摘 要

药物重定位是指在原有药物适应症范围之外，对已批准上市或正在研究的药物进行新用途识别的一种策略，它具有风险低、时间和费用开销少等优势。药物的解剖治疗化学(Anatomical Therapeutic Chemical，ATC)分类系统是世界卫生组织对药物的官方分类系统。正确预测药物的解剖治疗化学类别，将给定化合物识别到 ATC 系统中，有助于更好地研究该药物的潜在活性成分及其治疗、药理和化学性质，为进一步发现药物的新适应症奠定了基础，对药物研发工作具有重要意义。传统的药物 ATC 分类主要由人工药理学实验来完成，不仅耗费资源，也存在无法对药物进行快速、批量分类等局限性。随着生物信息技术的发展，越来越多的生物医药数据被累积，为药物 ATC Code 的预测提供了数据支撑，基于计算方法的药物 ATC Code 预测研究成为热点。近年来，基于深度学习的计算方法在药物 ATC Code 预测研究上表现出不错的性能。然而，已有的药物 ATC Code 预测方法，大多数从单一的药物自身信息入手，较少考虑药物与其相关生物实体间的关联信息及药物 ATC Code 标签间信息对预测性能的影响。如何有效的整合药物及其相关生物实体间的链接关系，充分考虑药物 ATC Code 标签提供的信息，设计出合理高效的方法对药物 ATC Code 进行预测是极其重要的。为此，本文基于多源信息和图转换网络，在以下两方面进行研究：

（1）针对以往预测方法数据源单一，未考虑药物及其相关生物实体间链接关系的问题，提出一种基于图转换网络的药物 ATC Code 预测方法 DACPGTN。DACPGTN 通过应用不同生物实体的相似性特征和药物、疾病、靶标蛋白质之间的相互作用，预测给定药物的第一级 ATC Code。首先，整合已知信息构建代表药物、疾病、靶标蛋白质特征的复合矩阵。在此基础上，引入药物-靶标蛋白质、药物-疾病、靶标蛋白质-疾病间关联信息，构建异构网络集合；并使用图转换网络从多个异构网络中学习药物-靶标蛋白质-疾病生物实体之间的潜在多重关联。最后，利用获取的潜在关联信息图结构与构建的复合特征矩阵，对药物做出最终 ATC Code 预测。在基准数据集上的实验结果表明，该方法相比于其他药物 ATC Code 预测方法在性能上有较大的提升。此外，案例分析与新药测试也证明了该方法的可靠性与稳定性。

（2）针对图转换网络无法将生物实体异构网络中独立存在的节点进行链接及未考虑

药物 ATC Code 标签信息的问题，提出一种基于改进的图转换网络结合标签相关性的药物 ATC Code 预测方法 DACPGTN_L。DACPGTN_L 方法简化图转换网络的实现过程，并将图转换操作扩展到包含节点特征的非局部操作。在每一层的异构网络集合中添加基于生物实体节点间相似性构建的新网络。在学习不同生物实体间潜在关联时，可以对异构网络中独立节点进行有效链接。其次，获取药物 ATC Code 标签特征，并从条件概率、标签间先验知识两个维度构建 ATC Code 标签共现关系图结构。然后，通过 GCN 网络来获取 ATC Code 多标签间的标签信息作为预测信息的补充。最后，将标签信息与图转换网络获取的药物节点嵌入表示相结合，对给定药物做出 ATC Code 预测。在基准数据集上的实验结果表明，该方法优于其他药物 ATC Code 预测方法。此外，案例分析以及新药测试均表明了该方法的可靠性和稳定性。

关键词：药物 ATC Code，深度学习，多标签分类，药物重定位，图转换网络

目 录

第 1 章 绪论

1.1 研究背景及意义

近几十年以来，随着基因组学、生命科学、蛋白质组学等领域技术的发展，积累了海量的生物学、医学数据[1] [2]，药物研发技术也取得了巨大的进步。越来越多的政府、研究人员、医药公司等，在政策支持和资本投入上加大了力度，助力新药的研发。但是，药物研发的周期依旧很长[3]。现实世界中，化学元素种类众多。在实验室中，通过现有技术合成的化合物更是数以亿万计。开发一种药物，首先需要从这些潜在的化合物中进行筛选，筛选后的药物还需要经过严格的新药开发流程，包括但不限于动物实验，毒理性分析、副作用记录等。同时，候选药物在实验动物及临床病人这些不同的物种上的安全性和耐受性测试也存在不同。由于各种限制和阻碍，一种药物从设计到批准投入临床使用的时间往往长达十年以上。另外，开发一种全新的药物，通常投资也是巨大的[4] [5]。乐观估计，平均每种新药需要高达 8 亿美元以上的金钱投入[6]。尽管药物研发有着如此巨大的时间花费以及金钱投入，但实际产出与此是极度不对等的。在新药研发上，投资产生的回报率反而出现不增反减的趋势。现有的数据显示，仅有不到十分之一的初筛候选药物，可以通过 FDA 评估实验[7]。传统的研究策略，即寻找某种全新化合物来解决特定疾病，相对而言过于保守且效率低下。一方面，这种策略存在成本高、研发周期长以及风险性大等问题。另一方面，即使初步实验获得了成功，新药上市前也缺乏系统性评价。这些问题使得药物研发的效率无法得到有效提升。因此，如何找到简单、快速、高效的方法，提高药物研发的成功率是当前迫切需要解决的问题。

利用现有已知安全可靠的药物，挖掘其新的适应症，提高药物利用价值成为当前一个研究热点[8]。药物重定位技术应运而生，成为药物发现领域中极具前景的研究方向[9]。相关统计数据显示，药物重定位方法的使用，相比于传统的药物开发策略，可以将药物研发的时间从 10-17 年缩短到 3-12 年[11]。药物重定位又被称为“老药新用”，即对现有已知的药物进行研究，利用现有技术及药物的化学、药理学、适应症等相关信息，进行特定效用的筛选和组合。在这一过程中，药物需要经过一系列的改造、筛选组合，以达到治疗新疾病和扩展新应用价值的目的。药物重定位研究涉及的范围主要是已经获得上市批准或经过多个阶段的临床试验和毒理学分析的药物[10]，相对于从无到有的开发一种

新药，这种策略成本低且成功率更高。此外，药物重定位研究还可以获得更高的投资回报率，使药物更加快速地通过监管部门的审查并投放到市场中，从而更快地达到缓解患者病痛的目的。目前，已有很多成功重定位的药物在新的适应症上发挥作用，引起制药领域和研究领域投入更多的关注。例如，非那雄胺（Proscar）[12]，这种药物上市初期针对前列腺增生疾病，用于降低前列腺癌的发生。新的发现表明，其可以降低双氢睾酮含量，达到抑制毛囊变小，逆转脱发的目的，已广泛应用于治疗雄性激素性脱发。西地那非（sildenafil）[13]，该药物研发初期作为 5-磷酸二酯酶抑制剂，用于治疗心血管疾病。但由于疗效不佳，却偶然发现其在治疗阴茎勃起功能障碍（ED）治疗上具有极佳的效果，并最终成为主要治疗适应症，带来了极高的经济收益。诸如此类成功重定位的情况，还有很多，但依旧存在一些问题。现有重定位成功的药物，一部分来自实际临床应用或药理性实验分析中的偶然性发现。一部分来自对此种药物在人体内作用机制充分透彻的了解，开辟出新的研究思路，并最终模糊性设计获得该药物的新用途。因此，虽然药物重定位取得了一定的成果，但系统性、批量性的对现有药物进行成功重定位，依旧是一件难事。

随着技术进步，海量基因组学数据和生物医学数据涌现，信息的不充分已不再是当前药物研究中的困难所在，而是如何更好地解读和充分利用这些已有的数据。利用现有数据，开发出系统、高效、合理、稳定的药物重定位方法，是每个从业研究人员和企业的共同目标。现有药物数据库的建立，例如：DrugBank、KEGG、PubChem 等，为药物重定位的快速发展提供了有力支持。计算方法和计算机技术的快速发展，使基于计算分析、机器学习等相结合的精准生物信息学方案被进一步建立[14]。生物信息学方法的逐步应用，进一步降低药物研发成本，使药物重定位进入理性设计并大规模应用的新篇章。具体的，药物重定位利用计算机技术及数学计算分析，对已上市或经过多轮安全性测试的药物进行重定位，加速药物的开发进程，主要有以下几种常见思路：

（1） 分析当前重定位药物的药理学性质，根据已有数据，查找并分析其作用的靶标。对当前药物作用的靶标进行分析，基于已知信息，预测当前药物与其他性质类似的靶标的作用关系，从而做到使药物定向作用于新的靶标，即药物-靶标相互作用预测[15]。

（2） 对当前重定位药物的元素组成、化学性质、结构稳定性等进行分析，从多个维度定义不同药物间的相似性。然后，在药物数据库中，查找其相似的药物

进行推理并验证，探寻相似药物之间的治疗特性或联合用药推荐，即药物-药物相互作用预测[16]。

（3） 根据当前重定位药物的配体蛋白质，进行整体性分析，整合药物相关靶标蛋白质、药物-药物相互作用信息等多源生物医学数据，预测出药物与某些疾病之间潜在的关联关系，最大限度的对药物进行有效利用，即药物-疾病相互作用预测[17]。

（4） 整合现有生物医药信息，对给定药物预测其所属的解剖治疗化学（Anatomical Therapeutic Chemical，ATC）类别，推断其有效成分、治疗效用、药理和化学性质，指导该药物在临床上的应用，挖掘其化学空间特征，为构建其他药物发现模型提供帮助加快药物开发过程，即药物 ATC Code 预测[18]。

本文从药物 ATC Code 预测的角度来进行药物重定位研究，即给定一个药物，对该药物的所属解剖治疗化学（ATC）类别进行预测。世界卫生组织（WHO）提出的解剖治疗化学分类系统[19]作为药品的官方分类系统，是药物再利用和发现的重要信息来源。ATC 编码系统将药物分为五个级别，分别为：解剖学分类、治疗学分类、药理学分类、化学分类、化合物分类，每个药物具有专属的 ATC Code。ATC 系统中标准 ATC Code 的引入，极大的方便了治疗阶段的药物使用[20]。根据 ATC Code 的第一级代码，药物被细分为 14 个解剖学类别，在这 14 个规定的第一级类别中，药物可能由于其解剖学特性同时被划分到多个类别中。通过预测药物的 ATC Code，可以推断药物有效成分、治疗效用、药理和化学性质，帮助正确使用该药物，挖掘药物新的适应症，发现潜在的毒副作用，加快药物开发过程。

1.2 国内外研究现状

随着现有药物研究的不断迭代发展，积累的数据未能进行快速充分解读，使得对药物分类存在一定困难。在目前使用广泛的药物信息数据库中，存在大量没有 ATC Code 的药物，应用传统的实验方法对新药或已有药物进行 ATC Code 分类，费时费力。运用现有技术，实现药物 ATC Code 的快速高效预测是必不可少的。在过去的研究中，已经开发出一些计算方法来对药物进行 ATC Code 预测。

利用化合物的物理化学性质和分子指纹，Dunkel 等[21]人提出了第一种药物 ATC Code 预测方法，来预测药物的单一标签。但随着研究的逐渐深入，对药物 ATC Code 分类的认识进一步清晰，将药物 ATC Code 预测从简单的多分类任务转化为多标签分类任

务是更加合理的做法[22]。多标签分类问题相对于多分类预测难度更大，对药物 ATC Code 预测模型提出了更高的要求。机器学习的发展与应用，为实现药物 ATC Code 快速多标签分类提供了可能。近年来，一些针对药物 ATC Code 的多标签分类方法被提出。Chen 等[23]，首先提出了通过对药物化学-化学相互作用信息和化学-化学相似性信息进行整合，开发一种分类方法对药物 ATC Code 进行分类，并构建了药物 ATC Code 一级代码基准数据集。在此基准数据集的基础上，陆续提出了一些集成多种药物相关信息的分类方法预测药物的 ATC Code。Cheng 等[24]提出了一种多标签的 Gaussian 核回归分类器 iATC-mISF，基于药物化学-化学相互作用、结构和指纹相似度将药物分配到 14 个 ATC Code 第一级类别中。在此之后，Cheng 等[25]通过进一步整合基于药物本体信息[26]的预测因子 iATC-mDO，在此基础上将 iATC-mISF 改进为 iATC-mHyb，提升了分类器的性能。Nanni 和 Brahnam[27]开发了一种基于梯度直方图算法的多标签分类器 EnsLIF，将药物化合物的一维特征向量构建成为二维矩阵，相比于之前的研究，在分类性能上有一定程度的提升。Zhou 等[28]构建多个药物相互作用网络，并通过网络嵌入算法 Mashup[29]提取网络中的药物特征，采用 Random k-labelsets(RAKEL)算法[30]将原始的多标签分类问题转化为多个单标签分类问题，在分类阶段采用经典的机器学习算法支持向量机[31]构建分类器 iATC-NRAKEL，取得较好的效果。在 iATC-NRAKEL 分类器的基础上，Zhou 等[32]简化分类器的输入，提出了一种仅使用药物指纹信息作为特征输入的多标签分类器 iATC-FRAKEL，用于识别药物的 ATC Code，并提供了 web 服务。Wang 等[33]提出了一种预测药物第一级别 ATC Code 的方法 ATC-NLSP，ATC-NLSP 使用机器学习框架，结合药物-药物相互作用信息、结构相似性和指纹相似性，并采用 NLSP 方法[34]来探讨标签之间的相关性，提供更好的预测结果。

随着深度学习的发展，这些技术在药物 ATC Code 预测任务上也被成功应用。Nanni 等[35]提出了一种基于深度学习方法集成的第一级 ATC 类别多标签分类器系统 FUS3。该模型利用卷积神经网络(CNN)和长短期记忆网络（LSTM）[36]提取特征，在两个通用分类器上进行训练，取得了较好的效果。Zhao 等[37]提出了一种新的药物 ATC Code 端对端预测模型 CGATCPred，其使用多层 CNN 从 7 个药物关联得分矩阵中提取复合特征。并建立了 ATC Code 标签间的关联图，结合词嵌入信息通过两层 GCN 网络学习标签信息。利用复合特征与生成的标签特征矩阵之间的点积来构造新的特征，将生成的新特征与 CNN 层提取的复合特征拼接到全连接神经网络层中，预测药物的 ATC Code。Wang 等

[18]提出了一种深度融合学习框架来构建 ATC Code 预测模型 DeepATC。该模型采用随机行走(RWR)和节点降维方法从分子异构生物网络中获得药物的低维表示。然后利用多层感知器(MLP)进行进一步的特征映射变换。利用图卷积神经网络和分子图提取药物拓扑信息。然后，采用多模型关注融合网络对各子模块的高阶特征进行融合，对药物做出最终预测。Cao 等[38]人提出了一种轻量级深度学习模型 ATC-CNN 来对药物 ATC Code 做出精准预测，其简化了模型的数据输入，有效的降低了模型的复杂度。在模型的输入阶段仅使用 SMILES 数据[39]作为唯一的数据源，消除对其他实验数据的依赖。在模型的预测阶段，使用 CNN 网络与全连接层相结合，对药物最终 ATC Code 做出预测。ATC-CNN 方法在仅使用结构信息的情况下，取得了不错的性能。

综上所述，对于药物 ATC Code 预测问题，现有预测方法大多数从药物自身的物理化学性质入手，考虑药物不同维度的特征表示，或结合标签之间的相关性作为预测的主要信息来源。这些方法，一方面较少考虑药物相关生物实体间关联信息在药物 ATC Code 预测中的潜在影响。另一方面，对药物 ATC Code 标签间关联信息的利用不够充分，存在改进空间。因此，解决以上问题，综合考虑药物相关生物实体的关联信息及药物 ATC Code 标签间信息，对预测给定药物的 ATC Code 至关重要。

1.3 本文主要工作及创新

本文的研究是对给定药物的所属 ATC 系统第一级进行精准预测，对先前药物 ATC Code 预测方法中存在的不足进行改进，开发出新的方法解决药物 ATC Code 预测问题。现有研究表明，元素组成、化学性质类似的药物在药理学、毒副作用等性质上可能存在相似性[40] [41]，此类性质不仅仅可以应用在药物-药物链路预测、药物-疾病链路预测、药物-靶标链路预测等药物重定位领域，也适用与药物 ATC Code 预测研究。当两种药物作用于同种靶标蛋白质或疾病，或两种药物与靶标蛋白、疾病中间存在多重关联时，两种药物的 ATC Code 类别有可能相同。此外，随着异构网络的研究不断深入[42]，许多异构网络学习框架被提出，例如 HERec[43]、HAN[44]、GTN[45]等。异构网络通过学习不同类型节点间的多跳元路径（meta-path）对目标节点进行分类，在实际应用中表现优秀。利用不同生物实体与药物之间存在的关联关系，对给定的药物进行 ATC Code 类别预测，为药物重定位中药物 ATC 分类的研究提供了新的思路[46]。另外，在现有多标签预测任务研究中[47] [48]，标签间的多重关联信息[49]及类别特征越来越多的被考虑，多标签间关系提供的辅助信息对模型预测性能的提升有一定帮助。为此，需要从不同的维度对药物

ATC 标签间关系及类别特征进行建模，更加完备的考虑标签间信息对预测任务带来的影响。

本文针对以上问题，对药物 ATC Code 预测任务，给出两种可行的解决方案。具体研究内容如下：

（1）针对以往预测方法数据源单一，未考虑药物及其相关生物实体间链接关系的问题，提出一种基于图转换网络的药物 ATC Code 预测方法 DACPGTN。DACPGTN 通过应用不同生物实体的相似性特征和药物、疾病、靶标蛋白质之间的相互作用，预测给定药物的第一级 ATC Code。首先，整合已知信息构建代表药物、疾病、靶标蛋白质特征的复合矩阵。在此基础上，引入药物-靶标蛋白质、药物-疾病、靶标蛋白质-疾病间关联信息，构建异构网络集合；并使用图转换网络从多个异构网络中学习药物-靶标蛋白质-疾病生物实体之间的潜在多重关联。最后，利用获取的潜在关联信息图结构与构建的复合特征矩阵，对药物做出最终 ATC Code 预测。在基准数据集上的实验结果表明，该方法相比于其他药物 ATC Code 预测方法在性能上有较大的提升。此外，案例分析与新药测试也证明了该方法的可靠性与稳定性。

（2）针对图转换网络无法将生物实体异构网络中独立存在的节点进行链接及未考虑药物 ATC Code 标签信息的问题，提出一种基于改进的图转换网络结合标签相关性的药物 ATC Code 预测方法 DACPGTN_L。DACPGTN_L 方法简化图转换网络的实现过程，并将图转换操作扩展到包含节点特征的非局部操作。在每一层的异构网络集合中添加基于生物实体节点间相似性构建的新网络。在学习不同生物实体间潜在关联时，可以对异构网络中独立节点进行有效链接。其次，获取药物 ATC Code 标签特征，并从条件概率、标签间先验知识两个维度构建 ATC Code 标签共现关系图结构。然后，通过 GCN 网络来获取 ATC Code 多标签间的标签信息作为预测信息的补充。最后，将标签信息与图转换网络获取的药物节点嵌入表示相结合，对给定药物做出 ATC Code 预测。在基准数据集上的实验结果及案例分析表明，该方法优于其他药物 ATC Code 预测方法。此外，案例分析以及新药测试均表明了该方法的可靠性和稳定性。

1.4 章节内容安排

本文章节内容安排如下：

第一章：绪论。本章节首先介绍了药物开发的现状及真实需求，并引出药物重定位的研究背景及意义。然后，介绍本文关于药物重定位研究的具体方向即药物 ATC