，从而用于药物-疾病关联预测和药物重定位[36]。

为了提高模型的泛化能力和预测准确性，有研究者提出了基于多源相似度和多视图学习的药物重定位算法。这些算法通过融合不同生物数据源的信息，有效缓解了药物-疾病关联矩阵的数据稀疏性问题。

在本章中，将主要介绍以语义分析、偏随机游走以及扩散模型为主的几个综合提升模型可解释性的方案，如多尺度相互作用组与 DREAMwalk 模型。

6.2.1 可解释性的基本算法

鉴于语义分析、偏随机游走以及扩散模型在处理生物医学大数据、揭示药物与疾病之间复杂关系的能力，这些方法常在药物重定位可解释性提升中被采用。

6.2.1.1 语义分析在模型可解释性方面的优势

在药物重定位领域，语义分析方法的原理基于自然语言处理（NLP）技术，旨在使计算机能够理解和解释生物医学文献中的语言含义。此方法的核心在于从大量的医学文献和临床数据中提取关键信息，构建药物与疾病之间的语义网络，从大量的医学文献和临床数据中提取关键信息，构建药物与疾病、药物与靶标之间的语义网络[35]。

语义分析方法通常包括以下几个关键步骤：1.通过词法分析对文本进行分词，识别出名词短语、专业术语以及关键词。2.进行句法分析，确定文本中的语法结构，包括短语结构和依存关系，这有助于理解句子成分之间的逻辑和语义联系。3.实体识别 NER 用于识别文本中的特定名词短语，如药物名称、疾病名称、基因和蛋白质等。4.关系抽取识别和解析文本中实体之间的关系，揭示药物作用的生物学机制和疾病关联。

此外，语义角色标注 SRL 有助于进一步细化实体间的关系，其通过识别谓词及其对应

的语义角色，明确具体的作用对象以及相关影响。此步骤对于理解药物作用机制尤为重要，因为它能够揭示药物如何影响特定的生物学过程或疾病状态，为药物重定位提供了坚实的科学基础与可解释性。

所以，语义分析方法在药物重定位中的应用，通过综合运用 NLP 技术，从生物医学文献中提取深层次的语义信息，构建药物与疾病之间复杂的关系网络，能够提供药物作用机制和药物与目标靶点之间复杂相互作用的深入理解，为药物重定位提供了坚实的科学基础和可解释性[35]。

6.2.1.2 扩散模型在模型可解释性方面的优势

在药物重定位的实践中，扩散模型方法的原理基于复杂网络理论，通过模拟信息在网络中的传播和扩散来揭示药物与疾病之间的潜在联系[37]。扩散模型的核心是通过药物和疾病在生物医学网络中的拓扑位置和连接模式来表征它们之间的相互作用。这些网络通常由药物、基因、蛋白质和疾病等节点组成，节点之间的边代表它们之间的生物学关联，如药物-靶点相互作用、基因-疾病关联等。

扩散模型是通过在网络中模拟信息的传播过程来学习节点的嵌入表示。在此过程中，信息可以从一个节点传播到邻居节点，并且随着时间的推移，这种传播能够覆盖整个网络。模型通过这种方式捕捉到药物和疾病之间的间接联系，从而揭示它们之间可能的相互作用[38]。因此，如果两种疾病在生物学通路或基因表达模式上具有相似性，二者相关药物也可能具有潜在的交叉治疗效果，从而可以借此推断出两者解释性上的联系。

不仅如此，扩散模型还通过引入注意力机制或自适应学习模块来增强其预测能力，这些机制能够区分不同节点和边的重要性，从而提高模型对关键生物学信息的敏感性[37]。通过该方式，扩散模型不仅能够预测已知药物的新适应症，还能够为缺乏有效治疗选项的疾病发现潜在的药物候选。

扩散模型以类似于生物医学知识图谱中实体间复杂的相互作用的方式，模拟和捕捉信息生物医学网络中的传播和扩散[39]，有助于探索药物与疾病及靶标之间的复杂关系，从而提高可解释性。

6.2.1.3 偏随机游走在模型可解释性方面的优势

在生物医学知识图谱的框架内，有偏随机游走 biased random walks 算法能够在不同层次的网络结构中实现信息的整合。该机制使得模型能够同时考量药物和疾病的生物学功能

及其关联，从而生成更为全面和丰富的节点表示，为药物重定位提供了多维度的信息支持。（图 6-1）有偏随机游走产生的路径直观地映射了药物与疾病之间的潜在联系，揭示了药物通过特定生物学过程或分子途径与疾病之间的关联，能够更深入地理解药物的潜在治疗效果和作用机制原理[35]。

偏随机游走算法还能揭示网络中的结构特征，包括社区结构、网络中心性节点和关键路径。这些结构特征为理解药物和疾病在生物医学网络中的位置及作用提供了直观视角，有助于识别关键生物学实体和过程[40]。通过对有偏随机游走生成的路径进行分析，研究人员可以提出新的生物学假设，例如药物的新适应症、疾病的新生物学标志物或药物与疾病之间的新关联。这些假设可进一步通过实验研究进行验证，推动生物医学知识的发现。

而且，偏随机游走的机制相对直观且易于解释，模型的每一步决策过程都是透明的。该算法通过模拟生物医学网络中的复杂交互，提供了一种强有力的方法来揭示和解释药物和疾病之间的潜在联系，该方法的直观性和灵活性使其成为提高生物医学知识图谱模型可解释性的重要工具。

Algorithm 1 Biased Random Walk Algorithm

```
procedure BiasedRandomWalk(G, p, q, d, r)
    Input:
    G: Graph, the network on which the walk is performed
    p: Return parameter, controls the likelihood of immediately revisiting a node
    q: In-out parameter, controls whether the walk is more likely to move towards leaf nodes (out) or hub nodes (in)
    d: Number of dimensions of the node embeddings
    r: Number of random walks per node
    Output:
    Embeddings for each node in the graph
    for each node v in G do
        for i ← 1 to r do
            walk_v ← Start a random walk from node v
            walk_v ← Perform a biased random walk using parameters p and q
            Embed(walk_v) ← Perform dimensionality reduction on the walk sequence
        end for
    end for
    return Embeddings
end procedure
```

（图 6-1）有偏随机游走算法伪码

6.2.2 结合偏随机游走的综合提升模型可解释性的方案

6.2.2.1 偏随机游走结合其他方法的基础

因为偏随机游走是基于随机游走的，因此其也保有了随机游走的大部分特性。而且，偏随机游走算法相较于传统的随机游走方法，提供了一种更为精细化的网络探索机制。在传统的随机游走中，游走者在每一步都以等概率随机选择一个相邻节点进行转移，这一过程不依赖于任何外部信息或特定策略。相比之下，偏随机游走则在游走者的选择中引入了一定的偏好，这些偏好可以基于节点的属性，如权重、中心性或连接度，或是根据特定任务的需求来定制。这种偏好的引入使得偏随机游走能够更加聚焦于网络中的特定区域或特征，在生物医学知识图谱等应用中，可以沿着与特定疾病相关的生物学路径进行有针对性的探索，这使得偏随机游走能够与扩散模型进行结合使用[39]。

偏随机游走算法能够整合网络之外的额外信息，如节点的语义信息和拓扑结构信息，以指导游走过程。这种方法在任务驱动的应用中显示出其独特的优势，如节点分类、链接预测和社区检测等，其中特定的游走策略有助于提升任务的性能。与随机游走相比，偏随机游走展现出更高的可控性和灵活性，研究者可以通过调整节点的语义的权重来调整游走策略和偏好来优化游走过程，以适应不同的研究需求。因此，偏随机游走也通常会与语义分析方法结合使用[41]。

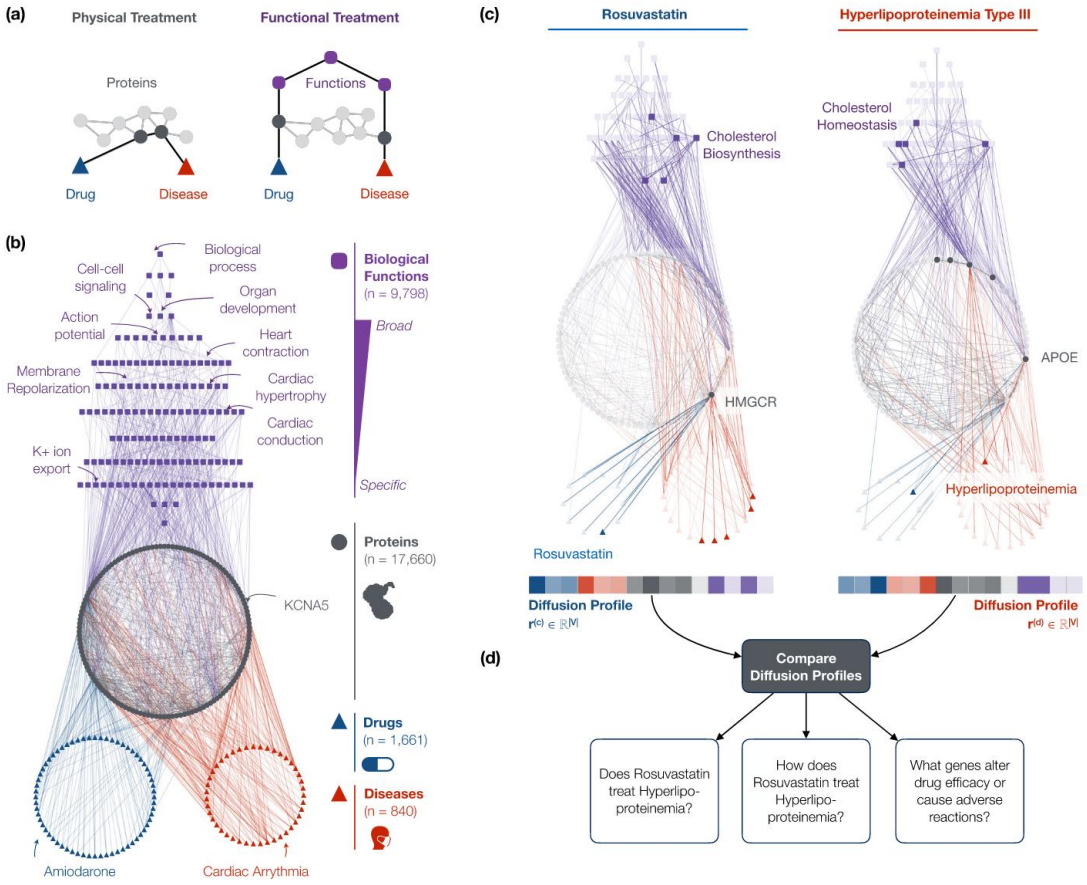
当偏随机游走与扩散模型或者语义分析方法相结合后，由于游走过程受到特定策略的指导，这些策略与网络的语义和结构特征紧密相关，因此偏随机游走的结果通常会更易于解释尤其是在需要深入解释和理解网络中复杂模式的场景中，例如在生物医学知识图谱分析和社交网络分析中，偏随机游走能够提供更直观的网络探索方式和更丰富的任务性能，综合提升模型可解释性。

6.2.2.2 多尺度相互作用网络：有偏随机游走与扩散模型相结合

多尺度互作网络 multiscale interactome 是提高药物重定位模型的可解释性的一种新的途径。该网络通过整合疾病扰动蛋白、药物靶点以及生物功能，构建了一个多层次的生物医学知识图谱。在该网络中，药物和疾病的效应不仅通过蛋白质之间的物理互作传递，还通过生物功能的层次结构进行传播。此方法的核心在于，其不仅考虑了药物直接作用的靶点，还考虑了药物通过影响生物功能而间接作用的更广泛的生物学过程。

多尺度互作网络采用有偏随机游走的方法，模拟药物效应如何通过生物功能的层次结构和蛋白质-蛋白质互作网络进行传播。有偏随机游走能够捕捉到药物作用的扩散过程，并

生成一个表示药物或疾病对网络中每个节点（如蛋白质和生物功能）影响的扩散谱 diffusion profile。通过比较药物和疾病的扩散谱，可以识别出与治疗相关的蛋白质和生物功能，从而为药物重定位提供了一个可解释的生物学基础。（图 6-2）



（图 6-2）多尺度互作网络有偏随机游走传播扩散谱模拟药物治疗

此外，多尺度互作网络还能够预测哪些基因变异会影响药物的疗效或引起严重的不良反应。通过分析药物和疾病扩散谱中的基因的网络重要性，可以预测这些基因是否会改变特定治疗的效果。这种方法不仅提供了预测药物疗效的潜力，而且还能够识别和解释影响治疗效果的生物学机制[39]。

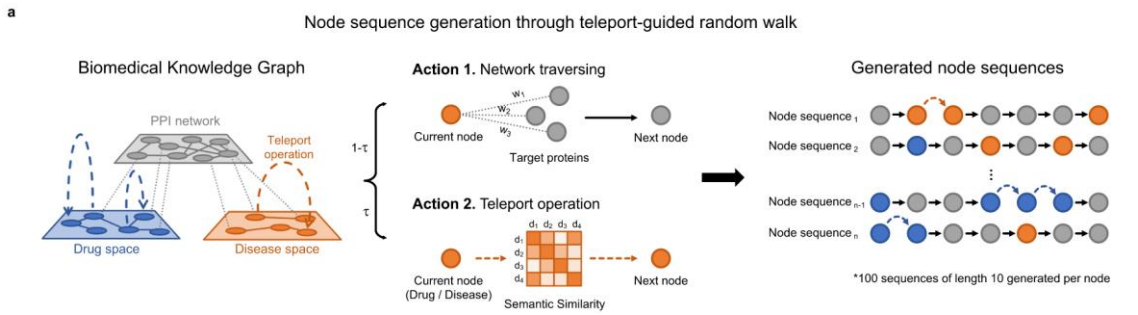
多尺度互作网络通过整合物理互作和生物功能，提供了一个更为全面和深入的视角来理解药物作用的机制。这种方法不仅能够预测药物与疾病的关联，还能够揭示这些关联背后的生物学原理，从而显著提高了药物重定位模型的可解释性，有助于推动药物重定位的临床应用和药物新适应症的发现。

6.2.2.3 DREAMWalk：结合语义分析方法的有偏随机游走

在生物医学知识图谱的学习与分析中，传统的随机游走算法往往受到网络结构偏差的

影响，导致药物和疾病实体的嵌入表示不够精确，这在一定程度上限制了药物重定位模型的性能；而且传统的随机游走算法生成的图由于其完全随机的过程，其意义其实较难以解释。而 DREAMWalk 模型创新性地将语义信息引导的远程传输机制与随机游走算法相结合（图 6-3），显著提升了药物重定位的准确性和模型的可解释性。

具体而言，DREAMWalk 模型通过引入一种语义引导的远程传输策略，实现了在随机游走过程中对药物和疾病节点的频繁和有偏访问。该策略依据药物和疾病在生物医学本体中的语义相似性度量，利用一个用户定义的传输因子 τ，控制随机游走者在到达药物或疾病节点时，以一定的概率跳转至语义上相似的节点。这种机制有效地平衡了网络中各节点的表示，尤其是在基因节点占主导的知识图谱中，显著增强了药物和疾病节点的嵌入质量[35]。



（图 6-3）语义信息引导的远程传输

DREAMWalk 模型的这一创新策略，不仅在生物医学知识图谱的表示学习中实现了药物和疾病实体的更准确映射，而且通过生成的嵌入向量空间，提供了对药物和疾病之间潜在关联的深入洞察。通过这种方法，DREAMWalk 模型在药物重定位任务中取得了显著的性能提升，其在预测药物-疾病关联的准确性上比现有的最先进模型提高了最多 16.8%。此外，通过探索嵌入空间，Bang, Kim 等研究者发现该模型能够揭示生物和语义上下文中的对应关系，这为药物重定位提供了新的视角和潜在的靶点。

在实际应用中，DREAMWalk 模型已经通过乳腺癌和阿尔茨海默病的重定位案例研究，展示了其在生物医学知识图谱上应用多层 guilt-by-association 原则的潜力[35]。这些案例研究不仅验证了 DREAMWalk 模型在药物重定位上的实用性，也强调了在生物医学领域中，将多层语义信息与网络结构相结合的方法对于发现新的药物-疾病关联的重要性。因此，DREAMWalk 模型作为药物重定位的一种新的、高效的计算框架，有助于将分子层面的信息转化为临床应用中的洞见。

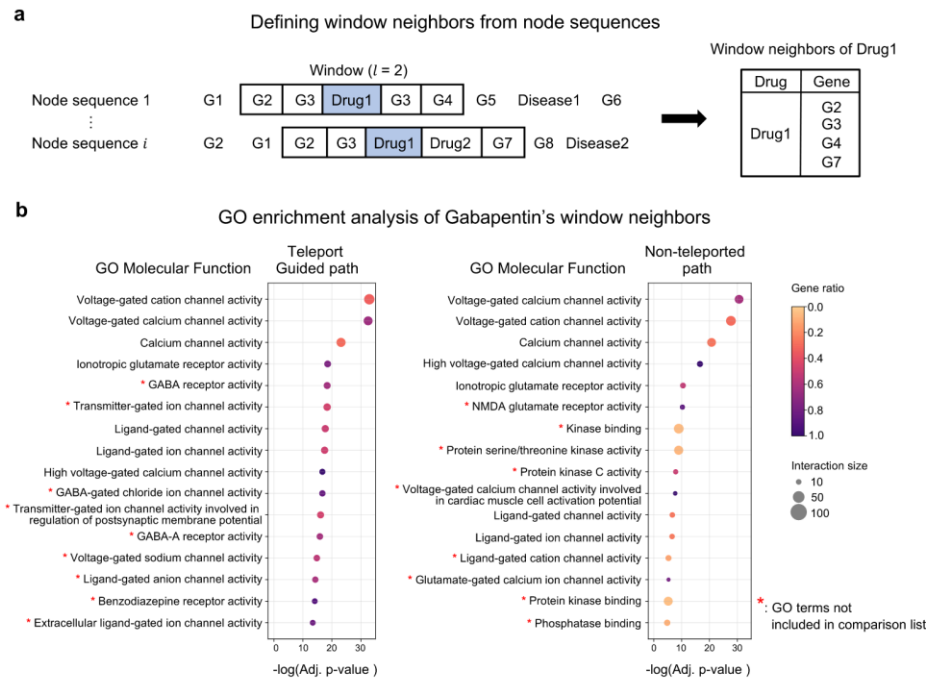
6.3 模型可解释性的具体案例展示

6.3.1 Baricitinib 的 COVID-19 重定位案例分析

Baricitinib 最初被批准用于治疗类风湿性关节炎。在 COVID-19 大流行期间，通过计算模型的分析，发现 Baricitinib 能够通过抑制 Janus 激酶（JAK）通路来减少细胞因子风暴，这是 COVID-19 重症患者中常见的一种病理现象。多尺度相互作用组的方法揭示了 Baricitinib 的分子靶点与 COVID-19 病理生理学之间的联系，为 Baricitinib 在 COVID-19 治疗中的潜在应用提供了科学依据[39]。可解释性在这一过程中发挥了至关重要的作用。通过理解 Baricitinib 如何通过 JAK 通路抑制炎症反应，医生和研究人员能够快速提出将其用于 COVID-19 治疗的假设。这种深入的理解帮助监管机构和医疗专业人员对使用 Baricitinib 治疗 COVID-19 的安全性和潜在效果做出快速决策。最终，Baricitinib 被美国食品药品监督管理局（FDA）批准与瑞德西韦（Remdesivir）联合用于治疗 COVID-19，这一决策的迅速性在很大程度上归功于模型的可解释性。

6.3.2 Gabapentin 在阿尔茨海默病（AD）中的潜在应用案例分析

Gabapentin 主要用于治疗癫痫和神经病理性疼痛。DREAMwalk 模型通过分析 Gabapentin 的语义邻居和生物学邻居，揭示了其可能对 AD 有治疗效果。模型通过随机游走生成的路径分析显示，Gabapentin 与 GABA 相关基因的富集，尽管它不直接作用于 GABA 受体[35]。Gabapentin 可显著增加大脑中的 GABA 水平。但该药物似乎并不直接影响 GABA 特异性酶或受体。药物靶标数据库表明 Gabapentin 不直接与 GABA 受体结合。通过对窗口邻居中的 gene 富集，可以看出，经过瞬移的可以富集到 GABA 受体，GABA 门控氯离子通道，GABA-A 受体。而不经过瞬移的富集结果和 GABA 无关。（图 6-4）



（图 6-4）邻居窗口及加巴喷丁作用机理

可解释性在这一探索中同样发挥了关键作用。这种分析提供了 Gabapentin 可能通过影响 GABA 途径来改善 AD 症状的解释。这为进一步的实验研究和临床试验提供了合理的生物学假设。虽然 Gabapentin 尚未被批准用于 AD 治疗，但这种分析为未来的研究提供了方向，并可能加速新适应症的开发[35]。

6.4 本章小结

本章深入探讨了药物重定位方法的可解释性问题，并分析了提升模型可解释性的多种方案。随着人工智能预测模型在药物重定位领域的广泛应用，模型的透明度和可解释性成为了关键因素。概述了可解释性在提高模型预测透明度、增强决策者信任度、监管审批、药物研发流程优化、后续研究指导以及新生物学知识发现中的重要作用。随后又分析了现存的综合提升模型可解释性的方案，包括多源数据融合技术、基于语义多层关联推断的知识图谱学习模型、基于异构图神经网络的药物重定位方法以及基于多源相似度和多视图学习的药物重定位算法。这些方案涵盖了从数据预处理到模型构建的各个阶段，旨在通过整合多源信息、提升嵌入表示的准确性、构建生物医学网络以及融合不同生物数据源的信息，来增强模型的可解释性。

本章也详细介绍了可解释性的基本算法，包括语义分析、扩散模型和偏随机游走。这些方法在处理生物医学大数据、揭示药物与疾病之间复杂关系方面具有显著优势。语义分

析方法通过自然语言处理技术，从医学文献和临床数据中提取关键信息，构建药物与疾病之间的语义网络。扩散模型基于复杂网络理论，通过模拟信息在网络中的传播和扩散来揭示药物与疾病之间的潜在联系。偏随机游走算法则在生物医学知识图谱的框架内，通过有偏的网络探索机制，揭示药物与疾病之间的潜在联系。

本章接下来探讨了结合偏随机游走的综合提升模型可解释性的方案，包括多尺度相互作用网络和 DREAMWalk 模型。多尺度互作网络通过整合物理互作和生物功能，提供了一个更为全面和深入的视角来理解药物作用的机制。DREAMWalk 模型则通过引入语义引导的远程传输策略，显著提升了药物重定位的准确性和模型的可解释性。

最后，通过 Baricitinib 的 COVID-19 重定位案例和 Gabapentin 在阿尔茨海默病中的潜在应用案例，展示了模型可解释性在实际药物重定位中的应用和重要性。这些案例分析进一步证实了可解释性在药物重定位过程中的关键作用，以及该特性在推动药物新适应症发现和临床应用中的潜力。

本章全面地分析了药物重定位方法的可解释性，并提出了多种提升模型可解释性的方案。这些方案不仅为药物重定位领域提供了新的视角和工具，也为整个医药行业的进步提供了重要的驱动力。未来的研究应继续探索和优化这些方法，以进一步提高药物重定位的成功率和模型的可解释性。

7 总结与展望

随着人工智能和机器学习技术的飞速发展，药物重定位已成为新药研发的重要途径。本研究报告全面探讨了药物重定位的现状与进展，特别关注于如何利用基于计算的手段来提高药物重定位的效率与准确度。通过深入分析药物重定位模型的总体框架、异构网络的构建、基于图表示学习的药物重定位策略以及基于机器学习的策略，本文为现有的药物重定位研究进行了初步的探讨。

本文首先总结了药物重定位的背景和意义，强调了新药研发面临的挑战以及药物重定位在缩短研发周期、降低成本方面的重要性。接着，文章详细介绍了生物异构网络的构建过程，包括相关概念、数据库资源和网络构建的具体方法。在此基础上，本文进一步探讨了基于图表示学习的药物重定位策略，包括图神经网络、图卷积网络和图池化网络等技术，并分析了这些技术在药物重定位中的应用。同时，本文还探讨了基于机器学习的策略，如极限学习机和集成学习在药物重定位中的应用。最后，文章强调了模型可解释性的重要性，并提出了提升模型可解释性的多种方案，包括有偏随机游走和基于语义多层关联推断的知识图谱学习模型。

尽管药物重定位领域已经取得了显著的进展，但仍存在一些挑战和未来的研究方向。首先，随着生物医学数据的爆炸性增长，如何有效地整合和利用这些数据，提高药物重定位的准确性和效率，仍是一个重要的研究课题。其次，模型的可解释性是药物重定位研究中的关键因素，需要进一步探索和优化，以提高模型预测的透明度和可信度。此外，随着人工智能技术的不断进步，如何将最新的技术应用于药物重定位，如深度学习、自然语言处理等，也是未来研究的重要方向。最后，跨学科合作也是推动药物重定位研究的关键，需要生物学家、计算机科学家、药理学家等多方共同努力，以实现药物重定位的最终目标——发现新的药物适应症，加速新药的研发进程。

药物重定位是一个充满挑战和机遇的领域，随着技术的不断发展和研究的深入，我们有理由相信，未来药物重定位将为人类健康带来更多的希望和突破。

参考文献

[1] 李 坤 . 基 于 知 识 增 强 的 深 度 学 习 药 物 重 定 位 方 法 研 究 [D]. 浙 江 大 学,2023.DOI:10.27461/d.cnki.gzjdx.2023.000715.

[2] 张 迎 雪 . 异 构 网 络 动 态 表 示 学 习 技 术 研 究 [D]. 东 南 大 学,2022.DOI:10.27014/d.cnki.gdnau.2022.001138.

[3] 朱 思 怡 . 基 于 多 组 学 生 物 异 构 网 络 的 药 物 重 定 位 方 法 研 究 [D]. 厦 门 大 学,2020.DOI:10.27424/d.cnki.gxmdu.2020.002289.

[4] 陈 若 兰 . 基 于 异 构 生 物 网 络 的 协 同 过 滤 药 物 - 靶 点 预 测 方 法 研 究 [D]. 厦 门 大 学,2020.DOI:10.27424/d.cnki.gxmdu.2020.001864.

[5] Xuan P, Ye Y, Zhang T, et al. Convolutional neural network and bidirectional long short-term memory-based method for predicting drug-disease associations[J]. Cells, 2019, 8(7): 705.

[6] Moridi M, Ghadirinia M, Sharifi-Zarchi A, et al. The assessment of efficient representation of drug features using deep learning for drug repositioning[J]. BMC Bioinformatics, 2019, 20(1): 1-11.

[7] Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing[J]. Bioinformatics, 2020, 36(Supplement_1): i525-i533.

[8] Wang X, Xin B, Tan W, et al. DeepR2cov: deep representation learning on heterogeneous drug networks to discover anti-inflammatory agents for COVID-19[J]. Briefings in Bioinformatics, 2021, 22(6): bbab226.

[9] Cai L, Lu C, Xu J, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network[J]. Briefings in Bioinformatics,2021, 22(6): bbab319.

[10] 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.

[11] 孙 鑫 亮 . 基 于 图 神 经 网 络 的 药 物 重 定 位 方 法 研 究 [D]. 中 南 大 学,2023.DOI:10.27661/d.cnki.gzhnu.2023.002475.

[12] 焦 齐 齐 . 基 于 子 图 聚 合 的 药 物 疾 病 关 联 预 测 [D]. 哈 尔 滨 工 业 大 学,2023.DOI:10.27061/d.cnki.ghgdu.2023.001993.

[13]SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS Thomas N. Kipf University of Amsterdam T.N.Kipf@uva.nl Max Welling

University of Amsterdam Canadian Institute for Advanced Research (CIFAR) M.Welling@uva.nl

[14] Ruolan Chen、Xiangrong Liu、Shuting Jin、Jiawei Lin. Machine Learning for Drug-Target Interaction Prediction, 2018, 23(3): 2208-2216

[15] 胡苓芝、傅城州、蔡永铭、杨进、唐德玉.球形演化极限学习机在药物-靶标相互作用智能预测中的应用. 华南师范大学学报, 2023, 55(1): 121-128

[16] 古万荣、谢贤芬、何亦琛、张子烨. 基于 AdaBoost 算法的药物—靶向蛋白作用预测算法.生物医学工程学杂志.2018,35(6):935-94

[17] 卢艳峰, 杨思瀚, 莫鸿仪, 侯凤贞. 基于知识图谱嵌入的阿尔茨海默病药物重定位研究. 中国药科大学学报, 2023, 54(3): 344-354

[18] 图表示学习综述. 2023. https://d.wanfangdata.com.cn/Periodical/bjsfdxxb202305005

[19] J. Tayebi, B. BabaAli. EKGDR: An End-to-End Knowledge Graph-Based Method for Computational Drug Repurposing. Journal of Chemical Information and Modeling, 2024, 64(6): 1868-1881

[20] X. Jia, X. Sun, K. Wang, M. Li. DRGCL: Drug Repositioning via Semantic-enriched Graph Contrastive Learning. 2024: 1-12

[21] B. Perozzi, R. Al-Rfou, S. Skiena. DeepWalk: online learning of social representations, in: KDD '14: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York New York USA, ACM, 2014: 701-710

[22] 基于深度学习的药物-靶标相互作用预测研究--《中南大学》2023 年博士论文. 2024. https://cdmd.cnki.com.cn/Article/CDMD-10533-1024366459.htm

[23] Y. Zhu, C. Ning, N. Zhang, M. Wang, Y. Zhang. GSRF-DTI: a framework for drug-target interaction prediction based on a drug-target pair network and representation learning on a large graph. BMC Biology, 2024, 22(1): 156

[24] N. Zong, H. Kim, V. Ngo, O. Harismendy. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. Bioinformatics, 2017, 33(15): 2337-2344

[25] B. Liu, D. Papadopoulos, F. D. Malliaros, G. Tsoumakas, A. N. Papadopoulos. Multiple similarity drug-target interaction prediction with random walks and matrix factorization. Briefings in Bioinformatics, 2022, 23(5)

[26] Y. Zhang, Y. Wang, C. Wu, L. Zhan, A. Wang, C. Cheng, et al. Drug‐target interaction

prediction by integrating heterogeneous information with mutual attention network. BMC Bioinformatics, 2024, 25(1): 361

[27] J. Ali, R. Khan, N. Ahmad, I. Maqsood. Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI), 2012, 9

[28] G. Louppe. Understanding Random Forests: From Theory to Practice. arXiv, 2015

[29] S. Ahn, S. E. Lee, M. hyun Kim. Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence. Journal of Cheminformatics, 2022, 14(1): 67

[30] A. Cakir, M. Tuncer, H. Taymaz-Nikerel, O. Ulucan. Side effect prediction based on drug-induced gene expression profiles and random forest with iterative feature selection. The Pharmacogenomics Journal, 2021, 21(6): 673-681

[31] J. L. Yu, Q. Q. Dai, G. B. Li. Deep learning in target prediction and drug repositioning: Recent advances and challenges. Drug Discovery Today, 2022, 27(7): 1796-1814

[32] F. Vazquez-Novoa, J. Conejero, C. Tatu, R. M. Badia. Scalable Random Forest with Data-Parallel Computing, 见: 29th International European Conference on Parallel and Distributed Computing, Euro-Par 2023, August 28, 2023 - September 1, 2023: 14100 LNCS, Limassol, Cyprus, Springer Science and Business Media Deutschland GmbH, 2023: 397-410

[33] Learning characteristics of graph neural networks predicting protein–ligand affinities | Nature Machine Intelligence. 2024. https://www.nature.com/articles/s42256-023-00756-9

[34] 基于多源相似度和多视图学习的药物重定位算法研究--《中南大学》2023 年硕士论文. 2024. https://cdmd.cnki.com.cn/Article/CDMD-10533-1024366516.htm

[35] D. Bang, S. Lim, S. Lee, S. Kim. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. Nature Communications, 2023, 14(1): 3570

[36] 基于深度学习的药物虚拟筛选及重定位模型构建研究--《北京协和医学院》2023 年硕士论文. 2024. https://cdmd.cnki.com.cn/Article/CDMD-10023-1023124117.htm

[37] K. Huang, P. Chandak, Q. Wang, S. Havaldar, A. Vaid, J. Leskovec, et al. A foundation model for clinician-centered drug repurposing. Nature Medicine, 2024: 1-13

[38] Z. Lin, Y. Gong, Y. Shen, T. Wu, Z. Fan, C. Lin, et al. Text Generation with Diffusion Language Models: A Pre-training Approach with Continuous Paragraph Denoise, 见: Proceedings of Machine Learning Research, 2023: 21051-21064

[39] C. Ruiz, M. Zitnik, J. Leskovec. Identification of disease treatment mechanisms through the multiscale interactome. Nature Communications, 2021, 12(1): 1796

[40] Z. Sun, W. Dong, J. Shi, Z. Huang. Interpretable Disease Progression Prediction Based on Reinforcement Reasoning Over a Knowledge Graph. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(3): 1948-1959

[41] A. Grover, J. Leskovec. Node2vec: Scalable feature learning for networks, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 855-864

附录-小组分工

鲍弈慎：研究背景和基于生物异构网络的药物重定位现状

冯子益：摘要+基于图表示学习的药物重定位策略-研究现状

陈锦身：基于机器学习的药物重定位研究现状

曹炜杰：基于图嵌入结合机器学习的药物重定位方法的补充+对于药物重定位方法可+解释性的探讨