

**国家级大学生创新创业训练计划**

**创业类项目申报书**

**项目名称： 融合BERT表示学习的抗新冠病毒药物预筛 选技术**

**项目负责人： 马书晗**

**所在学院： 生物医学工程学院**

**专业年级： 2020级**

**学 号： 2019141660106**

**手 机： 13891579783**

**电子邮箱： shuhan\_pp@outlook.com**

**指导教师： 吕志彬**

**项目起止年月： 2022年11月至2023年10月**

**项目参与学生人数： 4**

**四川大学教务处制**

2022 年 11月

填写说明

一、凡申报**四川大学“国家级大学生创新创业训练计划”**必须填写本申报书。创新类项目是本科生个人或团队，在导师指导下，自主完成创新性研究项目设计、研究条件准备和项目实施、研究报告撰写、成果（学术）交流等工作。

**二、“项目所属一级学科和代码”**参考《普通高等学校本科专业目录和专业介绍（2012年）》。

三、**“项目开展支撑平台”**指支撑本项目开展的国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台、教学实验中心（实验室）、企业、事业或其他单位等，表中填写平台名称，可以多个。

四、**“项目组成员”**人数原则上不超过五人，应排序。

五、**“项目成熟度**”请参考附件《项目成熟度量表》。

六、本书应该填写完整、内容详实、表达准确，数字一律填写阿拉伯数字。

七、报送申报书的电子文档至负责人所在学院。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 项目名称 | 融合BERT表示学习的抗新冠病毒药物预筛选技术 | | | |
| 项目属性 | √面上项目 □2035特区子计划项目 □交叉学科子计划项目 | | | |
| 申请类别 | √科学探索与工程技术类 □人文艺术与社会科学类  □软件信息与文创类 □智能装备与医疗器械类  □生物医药与新材料类 | | | |
| 申请经费 | 6500元 | | 起止时间 | 2022年11月至2023年10月 |
| 项目所属 一级学科和代码 | 生物医学工程  0831 | | | |
| 项目开展 支撑平台 |  | | | |
| 项目来源  （可多选） | √十大重点支持领域的项目  √进课题组、进实验室、进科研团队参与的项目  □国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台支持申报项目  □“2035特区子计划”命题的项目（2035特区子计划）  □交叉学科创新项目  □“青年红色筑梦之旅”计划项目  □基于前期研究实践成果、继续深入研究实践的创新项目  □其他 | | | |
| 2035特区子计划项目情况（非2035特区子计划项目可不填） | 命题名称 |  | | |
| 校内指导老师姓名（非交叉学科子计划项目一般仅允许一位指导老师） |  | | |
| 所属重点支持领域（可不选） | 选择1项：E  A.不填  B.泛终端芯片及操作系统应用开发  C.重大应用关键软件  D.云计算和大数据  E.人工智能  F.无人驾驶  G.新能源与储能技术  H.生物技术与生物育种  I.绿色环保与固废资源化  J.第五代通信技术和新一代IP网络通信技术  K.社会事业与文化传承 | | | |
| 负责人之前参与大创项目情况 | 1.张艺千/马书晗，冷欣悦，梁秋莹，向宏，202110610548，抗病毒药物BCX4430的绿色合成研究，省级，创新训练，2021，合格  2.侯垚/杨帆，马书晗，魏晓玉，耿旭晴，C2022120006，基于Unity 3D 的软体组织变形仿真探索，省级，创新训练，2022，良好 | | | |
| 项目成员之前参与大创项目情况 | 1.李嘉钰，蔡洁/曹彧恒，郁禾，李嘉钰，李冬昊，C2022121776，运动损伤管理与智能平台初步探索及构建，校级，创新训练，2022，合格 | | | |
| 项目负责人基本信息 | | | | |
| 姓名 | 学号 | 专业年级 | | 所在学院 |
| 马书晗 | 2019141660106 | 2020级 | | 生物医学工程学院 |
| 性别 | 手机 | 电子邮箱 | | 身份证号 |
| 女 | 13891579783 | shuhan\_pp@outlook.com | | 61240120010820038X |
| 项目组成员基本信息 | | | | |
| 序号（含排序） | 1 | 2 | | 3 |
| 姓名/性别 | 李嘉钰/女 | 孙家盛/男 | | 郑博文/男 |
| 学号 | 2020141240149 | 2020141410054 | | 2021141670012 |
| 专业年级 | 2020级 | 2021级 | | 2021级 |
| 所在学院 | 生命科学学院 | 计算机学院 | | 生物医学工程学院 |
| 手机 | 17208274585 | 17208280800 | | 17313175425 |
| 电子邮箱 | 474313717@qq.com | 2651472193@qq.com | | zhengbowen101@foxmail.com |
| 身份证号 | 12010320010927512X | 321322200307053657 | | 430681200310207718 |
| 签名 |  |  | |  |
| 指导教师1 基本信息  （非交叉学科子计划项目一般仅允许一位指导老师） | | | | |
| 姓名 | 所在学院或单位 | 研究方向 | | 职称/职务 |
| 吕志彬 | 生物医学工程学院 | 生物医学信息 | | 副研究员 |
| 性别/年龄 | 手机 | 电子邮箱 | | 签名 |

|  |  |  |  |
| --- | --- | --- | --- |
| 男/36 | 18384181480 | [Lvzhibin2016@163.com](mailto:Lvzhibin2016@163.com) |  |

|  |  |  |
| --- | --- | --- |
| **项目内容概述（限200字以内）** | | |
| 在新冠疫情的持续发展下，为推动抗新冠病毒药物的开发设计，本项目拟开发融合BERT表示学习的抗新冠病毒药物预筛选模型。运用Transformer机制对药物和靶蛋白序列信息编码并进行节点表达，利用图卷积神经网络自动学习异构网络的邻域信息进行节点嵌入，最终筛选出对新冠肺炎治疗有用的药物，以期达到药物预测更高的准确性，为抗新冠病毒药物研发领域提供更加先进的工具与研究思路。 | | |
| **项目特色创新点概述（限100字以内）** | | |
| 1.利用BERT模型的Transformer编码药物和靶点分子，提高全局性，减小计算复杂度。  2.图卷积网络结合注意力机制，充分利用分子领域信息，获得更多可解释的节点嵌入。 | | |
| **项目成熟度评估（项目成熟度请参考附件《项目成熟度量表》）**   |  | | --- | | 目前项目成熟度自评估为 3 级；  预期结题时项目成熟度达到 9 级。 | | |  |
| **项目组成员分工** | | |
| **姓名** | **承担工作内容** | |
| **马书晗** | **申报书撰写，技术路线算法设计，深度学习模型建立、模型训练与验证** | |
| **李嘉钰** | **查找文献资料，技术路线算法设计，模型训练与验证** | |
| **孙家盛** | **查找文献资料，深度学习模型建立，数据收集及处理** | |
| **郑博文** | **查找文献资料，数据收集及处理，模型验证** | |

|  |
| --- |
| **一、项目简介（研究内容、目的意义、具体目标、国内外研究现状分析及评价等）** |
| 1. 研究背景   自从2019年底新型严重急性呼吸系统综合征冠状病毒（SARS-CoV-2）被发现，新型冠状病毒肺炎（COVID-19）已持续至今并导致全球持续性的大流行，奥密克戎（Omicron）等变异株的不断出现，全球疫情防控仍持续面临高压力。由于COVID-19的高感染率，病例和死亡人数仍在不断增加。目前累计已有超过6亿人次确诊感染，死亡人数超过650万。  SARS-CoV-2由基因组和包裹它的膜组成。它大致呈球状，突出的棒状表面突起代表其尖峰蛋白（S蛋白），具有“冠状”外观。这些尖峰是病毒入侵人体的工具。当刺激物接触到靶细胞时，S蛋白会附着到呼吸道细胞表面的血管紧张素转换酶2（ACE2）上，ACE2是一种参与血压调节的重要调节剂，触发SARS-CoV-2病毒的内吞作用并形成内体。内体囊使病毒能够逃脱免疫系统。在内体内，S蛋白可以被S1/S2位的原蛋白转换酶Furin和S20位的跨膜丝氨酸蛋白酶2（TMPRSS2）切割，从而使蛋白质能够重新折叠并改变其构象。新暴露的部分与内体膜融合并释放。病毒RNA在3CL蛋白的作用下被翻译成多肽，合成RdRp。RdRp是一种RNA聚合酶，允许RNA在体内复制。与此同时，随着病毒基因组开始复制，入侵的宿主细胞现在活跃地生产病毒蛋白，而不是细胞蛋白。这些新产生的蛋白质和病毒基因组很快被组装成新的病毒，然后病毒从宿主细胞逃逸到入侵其他细胞。  SARS-CoV-2突变株具有传播速度快、毒力强和容易免疫逃逸等特点[[1](#_ENREF_1)]。以奥密克戎变异株为例，其传播速度是德尔塔（Delta）变异株的2倍，已成为传染性最强突变株。面对变异株，新冠肺炎疫苗保护率下降，抗体类药物的活性降低。研究数据显示，2剂阿斯利康公司AZD1222疫苗对德尔塔变异株的保护率为44%，对奥密克戎变异株的保护率仅5%[[2](#_ENREF_2)]；SARS-CoV-2中和抗体Bamlanivimab/Etesevimab与Casirivimab/Im⁃devima治疗方案均对奥密克戎变异株无效[[3](#_ENREF_3)]。临床亟需对各种变异株均能保持活性的抗病毒药物，高效的抗新冠病毒药物设计对疾病的治疗具有重要意义。  面对新冠肺炎疫情，全世界研究人员积极投身于对抗新冠病毒特效药的研发中。目前，国际抗新冠药物研发主要聚焦于“小分子创新药物研发”、“老药筛选”及“抗体药物筛选”三个方向。  新药研发主要针对新冠病毒生命周期中的关键蛋白质开展基于结构的新型小分子药物设计与发现工作，研究人员已提出多种药物可以用于COVID-19的治疗中[[4](#_ENREF_4)]。2021年11月4日，全球首个COVID-19口服药莫那匹韦（Molnupiravir）在英国获准上市，它是一种核苷类RdRp抑制剂，针对轻、中症非住院患者的Ⅲ期临床试验（MOVe-OUT）结果显示，与对照组相比，莫那匹韦可将新冠肺炎非住院患者的临床或死亡相对风险降低30％。但其对于SARS-CoV-2关注变种（VOC）的疗效仍有尚未清楚之处，Carolin M. Lieber等人[[5](#_ENREF_5)]研究概括了Molnupiravir在人体试验中的混合功效，并提出随着VOC的进化，Molnupiravir的疗效需要被重新评估的建议。辉瑞研发的帕罗维德（Paxlovid）是奈玛特韦（PF-07331332）与利托那韦（Ritonavir）的组合，EPIC-HR（非住院、严重疾病风险患者Ⅱ／Ⅲ期临床试验）数据显示，与对照组相比，帕罗维德可将新冠肺炎非住院患者住院或死亡率降低89％。但帕罗维德不断被曝出“复阳”问题，已促使美国国立卫生研究院（NIH）开始做相关研究。  由此可见，抗新冠病毒药物开发设计并不完善，但仍具有紧迫性、重要性。但是新药开发是一个研发周期长、成本高昂且效率低下的过程[[6](#_ENREF_6)]。采用“老药新用”的研发策略，可以节省临床前和部分临床研究时间，从而大大缩短药物研发的时间成本，是应对突发性烈性传染病等疫情暴发的有效手段之一。瑞德西韦早先的适应证是抗埃博拉病毒（ Ebolavirus，EBOV），随着对瑞德西韦研究的深入，研究人员发现瑞德西韦的抗病毒作用并不仅限于埃博拉病毒，其对MERS-CoV和SARS-CoV等多种冠状病毒也有抑制效果。2022年10月22日，美国食品药品管理局（Food and Drug Administration, FDA）已通过瑞德西韦的审批，令其成为了美国市场上首款用于治疗新冠肺炎的药物。瑞德西韦虽已成功上市，但其临床试验数据仍存在争议。  在此基础上，药物的重新定位成为目前新兴的药物研发策略之一。药物重新定位是将现有药物应用于新的疾病的过程。在已批准上市药物的耐药性及其动力学的基础上挖掘其新的用途，实现利用旧药物治疗新疾病，从而可以有效降低药物的使用安全风险，节省药物设计和筛选成本[[7](#_ENREF_7), [8](#_ENREF_8)]。  加速药物发现以及药物重新定位的一个关键基础是准确高效地识别药物与靶点的相关性[[9](#_ENREF_9)]。药物-靶点相互作用（drug-target interaction, DTI）是药物通过作用于靶蛋白并与其发生相互作用，从而影响靶蛋白的药理作用以达到表型效应，这是药物产生药效的前提[[10](#_ENREF_10)]。由此强调了DTI在药物开发中的关键作用，对于为包括 SARS-CoV-2 在内的新病原体找到有效且安全的治疗方法尤其重要。  正确识别和验证DTI是药物发现以及药物重定位流程的第一步。到目前为止，还有许多潜在的DTI尚未被发现。通过计算机预测研究药物靶点的相关性，不仅有助于更好地理解复杂的生物相互作用和重要的生物过程，而且可以缩小候选实验药物的搜索空间，减少相应的时间和成本。使用计算方法进行虚拟筛选，找出一些对新冠肺炎有效的潜在药物，可以指导实验科学家，并在更短的时间内验证能够对抗病毒的分子。其中，基于人工智能的模型显示出其强大的能力[[11](#_ENREF_11)]。许多机器学习模型应运而生，从而提高预测性能。大多数都将预测任务视为一个二分类问题，即预测是否存在药物与靶的相关性。然而，在一些机器学习方法中，通过将模型限制为简单的形式，如双线性或对数双线性函数等，可能不足以获得异构数据背后的复杂隐藏特征。深度学习（Deep Learning）在识别、处理和推断分子数据中的复杂模式方面提供了更强的表达能力，从而改进了传统的化合物计算建模[[12](#_ENREF_12), [13](#_ENREF_13)]。随着深度学习的兴起和发展，研究人员可以构建性能更好的更深层次的学习模型，以便更好地从大规模关联网络数据中提取有用和复杂的信息，提高DTI预测的准确性。      图 1 药物和靶点的对接模拟图及其网络图[[14](#_ENREF_14)]。（A）药物分子和靶点结构图；（B）药物靶点网络可视化图。  Transformer模型中抛弃了传统的卷积神经网络（CNN）和循环神经网络（RNN），整个网络结构完全是由注意力（attention）机制组成[[15](#_ENREF_15)]，对于长序列信息具有很好的提取能力。BERT（Bidirectional Encoder Representation from Transformers）是基于双向Transformer编码器的预训练模型，能够同时获取序列前后两个方向上的信息。该模型的主要创新点是运用了Transfromer机制和无监督学习的预训练方法，易于使用，稳定性强。  为了处理复杂的图形数据，许多研究人员借鉴了图像深度学习的思想来设计神经网络结构。图神经网络（GNN）的概念、操作和模型一直在不断地进化和发展。图神经网络在提取图形嵌入方面表现出强大的性能，将深度学习的应用扩展到了图结构领域，并在药物设计和发现的各个步骤中取得了良好的性能[[16-18](#_ENREF_16)]。其快速发展促使我们将其与DTI预测相结合，从复杂网络中提取有用的信息以更好地进行DTI预测。图形卷积神经网络在提取嵌入图形方面表现出很强的性能，其与药物靶点预测相结合，从复杂网络中提取有用信息可以更好地预测药物靶点。  一些基于GNN的方法利用异构生物数据预测DTI。其方法是将异构性的生物数据构建成多个与药物和靶点相关的网络，其中节点是各种生物实体（即药物、靶点、疾病、副作用等）,而边缘是这些生物实体之间的相互关系。然后将DTI预测作为一个链路预测问题。然而，这些方法在以下方面存在缺陷:   1. 以相同的方式分别考虑每个关系网络（例如，药物-疾病网络），忽略了不同关系网络之间的结构依赖性，以及跨网络中不同的语义关系，如药物-靶点-疾病，这限制了它们区分不同生物实体的结构作用的能力。 2. 现有的工作，例如NeoDTI，局限于一阶相邻节点的接受域，并且不能在单个GNN层中捕获远程依赖关系。虽然叠加多个GNN层增加了接受域，但这种深层的模型存在特征过平滑的瓶颈。 3. 为齐次图建立了图卷积神经网络和图注意力神经网络。它们的操作是基于同质性的假设，即连接的节点往往具有相似的标签。异构生物数据具有复杂的结构、不同的节点类型和边缘类型，这给生物实体节点表示学习带来了挑战。因此，直接将GCN/GAT应用于生物关系网络不可避免地限制了图结构的学习能力。   基于上述抗新冠病毒药物开发中出现的问题与缺陷，本文基于对相关工作的分析，提出了一种应用BERT表示学习和图神经网络的抗新冠病毒药物预筛选模型。从而对抗新冠病毒药物的预测技术进行优化。   1. 研究内容   本课题主要进行抗新冠病毒药物预筛选技术的优化，我们期望开发一种精确度高、预测性能好的基于BERT表示学习的抗新冠病毒药物预筛选优化模型。采用BERT模型的Transformer编码器进行节点表达，形成由12种不同类型的关系网络组成的异构网络，利用图卷积神经网络自动学习复杂异构关系网络的邻域信息，并结合注意力机制反映不同类型邻域信息之和的重要性，得到每个节点的表示特征。应用网络拓扑重构提取关系异构网络中的特征表示，探究12种关系网络不同组合对于模型的影响。基于SARS-CoV-2相关靶点，筛选出可能对新冠肺炎治疗有用的药物并结合现有药物验证抗新冠病毒药物筛选结果正确性。以端到端的方式训练模型，有效更新模型参数，形成更先进的抗新冠病毒药物的优化筛选方案。   1. 目的意义   目前已有多种针对抗新冠病毒DTI预测模型的研究，但仍缺乏一款操作简便，性能优异，准确度高的抗新冠病毒药物预筛选优化模型。本项目针对存在抗新冠病毒药物预测的不足之处，采用BERT表示学习和图神经网络，实现抗新冠病毒药物预筛选技术的优化，从而提高预筛选的准确性，科学性和简便性，提升模型适用性和可扩展性，从而作为药物开发的良好辅助工具。  本项目通过完善抗新冠病毒的药物预筛选技术，辅助COVID-19的医学研究，优化现有的DTI预测模型，为抗新冠病毒的药物研究提供更为精确的预测方案，降低经济损耗，加速COVID-19药物研发进程，改善研究人员对于预测模型缺知识、难使用的操作问题，将改进后的融合BERT表示学习的药物预筛选技术推广至其他疾病的DTI预测中，为药物研发领域提供更加先进的工具与研究思路，为健康中国2030战略打下坚实的基础。     1. 具体目标   本项目拟在构造DTI基本预测模型的基础上，完善药物分子和靶蛋白的编码模块，拓展探究异构网络的节点嵌入，对抗新冠病毒药物预筛选技术进行优化，探究出精确度高、性能优异、操作简便的预筛选模型。具体目标如下：   1. 对药物分子、靶蛋白运用BERT表示学习进行编码，探究BERT节点表示对于模型的精度影响。 2. 利用图神经网络学习异构关系网络的邻域信息，得到节点嵌入特征。探究12种关系网络组合形成的异构网络对于节点嵌入的优劣影响。 3. 训练模型，基于COVID-19相关靶点筛选候选药物，评估验证模型的准确性、先进性。 4. 国内外研究现状分析及评价   近几年，由于DTI预测研究在药物研究、开发领域充分展示出它的重要性，因此国内外研究人员利用多种技术预测DTI，从而开展了大量工作并取得多项研究成果[[19](#_ENREF_19)]。目前，DTI预测的计算方法主要有三类[[20](#_ENREF_20)]：基于配体的方法、对接模拟方法和化学基因组学方法[[21](#_ENREF_21)]。  基于配体的方法主要是根据药物-配体相似性从而形成药理特征以及与靶蛋白之间的联系[[22](#_ENREF_22)]。然而已知配体的数量不足、蛋白质的三维结构未知会导致模型性能不佳。基于分子建模技术，分子对接模拟可以利用分子建模技术，通过动态模拟识别DTI。但是此方法受限于靶点的三维结构[[10](#_ENREF_10)]。  化学基因组学方法可进一步分为几类，如基于机器学习的方法和基于相似性的方法[[23](#_ENREF_23)]。其中，基于机器学习的模型因其可靠的预测结果而备受关注。此方法克服上述两种方法的局限性，通过将药物分子的化学空间与靶点蛋白质的基因空间结合并利用机器学习方法来预测DTI。  在利用机器学习方法预测DTI时，首先要获得表示药物和靶蛋白的特征信息。蛋白质常用的特征是基于序列的特征，例如氨基酸组成[[24](#_ENREF_24)]。此外，还有一些基于蛋白质物理化学性质的特征，其中包括自相关因子等。药物化合物特征的提取有简化分子线性输入规范（SMILES）、利用 DRAGON计算各种类型的描述符[[25](#_ENREF_25)]等。  在基于机器学习的DTI预测中，针对同一问题采用不同的分类算法，或将不同的特征提取方法和分类算法的联合使用，都取得了较好的预测结果。Yamanishi等人[[26](#_ENREF_26)]提出了一种基于核回归的方法，构建了包含人体中四类DTI的黄金标准数据集，并开发一种基于二分图的监督学习方法，将化学和基因组空间中的药物和靶点映射到一个统一空间来有效预测DTI网络。在此基础上，Bleakley 和 Yamanishi [[27](#_ENREF_27)]开发了二分局部模型（BLM），通过将边缘预测问题转化为二元分类问题来预测未知的药物靶点相互作用。Ding 等人[[28](#_ENREF_28)]为了提高DTI的预测性能，分别利用分子子结构指纹、多元互信息和网络拓扑结构来表示药物、靶点及其之间的关系，最后使用支持向量机（Support Vector Machine, SVM）完成DTI的预测。Mahmud等人[[29](#_ENREF_29)]利用 PreDTIs对DTI进行预测，使用MACCS指纹和氨基酸序列分别对药物化合物和靶蛋白编码。结合Fast US和Mo IFS算法处理样本特征后输入到Light GBM分类器中预测DTI。    图2：PreDTIs识别 DTI 的工作流程[[29](#_ENREF_29)]  机器学习可能不足以获得异构数据背后的复杂隐藏特征。作为机器学习的一个发展领域，深度学习（Deep Learning）已经成功运用于DTI预测研究，且具有强大的能力。深度学习作为一种拥有多个隐含层的神经网络，可以在一些复杂的大型数据集上的分类任务中深度挖掘数据中的非线性关系。Ozturk等人[[13](#_ENREF_13)]提出了一种基于深度学习的模型DeepDTI，该模型仅使用目标和药物的序列信息来预测 DTI。通过 CNN 构建药物和靶点的高级表示。DeepDTI优于 KronRLS 算法和 SimBoost，有更优秀的性能。  Huang等人[[30](#_ENREF_30)]提出了一种基于分子交互的Transformer （MolTrans） ，研究者使用FCS 挖掘模块将药物与蛋白质分解为明确的子结构序列。通过增强的Transformer编码器为每个子结构获得增强的上下文嵌入，在交互图上应用 CNN 层以捕获高阶交互，最终完成DTI预测。Liu等人[[31](#_ENREF_31)]提出分子间图的Transformer（IGT）该方法对受体图、配体图和复杂图进行单独处理，采用分子间注意力机制通过三向转换器对分子间信息进行建模。以此为启发，可以将药物与蛋白质序列信息通过Transformer进行编码，从而解决DTI 的子结构性质和未标记分子数据问题。Wang等人[[32](#_ENREF_32)]提出一种多粒性靶点-配体相互作用模型（MGPLI），采用Transformer编码器表示字符级特征和片段级特征，模拟残基与原子或其片段之间可能的相互作用。使用卷积神经网络根据Transformer编码器输出提取更高级的特征并融合蛋白质和药物特征。    图3:MGPLI的模型结构[[32](#_ENREF_32)]  在深度学习中，图神经网络（GNN）的快速发展将深度学习的应用扩展到了图结构领域，相关方法也应用于药物发现[[17](#_ENREF_17)]。Nguyen等人[[33](#_ENREF_33)]提出了一种名为 GraphDTA 的模型，该模型将药物表示为图形，输入药物-靶点对，输出该对的结合亲和力的连续测量，使用图神经网络来预测药物与靶点的亲和力。Shao等人[[34](#_ENREF_34)]提出了一种基于具有注意力机制的异构图的端到端模型（DTI-HETA）。该模型在药物靶点信息构造的异构图上利用图卷积神经网络获得药物和靶点的嵌入表示，并在节点的信息聚合过程中引入了图注意力机制（GAT），最后应用内积解码器来预测DTI。Li等人[[35](#_ENREF_35)]提出了一种基于特定尺度和特定原子分子图的多物理图神经网络模型（MP-GNN），从特定尺度和特定元素去衡量相互作用的类型，并建立了一系列表示相互作用的图表示。用特定权重共享结构构造图卷积网络（GCN）模型，并将单尺度和多尺度集成学习方案进一步整合。Cheng等人[[36](#_ENREF_36)]提出了IIFDTI，利用双向编码器-解码器架构，分别通过图神经网络和卷积神经网络提取药物和靶点的独立特征，将其融合并输入到下游任务中的完全连接的密集层中，实现DTI预测。    图4：DTI-HETA详细工作流程，包括图构建、图嵌入和链接预测[[34](#_ENREF_34)]。  基于机器学习的DTI预测研究不仅可以弥补传统实验预测的不足，减少了资源的消耗，还可以拓宽实验方法的思路，并为药物研发、寻找药物作用的新靶点提供了一种有力的工具。但是仍然存在一些问题未得到解决，药物和靶点之间的内在联系还未挖掘充分，因此DTI的预测研究仍有很大的提升空间和重要的研究价值。  新冠药物的的研发也取得长足进展，在针对SARS-CoV-2的最新计算药物定位策略中，研究人员采用了基于基因组序列的计算技术，如流量平衡分析（FBA），这种方法在病毒宿主细胞识别、代谢模型的使用上具有局限性。基因表达计算技术中，签名匹配专门应用于新冠肺炎，并被证明具有潜在的治疗益处，其常用方法为基因集浓缩分析。然而摄取数据的数量限制了这些方法的应用。  基于化学结构的计算技术中，研究人员使用神经网络来预测分子与蛋白质之间的联系。Ke等人[[37](#_ENREF_37)]使用深度神经网络（DNN）从扩展连接性指纹（ECFP）、官能级指纹（FCFP）和辛醇-水分配系数中识别最重要的分子描述符，以分配不同的权重。该模型经学习后可以用于搜索候选药物。最后，人工智能系统确定了80种上市的潜在药物。其中，8种药物具有体外抗FIP活性。Beck等人将SARS-CoV-2的氨基酸序列和药物的SMILES表示视为自然语言，并使用基于深度学习的分子转换器-药物靶点相互作用（MT-DTI）来预测药物-靶点相互作用（靶点是SARS-CoV-2的3CL蛋白酶）。阿扎那韦、雷米西韦都被预测可以抑制SARS-CoV-2。  单独的蛋白质并不是孤立于复杂的系统、相互作用、网络和结合了许多其他蛋白质参与者的复杂的分子结构的路径中运作的。因此，每一种药物-靶点相互作用都需要在其综合背景下进行测试。基于交互的计算技术中，研究人员使用图神经网络进行预测。图中的每个节点由于其邻居和其他节点的影响而不断改变其状态，直到达到最终均衡。距离越近，影响就越大。Ge等人[[38](#_ENREF_38)]使用基于GCN的CoV-KGE对COVID19进行药物的初步筛选。GCN迭代地更新所有节点的隐藏状态，以基于节点邻域的聚合来产生有用的特征表示。经过计算，他们得出了每个病毒-靶点-药物对的置信度分数，并通过z检验获得了相应的P值。选择P值<0.05的药物作为各病毒蛋白的候选药物。在网络实验室验证之后确定多聚ADP-核糖聚合酶1（PARP1）抑制剂CVL218是一种潜在的有效药物，目前正处于临床试验的第一阶段。虽然这项研究为药物重新定位提供了一个有价值的综合管道，但它仍然存在一个限制，即它只能针对病毒蛋白。Collins等人[[39](#_ENREF_39)]提出了一种神经网络架构ComboNet，可以联合学习药物-靶点相互作用和药物-药物协同作用。该模型能够利用药物—靶点相互作用数据和单药抗病毒活性数据，提高协同药物组合预测准确性，并有效解决了药物组合训练数据不足的问题。实验结果表明，该模型发现了两种有效药物组合，它们在体外实验中显示出强大的抗SARS-CoV-2 病毒协同作用。其缺陷是仍需要纳入更多生物信息。    图5：ComboNet模型架构  综上所述，目前基于计算方法来预测潜在的药物-靶点相互作用和抗新冠药物预筛选的一些方法还存在不足，因此开发出更加准确且高效的预测方法对于抗新冠病毒药物的研发和药物重定位等方面的研究起到关键的作用。    参考文献：  [1] Davies N G, Abbott S, Barnard R C*, et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England [J]. Science, 2021, 372 (6538):eabg3055.  [2] Kokic G, Hillen H S, Tegunov D*, et al.* Mechanism of SARS-CoV-2 polymerase stalling by remdesivir [J]. Nat Commun, 2021, 12 (1):279.  [3] Yu X, Wei D, Xu W*, et al.* Reduced sensitivity of SARS-CoV-2 Omicron variant to antibody neutralization elicited by booster vaccination [J]. Cell Discov, 2022, 8 (1):4.  [4] Marlin R, Desjardins D, Contreras V*, et al.* Antiviral efficacy of favipiravir against Zika and SARS-CoV-2 viruses in non-human primates [J]. Nat Commun, 2022, 13 (1):5108.  [5] Lieber C M, Cox R M, Sourimant J*, et al.* SARS-CoV-2 VOC type and biological sex affect molnupiravir efficacy in severe COVID-19 dwarf hamster model [J]. Nat Commun, 2022, 13 (1):4416.  [6] DiMasi J A, Hansen R W, Grabowski H G*, et al.* Cost of innovation in the pharmaceutical industry [J]. Journal of Health Economics, 1991, 10 (2):107-142.  [7] Jourdan J P, Bureau R, Rochais C*, et al.* Drug repositioning: a brief overview [J]. Journal of Pharmacy and Pharmacology, 2020, 72 (9):1145-1151.  [8] Liu R, Wei L, Zhang P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data [J]. Nature Machine Intelligence, 2021, 3 (1):68-75.  [9] Nunez S, Venhorst J, Kruse C G. Target-drug interactions: first principles and their application to drug discovery [J]. Drug Discovery Today, 2012, 17 (1-2):10-22.  [10] Bagherian M, Sabeti E, Wang K*, et al.* Machine learning approaches and databases for prediction of drug-target interaction: a survey paper [J]. Briefings in Bioinformatics, 2021, 22 (1):247-269.  [11] Zhang L, Tan J, Han D*, et al.* From machine learning to deep learning: progress in machine intelligence for rational drug discovery [J]. Drug Discovery Today, 2017, 22 (11):1680-1685.  [12] Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences [J]. Plos Computational Biology, 2019, 15 (6):e1007129.  [13] Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction [J]. Bioinformatics, 2018, 34 (17):821-829.  [14] Luo Y, Zhao X, Zhou J*, et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information [J]. Nature Communications, 2017, 8:573.  [15] Vaswani A, Shazeer N, Parmar N*, et al.* Attention is all you need [C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA, 2017.  [16] Zhang Z, Chen L, Zhong F*, et al.* Graph neural network approaches for drug-target interactions [J]. Curr Opin Struct Biol, 2022, 73:102327.  [17] Abbasi K, Razzaghi P, Poso A*, et al.* DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks [J]. Bioinformatics, 2020, 36 (17):4633-4642.  [18] Sun M, Zhao S, Gilvary C*, et al.* Graph convolutional networks for computational drug development and discovery [J]. Briefings in Bioinformatics, 2020, 21 (3):919-935.  [19] Chen X, Yan C C, Zhang X*, et al.* Drug-target interaction prediction: databases, web servers and computational models [J]. Briefings in Bioinformatics, 2016, 17 (4):696-712.  [20] Chen R, Liu X, Jin S*, et al.* Machine Learning for Drug-Target Interaction Prediction [J]. Molecules, 2018, 23 (9):2208.  [21] Ezzat A, Wu M, Li X-L*, et al.* Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey [J]. Briefings in Bioinformatics, 2019, 20 (4):1337-1357.  [22] Ding Y, Tang J, Guo F. Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion [J]. Knowl. Based Syst., 2020, 204:106254.  [23] Sachdev K, Gupta M K. A comprehensive review of feature based methods for drug target interaction prediction [J]. Journal of Biomedical Informatics, 2019, 93:103159.  [24] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies [J]. Journal of Molecular Biology, 1994, 238 (1):54-61.  [25] Yap C W, Chen Y Z. Prediction of cytochrome p450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines [J]. Journal of Chemical Information and Modeling, 2005, 45 (4):982-992.  [26] Yamanishi Y, Kotera M, Kanehisa M*, et al.* Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework [J]. Bioinformatics, 2010, 26 (12):i246-i254.  [27] Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models [J]. Bioinformatics, 2009, 25 (18):2397-2403.  [28] Ding Y, Tang J, Guo F*, et al.* Identification of drug-target interactions via multiple kernel-based triple collaborative matrix factorization [J]. Briefings in Bioinformatics, 2022, 23 (2):bbab582.  [29] Mahmud S M H, Chen W, Liu Y*, et al.* PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques [J]. Briefings in Bioinformatics, 2021, 22 (5):bbab046.  [30] Huang K, Xiao C, Glass L M*, et al.* MolTrans: Molecular Interaction Transformer for drug–target interaction prediction [J]. Bioinformatics, 2020, 37 (6):830-836.  [31] Liu S, Wang Y, Deng Y*, et al.* Improved drug–target interaction prediction with intermolecular graph transformer [J]. Briefings in Bioinformatics, 2022, 23 (5):bbac162.  [32] Wang J, Hu J, Sun H*, et al.* MGPLI: exploring multigranular representations for protein–ligand interaction prediction [J]. Bioinformatics, 2022:btac597.  [33] Nguyen T, Le H, Quinn T P*, et al.* GraphDTA: predicting drug–target binding affinity with graph neural networks [J]. Bioinformatics, 2020, 37 (8):1140-1147.  [34] Shao K, Zhang Y, Wen Y*, et al.* DTI-HETA: prediction of drug–target interactions based on GCN and GAT on heterogeneous graph [J]. Briefings in Bioinformatics, 2022, 23 (3):bbac109.  [35] Li X-S, Liu X, Lu L*, et al.* Multiphysical graph neural network (MP-GNN) for COVID-19 drug design [J]. Briefings in Bioinformatics, 2022, 23 (4):bbac231.  [36] Cheng Z, Zhao Q, Li Y*, et al.* IIFDTI: predicting drug–target interactions through interactive and independent features based on attention mechanism [J]. Bioinformatics, 2022, 38 (17):4153-4161.  [37] Yi-Yu K, Tzu-Ting P, Teng-Kuang Y*, et al.* Artificial intelligence approach fighting COVID-19 with repurposing drugs [J]. Biomedical Journal, 2020, 43 (4):355-362.  [38] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [J]. Journal of Chemical Information and Computer Sciences, 1988, 28 (1):31-36.  [39] Jin W, Stokes J M, Eastman R T*, et al.* Deep learning identifies synergistic drug combinations for treating COVID-19 [J]. Proceedings of the National Academy of Sciences, 2021, 118 (39):e2105070118. |

|  |
| --- |
| **二、研究技术路线及可行性分析** |
| 1. 研究技术路线     图6：研究技术路线  本课题在药物-靶点相互作用的基础上进行拓展，该模型构建了药物、靶点、副作用和疾病之间的异构关系网络，根据药物与药物、靶点、副作用和疾病的相互作用来预测药物靶点。   1. 基础数据集   本项目以Binding Database（BindingDB）、DAVIS、KIBA、Universal Protein（UniProt）、Drug Bank、SIDER、HPRD、CTD为主的公共药物靶向结合基准数据集为支撑，以多个相互作用及相互关系网络数据集、美国国家生物技术信息中心（NCBI）、COVID-19生物测序数据AID1706及靶点SARS-CoV2 3CL Protease、SARS\_CoV2 RNA Polymerase、SARS-CoV2 Helicase等信息为主要研究对象，构建基础数据集。  BindingDB是加州大学圣地亚哥分校 Michael K.Gilson实验室发布的一个可公开访问的主要收集药物靶点蛋白质和类药小分子之间相互作用亲和力的数据库。目前数据库包含超过920703个药物分子与8185种蛋白质的2096653个相互作用关系数据，以及100%序列匹配的2823种和85%序列匹配的8263受体/配体晶体结构数据，并不断更新中。  DAVIS包含 72 种药物和 442 种蛋白质的30056个相互作用关系数据结合亲和力，以 Kd 常数测量，范围从 5.0 到 10.8。  KIBA包含 2116 种药物和 229 种蛋白质的118254个相互作用关系数据，以 KIBA 分数衡量，范围从0.0到17.2。  UniProt由Swiss-Prot、TrEMBL和PIR-PSD三大子数据库构成的蛋白质数据库，数据主要来自于各物种基因组测序完成后得到的全基因蛋白质序列，并包含了很多来自文献中的蛋白及其功能信息。是资源最广、信息最丰富的蛋白质数据库。  美国国家生物技术信息中心（NCBI）：具备关于分子生物学，生物化学，和遗传学知识的大型数据库，可以提供新冠病毒基因组、蛋白质结构等一系列相关数据。   1. BERT节点表示   2.1 药物信息  使用简化的分子线性输入系统（SMILES）表示，表达成计算机可读的SMILES字符串形式，实现快速检索和子结构搜索等应用。可以使用自然语言处理（NLP）技术来表征字符串，或者直接在卷积神经网络（CNN）中使用它们，从SMILES字符串中推断出重原子数或价电子数等药物描述符，可以作为DTI的药物特征。  图7：药物化学结构到SMILES字符串示意图  **2.2 靶蛋白信息**  使用一串代表氨基酸的 ASCII 字符表示蛋白质序列信息。每种氨基酸类型都根据其相关的字母符号用一个整数编码从而表示为整数序列。例如：丙氨酸（A）为 1，胱氨酸（C）为 3，天冬氨酸（D）为 4。所有蛋白质序列剪切或填充为 固定残基长度序列。如果序列较短，则用零值填充。  **2.3 Transformer编码器**  编码器是通信领域的一个名词，是指将信号进行编制转换成容易传输的形式。在自然语言处理领域中是指将句子编码成一个能够映射出句子内容的固定长度的向量。因此，编码器可以视为一个特征提取器，主要作用是对输入序列进行特征提取。在DTI预测中，编码器主要是对分子信息进行特征提取。  在对分子信息进行编码时，本项目采用BERT模型中的Transformer编码器。Transformer是基于自注意力机制构建的模型，其核心模块为自注意力（Attn），同时还包括位置表达、层归一化（LN）、残差连接、前馈神经网络（FFN）。Transformer编码器（Encoder）共有6个相同的编码器层（layer），每个layer有两个子层，为多头注意力层（Multi-Head Attention）和前馈神经网络层（Feed Forward Neural Network, FNN）。    图8：Transformer编码器模型  多头注意力层是由自注意力机制构建，其主要目的是使当前节点不只关注当前信息，而是对整体分子中的序列信息进行关注。自注意力机制可以使得神经网络具备专注于其输入（或特征）子集的选择特定的输入的能力。在计算能力有限的情况下，注意力机制有助于将计算资源分配给更重要的部分。其表示形式为：    前馈网络层由两个线性变换函数和一个激活函数构成，其目的是更新多头注意力层中输出向量的每个状态信息。编码器层中的两个子层由残差网络和归一化层进行连接。残差网络主要对原始向量和经过多头注意力层的输出向量进行相加操作，其主要目的是防止经过多头注意力层的向量丢失一些重要信息。归一化层主要控制输入向量长度的动态变化，防止神经网络层的参数变化导致其输入的分布产生较大的差异。  Transformer采用多头注意力机制（Multi-head attention），自注意力机制可以描述为将查询和一组键-值对映射到注意力机制的输出中，其中查询（Query, Q）、键（Key, K）、值（Value, V）和输出都是向量。当使用自注意力机制进行实验时，使用不同的线性函数将查询、键和值分别映射到 h 次向量中能提取到不同的关注内容信息。因此提出多头注意力机制，其放缩点乘注意力（Scaled Dot-Product Attention）需做次，即多头。头之间参数不共享，即每次Q,K,V进行线性变换的参数W不一样，然后将次的放缩点乘注意力结果进行拼接后再进行一次线性变换得到的值作为多头注意力的结果，从而允许模型在不同表示的子空间里学习到相关的信息。    图9：自注意力机制的计算过程  具体来说，给定一个药物SMILES和蛋白质序列信息，将分子序列信息分解为子结构，特征矩阵，每个头的多头自注意力模块使用放缩点乘注意力，计算为：    其中，代表第l层k头的可训练投影权值矩阵。每个子结构有三个不同的向量Q, K, V，它们是由嵌入向量X乘以得到。代表每个自注意力头的维度，代表自注意力头的数量。  为缩放因子，从而避免较大时，存在某个Ｋ与Ｑ计算出来的对齐分数远大于其他的Ｋ与该Ｑ算出的对齐分数，这时softmax函数对各个的偏导数都趋于0。  将次的放缩点乘注意力结果拼接并进行一次线性变化得：    输出后将其传递给前馈神经网络（FFN），这个全连接包含两个线性转换，并使用函数ReLU进行激活。可以表示为：      图10：（a）放缩点乘注意力机制（Scaled Dot-Product Attention）  （b）多头注意力机制（Multi-Head Attention）  Transformer 的全局性使每个子结构信息能够以更近的距离关注 Transformer 层中任何位置的信息。同时，Transformer具有很高的并行性。   1. 药物-靶点生物信息异构网络   **3.1 异构网络构建**  将具体的实体看作网络图结构中的节点，实体之间的复杂的关系看作是节点间的互作用关系，即为网络图中的边，这样实体对象就可以称为信息网络。异构信息网络是一种特殊的信息网络，其包含了多种类型节点或多种类型的边。  一个信息网络可以定义为一个图𝐺 = （𝑉, 𝐸），其中𝑉代表图𝐺中的节点集，𝐸表示图𝐺中的边集合。运用映射函数𝜑: 𝑉 → 𝐴来表示每一个节点𝑣 ∈ 𝑉的特定节点类型𝜑（𝑣） ∈ 𝐴 ；同样，用映射函数𝜓: 𝐸 → 𝑅来表示图中每条边𝑒 ∈ 𝐸的特定边的类型𝜓（𝑒） ∈ 𝑅。相应的，若在信息网络图𝐺中节点的类型|𝐴| > 1或节点间边的类型|𝑅| > 1 ，这样的网络便称为异构信息网络。反之，这样的网络就称为同构信息网络。  以本文所研究的药物-靶点异构网络为例，下图介绍了一个简略的药物-靶点异构网络，该异构网络包含三种类型的节点，分别是药物、靶点、疾病。其还包含两种类型的边，分别是药物-靶点的互作用关系，药物-疾病的关联关系。在异构网络中，不同类型的节点通过代表不同语义路径的边连接，这样的路径通常称为元路径。假设一个异构信息网络图𝐺 = （𝑉, 𝐸），其节点类型和边的类型分别表示为𝐴和𝐸。异构网络中的一条元路径 ϕ可以定义为：    可以简单地表示为，复合关系。如下图所示，两个药物可以通过多个元路径连接，例如药物-疾病-药物和药物-靶点-药物。不同的元路径代表不同的语义信息。给定一个元路径ϕ，每个节点存在一组基于元路径的邻居，这些邻居可以在异构图中揭示各种结构信息和丰富的语义信息。在异构信息网络中给定一个节点𝑖和元路径ϕ，节点𝑖的基于元路径的邻居定义为通过元路径ϕ与节点𝑖 连接的节点的集合。其中节点𝑖的邻居包含其节点本身。 以下图为例，在给定元路径药物-靶点-药物的情况下，基于该元路径的dr1的邻居包括dr1，dr2，dr3。显然，基于元路径的邻居可以利用异构图中的结构信息的不同方面。    图11：简单异构信息网络示例图    本模型中，图表示为，其中V是节点集，E是边集。该模型中药物、靶点等被表示为节点，节点之间的相互作用或关联表示为边。因此，对象类型设置为O [药物、靶点、疾病、副作用]，共四种。由前述，我们把各节点用药物分子、靶蛋白的Transformer表示。关系类型集合R[药物相互作用、药物结构相似性、药物-疾病关联、药物-副作用关联、疾病-药物关联、副作用-药物关联、药物-靶点相互作用、靶点-药物相互作用、靶点-靶点相互作用、靶点-靶点序列相似性、靶点-疾病关联、疾病-靶点关联]，共12种。所有边缘类型都是无向的，边缘值都是非负的，相同的两个节点可以具有多种不同类型的边连接。  本模型将组成12种关系网络，构造药物-靶点异构网络。   1. 药物-靶点相互作用网络/靶点药物相互作用网络：通过Drug Bank Version 3.0 下载中获得。其代表已知的药物-靶点互作用关联网络。Drug Bank 数据库将药物分子（包括生物技术药物）的结构和药理数据与其药物靶点的蛋白序列，结构和作用模式相结合，同时整合了药物的化学结构，药理作用，作用蛋白靶点，作用的生理通路，药物间互作用等信息并链接 KEGG 数据库来分析药物的详细信息。 2. 药物-药物相互作用网络：通过Drug Bank Version 3.0 下载获得。其包含多种药物-药物相互作用关系。 3. 药物-疾病关联网络/疾病-药物关联网络/靶点-疾病关联网络/疾病-靶点关联网络：通过比较毒理基因组学数据库（Comparative Toxicogenomics Database, CTD）获得药物-疾病的关联关系和靶点-疾病之间的关联关系。CTD 是一个有效的、开源的数据库，其目的在于加强人们对环境暴露如何影响人类健康的了解。它提供了关于化学基因/蛋白质互作用、化学疾病和基因疾病关系的有关信息。这些数据与功能和路径数据相结合，可以帮助验证关于环境影响疾病的机制假设。 4. 药物结构相似性网络：通过 RDkit库以半径为2的摩根指纹的计算方法来计算药物-药物化合物之间的戴斯（Dice）相似系数，从而确定药物化合物之间的结构相似性。该网络描述了药物与药物之间的结构相似关系。 5. 药物-副作用关联网络/副作用-药物关联网络：通过SIDER database Version 2获得药物与有关副作用的关联网络。 6. 靶点-靶点相互作用网络：由HPRD database Release 9获得。HPRD （ Human Protein Reference Database）存储了大量的人类蛋白质相互作用信息。和其他同类数据库相比，该数据库中存储的大量靶点之间的相互作用信息已经过实验验证。 7. 靶点-靶点序列相似性网络：基于Smith-Waterman算法计算靶点-靶点之间的基于序列顺序的相似性得分，从而获得靶点-靶点相似网络。     根据这些网络，我们可以整合构建出一个完整的异构网络。12种关系网络结合在一起，构建了一个包含708种药物、1512个靶点、5603种疾病和4192种副作用的异构网络，其中包括12,015个节点和1,895,445条边。网络异构性丰富，为预测潜在的药物-靶点互作用关系提供了大量的信息。  表1：异构网络数据统计表   |  |  |  |  | | --- | --- | --- | --- | | 药物 | 靶点 | 疾病 | 副作用 | | 708 | 1512 | 5603 | 4192 | | 药物-药物 | 药物-靶点 | 药物-疾病 | 药物-副作用 | | 10036 | 1923 | 199214 | 80164 | |  | 靶点-靶点 | 靶点-疾病 |  | |  | 7363 | 1596745 |  |   **3.2 节点嵌入**  该模块分为三部分：关系图卷积神经网络、注意力机制和网络拓扑重构。在给定一个异构网络的情况下，该将网络的拓扑信息和节点自身的唯一信息映射到其对应的特征空间，并尽可能地保留其原始的拓扑信息。本项目应用图卷积神经网络学习异构网络的节点级嵌入，并且集成了注意力机制，使得节点嵌入能够更好地表达不同关系类型对节点的影响程度，嵌入的节点更多。  **3.2.1关系图卷积神经网络**  图卷积神经网络（Graph Convolutional Network，GCN）是图神经网络的一个重要组成部分。与图嵌入中嵌入表示和下游任务的分离相比，GCN得到节点的低维向量嵌入，然后根据不同的任务进行端到端的训练，如节点分类、图分类和链接预测。  如图所示，四个不同计算图：（a）、（b）、（c）和（d）。在（a）中，中心节点是药物d1，其所有相邻节点都是药物；在（b）中，中心节点d3是药物，并且有两种类型的相邻节点：药物d1和d5以及靶点t4。（c）和（d）是目标节点位于中心的另外两种情况。将由（a）和（b）计算的同一药物节点的特征相加，以获得其嵌入表示。类似地，根据（c）和（d），可以获得目标节点的特征表示。节点嵌入量计算如下：    其中表示药物节点d的嵌入表示；和分别表示节点d在局部图（a）和（b）中的状态；表示目标节点t的嵌入表示，和分别表示节点t在局部图（c）和（d）中的状态。    图12：节点嵌入计算例图  给定一个异构网络，图的卷积神经网络嵌入到一个结点i的l层，其邻域嵌入为，则可以表示结点i的l+1层嵌入为：    其中表示节点i在关系r下的邻域集合，表示节点i与其具有边类型r的邻域之间的边界值之和，s（e）表示节点i和节点j之间的边界值，W表示神经网络中的权重，是ReLu激活函数。对于异构网络的图卷积神经网络，我们引入了特定于关系的变换，即依赖于边的类型和方向。对于每条边类型r，首先利用非线性变换进行节点i的邻域嵌入，然后通过归一化边权重对其进行平均，从而得到节点i关于每条边类型r的邻域信息聚合；在对节点i的所有边类型的邻域信息求和后，得到节点i的邻域信息之和；最后，将该邻域信息之和与节点自身嵌入的变换特征向量相结合，使得节点i在下一层的嵌入表示既包含节点的邻域信息，又包含节点自身的特殊信息。此外，将多层图卷积神经网络进行堆叠，使神经网络更深入，更好地学习异构网络的节点级嵌入。    图13：异构网络例图  **3.2.2图注意力机制**  在中心节点信息汇聚的过程中，图注意力机制（Graph attention mechanism，GAT）为领域节点分配不同的权重。该模型的注意机制与图卷积神经网络的最后一层相结合。在最后一层图卷积神经网络得到不同关系类型r的邻域信息之和后，利用注意力机制将不同关系类型的邻域信息分配给不同的重要程度，并指导最终节点的嵌入。    其中，是关系类型为r的节点i的邻域信息之和，用Softmax函数对其进行处理，得到关于节点i和关系类型r的关注系数，乘以原始嵌入的，得到添加注意力机制后不同关系类型的领域信息之和。    图14：结合注意机制的图卷积神经网络示意图  此外，将处理后的域信息的总和乘以不同关系类型的数量，再加上嵌入在前一层中的节点的映射，并由激活函数进行处理，以获得最终的节点嵌入。模型采用了注意力机制，可以更好地反映不同关系类型的信息对节点的影响，提高模型的可解释性，使模型更好地提取节点的拓扑信息，提高模型的性能。  **3.2.3 网络拓扑重建**  在给定多层图卷积神经网络处理的节点嵌入和注意力机制的情况下，边界值和网络拓扑重构公式可以定义为：    其中是特定边缘类型r的投影矩阵。如果边缘类型r是对称的，如药物-药物相关性、靶点-靶点相关性，则。将边缘分别用和投影表示为和后，求出两个投影向量的内积。让这个内积结果尽可能地重建原始边值s（e）。所有这些定义中使用的所有权重和投影矩阵都是用高斯初始化。通过网络拓扑重构过程，可以得到不同节点之间关系的强度。   1. 筛选抗新冠病毒药物   本模型将数据集分为10个同质和互斥的子集，每个子集都是通过随机分层抽样获得的。然后将9个互不相容的子集分别作为训练集和测试集，从而将数据集划分为10个子集，并且从训练集中随机选择5%作为验证集。因为有很多方法可以分成10个子集，为了缩小差异，我们将重复10次10折交叉验证，对最终的评估结果进行平均，找到模型泛化性能最优的超参值，减少数据划分的负面影响，并在一定程度上降低过拟合度。与训练集和测试集相比，验证集相对较小。如果损失函数的梯度下降在验证集中无效，则测试集不能用于该迭代，从而减少了不必要的计算。  本模型使用ADAM优化器来优化模型，ADAM优化器同时考虑一阶梯度矩估计和二阶矩估计，并且计算更新的步长。损失函数的数学构造可微，满足ADAM优化器的应用要求。ADAM优化器在超级参数的选择上具有鲁棒性，并且可以控制学习速率。更新步长不受渐变大小的影响，因此系统能够更好地处理稀疏渐变。因此使用ADAM优化器寻找全局最小损失函数。  我们使用网格搜索来确定模型的参数，本模型采用3层神经结构，节点嵌入维度为1024，图形卷积神经网络和注意力层的隐含单元为1024，网络拓扑重构层的隐含单元为512。  我们依托前文所述基准数据库、相互作用网络数据集、麻省理工学院AI Cures提供的COVID-19生物测序数据AID1706，以及靶点SARS-CoV2 3CL Protease、SARS\_CoV2 RNA Polymerase、SARS-CoV2 Helicase等列出与COVID-19密切相关的靶点。选取出现在异构网络数据集中的靶点。基于训练好的模型和数据集中的这些目标，我们可以筛选出可能影响COVID-19的候选药物。通过这种方法，我们可以更快、更具体地找到后续湿实验的候选药物。   1. 模型评估   本项目基于基础数据集，以端到端的方式训练模型，模型参数可以更有效地更新。分析模型中涉及的参数，获得一组最优模型参数。将本文提出的模型在两个数据集上与其他方法进行比较，以验证其优越性。  采用二元交叉熵作为损失函数，以接收器操作特征下的面积（ROC-AUC）、精确召回下的面积（PR-AUC）和 F1评分作为性能指标。  预测未知的药物-靶点互作用关系看作为一个分类问题，即若某种药物和某个靶点之间存在已知的互作用关系，可以用类标签“1”来代表，反之，若它们之间的互作用关系为未知的，可以用类标签“0”来代表。而在进行二元分类时，预测结果与真实值的对比可以产生四种结果：（1）真阳性（True Positive=TP）：真实值是正例，预测结果为正例。即分类器将正例分类正确。（2）假阴性（False Negative=FN）：真实值是正例，预测结果为负例。即分类器将正例错误分类成负例。（3）假阳性（False Positive=FP）：真实值是负例，预测结果为正例。即分类器将负例错误分为正例（4）真阴性（True Negative=TN）：真实值是负例，预测结果为负例。即分类器正确的分类了负例。  为了评估抗新冠病毒药物预筛选模型的泛化能力，该模型为测试样本产生一个预测分数，然后设定一个阈值（threshold）与预测得分进行对比，若样本的预测分数大于阈值则将该样本分类为正例，反之将其分类为负例。根据这个预测得分，可以将测试样本进行排序，“最可能”是正例的排在最前面，“最不可能”是正例的排在最后面。这样，分类过程就相当于在这个排序中以某个“截断点”（cut point）将样本分为两部分，前一部分视作正例，后一部分则视作负例。然后根据预测结果对样例进行排序，按此顺序逐个把样本作为正例进行预测，每次都计算出两个重要的值的真正例比率（True Positive Rate，TPR）和假正例比率（False Positive Rate，FPR）    分别以 TPR 和 FPR 为横、纵轴作图，便可以得到 ROC 曲线。利用 ROC 曲线比较不同的预测模型的预测效果时，若一个模型的 ROC 曲线被另一个模型的 ROC曲线盖住，则可直接判定后者的预测性能比前者强；若两个模型的 ROC 曲线互有交叉，则不能简单地分辨两个模型性能的强弱。此时，可以通过另一个指标ROC 曲线下方的面积 AUC（Area Under ROC Curve）来进行比较。通过计算 AUC曲线下方各个部分的面积并求和，即可得到相应的 AUC 值。假定ROC曲线是坐标为的点按序连接形成的，则ROC-AUC估算为：    PR 曲线与 AUC 曲线类似，将模型的预测结果进行排序，排名靠前的是模型认为“最大可能”是正例的样本，排名最后的则是学习器认为“最不可能”是正例的样本。按此顺序逐个把样本作为正例进行预测，则每次可以计算出当前的准确率（precision）和召回率（recall）。以召回率为横轴、准确率为纵轴作图，就得到了 PR 曲线。根据上述描述，准确率和召回率分别定义为：    其中，TP是真正例，FN是假反例，FP是假正例。PR-AUC为接收器操作特征下的面积，即PR曲线下的面积。  F1评分是查准率和查全率的调和平均值，公式为：    我们把药物靶点预测和药物发现作为链路预测的任务。由边连接的节点指示这是已知的相互作用或关联，而未由边连接的节点指示未发现的动作关系或没有关系。这是一个二元分类问题，其中已知的是正样本，未知的是负样本。我们的模型是在复杂异构网络中基于现有关系发现潜在未发现的动作关系。在进行对比实验时，首先考虑将所有未知对作为负样本，并采用十重交叉验证法划分训练集和测试集，取训练集样本的5%作为验证集。最后，对十个测试集的PR-AUC值求平均。并将其与其他基于异构网络的先进药物重新定位方法进行性能比较。  具体来说，我们设计如下对比实验：   1. 测试模型的节点表达。通过保留/去除模型的BERT对节点的表达，观察模型的性能指标的变化。 2. 测试异构网络组成对模型性能的影响。通过改变组成异构网络的12种关系网络的组成及个数，测试模型的精确性，找到最优组合。 3. 测试训练数据量对预测精度的影响。通过改变交叉验证的次数来改变训练集的大小，并使用这些训练集来训练模型以比较预测精度。 4. 测试模型的参数及性能。利用同样的数据集，与其他基于异构网络的先进药物预测模型进行了性能比较；使用其他模型的数据集，得到本模型的预测结果并进行比较。    1. 可行性分析 5. 项目理论基础方面   目前对于深度学习在药物-靶点相互作用预测的应用已有丰富的理论成果，如NeoDTI模型、ComboNet模型、DeepConv-DTI模型、MolTrans模型等等，为项目提供理论支撑。数据库具有成熟性、广泛性，Python、Anaconda在网络上有丰富的学习资料，同时依托指导老师团队的应用经验，能够顺利完成模型的搭建与训练。   1. 前期准备方面   项目组成员均具有编程基础，已经在指导老师以及学长学姐的帮助下，通过课余时间自学，掌握Python、Anaconda等的使用。此外，项目组成员员已具备外文资料查阅能力，已广泛查阅了大量与本项目相关的国内外资料，对基于深度学习DTI预测知识、抗新冠病毒药物预筛选技术发展了解较为全面，整理并讨论项目技术路线。   1. 参与人员方面   本项目的参与人员来自生物医学工程学院、生命科学学院、计算机学院，具备良好的学习与沟通能力，对机器学习的跨学科研究有着很大的热情，可以保证项目的顺利进行。团队成员已修读药学分子生物学，细胞生物学，生物化学等课程，对药物靶点研究等相关知识均有学习；已修读C语言程序设计、面向对象程序设计、数据结构与算法、计算机网络、计算机组成原理等课程，具备 C++、Python等编程基础知识以及数据结构知识，能够进行算法分析以及代码编写。   1. 指导老师团队方面   项目指导老师有着丰富的指导经验，能够给予及时且专业的指导。老师长期从事材料/化学/生物/机器学习的跨学科研究，有着深入的研究基础。同时项目指导实验平台也为大创项目提供了深厚的技术支持。  综上所述，本项目有着人员、技术、理论层面上的保证，能够保证项目顺利进行，完成预期目标。 |

|  |
| --- |
| * 1. **研究基础（对项目的参与动机、已有知识储备、相关研究和训练基础）** |
| 1. 对项目的参与动机   团队成员来自生物医学工程学院、计算机学院、生物科学学院，皆有基础专业知识储备，对生物信息方向抱有极大的兴趣。在与老师的多次沟通下，将发展迅速的机器学习领域与药学、生物学进行结合，确定了项目的选题方向。团队成员均阅读大量文献并积极展开讨论，本着科研探索的热情与科学研究的严谨态度，参与研究此课题的相关知识及相关发展状况，旨在对基于深度学习的药物-靶点相互作用预测模型进行搭建并优化，探索完善具有更好精确性、科学性的抗新冠病毒药物的预筛选技术。我们对该项目抱有极高的期待，也有信心探索，并会认真完成该项目。从而促进人工智能在生物医药等交叉领域的深化发展，发挥新型交叉学科的重要作用。   1. 已有知识储备   本课题组成员由生物医学工程学院、计算机学院、生物科学学院的4名同学组成，团队成员有省级大创、互联网+、四川省生物医学工程创新设计大赛、全国大学生数学建模竞赛等项目比赛获奖经历。团队成员已有数据结构、面向对象程序设计、药学分子生物学、生物化学、等专业知识储备，具有良好的编程能力及文献阅读能力。 团队成员围绕项目进行多次讨论，广泛查阅文献并咨询老师，对图神经网络、多层感知机、支持向量机等机器学习与深度学习相关模型知识均有认真钻研，对相关实验操作有充分了解，对优化方案进行深入探讨，制定了详实可靠的实验方案。   1. 相关研究和训练基础   项目成员已经共同阅读过大量英文文献，绝大部分是最新出版的文章，代表了此领域的最高水平，目前已对新冠病毒药物筛选、药物-靶点相互作用预测的深度学习模型有深入了解。因此，团队对该领域最新的发展都有一定的了解和认识，前期准备工作充足。为项目开展提供理论支持。我们将保证本项目研究计划的完成，并达到预期目标。 |
| **四、研究计划和进度（就文献查询、社会调查、方案设计、实验研究、数据处理、研制开发、撰写论文或研究报告、结题和答辩、成果推广、论文发表、专利申请等工作逐项计划时间，时间节点精确到月份）** |
| 1. 前期科研准备（2022年10月——2022年11月）   （1）与指导老师交流，对项目方向进行讨论，对研究课题进行完善，确定研究方向以及基础研究路线。  （2）收集研究数据，对数据进行整理、预处理。准备立项答辩。   1. 抗新冠病毒药物预筛选模型的构建与测试（2022年11月——2023年6月） 2. （1）探究BERT对分子序列信息的准确提取，使用BERT的Transformer编码器对药物分子、靶蛋白进行编码。 3. （2）组建12种关系网络，构建药物-靶点生物信息异构网络。利用图神经网络学习复杂异构关系网络的邻域信息，应用网络拓扑重构的方法来提取特征表示，得到节点嵌入特征。 4. （3）对模型进行训练，优化确定最优参数，实现端到端模型。 5. （4）根据已知的COVID-19靶点预测可能影响新冠肺炎治疗的候选药物，检测该模型的可靠性、准确性，并不断在实际测试中完善模型。 6. （5）整合实验数据，编写代码，建立抗新冠病毒的药物预筛选模型。 7. **药物预筛选**模型的分析与评估（2023年6月——2023年8月）   （1）选取恰当的性能指标，对模型进行评估，分析模型的优势与不足。  （2）设计比较实验，对模型优越性进行评估。   1. 数据处理、项目结题（2023年8月——2023年10月）   （1）依据项目评估结果，对模型做最终优化工作。  （2）收集整理相关实验数据，发表相关学术论文。  （3）梳理项目流程，在各项研究完成的基础上撰写结题报告，完成结题。 |

|  |  |
| --- | --- |
| **五、项目研究支撑条件（项目所依托的重点实验室（中心、平台）、双创平台、课题组等各类单位能提供的直接支持项目开展的软硬件设施和其他校内外资源）** | |
| 1. 指导老师团队支撑   本项目指导老师为吕志彬。现为副研究员，理学博士，担任Frontiers in Genetics Associated Editor同时担任Frontier in Bioengineering and Biotechnology 和Frontier in Bioinformatics Guest Editor。  吕志彬老师2013年博士毕业于北京大学。2013-2021先后在中国工程物理研究院、中国化工集团、电子科技大学从事跨学科的基础应用科学研究，并先后担任助理研究员、高级工程师、事业部副总经理。  吕志彬老师长期从事材料/化学/生物/机器学习的跨学科研究。迄今发表SCI收录27篇，其中中科院一区论文9篇，高被引论文2篇，单篇第1作者最高影响因子超过30，个人h因子19，文章累计引用超过1700次。主要一作或者通讯作者期刊有: Energy & Environmental Science、 Nanoscale、Nano Energy、Bioinformatics、Briefings in Bioinformatics等。主持国家自然科学基金、四川大学人才基金等项目累计4项，累计经费超过150万；参与的项目有973项目、国家自然科学基金重点项目、教育部重点项目等。获得教育部自然科学奖二等奖1项。  吕志彬老师对待学生热情、严谨、有责任心，相信在吕志彬老师的带领下，我们的项目可以有条不紊的进行。   1. 数据来源   本项目所需的药物、靶蛋白等分子信息数据，来自于BindingDB、DAVIS、KIBA、Durg Bank、SIDER、HPRD、CTD为主的公共药物、靶点基准数据集，蛋白质数据库Universal Protein（UniProt），开源的药物重利用数据集Curated Antiviral Drugs Library、Broad Repurposing Hub，麻省理工学院AI Cures提供的COVID-19生物测序数据AID1706，以及靶点SARS-CoV2 3CL Protease、SARS\_CoV2 RNA Polymerase、SARS-CoV2 Helicase等，数据具有广泛性、准确性、科学性，认可度高。 | |
|  |  |
| **六、预期成果形式（可多选）** | |
| （1）□SCI论文 篇  （2）√核心期刊论文 1 篇  （3）□会议论文 篇  （4）□内部编印期刊论文 篇  （5）□授权发明专利 项  （6）□申请发明专利 项  （7）√创新类竞赛获奖  （8）□创业类竞赛获奖  （9）□其他 名称： | |
|  |  |
| **七、项目经费概算（按申报项目目标任务需要进行预算，经费执行情况将与结题考核成绩挂钩）** | |
| **1.经费来源（单位：元）**  申请项目专项经费 6500  **2.经费支出（单位：元）**  （1）仪器设备费 3000  （2）耗材费 1500  （3）测试加工费 800  （4）国内会务及差旅费 0  （5）国外会务及差旅费 0  （6）文献/知识产权事务费 500  （7）办公费（含文印、办公用品等） 200  （8）其他费用 500  **合计 6500** | |

|  |
| --- |
| **八、评审情况** |
| **指导教师意见：** |
|  |
| **指导教师（签名）： 年 月 日** |
| **学院推荐意见：** |
|  |
| **主管院长签名： 年 月 日** |
| **学校专家评审意见：** |
|  |
| **组长签名： 年 月 日** |
| **学校认定意见及批准经费：** |
|  |
| **学校负责人签名： 年 月 日** |