

开 题 报 告

毕业设计题目：

基于卷积神经网络的药物-靶标相互作用预测

浙江理工大学本科毕业设计（论文）开题报告

[illegible]

基于卷积神经网络的药物-靶标相互作用预测 开题报告

1 选题背景与意义

1.1 选题背景

近年来，全球药物研发经历改革和发展。从 2021 年至 2024 年期间，受到新冠疫情影响，医药行业加速创新。截至 2023 年底，国家药监局累计批准创新药品 142 个、创新医疗器械 235 个。随着人口老龄化加剧、居民消费水平及医疗需求不断攀升，预计相关医疗卫生支出将保持持续增长，我国医药市场规模将以 10%-12% 速度增长，预计到 2025 年，行业规模将超过 3 万亿元。如何充分的利用机器学习的资源，提高药物靶标相互作用预测的准确率是当前药物发展的重要问题。DTI (Drug Target Interaction) 药物靶标相互作用是指药物分子与生物体内特定的靶标之间结合和相互作用。靶标主要是指体内具有药效功能并被药物作用的生物大分子，如蛋白质，核酸等等。常见的药物靶标类型有 G 蛋白偶联受体 (GPCR)，以此为靶点的药物约占 44%；酶类，以此为靶点的药物约占 29%；载体蛋白 (又称转运体)，以此为靶点的药物约占 15%；以及离子通道，以此为靶点的药物约占 7%。

靶标相互作用的强度称为结合力 (Affinity)，较高的亲和力可以使药物在低剂量的情况下产生预期的疗效以降低副作用的风险。^[1]亲和力衡量药物和靶标之间的相互作用有常用抑制常数 (K_i)、解离常数 (K_d)、半数抑制浓度 (IC_{50}) 来衡量药物与靶标之间的相互作用。^[2]抑制常数 (K_i) 表示抑制剂与酶或受体结合的亲和力，数值越小，亲和力越强；解离常数 (K_d) 衡量配体与靶标之间结合的稳定性，数值越小，表示结合越紧密；半数抑制浓度 (IC_{50}) 表示药物抑制靶标活性的浓度，数值越小，说明药物的抑制效果越好。药效学研究中发现，当药物与靶标相互作用的时候，亲和力直接反应了临床疗效。^[3]总结, K_d 和 K_i 都用于衡量配体或抑制剂与靶标的亲和力，数值越小，亲和力越强。

传统的药物主要是基于配体的方式或基于结构的方法。^[4]基于配体的方式是利用已知活性分子的特征，预测新的分子与靶标相互作用。常用的基于配体的预测方法为 QSAR (Quantitative Structure-Activity Relationship)，通过化合物的分子结构和其生物活性之间的数学关系，来预测新分子活性的方法，但数据依赖性极强，需要可靠的实验验证数据来建立和验证模型。基于结构的方法是根据药物和靶标的三维空间结构对二者进行预测，常用的分子对接是通过模拟药物分子如何结合到靶标的活性位点。但靶标结构信息不足，并不是所有的药物靶标都有三维结构，典型的例子是 G 蛋白偶联受体 (GPCRs)。GPCRs 里的 2 肾上腺素受体 (2-adrenergic receptor, 2AR) 就难以用传统

的 X 涉嫌金题学方法解析，这是与治疗哮喘和高血压有关的疾病。传统的药物研发遵循了分子靶病的范式，目的是寻找一个能够作用于一种疾病单个靶标蛋白质的药物，这类药物设计方法只从疾病的单一可能作用靶标出发，没有考虑到疾病趋向与多个靶标相互关联。研究表明，大多数药物能够与多个靶标相互结合，而单一靶标也可以与多个药物结合。

相对于传统的药物研发成本高，研发周期长，失败率高，难进行大规模的药物靶标相互识别工作。从 1993 年至今，FDA (United States food and drug administration) 每年批准的新药数量增长十分缓慢。《FDA 罕见用药法案》定义的罕见疾病(低于 20 万美国患病率的疾病)超过 7 000 种，95% 以上的疾病没有相应的治疗药物。^{[5][6]}

在机器学习中，药物靶标相互作用预测能减少药物研发的周期和成本，虽然药物靶标相互作用预测有许多优点，但依然存在以下问题：

1. **少样本的问题** 在药物靶标相互作用的预测，由于实验验证的样本数据有限导致了正样本稀缺以及数据来源有限，导致模型训练容易过拟合。这些数据来源主要是 DrugBank, ChEMBL, BindingDB 等等。
2. **特征提取不足** 传统方法在提取药物分子和靶标蛋白质的特征时，可能无法全面捕捉其化学和生物特性，特别是在建模化学空间与基因组空间的关系时存在困难。

以上原因，减慢了药物靶标相互作用预测的情况。

基于以上背景，我们致力于设计和开发基于卷积神经网络的药物-靶标相互作用预测并进行可视化展示。

1.2 选题意义

(1) 现实意义

选题背景和意义

- **发现新靶标** 通过预测现有的药物与其他未被识别的靶标之间的特异性结合，可以挖掘药物的新功能
- **减少实验筛选的工作量** 相对于传统的药物研发成本高。基于预测模型的方法可以大规模，快速的筛选潜在的相互作用，压缩药物的研发周期
- **药物重定位** 治疗罕见疾病或突发疾病，如果对罕见疾病进行新药研发，可供研究样本有限而且研发成本不可估量。

(2) 技术意义

- **跨学科融合** 通过药物靶标相互作用预测中，涉及到生物信息学、计算机科学、化学等多个学科的交叉，探讨如何将这些学科的知识融合，进一步推动计算机辅助设计的发展。

2 国内外研究现状与分析

在 20 世纪初期，AutoDock 成了分子对接的重要工具。基于三维结构模拟药物分子与靶标结合。它为药物靶标相互作用的预测提供了计算化学和分子模拟的方法。QSAR 是 2000 年随着生物信息学和计算机科学的快速发展，成为现代药物发现中不可或缺的工具之一。随着机器学习方法的逐步成熟，药物靶标相互作用的预测与机器学习算法结合。Yamanishi et al.^[7]在 2008 年提出的基于相似性的方法，首次将药物与靶点的相似性纳入一个统一框架。

2008 年 Jacob et al.^[8]提出了使用 SVM 在蛋白质-配体相互作用预测。2012 年，Yu et al.^[9]利用了 SVM 和随机森林 (RF) 并基于药物特性和生物数据，构建了高效的预测模型。数据来源包括 DrugBank 等多个数据库，并对方法的性能进行了交叉验证分析。2014 Cao et al.^[10]提出的机遇特征提取的该并 SVM 方法。但 SVM 是以二分类为主，忽略了结合亲和力值。

2011 年，Cheng et al.^[11]基于网络推断 (NBI) 方法，用于预测化学-蛋白质相互作用 (CPI)。这种网络推断方法可以帮助快速筛选潜在药物靶标，降低实验工作量，并为药物多靶标设计和副作用预测提供参考。2012 年，Chen, H., et al.^{Chen2012}提出了异构信息的网络推断方法 (INDTI)，利用 INDTI 预测了 22 个与阿尔茨海默病相关的靶点，发现了多个潜在相互作用。然而稀疏数据仍然是这一方法的挑战之一。出自同一个期刊的 PLoS One 的是 2012 年 Cobanoglu et al.^[12]提出了基于协同过滤的矩阵分解方法，矩阵分解技术不需要药物或靶点的结构数据，能够处理大规模数据。但对于高维非线性关系建模有限。

随着深度学习的算法普及，2017 年 DeepDTA^[13]引入了 CNN，从原始数据 (SMILES 和序列) 中直接学习特征。特点是回归任务不会忽略亲和力值。在 2019 年，DeepConv-DTI^[14]和 DeepDTAF^[15]进一步的优化了 CNN 的架构。优点是不用手动提取，但对高噪声数据敏感是指模型在处理包含大量无关或错误信息的数据时，其性能会显著下降，甚至可能产生误导性的预测结果。

2019 年，GraphDTA^[16]使用图神经网络 (GNN) 结合分子和靶标特征，提高了预测性能，但依然存在以下问题随着数据规模的扩大，模型的训练时间和内存需求可能成

为瓶颈。2020 年, NeoDTI^[17]利用了异构图网络 (HGN) 进行预测, 它利用图神经网络 (GNN) 来整合多种类型的节点和边信息。尽管这一方法具有很大的优势, 能够捕捉多模态信息的关系, 但异构图的构建和训练需要处理大量节点和边, 计算成本较高。

随着强化学习的兴起, You, J., et al.^[18]通过强化学习的方式, 利用 Q-learning 强化学习算法进行药物-靶标相互作用 (DTI) 预测,。缺点是 QLDTI 的性能严重依赖于高质量和全面的数据集, 数据偏差可能导致预测偏差。在药物-靶标数据集较大的情况下, 模型训练时间可能会非常长, 影响其实用性, 融合多模型和多数据源可能导致计算成本和训练时间大幅增加。2020 年, Cao, X., et al.^[4] DeepDrug 结合了图神经网络 (GNN) 和强化学习 (RL) 来预测药物-靶标相互作用。药物分子和靶标蛋白质被建模为图, 节点表示分子或蛋白质的不同组成部分, 边表示它们之间的关系, 模型可以同时执行多个任务。但图神经网络和强化学习的结合在计算上通常非常昂贵, 尤其是在大规模药物数据集上训练时, 内存消耗和训练时间可能成为瓶颈。

3 研究的基本内容与拟解决的主要问题

3.1 研究的主要内容

3.2 输入与特征提取技术

在特征提取阶段, 我们将原始的药物和靶标数据转化为深度学习模型可以理解的高质量特征表示。特征提取的目标是捕捉药物和靶标的生物特性、化学性质以及它们之间的交互关系。在这一过程中, 采用了自注意力机制 (Self-Attention Mechanism)。通过这些步骤, 我们成功地将原始的药物和靶标数据转化为适合深度学习模型使用的高质量特征表示, 为后续的相互作用预测提供了坚实的基础。

3.3 模型设计

经过数据集和特征的提取后, 在模型设计阶段, 我们构建了一种基于 Transformers 和 CNN 的混合深度学习框架, 充分利用两种模型的特性来捕捉药物和靶标的多模态信息。对 DTI 任务进行微调, 以提升模型在小样本场景下的性能并充分利用数据的储备, 提高模型的泛化能力。模型通过全连接网络实现特征降维和映射, 最终输出药物与靶标相互作用的预测值。

3.4 模型评估

在模型评估阶段，我们使用了多种指标全面衡量模型的性能。具体包括回归指标（如 RMSE、MAE、R-squared）和分类指标（如 AUC 和 F1-Score），以确保模型的预测能力。在验证方法上，我们采用交叉验证和独立测试集来验证模型的稳健性和泛化能力。此外，为了防止过拟合，我们在模型训练过程中引入了正则化技术（如 Dropout 和 L2 权重衰减）。通过超参数调优进一步优化模型性能，同时结合注意力权重分析，验证模型在药物与靶标特征提取及交互建模中的表现。

3.5 拟解决的主要问题

3.6 少样本

少样本的问题主要实验成本高，数据稀缺等问题。少样本的后果会导致模型性能下降比如过拟合等问题。当训练数据不足时，模型可能记住训练样本而非学习通用规律，导致在测试集或新数据上的表现较差。以及模型训练困难在深度学习模型需要大量数据进行训练，少样本场景下模型可能难以收敛，甚至无法正确拟合数据。因此，解决少样本问题是提升 DTI 预测模型实用性和可靠性的关键，需要通过迁移学习和数据增强等方法来提高模型的泛化能力。

3.7 特征提取不足

特征提取不足是指在药物-靶标相互作用（DTI）预测中，无法从药物分子和靶标蛋白的数据中全面提取出能够反映其关键生物特性、化学性质和交互机制的有效特征。这种不足可能导致模型对药物和靶标的相互作用关系建模不准确。主要的因素是特征表示不充分，在传统手工设计的特征（如分子指纹、物理化学描述符）可能难以全面捕捉药物和靶标的深层次信息。以及交互信息缺失在药物和靶标之间的交互特征通常需要建模，但简单的特征连接或组合方法可能无法充分表达其复杂关系。这带来的后果会导致模型性能受限或交互关系建模不充分，例如准确率（Accuracy）、召回率（Recall）和 AUC 等指标下降。药物与靶标之间的复杂交互关系无法通过简单特征表示捕捉，模型可能忽略关键的化学基团或功能域。

4 总体研究思路及预期研究成果

4.1 流程图如图

流程图如图3所示，其中共包含三个步骤，分别为数据特征提取，模型算法，预测结果的可视化展示。

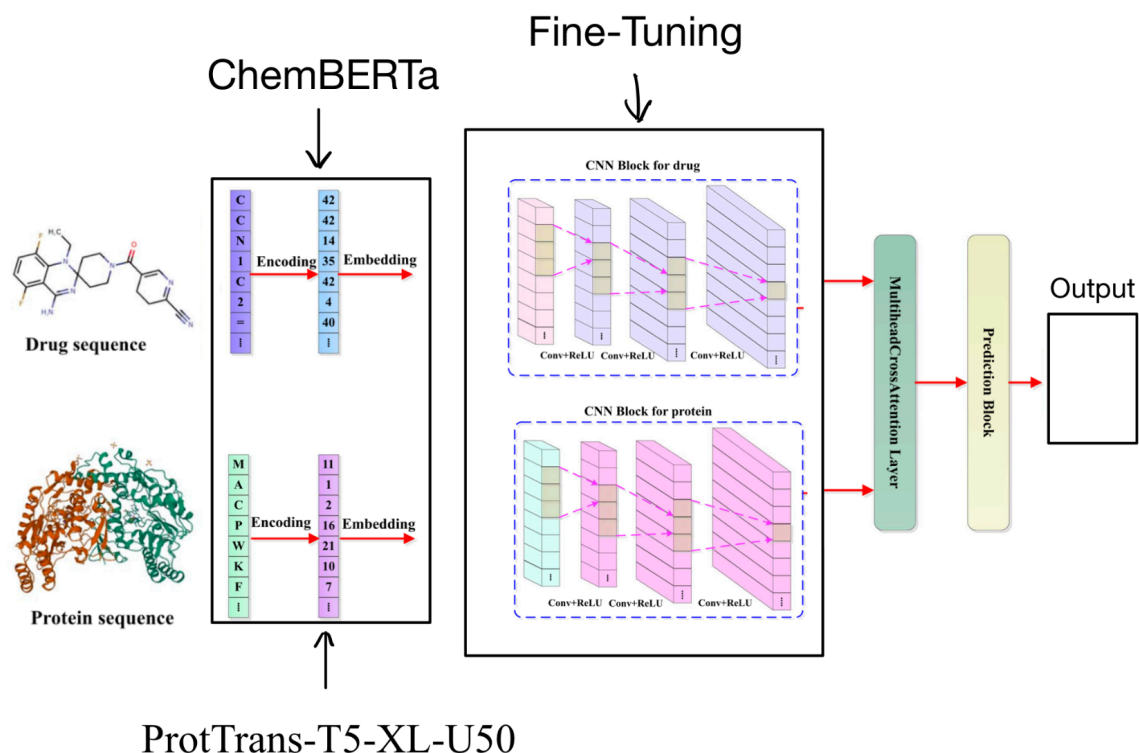


图 1 研究的主要内容

4.2 工具选型

(1) Pytorch

PyTorch 是一个广泛使用的深度学习框架，提供灵活的研究和开发环境，适合从快速原型到生产部署的多种场景。PyTorch 广泛支持的模块提供了丰富的神经网络模块（如 CNN、RNN、Transformer 等），以及常用优化器（如 SGD、Adam）和损失函数（如交叉熵、MSE），并且支持 GPU 加速（CUDA），利用多核并行计算提高运行效率。在本项目中，我们选择使用 PyTorch 实现自定义的深度学习模型，并便捷地处理高维输入输出，同时利用分布式训练在多个 GPU 上并行训练，为项目实现提供了高效可靠的解决方案。

(2) Dash

Dash 是基于 Python 的开源框架，提供了前后端一起，开发者无需掌握 HTML、CSS 或 JavaScript 的知识即可创建复杂的交互式界面。Dash 的核心组件包括 Dash Core Components 和 Dash HTML Components，支持多种数据可视化库（如 Plotly）和多种布局方式，同时提供灵活的回调机制，实现用户交互。本项目中，选择 Dash 为项目的可视化，展示了深度学习的训练过程，预测结果。

(3) Hugging Face Transformers

Hugging Face 的 Transformers 库提供了强大的自然语言处理系统。本项目使用的 ChemBERTa 和 ProtTrans-T5-XL-U50 模型来提取提取药物 SMILES 和蛋白质序列的特征。该项目中，我们将使用 Transformers 库加载预训练模型，提取药物和蛋白质的特征，为模型训练提供高质量的输入数据。

4.3 可行性分析

(1) 理论可行性

随着深度学习的发展，卷积神经网络在图像处理、自然语言处理等领域取得了巨大成功。在药物-靶标相互作用预测中，卷积神经网络也被广泛应用，取得了不错的效果。因此，基于卷积神经网络的药物-靶标相互作用预测具有理论可行性。

(2) 经济可行性

在本课题的研发过程中，不需要使用额外的数据集，因此在收集数据方面不需要开销。而且模型训练与评估所需的 GPU 可由导师实验室提供，项目开发成本较低，因此在经济方面具有可行性。

(3) 应用可行性

在药物研发领域，药物-靶标相互作用预测是一个重要的研究方向，对于药物研发具有重要意义。因此，基于卷积神经网络的药物-靶标相互作用预测具有较高的应用可行性，

4.4 研究思路

(1) 特征提取的算法实现

药物特征提取

药物分子的特征提取主要基于药物的 SMILES 表示。本研究采用预训练模型 ChemBERTa^[19]，从 SMILES 表达中自动捕获药物分子的全局上下文特征。具体实现过程如下：

1. 对于 SMILES 的预处理，
2. 使用 Hugging Face Transformers 加载 ChemBERTa 预训练模型
3. ChemBERTa 的分词器基于 (Byte-Pair-Encoder) BPE 方法和嵌入 (Embedding) 分词后的标记会被映射到一个连续的向量空间，能够有效地分解 SMILES 表达式，同时保留其化学结构的语义信息。分词后的标记通过嵌入矩阵映射为高维向量，ChemBERTa 使用一个嵌入矩阵 (embedding matrix)，不仅表示每个 token 的静态特征，还通过多头注意力机制捕获了分子全局的上下文关系，使得模型能够对 SMILES 表示中的重要子结构及其相互作用进行建模。
4. 将特征作为输入进卷积神经网络中，提取药物分子的特征表示。

蛋白质特征提取

蛋白质的特征提取利用预训练模型 ProtTrans-T5-XL-U50^[20]。具体实现过程如下：

1. 使用 Hugging Face Transformers 加载 ProtTrans-T5-XL-U50 预训练模型
2. ProtTrans-T5-XL-U50 对于蛋白质的预处理使用单个空格和 21 个词汇量进行标记。稀有氨基酸“U、Z、O、B”被映射到“X”。预处理步骤是动态进行的，通过切割和填充蛋白质序列直至 512 个标记。
3. 预训练过程中，随机选择 15% 的氨基酸进行掩蔽。80% 的被掩蔽氨基酸替换为特殊标记 [MASK]。10% 的被掩蔽氨基酸替换为随机的标准氨基酸（不同于原始氨基酸）。10% 的被掩蔽氨基酸保持不变。掩蔽策略的设计灵感来自于自然语言处理中的 BERT 模型，旨在帮助模型学习上下文信息，提高对被掩蔽氨基酸的预测能力。在蛋白质序列建模中，采用类似的掩蔽策略有助于模型捕捉序列中的生物学特征和结构信息。
4. 将特征作为输入到卷积神经网络中，提取蛋白质的特征表示。

(2) 基本神经网络基本算法

模型的设计与实现

在解决药物靶标相互预测问题时，我们需要设计一个基于卷积神经网络的模型和结合药物和蛋白质特征，Multi-Head Cross Attention 和迁移学习来预测药物-靶标相互作用. 这一算法的关键目标是准确预测药物与靶标之间的相互作用，高效的特征提取，从而辅助药物研发和生物医学研究。

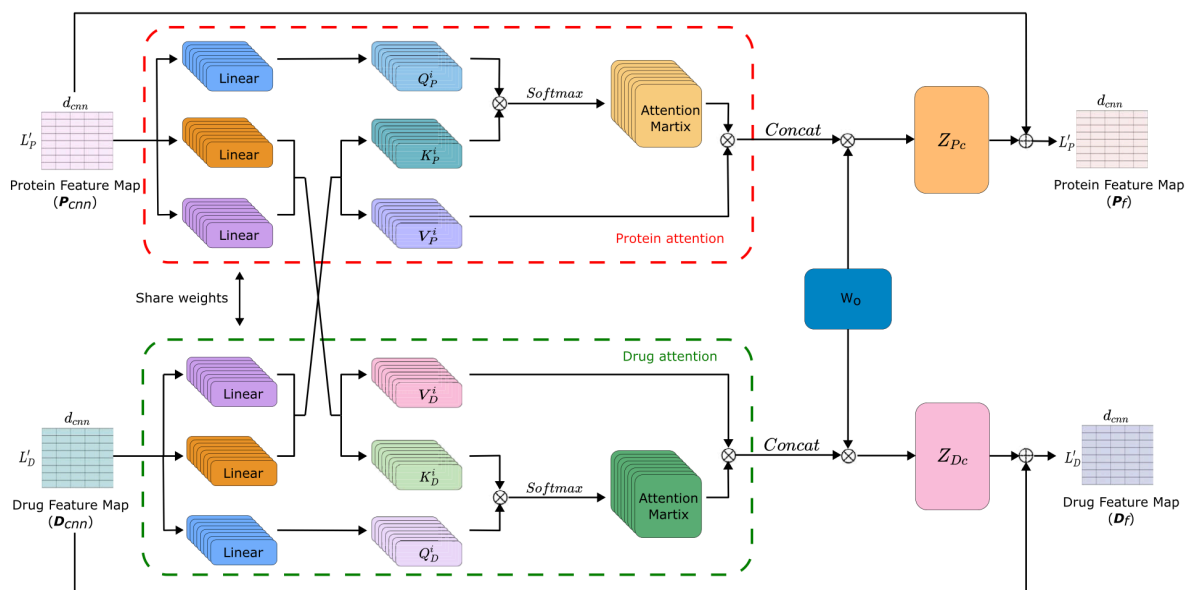


图 2 Multi-Head Cross Attention 的结构图

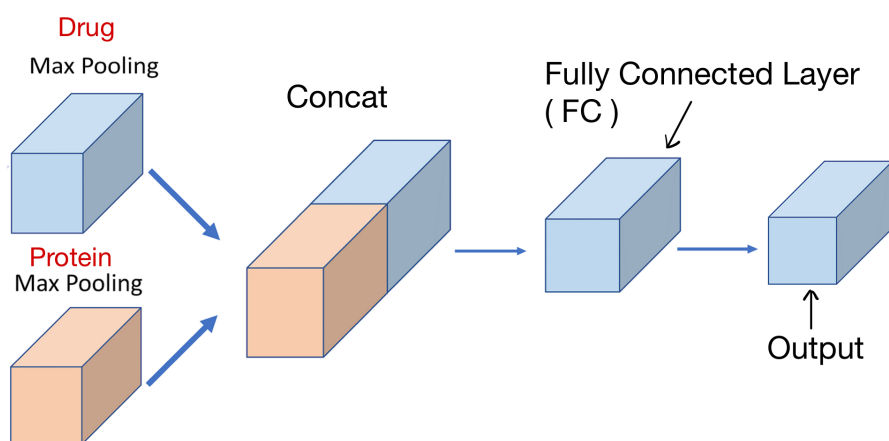


图 3 展示了 MaxPooling, Concat, FC, Output 的结构图

模型的设计：如果在机器学习中选择合适的算法是药物靶标相互作用预测最关键的问题。机器学习中常用的算法有 SVM，随机森林，神经网络等。这一领域的核心思想

是通过训练模型，从而做出预测。我们发现：

- 无监督学习：无监督学习能够利用大量未标记的数据，降低对人工标注的依赖，节省时间和成本。通过聚类、降维等技术可以发现新的药物-靶标相互作用。但无监督学习方法可能无法达到监督学习方法的性能，特别是在需要高精度预测的任务中。
- 监督学习：监督学习方法通过利用带标签的数据集，训练模型以预测新的药物-靶标对的相互作用。由于模型在带标签的数据上进行训练，能够学习输入与输出之间的直接映射关系，从而在预测未知数据时实现较高的准确性。但监督学习需要大量高质量的标注数据，而获取这些数据通常需要耗费大量的人力、时间和成本。
- 强化学习：强化学习是一种通过与环境交互，学习最优策略以实现特定目标的机器学习方法。强化学习注重策略的长期累积奖励，在复杂环境中，强化学习算法可能难以收敛到最优策略，或需要较长时间才能实现收敛，因此强化学习的应用尚处于探索阶段。
- 迁移学习：迁移学习能够利用在源领域预训练的模型，减少在目标领域从头开始训练的时间和计算资源。通过在相关任务中学习到的特征，模型能够更好地适应新的任务，提高泛化性能。但在迁移过程中，可能需要调整模型结构或参数。

在本项目中，我们选择基于卷积神经网络的模型，结合药物和蛋白质的特征以预训练模型的方式，多头交互注意力和迁移学习来实现药物-靶标相互作用的预测。

以下是多头注意力的公式：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \quad (1)$$

其中， Q 、 K 和 V 分别表示查询（Query）、键（Key）和值（Value）的矩阵。

模型的训练

模型的训练通常分为数据预处理，模型和预测的任务。在数据的预处理中，我们需要对药物和蛋白质的特征进行提取。我们将使用 ChemBERTa 和 ProtTrans-T5-XL-U50 预训练模型，从 SMILES 和蛋白质序列中提取特征。在模型的训练，我们选择使用监督学习的方式，提到过监督学习可以利用带标签的数据集，能够学习输入与输出之间的直接映射关系，从而提高准确率。在预训练作为输入，监督学习作为模型，迁移学习作为训练方式冻结大部分的 CNN 来提高模型的性能。

模型的评估

相对于无监督学习和监督学习，监督学习更适合于药物靶标相互作用的预测。所以使用监督学习作为主要的模型以及迁移学习来训练模型。模型的评估我们使用

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \tilde{y}_i)^2} \quad (2)$$

RMSE 量化回归模型整体误差，值越小，模型拟合效果越好。

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \tilde{y}_i| \quad (3)$$

MAE 衡量预测值与真实值的距离，值越小，模型拟合效果越好。

$$\text{CI} = \frac{1}{Z} \sum_{i,j,i>j} \sigma(y_i > y_j) \varphi(\tilde{y}_i - \tilde{y}_j) \quad (4)$$

$$\sigma(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (5)$$

$$\varphi(x) = \begin{cases} 1, & x > 0 \\ 0.5, & x = 0 \\ 0, & x < 0 \end{cases} \quad (6)$$

CI 衡量预测结果与实际结果一致的概率，值介于 0 1。

(3) 药物靶标相互预测的可视化

可视化操作在药物靶标相互预中扮演着一个非常重要的角色，它不仅帮助我们更直观的了解不同算法的效果，还提供我们一个更加方便的工具，将预测药物 - 靶标相互作用作为回归任务时，常用的模型评价指标包括均方根误差（root mean squared error, RMSE）、平均绝对误差（mean absolute error, MAE）、一致性指数（concordance index, CI）等等。

在实现药物靶标相互作用中，我们将使用 Dash 框架具体为以下几步：

用户输入接口：药物和蛋白质的输入文本框，显示药物图形结构基于 RDKit 和蛋白质的序列

实时返回数据：通过 Dash 来确保，MSE，MAE，CI 等指标的实时更新和准确率、精准率、召回率、F1、ROC 曲线、AUC 曲线等的预测结果的实时更新。

保存设计结果：提供下载按钮，允许用户将预测结果保存为 CSV 或 Excel 文件，便于后续分析和记录。

综上，使用了让用户输入后实现可视化的图形，还有实时结果的返回，以及最后的保存结果，这样的操作，使得用户更加方便的使用我们的模型。

(4) 评估算法指标并与现有研究比较

为验证我们新提出的药物靶标相互作用预测算法的有效性，我们将制定有参考性的对比评价指标，例如 RMSE, MAE, CI 等。通过使用现有算法验证过的数据集进行校验，我们将能够验证我们的算法相较于其他现有模型在药物靶标相互作用预测问题上的优越性。同时，我们将对新算法与现有研究算法进行深入研究比较。通过全面评估它们在各项指标上的表现，包括但不限于效率、准确性和实用性，我们将能够更全面地了解这些算法的优劣势。这一比较将为实际应用中的决策提供可靠的参考依据，确保所选算法能够在各方面胜任实际场景中的挑战。

4.5 非技术因素

法律方面：在进行药物靶标相互作用的预测研究时，本人仅用于个人学习与研究不涉及到商业的用途，也必须遵守相关法律法规，特别是在数据共享和使用方面。

数据隐私和安全：数据集都是网上公开的数据集，不涉及到个人隐私数据，在数据的处理过程中，遵守相关的数据隐私法规。

药物靶标的安全考虑 在药物靶标相互作用的预测的结果，需要确保预测的准确性和可靠性至关重要。错误的预测可能导致无效或有害的药物进入临床试验，带来安全隐患。因此，我们需要对模型的预测结果进行充分的验证和评估。

4.6 预期研究成果

一个基于卷积神经网络的药物-靶标相互作用预测：我们期望提出一种创新的基于卷积神经网络的药物-靶标相互作用预测。该算法将能够有效预测药物与靶标之间的相互作用，助力新药开发和药物重定位。通过利用 CNN 的强大特征提取能力，我们能够从药物和靶标的特征中学习到深层次的表示，从而提高预测的准确性和可靠性。

基于 CNN 药物-靶标相互作用预测和 Dash 可视化：通过与 Dash 的可视化让用户可以输入药物和蛋白质，实时数据的反馈，以及保存结果的功能。

对现有研究算法进行研究与比较：通过对现有研究算法的深入研究和全面比较，我们将提出的改进建议进行了优化以及验证我们的模型的有效性。这一比较将为实际应用中的

决策提供重要的参考依据。我们期望为药物研发和药物重定位提供更有效的计算工具，提升新药发现的效率和成功率。

5 研究工作计划

起止时间	课题实施计划
2024 年 11 月 09 日— 2024 年 12 月 14 日	查阅文献、撰写文献综述、外文翻译
2024 年 12 月 15 日— 2025 年 01 月 08 日	开题报告的撰写和答辩
2025 年 01 月 09 日— 2025 年 01 月 31 日	掌握基本神经网络基本算法及平台应用
2025 年 02 月 01 日— 2025 年 03 月 18 日	设计药物及靶标数据的特征提取算法
2025 年 03 月 19 日— 2025 年 03 月 19 日	中期检查
2025 年 03 月 20 日— 2025 年 04 月 13 日	实现并测试基于卷积神经网络的药物-靶标相互作用预测
2025 年 04 月 14 日— 2025 年 04 月 28 日	撰写毕业论文
2025 年 04 月 29 日— 2025 年 05 月 09 日	评阅老师评阅
2025 年 05 月 10 日— 2025 年 05 月 14 日	毕业设计答辩

参考文献

- [1] Moesser M A, Klein D, Boyles F, et al. Protein-Ligand Interaction Graphs: Learning from Ligand-Shaped 3D Interaction Graphs to Improve Binding Affinity Prediction[J/OL]. bioRxiv, 2022. eprint: <https://www.biorxiv.org/content/early/2022/03/07/2022.03.04.483012.full.pdf>. <https://www.biorxiv.org/content/early/2022/03/07/2022.03.04.483012>. DOI: 10.1101/2022.03.04.483012.

- [2] Abbasi K, Razzaghi P, Poso A, et al. Deep Learning in Drug Target Interaction Prediction: Current and Future Perspectives[J]. *Current Medicinal Chemistry*, 2021, 28(11):2100-2113. DOI: 10.2174/0929867327666200907141016.
- [3] 曾煜妮. 面向药物-靶标相互作用预测的多特征学习方法[D]. 四川大学, 2021. DOI: 10.27342/d.cnki.gscdu.2021.000780.
- [4] Yin Q, Cao X, Fan R, et al. DeepDrug: A general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction[J/OL]. *bioRxiv*, 2022. eprint: <https://www.biorxiv.org/content/early/2022/04/12/2020.11.09.375626.full.pdf>. <https://www.biorxiv.org/content/early/2022/04/12/2020.11.09.375626>. DOI: 10.1101/2020.11.09.375626.
- [5] Pushpakom S, Iorio F, Eyers P A, et al. Drug repurposing: Progress, challenges and recommendations[J/OL]. *Nature Reviews Drug Discovery*, 2019, 18(1): 41-58. <https://doi.org/10.1038/nrd.2018.168>. DOI: 10.1038/nrd.2018.168.
- [6] Sardana D, Zhu C, Zhang M, et al. Drug repositioning for orphan diseases[J]. *Briefings in Bioinformatics*, 2011, 12(4): 346-356. eprint: 2011Apr18. DOI: 10.1093/bib/bbr021.
- [7] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces[J/OL]. *Bioinformatics*, 2008, 24(13): i232-i240. <https://doi.org/10.1093/bioinformatics/btn162>. DOI: 10.1093/bioinformatics/btn162.
- [8] Jacob L, Vert J P. Protein-ligand interaction prediction: an improved chemogenomics approach [J/OL]. *Bioinformatics*, 2008, 24(19): 2149-2156. <https://doi.org/10.1093/bioinformatics/btn409>. DOI: 10.1093/bioinformatics/btn409.
- [9] Yu H, Chen J, Xu X, et al. A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data[J/OL]. *PLOS ONE*, 2012, 7(5): e37608. <https://doi.org/10.1371/journal.pone.0037608>. DOI: 10.1371/journal.pone.0037608.
- [10] Cao D S, Zhang L X, Tan G S, et al. Computational Prediction of DrugTarget Interactions Using Chemical, Biological, and Network Features[J]. *Molecular Informatics*, 2014, 33. DOI: 10.1002/minf.201400009.
- [11] Cheng F, Zhou Y, Li W, et al. Prediction of Chemical-Protein Interactions Network with Weighted Network-Based Inference Method[J/OL]. *PLoS ONE*, 2012, 7(7): e41064. <https://doi.org/10.1371/journal.pone.0041064>. DOI: 10.1371/journal.pone.0041064.
- [12] Cobanoglu E, Liu C, Hu J, et al. Prediction of Chemical-Protein Interactions Network with Weighted Network-Based Inference Method[J/OL]. *PLoS ONE*, 2012, 7(7): e41064. <https://doi.org/10.1371/journal.pone.0041064>. DOI: 10.1371/journal.pone.0041064.
- [13] Öztürk H, Özgür A, Özkirimli E. DeepDTA: deep drug - target binding affinity prediction [J/OL]. *Bioinformatics*, 2018, 34(17): i821-i829. <https://doi.org/10.1093/bioinformatics/bty593>. DOI: 10.1093/bioinformatics/bty593.

- [14] Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences[J/OL]. PLOS Computational Biology, 2019, 15(6): e1007129. <https://doi.org/10.1371/journal.pcbi.1007129>. DOI: 10.1371/journal.pcbi.1007129.
- [15] Wang K, Zhou R, Li Y, et al. DeepDTAF: a deep learning method to predict protein-ligand binding affinity[J/OL]. Briefings in Bioinformatics, 2021, 22(5): bbab072. <https://doi.org/10.1093/bib/bbab072>. DOI: 10.1093/bib/bbab072.
- [16] Nguyen T, Le H, Quinn T P, et al. GraphDTA: Predicting drug-target binding affinity with graph neural networks[J/OL]. bioRxiv, 2020. eprint: <https://www.biorxiv.org/content/early/2020/10/02/684662.full.pdf>. <https://www.biorxiv.org/content/early/2020/10/02/684662>. DOI: 10.1101/684662.
- [17] Wan F, Hong L, Xiao A, et al. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug - target interactions[J/OL]. Bioinformatics, 2019, 35(1): 104-111. <https://doi.org/10.1093/bioinformatics/bty543>. DOI: 10.1093/bioinformatics/bty543.
- [18] Gao J, Fu Q, Sun J, et al. QLDTI: A Novel Reinforcement Learning-based Prediction Model for Drug-Target Interaction[J/OL]. Current Bioinformatics, 2024. <https://www.x-mol.com/paper/1715466381594218496/t>. DOI: 10.2174/1574893619666230928112936.
- [19] Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction[J/OL]. Preprint, NeurIPS 2020, 2020. <https://arxiv.org/abs/2010.09885>.
- [20] Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing[J/OL]. bioRxiv, 2020. eprint: <https://www.biorxiv.org/content/early/2020/07/21/2020.07.12.199554.full.pdf>. <https://www.biorxiv.org/content/early/2020/07/21/2020.07.12.199554>. DOI: 10.1101/2020.07.12.199554.