

上海交通大学学位论文

**基于深度学习的蛋白质序列生成研究**

**姓 名： 张扬**

**学 号：** **519111910174**

**导 师： 魏婷**

**学 院**： **上海交通大学生命科学技术学院**

**学科/专业名称： 生物技术（生物信息试点班）**

**申请学位层次： 学士**

**2023年 4月**

A Dissertation Submitted to

Shanghai Jiao Tong University for Bachelor Degree

DISSERTATION TEMPLATE FOR BACHELOR DEGREE OF ENGINEERING IN

DEEP LEARNING-BASED PROTEIN SEQUENCE GENERATION RESEARCH

**Author:**  **Zhang Yang**

**Supervisor:** **Wei Ting**

School of Life Sciences and Biotechnology

Shanghai Jiao Tong University

Shanghai, P.R.China

April 28th, 2023

**上海交通大学**

**学位论文原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：

日期： 年 月 日

**上海交通大学**

**学位论文使用授权书**

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于 **：**

□**公开论文**

□**内部论文**，保密□1年/□2年/□3年，过保密期后适用本授权书。

□**秘密论文**，保密 年（不超过10年），过保密期后适用本授权书。

□**机密论文**，保密 年（不超过20年），过保密期后适用本授权书。

（请在以上方框内选择打“**√**”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

# 摘 要

蛋白质是生物的物质基础，在生物体中所有重要的组成部分都需要有蛋白质的参与，蛋白质不仅参与到细胞的运作，而且还可以激活细胞内的其他机制，从而促进细胞的正常运行，并且还可以通过信息的传输、调节和催化来实现各种生理过程。而蛋白质所发挥的功能主要通过特定的空间结构以及不同种类的蛋白质间的相互作用来发挥。因此蛋白质设计旨在设计出具有特定结构和功能的蛋白质，而蛋白质设计中存在着两个关键任务：蛋白质主链骨架设计和固定骨架蛋白质序列设计。而本文主要研究的是固定骨架的蛋白质序列设计，旨在设计出具有不同的氨基酸序列但能折叠成特定的蛋白质骨架，也就是具有相似空间结构的蛋白质序列。

在本研究中，将蛋白质视为一种层次化的结构模型，在提取蛋白质序列信息、蛋白质信息时不会采用更下层的信息，同时不会采集到蛋白质全原子的信息。为了所采集信息更加全面，本研究在已有数据集的基础上加入全新的数据集将数据量提升一个数量级增加训练精度。在处理得到各类特征信息之后将蛋白质作为图输入模型进行训练，模型主体采用循环图卷积神经网络并采用自注意力机制来提升网络对于蛋白质不同部分相关性的注意。同时为了防止过拟合等情况，本研究在模型中采用了随机失活、高斯噪音、随即掩码等方法。

通过将训练后收敛到最优的一次网络参数作为最后生成蛋白质序列的模型参数，可以在短时间内生成大量具有较好回复率的蛋白质序列。对于生成序列，还需要计算回复率、同一率、多样率以及表征空间相似程度的均方根偏差。在可视化方面为了更直观展示与原有序列的结构相似性，本研究还对生成序列与原始序列进行分子叠合后的结构进行展示。

**关键词**：蛋白质设计，蛋白质骨架，深度学习，生成模型，骨架预测

# ABSTRACT

Proteins are the fundamental building blocks of life, and their participation is required in all important components of an organism. Proteins play vital biological functions such as signaling, regulation, and catalysis in the life activities of an organism. The functions of proteins mainly rely on specific spatial structures and interactions between different types of proteins. Therefore, protein design aims to create proteins with specific structures and functions, with two key tasks: the design of the protein backbone and the fixed framework protein sequence. This article focuses on the design of fixed framework protein sequences to create proteins with different amino acid sequences but fold into specific protein backbones, i.e., protein sequences with similar spatial structures.

In this study, proteins are considered as a hierarchical structural model, and information from lower levels is not used when extracting protein sequence information and protein information. Moreover, information on full atomic details of proteins is not collected. To gather more comprehensive information, a new dataset is added to the existing dataset to increase the amount of data and improve training accuracy. After obtaining various feature information, proteins are treated as inputs to a graph model for training. The model mainly adopts a recurrent graph convolutional neural network and self-attention mechanism to strengthen the network's attention to the correlation between different parts of the protein. To prevent overfitting, this study uses methods such as dropout, Gaussian noise, and random masking in the model.

By using the converged network parameters obtained during training as the model parameters for generating protein sequences, a large number of protein sequences with good recovery rates can be generated in a short period of time. For the generated sequences, recovery rate, identity rate, diversity rate, and root mean square deviation representing the similarity in structural space are calculated. To provide a more intuitive demonstration of the structural similarity between the generated sequences and the original ones, this study also displays the superimposed structures of the generated and original sequences in visualization.

**Key words:** protein design, protein backbone, deep learning, generative model, backbone prediction

**目 录**

[摘 要 I](#_Toc134949517)

[ABSTRACT II](#_Toc134949518)

[第一章 绪论 1](#_Toc134949519)

[1.1 引言 1](#_Toc134949520)

[1.1.1 蛋白质设计 1](#_Toc134949521)

[1.2 现存蛋白质设计方法 2](#_Toc134949522)

[1.2.1 基于经典能量函数的蛋白质序列设计 3](#_Toc134949523)

[1.2.2 基于深度学习的蛋白质序列设计 3](#_Toc134949524)

[1.2.3 现存方法的主要缺陷 5](#_Toc134949525)

[1.3 深度学习技术 6](#_Toc134949526)

[1.3.1 独热编码 6](#_Toc134949527)

[1.3.2 图卷积神经网络 6](#_Toc134949528)

[1.3.3 多头自注意力机制 7](#_Toc134949529)

[1.3.4 循环神经网络 8](#_Toc134949530)

[1.4 本文研究主要内容 8](#_Toc134949531)

[1.5 本文研究意义 9](#_Toc134949532)

[1.6 本章小结 9](#_Toc134949533)

[第二章 数据集的建立 10](#_Toc134949534)

[2.1 数据来源 10](#_Toc134949535)

[2. 1. 1 AlphaFold数据库 10](#_Toc134949536)

[2. 1. 2 CATH 数据库 10](#_Toc134949537)

[2.2 数据清洗与筛选 11](#_Toc134949538)

[2. 2. 1 数据分布与统计 11](#_Toc134949539)

[2. 2. 2 蛋白质抽象类 13](#_Toc134949540)

[2. 2. 3 蛋白质剪切 14](#_Toc134949541)

[2. 2. 4 分层采样 15](#_Toc134949542)

[2.3 本章小结 16](#_Toc134949543)

[第三章 模型的构建与改进 17](#_Toc134949544)

[3.1 数据格式转换 17](#_Toc134949545)

[3.2 蛋白质表示 18](#_Toc134949546)

[3.3 模型的架构 19](#_Toc134949547)

[3.3.1 初始模型架构 19](#_Toc134949548)

[3.3.2 优化后的模型架构 21](#_Toc134949549)

[3.4 本章小结 23](#_Toc134949550)

[第四章 结果与分析 24](#_Toc134949551)

[4.1 模型评估指标 24](#_Toc134949552)

[4.2 使用ESMFOLD进行序列结构预测 25](#_Toc134949553)

[4.3 分子叠合计算RMSD与结果展示 25](#_Toc134949554)

[4.4 本章小结 25](#_Toc134949555)

[第五章 全文总结 26](#_Toc134949556)

[5.1 主要结论 26](#_Toc134949557)

[5.2 研究展望 26](#_Toc134949558)

[参 考 文 献 27](#_Toc134949559)

[致 谢 28](#_Toc134949560)

# 绪论

## 1.1 引言

对于一条长度为100个氨基酸的蛋白质序列而言，在不考虑其他影响因素的情况下，因为每一个位点可能的氨基酸种类都为20种，因此该蛋白质的可能序列会达到20的100次幂条，对于如此庞大的数字而言即使是使用计算机来进行计算难度也十分巨大。而对于自然界中的蛋白质而言，不仅存在序列上的差异，其折叠方式也可能存在着差异，同时蛋白质需要发挥其作用往往需要折叠成特定的空间结构。因此蛋白质设计在蛋白质相关课题中是十分重要的。

### 1.1.1 蛋白质设计

蛋白质是生物体中最基础、最重要的分子之一，是构成生命体系的重要组成部分。蛋白质具有复杂的结构和多样的功能，包括传递信号、催化反应、调节代谢等等。蛋白质的复杂性和多样性，来源于其特定的氨基酸序列和折叠成的空间结构。因此，为了更好地理解蛋白质的生物学功能和开发具有新功能的蛋白质，蛋白质设计逐渐成为一个热门的研究领域。

蛋白质设计是通过人工合成具有特定氨基酸序列和三维结构的蛋白质来实现特定的生物学功能。在蛋白质设计中，需要通过精确的氨基酸序列设计和优化，使得蛋白质能够折叠成具有特定功能的三维空间结构，从而实现特定的生物学功能。蛋白质设计可以通过多种方法实现，包括理论计算、基于结构的设计和基于演化的设计等等。其中，理论计算是通过计算机模拟来预测蛋白质的结构和功能，而基于结构的设计是通过改变已知蛋白质的结构来实现特定的功能，基于演化的设计则是通过改变已知蛋白质序列来实现特定的功能。

蛋白质设计的重要性在于，它可以为生命科学、医学和工业等领域提供具有广泛应用的新型蛋白质。通过蛋白质设计可以开发出具有特定生物学功能的蛋白质，如药物、酶和抗体等等。此外，蛋白质设计还可以为生命科学提供更好的工具，如用于解析蛋白质的结构和功能的分子探针、用于改变细胞代谢途径的催化剂等等。

在蛋白质设计的实践中，主要面临两个重要的挑战。一是如何设计出具有特定结构和功能的蛋白质，即如何实现蛋白质的主链骨架设计和固定骨架蛋白质序列设计。二是如何对设计出的蛋白质进行准确的结构和功能验证。这两个挑战在蛋白质设计的过程中非常关键。而第一个任务中的固定骨架蛋白质序列设计则是本文研究的主要内容，该任务旨在设计具有不同氨基酸序列但能折叠成特定蛋白质骨架的蛋白质序列。这种设计涉及到找到氨基酸序列与骨架之间的映射关系，以及寻找使蛋白质折叠成期望结构的最佳序列。为了解决这个问题，研究人员使用了机器学习和人工智能等技术来模拟和优化蛋白质折叠过程。通过这些方法，设计出的蛋白质可以用于制造新的药物、材料、催化剂等应用领域。

由于蛋白质设计可以产生具有特定结构和功能的蛋白质，因此在许多领域具有重要的应用。例如，设计新的药物靶点和药物分子，通过精确的蛋白质结构模拟和模拟药物-蛋白质相互作用可以优化药物的性能和效果，从而改善疾病治疗。此外，蛋白质设计还可以应用于纳米技术、生物传感器、基因工程、环境污染控制和可持续能源等领域。

蛋白质设计的成功离不开多学科的协作和技术的发展。不仅需要生物化学、分子生物学和生物信息学等学科的知识，还需要高性能计算和大数据分析等技术的支持。例如，蛋白质设计的机器学习方法需要建立庞大的训练数据集，并使用高效的算法进行训练和优化。此外，结构生物学和红外光谱学等技术可以用于确定蛋白质结构，从而更好地指导蛋白质设计。在未来，蛋白质设计将继续在各个领域发挥重要作用，并且预计会涌现出更多的新技术和应用。

## 1.2 现存蛋白质设计方法

蛋白质设计作为生物信息学的一个重要研究领域，旨在利用计算方法生成具有特定功能的新型蛋白质序列，从而探索蛋白质生物学的各个方面，如结构、功能、稳定性和相互作用。 从计算的角度来看，蛋白质设计方法可以大致分为两类：基于经典能量函数的蛋白质序列设计和基于深度学习技术的蛋白质序列设计。

### 1.2.1 基于经典能量函数的蛋白质序列设计

基于经典能量函数的蛋白质序列设计是一种基于物理化学原理的计算方法，是应用最广泛的蛋白质设计方法之一。该方法的核心思想是利用物理化学性质和理论基础构建一个能量函数，然后通过计算能量函数值来评价蛋白质序列的稳定性，从而预测出最优的蛋白质序列。该方法通常由两个主要组件组成：一是能量函数，用于计算蛋白质的稳定性和可折叠性；二是搜索算法，用于搜索最优的蛋白质序列。

该方法的计算过程主要分为两个阶段。首先，需要通过一些软件工具构建蛋白质结构的三维模型，一般是基于已知的蛋白质结构来构建的。然后，基于物理化学原理和理论基础，构建能量函数，该能量函数包括氢键、疏水性、电荷、范德华力等多个因素，每个因素都有相应的参数值。接下来，需要对生成的蛋白质序列进行能量计算，计算出对应的能量值，评估其稳定性。最后根据能量值的大小选择最优的蛋白质序列。而该方法的搜索算法可以使用各种优化算法，如蒙特卡洛模拟、分子动力学模拟、启发式搜索等，以找到具有最小总能量的蛋白质序列。

这种方法具有较多优点。首先，它依赖于物理化学原理和理论基础，可以准确地评估蛋白质序列的稳定性，有利于预测最佳蛋白质序列。 其次，该方法具有较高的可靠性、普适性、可重复性和计算效率，并且可通过各种不同的软件和工具进行实现和应用，适用于不同类型的蛋白质序列设计，对于尚未被解析的结构仍然适用。此外，该方法可用于设计多种目的，例如增强蛋白质稳定性、增加蛋白质亲水性或改变蛋白质的折叠状态等。

总体而言，基于经典能量函数的蛋白质序列设计是一种成熟经典的蛋白质设计方法，具有较高的可靠性和普适性。 尽管这种方法有一定的局限性，但在实际应用中仍然具有很大的价值和潜力。

### 1.2.2 基于深度学习的蛋白质序列设计

基于深度学习的蛋白质序列设计是一种新兴的方法，其利用深度神经网络学习蛋白质序列和结构之间的关系，来预测新的蛋白质序列，或改变已有的蛋白质序列实现蛋白质的设计。相比于基于经典能量函数的方法，基于深度学习的方法具有更高的精度和更大的灵活性。

要实现基于深度学习的蛋白质序列设计，需要训练一个模型，该模型可以根据输入的序列数据生成新的蛋白质序列。要训练一个好的模型，需要一个良好的数据集，其中包含足够数量和多样性的蛋白质序列及其结构信息。同时，还需要对模型进行优化，以获得更好的训练效果。

基于深度学习的蛋白质序列设计包括数据准备、模型构建、模型训练和模型预测四个步骤。首先，需要准备足够的训练数据，通常是从已知结构的蛋白质序列中获取。接着，需要设计一个深度学习模型，包括输入层、隐藏层和输出层。输入层接受蛋白质序列的信息，隐藏层学习序列之间的规律，输出层预测蛋白质序列的性质或结构。然后，使用训练数据对模型进行训练，以优化模型的参数。最后，使用已经训练好的模型对新的蛋白质序列进行预测与评估。

基于深度学习的蛋白质序列设计方法具有可拓展性，可以应用于各种蛋白质序列设计问题。它可以通过输入的训练数据学习到蛋白质序列的规律，并使用这些规律预测新的蛋白质序列。相比于基于经典能量函数的方法，基于深度学习的方法可以在不需要先验知识的情况下自动学习序列和结构之间的关系。而且，可以通过修改训练数据集来适应新的任务或目标，处理大量的数据从而提高设计的准确性和可行性。因此，能够处理更复杂的问题，并且能够通过大规模的数据训练提高预测准确度。此外，基于深度学习的方法还可以结合基于经典能量函数的方法，通过融合两种方法的结果来提高蛋白质设计的准确度。

具体而言，在基于深度学习的蛋白质序列设计中，最常用的方法是卷积神经网络（CNN, Convolutional Neural Network）和循环神经网络（RNN, Recurrent Neural Network）的方法。RNN可以处理变长的序列数据，并且能够通过长短时记忆单元（LSTM）来捕捉序列中的长期依赖关系。因此RNN常用于建模蛋白质序列中的时序信息。CNN则是一种针对固定长度的序列数据进行卷积和池化操作的网络，通过卷积核和池化层的组合可以提取序列中的局部特征，并生成新的蛋白质序列。因此CNN主要用于捕捉序列中的局部信息，例如氨基酸的相对位置和化学性质。另外，还有一些其他的深度学习模型被应用于蛋白质序列设计，例如变分自编码器（Variational Autoencoder, VAE）、生成对抗网络（Generative Adversarial Networks, GANs）等。这些模型可以通过无监督学习和生成对抗的方式来生成新的蛋白质序列，同时也可以进行蛋白质序列的修改和优化。这些方法通常涉及到的问题包括分类、回归和生成，它们可以预测蛋白质序列的二级和三级结构、蛋白质的热力学稳定性和其他性质。

基于深度学习的方法可以预测蛋白质序列的各种性质，例如结构、功能、稳定性等。其中，预测蛋白质结构是基于深度学习的蛋白质序列设计方法的重要应用之一。该方法可以通过学习大量已知结构的蛋白质序列，预测新的蛋白质序列的三维结构，从而为蛋白质的功能研究和药物设计提供帮助。近年来，基于深度学习的蛋白质序列设计方法已经在多个领域取得了成功，如蛋白质药物设计、生物能源生产等。

深度学习方法具有许多优点，例如可以自动学习复杂的特征、可以处理大规模数据集、可以减少人工特征工程的需求等。它们已经成功应用于蛋白质序列设计的许多方面，并在预测蛋白质结构和性质等方面表现出色。但是，这些方法需要大量的训练数据，并且由于其复杂性，很难解释它们如何做出预测。此外，与经典方法相比，它们的计算资源和时间成本更高。因此，深度学习方法在蛋白质序列设计中的实际应用仍然面临着挑战。

### 1.2.3 现存方法的主要缺陷

对于基于经典能量函数的蛋白质序列设计而言存在着一些缺陷。首先，由于能量函数的复杂性，可能陷入局部最优解，需要调整多个参数值，这可能会增加优化难度和计算成本。 其次，该方法无法捕捉蛋白质结构的动态性和灵活性，过于依赖初始结构和仅考虑静态结构，这可能会限制其在一些复杂蛋白质序列设计中的应用。 最后，该方法不能直接考虑蛋白质序列的生物学功能和具体应用需求，需要根据实际应用进行优化。虽然基于经典能量函数的蛋白质序列设计方法存在一些限制和缺陷，但其仍然是一种有用的工具，为蛋白质设计提供了有力的支持。

同时，基于深度学习的蛋白质序列设计方法也仍然存在一些挑战和限制。首先，它需要大量的训练数据和计算资源，才能达到较好的效果。因此其中一个主要的问题是数据不足和数据质量不高。由于蛋白质序列和结构的复杂性，需要大量的高质量数据来进行训练和验证。而其所耗费的大量计算资源以及时间也制约着模型的精度。其次，深度学习模型通常是黑盒模型，很难解释模型的决策和预测结果。当然这也是来源与深度学习本身的模型的可解释性和可靠性。因此，在实际应用中需要对模型进行解释和验证，以确保模型的可靠性和稳定性。最后，该方法还存在着所生成蛋白质序列的多样性不足、序列生成速度偏慢、生成序列折叠能力不足等问题。

## 1.3 深度学习技术

### 1.3.1 独热编码

独热编码（One-hot encoding）是机器学习和人工智能领域广泛使用的离散数据编码方法。它通过将每个类别映射到一个唯一的向量来表示数据，其中只有一个位置的元素是 1，其余元素是 0，因此也称为“one-hot编码”。

通过独热编码的方式可用于将蛋白质序列中的氨基酸序列编码为数值向量，以便于机器学习算法进行处理。具体来说，每个氨基酸可以被编码为一个大小为20的向量，其中只有一个位置的元素为1，代表该氨基酸的类型，其余位置元素均为0，这样蛋白质序列就能转化为一个向量矩阵。这样得到的序列向量矩阵可以用于各类任务，例如分类、回归和聚类等。

除此之外，独热编码还可以结合其他表示方法使用，例如结合蛋白质的物理性质特征或者是二级结构特征，形成一种更综合全面的序列表示方法。这种表示方法在蛋白质结构预测中也得到了广泛应用，能够为机器学习算法提供更多的信息。

因此，独热编码是一种简单有效的数据编码方法，在蛋白质序列表示和蛋白质结构预测等领域中具有重要的应用价值。

### 1.3.2 图卷积神经网络

图卷积神经网络（Graph Convolutional Network，GCN）是一种基于图的深度学习技术，特别针对图结构数据而设计的神经网络模型。在GCN中，每个节点代表一个数据点，而边则代表这些数据点之间的联系。例如，对于蛋白质分子，节点可以表示原子或氨基酸残基，而边则可以表示它们之间的相互作用。相比传统的卷积神经网络，GCN更适用于处理不规则的、非欧几里德结构的数据，例如图和点云数据等。因为蛋白质结构是由氨基酸残基之间的相互作用所决定的，因此，蛋白质可以被视为一个拥有复杂拓扑结构的图。利用GCN可以处理这种图结构数据，进而进行蛋白质结构的预测和优化。正因如此其在蛋白质设计中具有广泛的应用。

GCN的核心思想在于，利用邻接矩阵和节点特征矩阵来更新节点的特征表示。对于每个节点及其邻居节点，GCN将其视为一个局部图，并通过对节点的特征向量进行卷积操作，以实现节点特征的学习和提取。具体而言，GCN将邻接矩阵和节点特征矩阵相乘，得到一个新的特征矩阵，然后对其进行非线性变换，从而得到更新后的节点特征矩阵。通过这种方式，GCN可以利用节点之间的关系来推断节点的特征表示。

在蛋白质设计中，GCN已经被广泛应用于蛋白质结构预测、序列-结构-功能关系预测和蛋白质序列设计等方面。其中，GCN的使用主要是通过利用蛋白质分子中氨基酸残基之间的相互作用关系来预测蛋白质的稳定性和功能。具体来说，GCN可以将蛋白质分子中的氨基酸残基建立成图结构，然后利用GCN学习这个图的特征表示，并将学习到的特征表示用于预测蛋白质的稳定性和功能。

最近几年的研究表明，GCN在蛋白质序列设计方面具有巨大的潜力。这种方法可以通过将GCN与深度生成模型结合使用，从而实现高效而准确的蛋白质序列设计。此外，GCN与物理模拟相结合可以更精确地预测蛋白质分子的稳定性和构象。因此，GCN的发展为解决蛋白质结构和功能设计中的关键问题提供了强有力的支持，成为蛋白质序列设计领域一种全新的思路和方法。

### 1.3.3 多头自注意力机制

自注意力机制（Self-Attention）是一种用于处理序列数据的机制。它通过计算输入序列的每个位置与其他位置之间的相似性，以获得全局上下文信息。然后，此信息用于通过计算加权和来生成表示位置的向量。自注意力机制中的相似度计算和权重分配是通过学习得到的，可以自适应地适应不同任务和数据。

多头自注意力机制（Multi-Head Self-Attention）是自注意力机制的扩展。它使用多个注意力头来计算不同的相似性和权重分配，然后将不同头部的结果经过连接或者是取平均值，来获得更丰富的全局上下文信息。

在蛋白质结构预测中，自注意力机制和多头自注意力机制被广泛应用于处理蛋白质序列和结构数据。通过将蛋白质序列和结构视为序列数据，自注意力机制可以自主学习它们之间的关系，并获得全局上下文信息。多头自注意力机制的应用也非常广泛，它可以更充分地学习不同的关系和上下文信息，从而提高蛋白质结构预测的准确程度和鲁棒性。例如，当利用自注意力机制预测蛋白质二级结构时，使用多头注意力可以同时捕获序列中的局部和全局信息，能够更好地适应不同长度的序列。在使用自注意力机制预测蛋白质结构时，多头自注意力可以计算不同类型的关系，例如残基之间的空间距离和角度，从而更好地捕捉残基之间的相关性。

### 1.3.4 循环神经网络

循环神经网络（Recurrent Neural Network, RNN）是一种能够处理序列数据并且具有记忆能力的神经网络。在蛋白质结构预测中，循环神经网络可以用于处理蛋白质的序列信息，并预测其二级结构、三级结构等。与传统神经网络不同的是，循环神经网络通过使用递归的方式将之前所得到的信息与当前的输入相结合，并以此生成输出和更新隐藏状态。这种递归结构使得其能够捕获时间序列数据中的动态特征。循环神经网络的最大优点在于能够处理可变长度的序列数据，这使得它在蛋白质序列预测中得到广泛应用。

循环神经网络的核心是循环单元，每个循环单元都有一个隐藏状态，隐藏状态可以记忆之前所得的信息，并根据当前输入和之前的隐藏状态去计算当前隐藏状态。循环神经网络可以通过反向传播算法进行训练，以此来最小化预测输出与实际输出之间的误差。

在蛋白质结构预测中，循环神经网络可以提取蛋白质的氨基酸序列作为输入，通过学习氨基酸序列之间的关系，从而预测其二级结构、三级结构等。循环神经网络可以自动捕获序列中的长程依赖关系，并将这些信息编码为隐藏状态，从而更好地捕获蛋白质序列中的信息。

总之，循环神经网络是一种广泛应用于序列数据处理的神经网络，特别是在蛋白质设计中，该模型可以用于序列建模、特征提取和表示学习，从而提高序列的预测精度和生成新的蛋白质序列的能力。

## 1.4 本文研究主要内容

本文主要研究固定骨架蛋白质序列设计。该方法通常是通过对已知的蛋白质结构进行修改和优化，来获得所需的蛋白质序列。而在本文中所描述的固定骨架的蛋白质设计则指的是通过将蛋白质骨架与蛋白质的空间结构进行映射，达到能够生成与输入的原始蛋白质构象在空间结构上近似的蛋白质序列的目的。为了达到这一目的，本文在数据处理、模型构建、模型评估与改进进行研究。为了保证输入数据的数据的充足性与准确性，本文还在现有数据集中加入全新的数据集并对其进行筛选，使得输入给深度学习模型的数据足量且准确。在模型层面则采用了能够较好地表示空间结构的循环图卷积神经网络，而为了进一步提升模型的学习能力，在后续改进中本文还加入了自注意力机制以及图注意力网络。最后对于生成的蛋白质序列本文还使用了多个评估指标来对结果进行评价。总而言之，最终期望得到具有类似于目标蛋白的结构乃至于功能（因为蛋白质的功能在一定程度上取决于其空间折叠结构）的蛋白质序列。

## 1.5 本文研究意义

本文所研究的基于深度学习模型的蛋白质序列设计具有多方面的意义。首先，传统的基于物理学原理的蛋白质设计方法在解决某些问题时存在局限性，而基于深度学习模型的蛋白质序列设计方法可以更好地解决这些问题。其次，基于深度学习模型的蛋白质序列设计方法具有高度的自动化程度和效率，能够极大地缩短蛋白质序列设计周期，降低设计成本。此外，该方法还可以有效地扩展已知的蛋白质序列空间，帮助研究者理解和探索蛋白质的结构和功能。

具体来说，通过该方法有几率在保持蛋白质空间结构的前提下生成出出更稳定、更具活性的酶类蛋白质，提高其催化效率。同时，对于自然界中未出现过的蛋白质但其性质和功能更符合实际需要的序列，也有可能通过该方法扩展已知蛋白质序列空间来设计。更不必说这种方法可以通过大规模的蛋白质序列和结构数据的学习，提取出蛋白质序列中的特征，快速生成高效的蛋白质序列，深度学习模型自动学习优化的特性也避免了基于能量函数的序列生成中存在的手动调整大量的参数和参数组合的问题。

## 1.6 本章小结

本章介绍了蛋白质设计的内容及其意义，并且较为详细的介绍了两种不同的蛋白质序列设计思路：基于经典能量函数的蛋白质序列设计、基于深度学习的蛋白质序列设计以及他们的主要缺陷。除此之外还介绍了所使用的多种深度学习模型以及其在蛋白质设计领域中的应用和作用。在介绍背景之后，解释了本文的主要研究内容、研究思路以及研究意义。我相信本设计能够体现本科四年所学知识与实际应用的结合，并且为该方向研究提供积累。

# 第二章 数据集的建立

## 2.1 数据来源

### 2. 1. 1 AlphaFold数据库

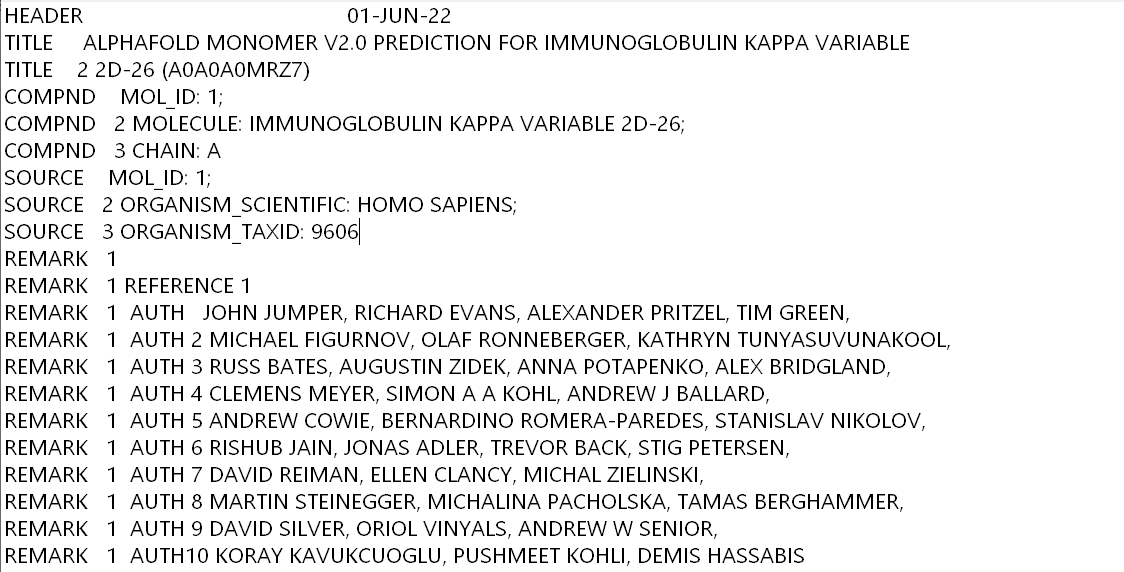
Alphafold数据库是一个包含大量蛋白质结构预测结果的数据库，由DeepMind公司开发。这个数据库中的数据基于DeepMind所开发的Alphafold模型，使用深度学习算法对已知的蛋白质序列进行预测，预测出它们的三维结构。Alphafold数据库包含了大量的蛋白质序列信息，这些序列可能来自于多种生物体，包括人类、动物、植物等48个物种的542,378条蛋白质序列。这些序列不仅包含已知序列，同时还包含未知序列。这些序列的长度也各不相同，从数十个氨基酸到数千个氨基酸不等。并且该数据库最显著特征是其包含的通过Alphafold模型预测得出蛋白质结构，这些预测结果包括了蛋白质的三维结构以及相关的结构信息，如二级结构、残基间距离等等。因此，使用该数据库可以使得深度学习模型所提取的序列特征更广泛，甚至包含一些主流数据库中暂时不存在明确结构的蛋白质，并以此来提升构想空间的搜索范围。

### 2. 1. 2 CATH 数据库

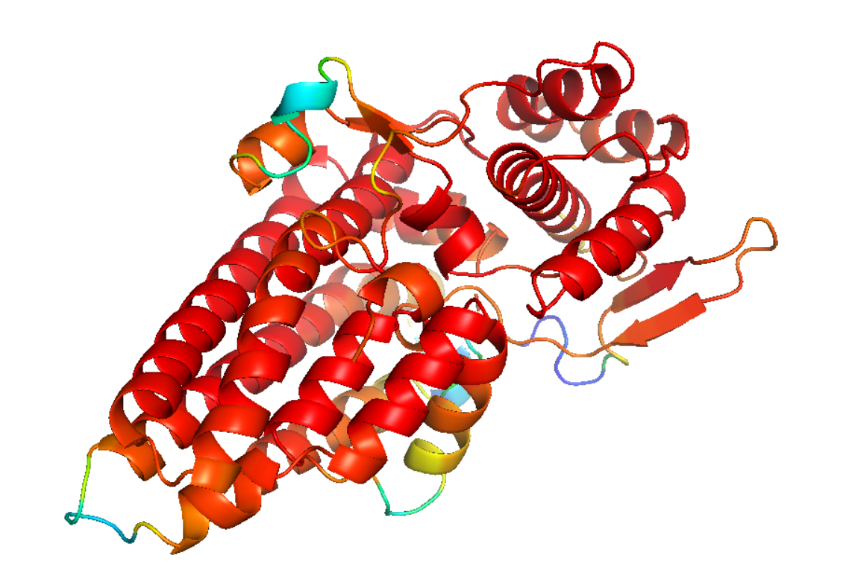
CATH（Class, Architecture, Topology, Homology）数据库是一个广泛使用的蛋白质结构分类数据库。它对蛋白质结构进行了细致的分类，是对现有结构分类方案的一种补充。CATH的数据集由蛋白质序列、结构及其分类信息组成。CATH数据集将所有已知的蛋白质结构划分为四级分类：类别（Class）、结构类型（Architecture）、拓扑类型（Topology）和同源型（Homology）。其中，类别对应蛋白质结构的总体折叠形式，包括四个类别：α（含有α螺旋的蛋白质）、β（含有β折叠的蛋白质）、α+β（含有α和β元件的蛋白质）和其他（结构不属于上述三种类型）。结构类型描述了蛋白质结构的三级结构元件如何连接起来。拓扑类型则指的是蛋白质结构的不同域以及域之间的连接方式。同源型是指拥有相似序列的蛋白质结构，包括同源超家族、同源家族和同源单体。因其蛋白质结构信息较为详细且含有大量经实验技术（如X射线晶体学、核磁共振等）所得结构便于研究，因此在本文提升数据集所包含天然蛋白质含量时加入该数据集。而在本文中所使用的是其中的剔除同源性高于40%的蛋白质后的单链蛋白质非冗余数据，其数量为28648条蛋白质。

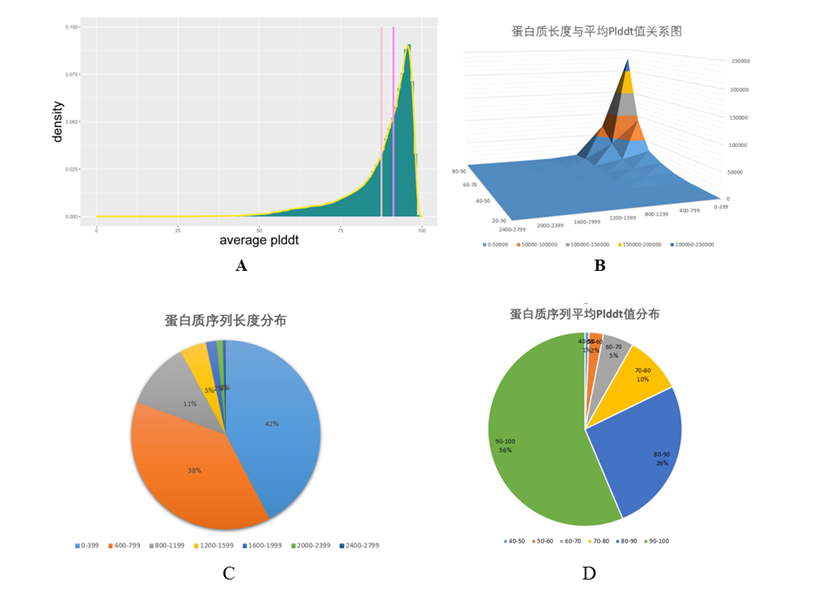
## 2.2 数据清洗与筛选

### 2. 2. 1 数据分布与统计

蛋白质的PDB文件是描述蛋白质三维结构的标准文件格式，通常由X射线晶体学、NMR等实验技术获得。每个PDB文件都包含了蛋白质分子的原子坐标、拓扑信息以及其他结构相关的注释信息等。而在本文中所使用的PDB文件则主要是由AlphaFold模型预测所得，因此对于所使用的蛋白质文件存在着预测精度的问题。在本文中是通过一个局部距离差检验，即Plddt（predicted local distance difference test）值来作为每个氨基酸位点预测精度的置信度。该值反映了对每个残基预测的三维空间结构准确性的平均置信度，其分数范围为 0 到 100，值越高表示对预测结构准确性的置信度越大。Plddt 分数高于 90 通常表示预测结构具有高精度，而分数低于 50 表示预测质量较低。同时，该筛选指标已经保存在数据库所存PDB文件的b-factor字段中，以供后续筛选。如图2-2所示，通过对蛋白质的氨基酸序列各位点plddt值所处不同范围进行染色后（红色由红色变为蓝色代表置信度由高到低），可以直观反映出蛋白质各部分预测精度的高低程度。

**图2-1 蛋白质PDB文件内容**

**图2-2 蛋白质氨基酸序列按照Plddt值进行染色**

为了了解数据库中蛋白质的特性，对它们的序列长度与其平均Plddt值进行统计并对其依赖性进行分析。如图2-3所示

**图2-3 从上到下从左到右依次为 ——**

**（A）所使用蛋白质平均Plddt值分布图及其均值、中位数（B）所使用蛋白质长度及其平均Plddt依赖图（C）蛋白质序列长度分布图（D）蛋白质序列平均Plddt分布图**

通过以上图2-3（C）可知所使用的蛋白质数据的蛋白质长度大量分布在0~800bp的区间，蛋白质序列长度分布从0~3000bp均包含。根据前文所述，图2-3（A）（B）反映了所选取数据的精度较高（这些蛋白质的plddt的平均值为91.23，中位数为87.55，甚至超过一半以上的蛋白质为平均plddt大于90的优质序列），并且对于各个长度区间的蛋白质其置信度分布是近似的，但总体上呈现出较短蛋白质具有较高预测精度的趋势。因此所使用的蛋白质数据的多样性以及精度都有一定程度的保证。当然这并不意味这所使用的数据不需要进行处理，因为即使在平均plddt值较高的蛋白质中仍然可能存在着plddt值较低的区域，当这部分数据被深度学习模型所使用到时，可能造成误差的增大，因此仍需进行剪切等处理。

### 2. 2. 2 蛋白质抽象类

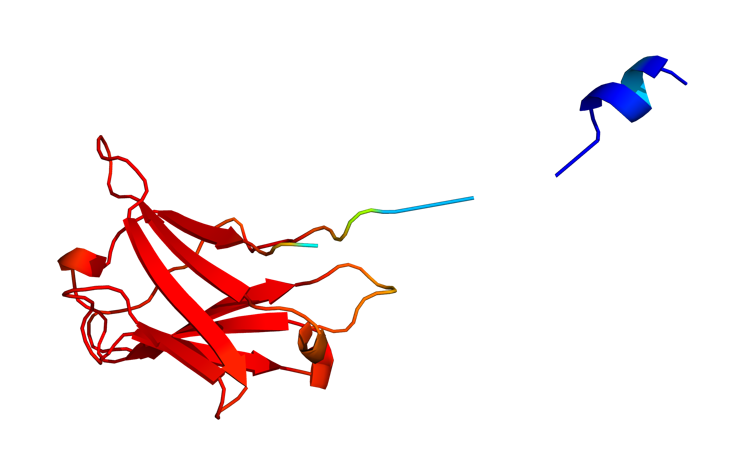
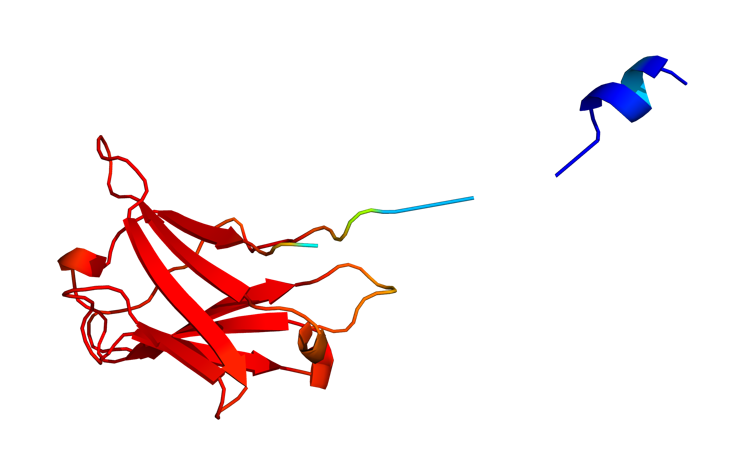
对蛋白质进行处理之前，首先应该用所学数据结构的知识将蛋白质分子抽象为数据结构的以便于后续的处理，具体地说，通过定义不同等级不同作用的类来表示不同的蛋白质分子结构信息，包括蛋白质序列，蛋白质二级结构，氨基酸的坐标位置，以及分子中的元素等信息。

首先最主要的需要先定义名为Protein的蛋白质父类，用于表示蛋白质分子的序列和二级结构信息。其中，序列信息以字符串的形式存储，二级结构信息以列表的形式存储，列表中的每个元素表示一个氨基酸残基，包括该残基的类型和所处的二级结构类型。并且该类需要包含其他的子类信息或是IO接口，以便于后续的信息的整合与处理。

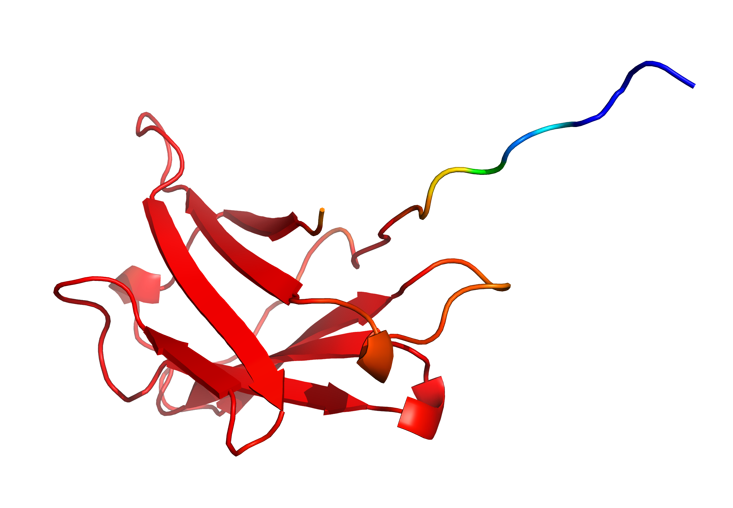
其次在本研究中定义了名为Molecule的分子类，用于表示蛋白质分子的三维坐标信息。该类包含多个属性，例如原子的坐标信息，原子的类型和序号等。此外，该类还包含一些方法，如计算两个原子之间的距离的函数等。而考虑到之后对于蛋白质文件的读写等操作的隔离需要定义了名为MoleculeIO的分子输入输出类，用于读取和写入蛋白质分子的结构信息。该类使用Python标准库中的pdb模块来解析PDB文件，并将其转换为MoleculeIO类的实例对象。当然最基础的同时也是这些类构建的根本，还需要用于表示分子中的元素信息的Dict的字典类。该类包含元素的名称，原子量，电子数量等信息，并且可以根据元素的名称获取相应的信息。

综上所述，通过这种数据结构的类的形式不仅将蛋白质分子的不同结构信息抽象为数据结构，而且提供了读取和写入分子结构信息的工具类。这样就将蛋白质转换为了可以操作的数据并且通过工具能分方便的操作这些蛋白质。

### 2. 2. 3 蛋白质剪切

显然，提高蛋白质数据准确性的最直观方法就是将2.2.1所述的蛋白质序列中预测置信度较低也就是plddt值较低的氨基酸位点进行剪切即可。而剪切的方法，一种直观的方法是使用上述2.2.2所述的蛋白质输入输出接口先输入蛋白质，之后将其实例化并针对其中的每个氨基酸位点进行筛选，筛选标准则是该位点的plddt值是否大于70，与此同时针对于整个蛋白质序列的所有氨基酸位点计算平均的plddt值，筛选标准则是该值是否大于90。之所以这么做是为了在保持整体蛋白质准确性较高的前提下，提高对于单个氨基酸位点的较低置信度包容程度，避免筛选过程将的蛋白质序列剪切的过于碎片化，在剪切完成后仍然使用接口类生成剪切后的蛋白质PDB文件。该方法可以简单理解为“局部位点剪切”+“总体蛋白筛选”。

**图2-4 对氨基酸单位点剪切后的蛋白质**

该方法生成的剪切后的蛋白质如图2-4所示，可以发现虽然蛋白质序列被成功进行了剪切，但整个蛋白质过于碎片化。一方面这种剪切方法过于直接忽略了相邻氨基酸位点对于其他位点的影响，另一方面这样的碎片化序列不利于深度学习模型进行特征提取与学习。因此联想到计算机网络中的滑动窗口模型，在原有的总体筛选指标不变的前提条件下，对剪切方法进行改进，将直接剪切的策略调整为对于计算单个氨基酸位点及其相邻的10个氨基酸位点的平均值（对于蛋白质头部与尾部的氨基酸仍采用之前的策略）后判断其是否大于70，并且随着窗口的移动，窗口内的氨基酸会不断发生变化直到完成全部剪切过程。实际上，改进后的方法能够将相邻氨基酸的预测置信度考虑在内来判断某一段而不仅仅是某一位点的plddt值，通过窗口增加了剪切算法的视野，同时能够让剪切后得到的蛋白质序列更加平滑而不是碎片化。

**图2-5 使用滑动窗口方法剪切后的蛋白质**

如图2-5所示，使用滑动窗口改进后的剪切算法得到的蛋白质序列主要是将头部与尾部的低置信度区域进行了剪切，而对中间的一些低置信度但局部平均置信度达到筛选标准的位点进行了保留，而这也使得整体的结构更加的平滑。通过该方法可将2.1.1所述的54万条蛋白质序列剪切并且筛选出36万条蛋白质序列作为后续深度学习模型的数据集。

### 2. 2. 4 分层采样

构建深度学习模型的数据集是直接影响模型训练性能的关键步骤，它直接决定了模型的训练效果。在本文中，我们需要将来自alphafold数据库和CATH数据集中的蛋白质进行合并，同时采用分层采样的方法来构建测试集和验证集。

首先对于alphafold数据库中的数据使用2.2.3中的筛选剪切方法进行剪切，并将CATH数据集筛选出同源性低于40%的序列。之后，我们需要将两个数据集进行合并。由于这两个数据集中的蛋白质序列是已知的，我们可以通过比对它们的序列来进行合并。具体来说，我们需要对每个蛋白质的序列进行比对，如果它们的相似性超过了一定的阈值，则将它们视为同一条蛋白质，并仅保留其中的一条，将其加入到最终的数据集中。最后得到的数据集中，两个数据集之间的差异性极大，几乎没有相似性较高的蛋白质序列。

接下来，我们需要使用分层采样的方法来构建测试集和验证集。分层采样是一种采样方法，它可以确保在采样过程中各个类别的样本数量比例与原始数据集中相同。在本文中，我们将数据集按照蛋白质所处数据库的种类进行划分，并按照一定的比例（7：1.5：1.5）从每个类别中随机选取一定数量的样本作为测试集和验证集。这样可以确保测试集和验证集中包含了不同类别的样本，从而使得模型在测试和验证阶段的效果更加准确。

分层采样在该过程中的作用是确保测试集和验证集中的样本在各个类别之间的比例与原始数据集中相同，从而避免因为样本数量的不均衡导致模型训练效果不稳定的问题，最终达到提升模型效果的目的。

## 2.3 本章小结

本章首先介绍了所使用数据的数据格式及其所包含的信息内容。之后对于所使用的数据库介绍了其中数据的数据特征和一些统计结果。在此基础上对于其中alphafold数据库中的蛋白质序列数据设计了单位点筛选剪切算法以及改进后的滑动窗口剪切筛选算法并进行了展示对比。而对CATH数据集则是采用同源性低于40%的筛选标准。最后将而这进行了合并，对合并后的数据集进行分层采样以此来构建深度学习模型中所使用的训练集、测试集。

# 第三章 模型的构建与改进

## 3.1 数据格式转换

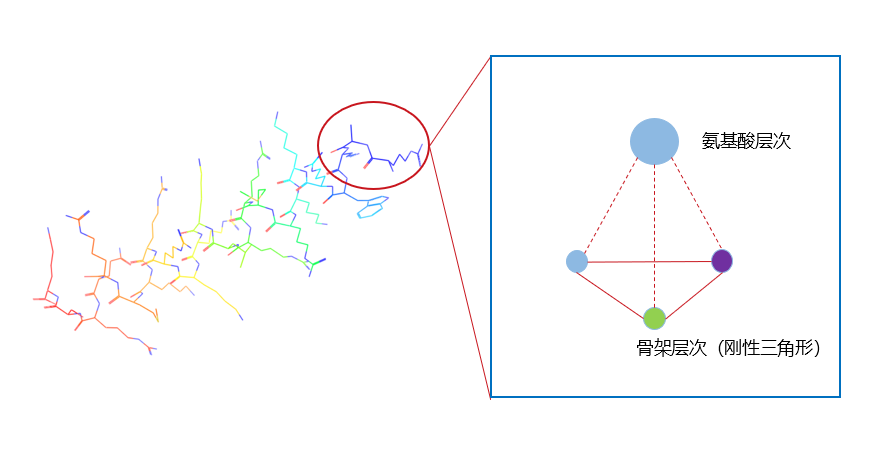
虽然在2.2.2节已经描述过所构建的工具类与方法，但对于深度学习模型而言其在训练过程中调用该方法来提取特征耗费时间过长，因此还需要先使用该方法将蛋白质所需特征数据先进行提取并保存为模型易于使用的格式，本文所选取的存储格式为HDF5文件。

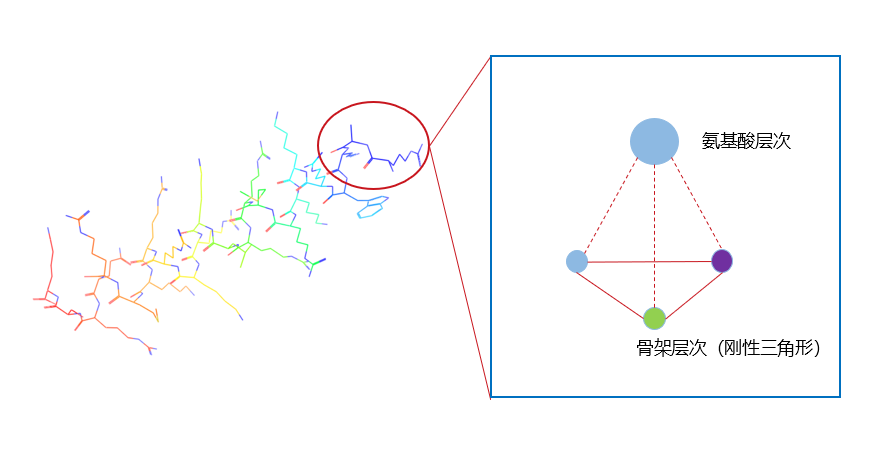
HDF5 是一种用于存储和管理大型科学数据集的文件格式。 它的主要优点是可以存储多种数据类型，包括多维数组、表格、图像和文本，并支持高效的读写操作。 因此，将蛋白质特征保存为HDF5格式，可以大大提高数据读写速度，也可以方便使用现有的PyTorch、TensorFlow等深度学习框架对数据进行处理和训练。 提取蛋白质特征并保存为HDF5格式的优点包括：（1）大大减少了数据集的存储空间； 蛋白质数据集通常非常大，甚至有时因数据集过大无法训练，因此使用 HDF5 格式可以将数据存储在单个文件中，而无需使用多个文件或将它们存储在内存中。 此外，HDF5格式具有良好的压缩性能，可以将数据集压缩到一个比较小的文件大小。 (2) 高效读取和处理大型数据集； 使用 HDF5 格式可以只在需要时读取部分数据集，而无需将整个数据集加载到内存中。 这使得处理大型数据集更加高效并减少了内存使用。 (3) 方便的数据访问和共享； 由于 HDF5 格式是一种通用格式，因此能够轻松共享数据集并使用不同的编程语言访问它们。

此外，针对于深度学习模型而言，HDF5格式还有一个重要的优势，即可以通过读取器（data loader）实现数据的并行读取和预处理。在训练深度学习模型时，需要加载大量数据，这可能会成为训练的瓶颈。使用HDF5格式和并行读取器可以加快数据加载和预处理的速度，从而提高训练效率。

在本文中该过程所使用的具体方法是使用蛋白质类下的IO接口子模块提取蛋白质特征，并将所有信息保存到HDF5文件中。首先，脚本从输入的PDB文件中提取蛋白质结构信息和序列信息，并将它们传递给蛋白质类。该类分别提取每个原子和氨基酸残基的特征向量。这些特征向量可以包含原子或残基的类型、位置、电荷等信息，以及其他一些自定义的特征。其将这些特征保存到一个字典中，并返回该字典。最后返回的蛋白质特征字典被转换为HDF5数据集，并保存到一个HDF5文件中。这个HDF5文件可以包含多个数据集，每个数据集包含蛋白质的不同特征信息（如氨基酸序列、原子位置、二级结构等）。在训练深度学习模型时，可以根据需要从这些数据集中读取数据，并将其用作模型的输入。

## 3.2 蛋白质表示

在本文中将蛋白质骨架视为层次化的空间结构。使用将蛋白质骨架的信息整合并嵌入到氨基酸这一层次的方式，以此来简化深度学习模型在训练时的蛋白质图中存在大量氨基酸内部子图的问题。具体而言，在本文中使用氨基酸中的Cα原子的位置来作为该氨基酸在空间中的三维坐标。而对于该Cα原子所处的氨基酸，将其视为刚性三角形来简化结构（使用Cα原子，氨基中的N原子以及羟基中的C原子作为该刚性三角形的顶点）。

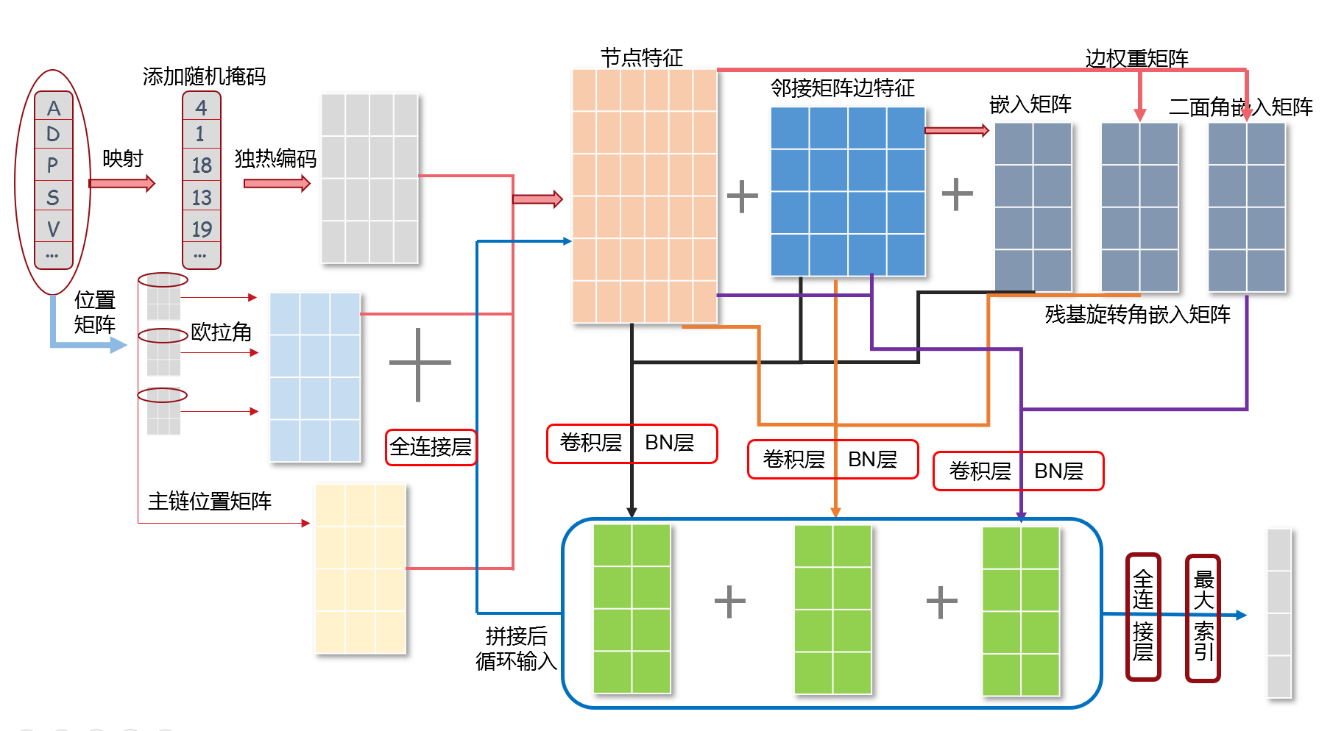


**图3-1 层次化的蛋白质空间结构**

如图3-1所示，在蛋白质的氨基酸序列基础上对于内部的结构使用嵌入的方式将蛋白质骨架这一层次的信息提供给模型。在分子结构中，每个层次具有不同的特点和作用，例如氨基酸序列反映了蛋白质的生物学特性，骨架结构则影响了蛋白质的构象。而当前的大部分蛋白质表示学习方法对于蛋白质内部的层次关系通常会采取忽略或是使用大量计算资源进行结构上的计算。因此本文实际上是采取了折中的方案，使用这种层次化的蛋白质表示方法更精细地捕捉蛋白质的特征的同时节省不必要的计算资源。对于蛋白质而言，通过这种结构能将其转化为一张规则的图，图中的每个节点表示氨基酸且每个结点的特征长度相同，这样也便于模型的训练。

## 3.3 模型的架构

### 3.3.1 初始模型架构

因为模型的输入数据蛋白质被转化为图的结构（3.2节），因此模型的主体采用图卷积神经网络的架构，模型的整体结构如图3-2所示。

**图3-2 所使用循环图卷积神经网络整体架构**

该模型以氨基酸作为基本节点，所使用的骨架信息会使用embbding的方法降维后嵌入到节点的信息中进行训练。蛋白质序列在该模型中的处理过程如下：

蛋白质序列信息作为最关键的信息在映射为离散值后需要先添加随机掩码，使用随机掩码的目的是因为我们期望最后训练的模型在对给定蛋白质结构生成近似空间结构的蛋白质时不使用到序列信息而仅使用其空间信息，因此在训练过程中模型需要添加一定比例的随机掩码将序列信息进行隐藏。不同于其他方法中添加固定比例的随机掩码的方法，本文采用的方法是使用softplus函数让随机掩码的比例随着训练轮次的增加期望提升，让训练数据逐渐接近测试数据不提供序列信息的情况。之后该序列信息经过独热编码从离散值转换为一个对应的连续型特征向量。因为经过独热编码之后每个取值都被转换成了一个独立的维度，而且这些维度之间没有大小关系，能够更好地捕捉特征之间的关系以此提升模型性能。

此外，通过蛋白质骨架（3.2节所述Cα原子，氨基中的N原子以及羟基中的C原子构成）的空间坐标矩阵（将每个氨基酸位点表征为3\*3的矩阵）后可以得到按照蛋白质氨基酸顺序的任意相邻两个氨基酸之间的三个平面欧拉角，虽然这种方法没有对于整体任意两氨基酸计算欧拉角，但对于可以旋转的主链序列方向而言并没有差别且能够降低特征维度。通过该方式将多个位置矩阵进行嵌入降维后与主链Cα原子坐标进行拼接之后，与上述经过独热编码的节点特征进行拼接，得到最终所使用到的节点特征。

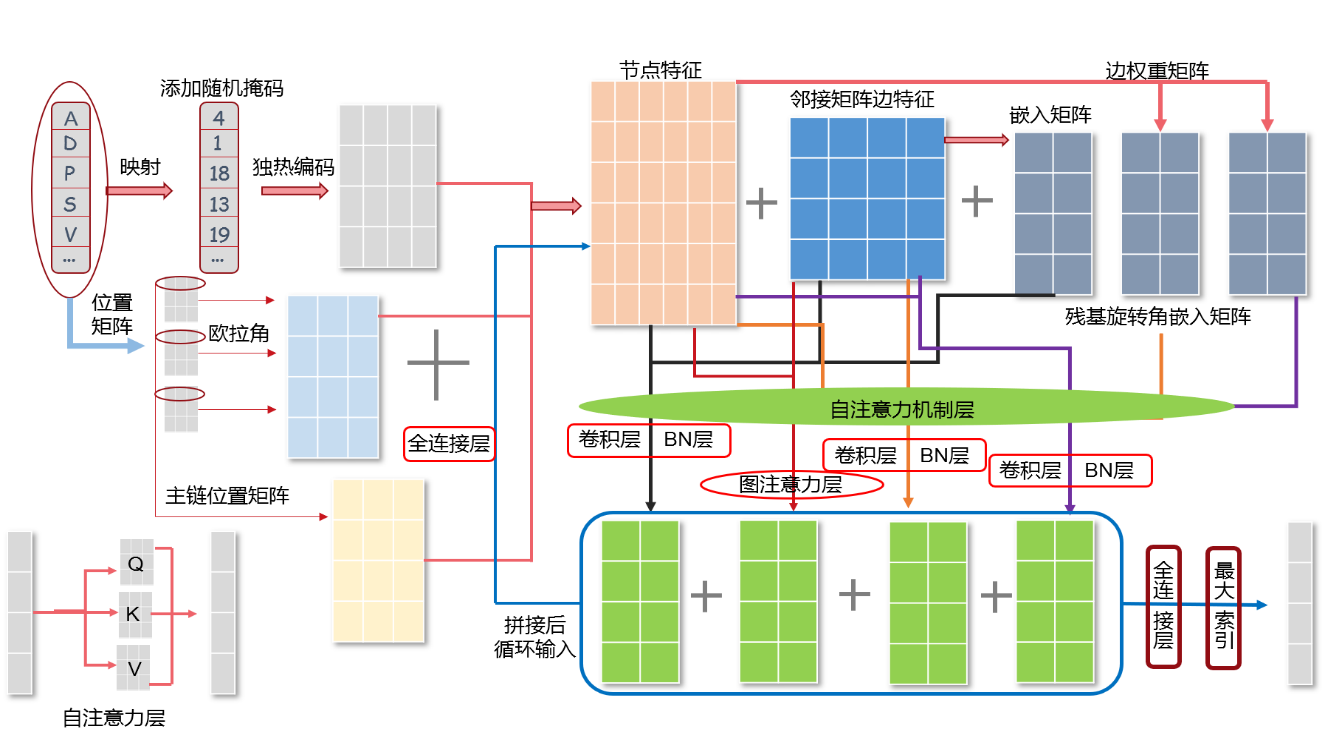
当然作为最终卷积神经网络的输入还需要氨基酸节点两两相互作用的邻接矩阵。在本文中使用两氨基酸位点在空间上相距是否超过10Å作为两氨基酸之间是否具有相互作用力的标准（一旦超过这个距离氨基酸分子之间的相互作用，包括静电相互作用、氢键相互作用、范德华力相互作用等都会变得十分微弱）。当达到这个标准时，在邻接矩阵相应位置上填入1，否则填入0以此得到该邻接矩阵。对于一般的图卷积神经网络而言，权重矩阵是用来进行节点特征的传播的。通常情况下，节点的特征是一个向量，而图卷积神经网络的目的是学习节点特征的表征，这个表征能够捕捉到节点在图中的结构和拓扑关系。具体来说，权重矩阵将节点的特征进行线性变换，然后对邻居节点的特征进行加权求和，得到新的节点表征。而在本文中对于权重矩阵采用的是使用一些先验信息来生成初始的矩阵再交给模型进行学习和进一步优化。当然使用不同方法生成的初始权重矩阵不同对应最后的模型参数也不相同，因此本文使用三种平行的方法生成该矩阵后统一为相同的张量形状后生成平行的结果。这三种方法分别是：将邻接矩阵嵌入降维为权重矩阵；使用氨基酸残基的三个旋转角（φ，ψ，ω）生成权重矩阵；使用氨基酸两两之间的二面角矩阵嵌入降维为权重矩阵。

在获得上述图卷积神经网络所需所有的矩阵之后对其进行卷积操作，卷积操作的作用是从邻居节点中提取特征，并将这些特征聚合为当前节点的表示，而对于本研究卷积操作的意义是让模型能够学习空间上相邻氨基酸的信息对于空间结构有更准确的表征。在进行卷积之后还需要使用BatchNorm技术来正则化（对网络中每一层的输入进行归一化处理，以使其均值为0、方差为1。这样可以使得网络的输出更加稳定，从而加速训练过程），并以此来缓解梯度消失和爆炸的问题，并提高模型的收敛速度和泛化能力。

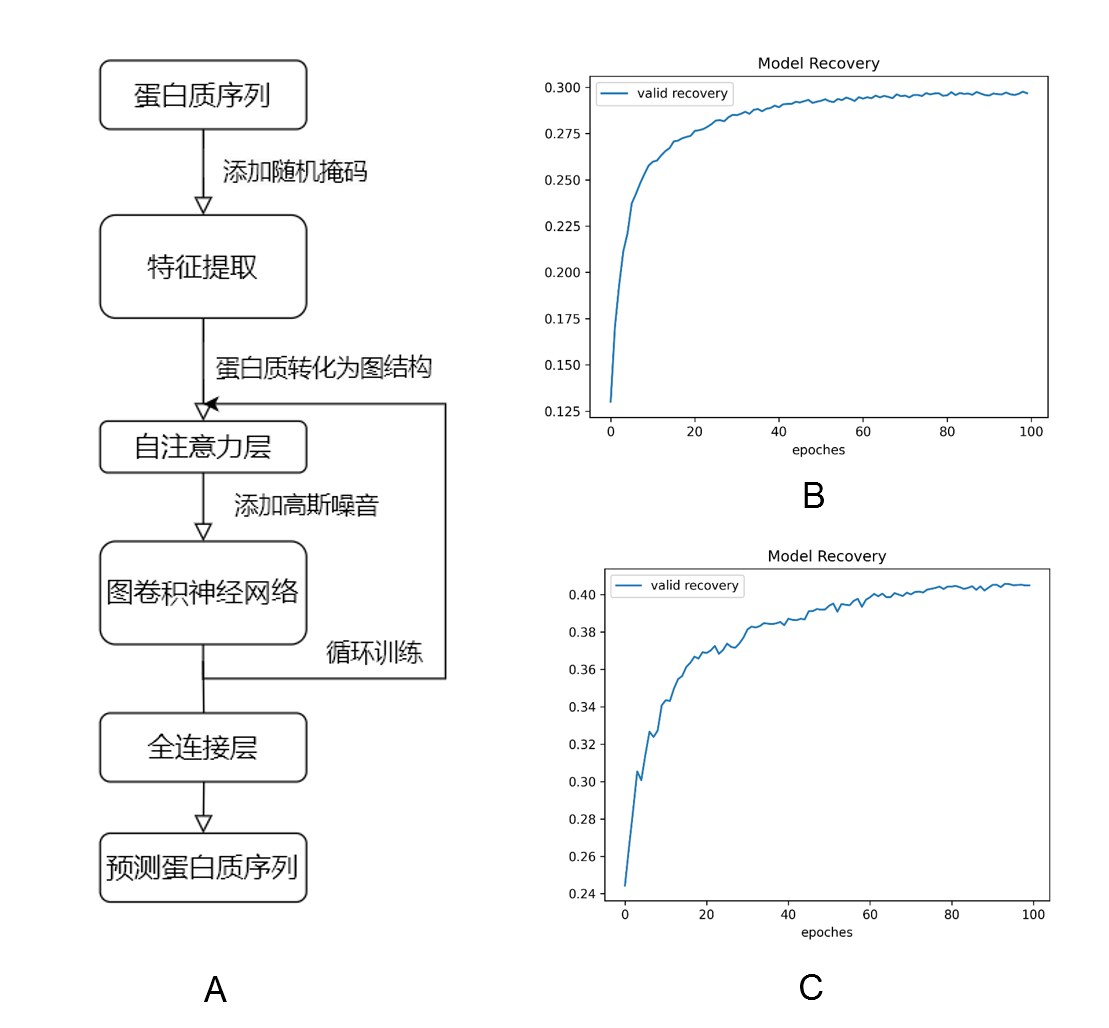
在使用三组不同的权重矩阵生成相同形状的结果之后将其拼接经过全连接层得到与节点特征形状相同的矩阵进行循环训练。该方法实际上是循环神经网络的方法，将输出结果作为输入特征进行循环训练，循环神经网络的优点在于，它能够处理变长的序列数据，并且可以利用历史信息对当前时刻的预测进行影响。同时，由于循环神经网络中存在循环连接，因此能够解决梯度消失和梯度爆炸的问题，从而有效地训练模型。对于本研究中的蛋白质序列的预测任务，时序信息可以帮助模型更好地捕捉序列中的局部特征和全局模式，从而提高预测的准确性。

在经过多次训练并经过全连接层之后，模型会生成一个与初始独热编码后得到的特征矩阵形状相同的结果矩阵，但该矩阵每个节点的信息为当前蛋白质的空间结构对应每个氨基酸位点的概率大小，因此对每个位点取概率最大值的索引根据字典进行反映射后得到模型推断出的预测蛋白质的氨基酸序列。

### 3.3.2 优化后的模型架构

虽然经过上述模型能够达到进行预测序列的目的，但对于该网络对于不同节点之间的重要程度没有区分，而是使用生成不同的初始化权重矩阵输入给图卷积神经网络，因此为了更好地捕捉节点之间的关系并提升模型的可解释性，在图卷积神经网络的三个平行子模块之外再加入图自注意力子模块来优化模型。除此之外，对于本研究中的蛋白质特征图可以使用自注意力机制来为输入蛋白质序列数据的不同氨基酸节点赋予不同权重，使得模型对蛋白质骨架中的关键部分更加关注，同时也减少计算量，提高模型的效率。改进后的模型架构如图3-3所示。

**图3-3 优化后的模型整体架构**

优化后的模型的模型训练过程如图3-4（A）所示流程图。同时为了证明优化思路的合理性，本文对于优化前后的训练过程进行了可视化，如图3-4（B）所示，优化前的模型随着模型训练轮次的增加，在验证集上的回复率这一指标（该值指标可以理解为越大越好）逐渐收敛到0.3。而在相同的轮次内，使用添加图自注意力子模块并使用自注意力机制优化后的模型在验证集上的回复率这一指标将会收敛到0.4。而这也是对模型性能提升最直观的展示。

**图3-4 从左到右、从上到下依次为——**

**（A）模型训练流程图（B）初始模型训练过程可视化（C）优化后模型训练过程可视化**

## 3.4 本章小结

本章首先简单介绍了如何使用前述蛋白质类将数据转化为深度学习模型所方便使用的格式，并且描述了使用层次模型将蛋白质骨架信息嵌入到氨基酸节点中并将其转化为图结构。最后，本章重点介绍了所使用的深度学习网络架构（主体为图卷积循环神经网络）。说明了如何使用自注意力机制和图自注意网络子模块对其进行优化，使模型性能以及可解释性得到提升，并通过优化前后训练过程的可视化证明该点。

# 第四章 结果与分析

## 4.1 模型初步评估指标

在蛋白质序列预测中，恢复率（recovery）、一致性（identity）和多样性（diversity）是较为常用的评估指标，这三个指标通常被用来评估蛋白质序列预测的质量和准确性。接下来详细介绍这些指标：

恢复率指标评估所预测的蛋白质序列与目标序列之间的相似程度。在本研究中，恢复率是由预测序列和目标序列之间的重叠长度（相同位点氨基酸类别相同）除以目标序列的长度得到的，表示预测序列中有多少部分与目标序列相同。恢复率指标的值越高，表示预测序列和目标序列越相似，预测的准确性越高。当然这一指标本身具有一定的局限性，因为在本研究中我们最终的预期是生成与原始蛋白质在空间结构上近似的蛋白质序列，而不仅仅是序列本身的相似性。但是对于模型初步的评估而言，使用恢复率这一指标能够较为快速且具有一定准确性的判别模型的预测能力。

一致性这一指标所衡量的是预测序列之间的相似性。在本研究中，该指标是通过计算模型所生成的预测序列在相同的位置上相同的残基数目除以序列长度来得到，其结果表示预测序列中有多少个氨基酸与其他预测序列完全相同。一致性的值越高，表示预测序列之间越相似，说明模型的序列设计能力不足。

多样性这一指标评估了预测序列和其他预测序列之间的多样性。在蛋白质序列预测中，如果预测的序列之间的差异很小，那么预测结果的多样性就很低。多样性指标可以通过计算预测序列之间的差异性得到。在本文中计算通过模型设计的预测序列之间的等效位置减去相同或相似氨基酸的比例得到该值。多样性的值越高，表示预测序列之间的差异越大，说明模型在对于不同序列的探索能力更强，对于所给定的的空间结构能给出更多合理的预测序列。

在研究中，使用两个独立的测试集来对于上述三个指标进行评估：分别是包含103个蛋白质的单链蛋白测试集以及包含14个蛋白质的denovo蛋白测试集（且这117个蛋白质均未出现在模型所使用的训练数据中）。通过不同模型在这两个测试集上的表现来说明所使用模型的性能。

**表4-1 在单链蛋白上不同模型上的表现**

|  |  |  |  |
| --- | --- | --- | --- |
| 模型名称 | 序列恢复率 | 序列一致性 | 序列多样性 |
| ProteinSolver | 0.191 | 0.214 | 0.526 |
| Structure Transformer | 0.195 | 0.742 | 0.138 |
| ESM-IF1 | 0.322 | 0.566 | 0.293 |
| ProteinMPNN | 0.260 | 0.634 | 0.237 |
| GPD | 0.279 | 0.608 | 0.280 |
| Original Model | 0.234 | 0.633 | 0.279 |
| Improved Model | 0.249 | 0.614 | 0.297 |

**表4-2 在denovo蛋白上不同模型上的表现**

|  |  |  |  |
| --- | --- | --- | --- |
| 模型名称 | 序列恢复率 | 序列一致性 | 序列多样性 |
| ProteinSolver | 0.246 | 0.222 | 0.498 |
| Structure Transformer | 0.441 | 0.813 | 0.074 |
| ESM-IF1 | 0.477 | 0.664 | 0.184 |
| ProteinMPNN | 0.487 | 0.688 | 0.168 |
| GPD | 0.462 | 0.658 | 0.219 |
| Original Model | 0.311 | 0.670 | 0.248 |
| Improved Model | 0.408 | 0.654 | 0.280 |

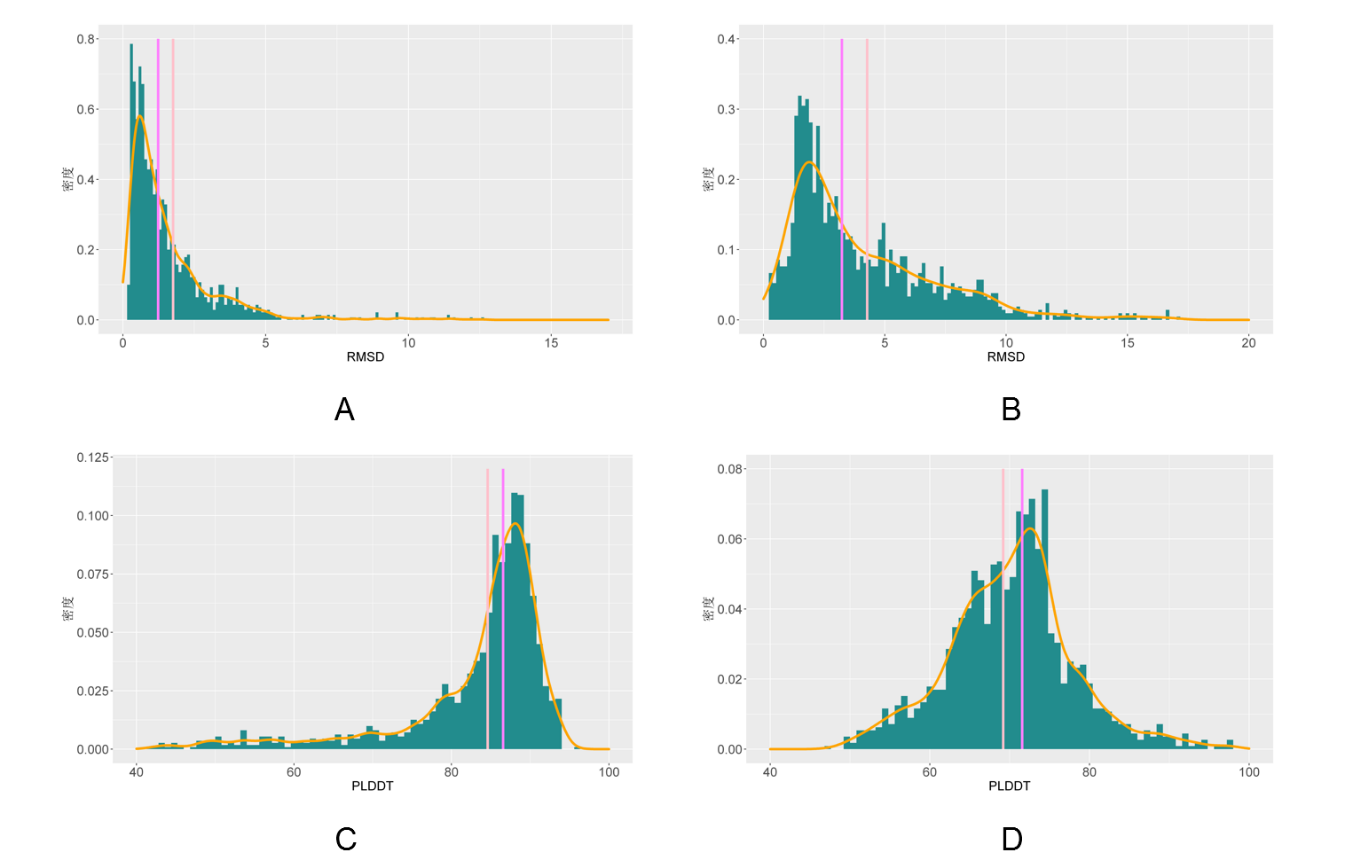
具体的实施方法是对于denovo蛋白和单链蛋白分别对每一条蛋白质生成对应的100条氨基酸预测序列，并计算上述三个指标。如上表4-1、表4-2所示，本研究中所使用的模型在序列恢复率这一指标上与目前在该数据集上表现最优异的模型仍具有差距。但是在不考虑本身序列恢复率最低的ProteinSolver模型的基础上，本文所使用的模型特别是优化之后的模型在多样性这一指标上具有领先的优势。且无论是优化前还是优化后的模型，都能在保持序列恢复率较高的前提下让多样性保持在较高水准。这也说明模型能够在保证所生成的蛋白质较为准确的条件下更多的拓宽序列搜索空间生成更多样化的蛋白质序列。相比于其他模型，我们的模型所生成的序列的恢复率以及多样性之间更加均衡。且如前文所述，序列恢复率并不是本研究所最终追求的目标，而只是初步的指标，还需要使用序列预测工具对于生成的氨基酸序列进行折叠使用折叠后的蛋白质的空间构想与原始构象进行比较来进一步评估模型。

## 4.2 使用ESMFOLD进行空间结构预测

ESMfold是一种新型的蛋白质结构预测方法，它是由Meta公司的研究人员开发的。与传统的基于物理学的模拟方法不同，ESMfold是一种基于神经网络的方法，它利用了大规模的蛋白质序列和结构数据来训练神经网络，从而预测蛋白质的三维结构。ESMfold不需要先验知识或者结构域库，可以在几分钟内预测出蛋白质结构。而在本研究中，如3.2节所述，所生成的结果是蛋白质的氨基酸序列也就是fasta文件，而不是结构pdb文件。因此要达到原始序列与设计序列在空间结构上进行对比的目的，还需要首先使用ESMfold对氨基酸序列进行折叠生成结构性文件。

当然并不是说使用ESMfold所得到的蛋白质结构一定是准确的，其生成的PDB文件中仍然可能存在一些结构不完整或者不准确的区域。因此需要对其生成的结构本身的准确性进行统计，所统计的数据采用2.2.1节所提到的PLDDT这一指标。因为通过我们的模型所生成的序列本身的质量会影响到ESMfold得到的结构信息。反之通过统计生成蛋白质pdb文件的plddt值分布能顾推断出我们模型本身的序列生成能力。除了该值之外，还需要计算使用ESMfold得到的蛋白质与所给的原始结构的平均距离平方根（RMSD，Root Mean Square Deviation）值的大小。该值是一种衡量两个结构在三维空间中结构相似性的常用指标。对于两个蛋白质分子，RMSD值越小表示它们在空间中的结构越相似。在将两个蛋白质分子进行分子叠合后，RMSD值反映的是它们的空间构象的差异程度。该值越小，说明二者之间的空间构象越相似，很可能具有相似的功能或生物学作用。

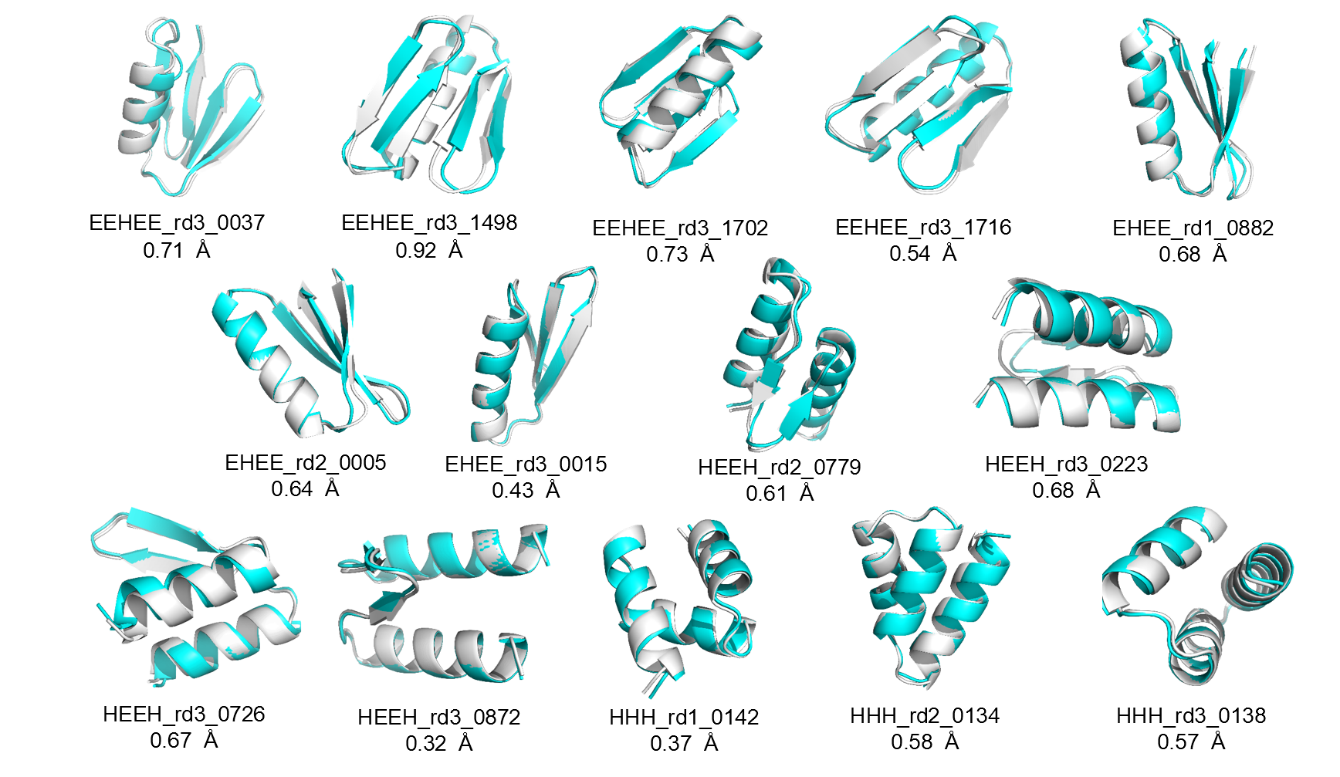
对4.1节所述生成的序列统计其PLDDT值以及生成构象与原始构象间的RMSD值，结果如下图4-1所示。在denovo蛋白上RMSD值大部分都在0-5Å的合理范围内，且均值为1.82Å，中位数为1.36Å，这也意味着生成序列中着大量RMSD值小于2 Å可以被认为是与原始构象非常接近的优质序列，同时其序列的plddt均值为84.6，中位数为86.5说明输入给ESMfold的序列质量较高。而对于单链蛋白而言，结果并不理想，在单链蛋白上RMSD值均值为4.08 Å，中位数为3.04 Å，只有一部分是RMSD值小于2 Å的优质序列。同时，其序列的plddt均值为69.4，中位数67.8也能反映所生成序列的质量不高。在所使用单链蛋白比denovo蛋白结构更复杂序列更长的条件下，这样的结果是合理的，但也说明了模型仍具有优化空间，需要进一步提高泛化能力。

**图4-1从上到下、从左到右依次为——**

**（A）denovo蛋白上RMSD密度分布（B）单链蛋白上RMSD密度分布**

**（C）denovo蛋白上PLDDT值分布 （D）单链蛋白上PLDDT值分布**

## 4.3 分子叠合计算RMSD与结果展示

为了更加直观的展示最终的结果，对于生成序列并用ESMfold得到的蛋白质蛋白质结构文件使用Pymol软件与原始蛋白质进行分子叠合。Pymol是一款常用的分子可视化软件，可以用于分子结构的展示、编辑和分析等。其中，分子叠合并计算RMSD是Pymol中一个常用的功能。将设计出的蛋白质与原始构象蛋白质分别导入后使用该软件的align功能即可实现分子叠合。在此我们挑选了一些分子叠合之后计算出的RMSD值小于1Å的可以认为与所给蛋白质空间结构非常接近的优质序列并进行展示。结果如下图4-2所示，其中浅蓝色为所给的天然结构，而灰色代表我们所使用模型预测出来的结构。

**图4-2 分子叠合后的优质结果展示**

## 4.4 本章小结

本章介绍了对于所生成序列的一些评估指标，例如恢复率、一致性、多样性，并在单链蛋白以及denovo蛋白两个独立测试集上使用这些指标对于本研究所使用模型与其他模型进行了评估。之后为了进一步评估生成序列是否具有与原始序列在空间结构上相似的性质，采用了ESMfold对生成序列进行了折叠得到蛋白质结构文件，并使用该文件与原始的结构文件进行分子叠合计算RMSD值以此来直观反映结构相似性，并统计了该RMSD值的分布以及序列的PLDD值分布。最后，挑选了一些与原始结构RMSD值小于1Å的优质序列与原始蛋白质进行了分子叠合来进行结果的展示。

# 第五章 全文总结

## 5.1 主要结论

本研究

## 5.2 研究展望

更深入的研究……

# 参 考 文 献

[1] 杨瑞林, 李力军. 新型低合金高强韧性耐磨钢的研究[J]. 钢铁, 1999(7): 41-45.

[2] 于潇, 刘义, 柴跃廷, 等. 互联网药品可信交易环境中主体资质审核备案模式[J]. 清华大学学报(自然科学版), 2012, 52(11): 1518-1523.

[3] SCHINSTOCK D E, CUTTINO J F. Real time kinematic solutions of a non-contacting, three dimensional metrology frame[J]. Precision Engineering, 2000, 24(1): 70-76.

[4] 温诗铸. 摩擦学原理[M]. 北京: 清华大学出版社, 1990: 296-300.

[5] 蒋有绪, 郭泉水, 马娟, 等. 中国森林群落分类及其群落学特征[M]. 北京: 科学出版社, 1998: 5-17.

[6] 贾名字. 工程硕士论文撰写规范[D]. 上海: 上海交通大学, 2000: 177-178..

[7] 张凯军. 轨道火车及高速轨道火车紧急安全制动辅助装置: 201220158825.2[P]. 2012-04-05.

[8] 全国信息与文献标准化技术委员会. 文献著录: 第4部分 非书资料: GB/T 3792.4-2009[S]. 北京: 中国标准出版社, 2010: 3.

（参考文献格式请参考GB/T 7714-2015《信息与文献 参考文献著录规则》）

# 致 谢

大学的四年时间飞快流逝，一转眼就到了面临毕业的日子，之前在交大的一幕幕场景仿佛就发生在昨日。在最后的这忙于毕业设计和毕业论文的大半年时间里，我曾经遇到了许多的困难与挫折，但是在各位老师和师兄师姐的帮助下我一一克服了这些困难顺利完成了我的论文，这也使得最后在交大的这段时间显得更加珍贵更加具有意义。

对于一年前的我而言，迷茫与困惑仿佛写满了脸庞，面对着考研前途未卜的巨大压力以及毕业设计的同时开展的任务压力，我几乎快被压得喘不过气来。在当时身边的朋友同学大部分都有了自己的去路，或保研或工作或出国，更是让当时决定考研的我感到孤独与疏离。对于当时的我而言，焦虑与失眠几乎成为了常态，更不用说在一年前还有着疫情封楼的状况。我当时对于自己的毕业设计持有一种反感厌烦的消极心态。但是，当我真正经历了一次这样的毕业设计从开始到结束的过程，完成了自己的论文，我认为自己是实实在在的经历了一次蜕变，让自己从自我内耗的泥淖中脱离出来并且能够自信勇敢的面对之后的不管是学习还是生活上的困难。在本次毕业设计中，我收获了很多，特别是在学术上和人际关系上。在此，我要对所有支持和帮助我的人表示最真挚的感谢。

首先，我要感谢我的本科生导师陈海峰教授。在我选择毕业设计课题时，他给了我很大的帮助。在我面临考研和毕业设计的双重压力时，他理解并支持我选择了自己感兴趣的方向，陈老师告诉我，在毕业设计中并不一定要做很多东西，最重要的是学会思考并且体会到科研的过程。陈老师对于我的边完成毕设边考研的历程给予了莫大的理解与包容。这不仅让我更加明确了自己的方向和目标，也让我更加坚定了自己的信心。

同时，我要感谢我的毕业设计指导老师魏婷老师。魏老师对我的毕业设计给予了近乎手把手的指导，在我遇到问题时，总是能够第一时间给予回复并给出指导意见，而魏老师的意见也经常为我指明方向。对于我能够顺利地完成毕业论文这件事，我起初是抱有怀疑态度的因为我对于自己的科研能力没有信心，是魏老师在毕设这一过程中帮我建立了这一信心。除此之外，在开题答辩、中期答辩以及论文撰写过程中，魏老师给出了详细且具有建设性的指导和修改意见，让我认识到了科研的严谨和深度。从我自己的感受来说，我认为魏老师对于我的耐心和细心对于我的毕业设计设计而言有着无法替代的作用。

除此之外，我还要感谢一些帮助过我的同学。我要感谢张博同学，他帮助我理清了我的毕业设计中的蛋白质设计整体思路，让我在这个方向上有了更加清晰的认识和思路。我要感谢司端淼同学，他帮助我使用R语言画出了令人满意的图片，让我的毕业论文更加美观。我还要感谢王钟逸同学和王喜同学，他们帮助我理解深度学习模型，让我对于深度学习的应用和理解更加深入。感谢他们的帮助和支持，让我感受到了友情的温暖和力量。

最后，我还要感谢我的家人。在我背负着考研和毕业设计的压力时，他们一直默默地支持着我，鼓励我，帮助我克服困难。感谢家人对我的支持和陪伴，让我能够坚定地走向我的目标，是他们让我在学业和生活中感受到了无尽的温暖和力量。

在此，我谨向所有在科研和生活中帮助过我给予过我温暖的老师们，同学们，家人们献上我最诚挚的谢意。

**DEEP LEARNING-BASED PROTEIN SEQUENCE GENERATION RESEARCH**

HCCI……***(英文大摘要正文)***

***英文大摘要单独编页码***