北京大学

# 本科生毕业论文

题目：

# 基于图神经网络优化蛋白质骨架的可设计性

# Protein scaffold optimization for protein design based on Graph Neural Network


姓　　名：王宇哲

学　　号：1800011828

院　　系：化学与分子工程学院

专　　业：化学

指导教师：来鲁华 教授 张长胜 副研究员


二〇二二 年 六 月

# 版权声明

# 摘 要

蛋白质计算设计是计算生物学领域的重要课题，而在基于骨架结构的蛋白质计算设计的过程中，蛋白质骨架，即主链结构的优化能够提高该骨架的可设计性，对后续的序列设计具有重要的意义。目前，以Rosetta软件为代表的结构优化方法使用主链与侧链原子相互耦合的能量函数，增加了计算复杂度，因此需要发展新的人工设计蛋白质骨架结构的高效优化方法。

本课题基于等变图神经网络和降噪分数匹配方法，在天然蛋白质结构数据集上训练了能够实现蛋白质主链结构模拟退火朗之万动力学优化的模型，并通过对高斯噪声扰动后蛋白质主链结构的恢复表现对优化模型进行了评估。计算测试表明，经高斯噪声扰动获得的与真实结构均方根偏差约为2.2 Å的主链结构，经优化后结构的均方根偏差平均下降约0.36 Å，显著提升了主链结构的质量，优化成功率达到100%。进一步对于模型进行了改进，增加了蛋白质疏水核心中残基侧链相互作用的信息，使均方根偏差平均下降 达到0.40 Å。还测试了模型在实际蛋白质设计中进行蛋白质主链优化的能力，使用SCUBA程序随机构建了10个初始骨架，使用模型对初始骨架进行优化，使用Rosetta程序进行固定主链的序列设计和精细优化，其中9个骨架优化成功、获得较低的打分，证明本课题开发的模型能够改善初始骨架结构的可设计性。

**关键词**：蛋白质计算设计，蛋白质骨架优化，图神经网络，可设计性

# ABSTRACT

Computational protein design is an important topic in the field of computational biology. With the continuous improvement of algorithms and the improvement of computing power, protein computational design is becoming more and more widely used in practical protein engineering, which greatly improves the efficiency of protein engineering design. In the process of protein computational design, the optimization of the protein scaffold can improve the designability of the initial backbone structure, which is of great significance for subsequent sequence design. Compared with searching the amino acid sequence space and the conformation space of a specific sequence at the same time, searching the main chain conformation space regardless of the specific type of side chain can greatly reduce the computational cost and avoid falling into local minimum in the energy minimization process. However, taking Rosetta software as an example, the existing scaffold optimization method uses the energy function that couples the main chain and side chain atoms, which increases the computational complexity. Therefore, it is necessary to develop a protein main chain structure optimization method that is independent of the side chain.

As an important branch of machine learning, deep learning extracts the hidden pattern features in the data through the connection of multiple layers of neurons, and has a strong learning ability. In recent years, it has achieved great success in a series of fields such as computer vision and natural language processing. Graph neural networks can process data represented in the form of graphs, and learn information related to the structure of graphs by training on datasets composed of a large number of graphs. The structure of chemical molecules can be represented in the form of Molecule Graph, which can save and transfer the information of chemical molecules intuitively and efficiently, so GNN has been widely used in cheminformatics.

Based on the Equivariant Graph Neural Network (EGNN) and the Denoising Score Matching method, we have developed a model that can optimize the structure of the protein main chain independent of the side chain on the natural protein structure dataset. Protein structure data in PDB format is converted to graph, where the Cα atom of each residue of the protein is converted to a node of the graph, and the interaction between a pair of residues in the protein corresponds to an edge. In the training process, Euclidean distances between Cα atoms of connected residues where edges exist are calculated to construct distance vector $\boldsymbol{d}$. A series of Gaussain noise with varying intensities $\{\sigma_i\}_{i=1}^{L}$ are used to perturb the distance vector $\boldsymbol{d}$. The score network $\boldsymbol{s_\theta}$ is trained to fit the gradient field that was defined as $\boldsymbol{d} - \widetilde{\boldsymbol{d}}$. When training on each protein structure data, the intensity of Gaussian noise $\sigma$ is gradually increased from small to large, so as to improve the robustness of the scoring network $\boldsymbol{s_\theta}$. In the protein scaffold optimization process, we adopt annealed Langevin dynamics sampling to generate optimized backbone structure. In the

experiment, we adjust the hyperparameters of the model based on the performance of the model on the validation set and the performance of the model on the actual optimization problem. During the training process, the loss function gradually decreases on both the training set and the validation set. The model was evaluated by its optimization performance on the protein backbone structure after Gaussian noise perturbation. Computational experiments have shown that for the perturbed structure with root mean standard deviation (RMSD) of $2.172 \pm 0.084$ Å from the real structure, the model achieved an average RMSD reduction of $0.363 \pm 0.076$ Å with a success rate of 100%, significantly improving the quality of the main chain structure. To further improve the performance of the basic model, the information based on the interaction of side chains of the residues around the hydrophobic core of the protein is added to the model, leading to a better average RMSD reduction of $0.403 \pm 0.076$ Å with a success rate of 92.2%. Finally, the model is applied to the actual protein backbone optimization problem, and the sequence design of the fixed backbone is carried out using Rosetta program. Ten initial backbone structures are generated randomly using SCUBA program according to a sketch of stacking of secondary structure elements. We apply Rosetta fixbb module to perform the fixed backbone sequence design procedure and Rosetta relax module to further refine the full-atom structure generated in the previous step. The optimized backbone gets a lower score on average, which indicates that the model we developed can improve the designability of the initial backbone structure.

In the model training process, Gaussian noise is used to perturb the three-dimensional coordinates of the structure in the training set, and it is hoped that the model will fit a gradient field with a noise reduction effect. This training process essentially captures the local information of the protein main chain, and the model fails to catch interactions between secondary structures, so large variations and optimizations of the initial protein backbone structure cannot be achieved. The limitations above will be overcome in the follow-up research.

In a nutshell, we have developed a model based on EGNN denoising score matching method to accomplish the optimization of the protein backbone independent of the side chain, so as to improve the designability of the initial protein scaffold in the *de novo* design of the protein, which is helpful for follow-up sequence design procedure. Specifically, the EGNN + denoising score matching model, which has shown excellent performance in several fields such as molecular conformation generation, is applied on the protein backbone structure optimization problem. And the graph construction method is designed according to the characteristics of the protein, embedding the secondary structure information, distances and interactions between residues and other features into the graph. We have carried out a series of experiments to prove that the model is capable of optimization of protein scaffold and improving the designability of the initial backbone structure. The model will be further improved in the future to gain global knowledge

about protein structure information.

**KEY WORDS**: Computational Protein Design, Protein Scaffold Optimization, Graph Neural Network, Designability

# 目 录

# 第一章 引言

## 1.1 蛋白质计算设计概述

蛋白质是生物体内执行催化、免疫、信号转导等大量生命活动的主要生物大分子，特定蛋白质的功能与其空间结构密切关联，而蛋白质的空间结构被其氨基酸序列所确定。在所有可能的氨基酸序列中，自然进化所产生的天然蛋白质的氨基酸序列所占比例很小，而天然蛋白质的结构也是极其有限的，难以满足生物医学等领域对于具有特定功能蛋白质的需求。因此，通过对天然蛋白质已知的氨基酸序列进行改造或从头设计全新的氨基酸序列、从而设计具有全新功能的蛋白质对于合成生物学有着极为重要的意义。

以定向进化（Directed Evolution）为代表的传统蛋白质工程技术通过物理或化学手段诱导基因突变，再对所产生的蛋白质结构进行实验筛选[1]。在不采用高通量筛选手段时，上述基于实验的方法效率很低，且对氨基酸序列的化学空间的采样极不完全。自 20 世纪 80 年代以来，蛋白质计算设计（Computational Protein Design）逐渐发展起来，这种方法根据已有的蛋白质结构信息，通过适当的算法确定蛋白质的氨基酸序列，既可以进行蛋白质的从头设计（*de novo* Design），也可以对已有的蛋白质进行改进，以实现特定的生物功能。随着算法的不断改进和计算能力的提高，蛋白质计算设计被越来越广泛地应用于实际的蛋白质工程中，极大地提升了蛋白质工程设计的效率，例如 Sun 等人运用计算和实验相结合的方法对蛋白-蛋白相互作用界面进行改造的研究表明，通过适当运用蛋白质计算设计的相关方法，可以将湿实验的工作量减小 3~4 个数量级[2]。

下文简要介绍蛋白质计算设计的基本原理。与蛋白质结构预测问题相类似，蛋白质设计基于蛋白质的折叠能量景观（Folding Energy Landscape），具体地，根据 Anfinsen 等[3]的假说，蛋白质折叠成其氨基酸序列所可以达到的最低能量构象；但蛋白质结构预测的目的是寻找给定氨基酸序列的最低能量构象，而蛋白质设计的目的是确定氨基酸序列使得所需的蛋白质结构稳定，也即寻找与所需蛋白质结构相近的低能量构象所对应的氨基酸序列，因此蛋白质设计也被称为逆向的蛋白质折叠问题[4]。

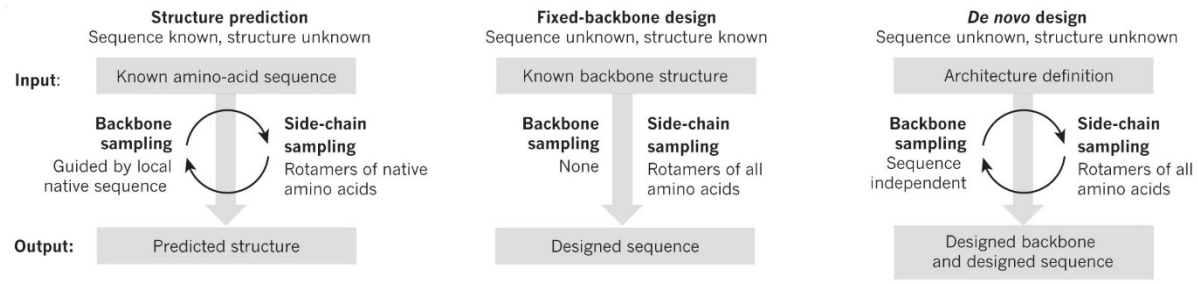


图 1.1 蛋白质结构预测、固定骨架设计和从头设计流程比较[5]

Baker 等[5]对蛋白质结构预测和计算设计的工作流程进行了比较，如图 1.1 所示，蛋白

质结构预测输入已知的氨基酸序列，通过基于已知结构的片段组装进行主链构象采样，再对侧链构象进行采样，不断进行上述过程，最终得到预测的蛋白质结构；而对于蛋白质设计问题，固定骨架设计（Fixed-backbone Design）固定蛋白质的主链结构，通过侧链构象采样确定氨基酸序列使得能量较低，而从头设计基于对于蛋白质二级结构的堆积方式也即折叠模式的粗略描述构建初始结构后，既需要进行侧链独立的主链构象优化，又需要在主链构象初步确定的前提下进行侧链采样以确定氨基酸序列，上述过程不断重复进行，从而设计出所需的蛋白质。可见，对于蛋白质结构预测和蛋白质设计问题，结构优化都是必要的步骤，因此需要确定蛋白质的能量函数。

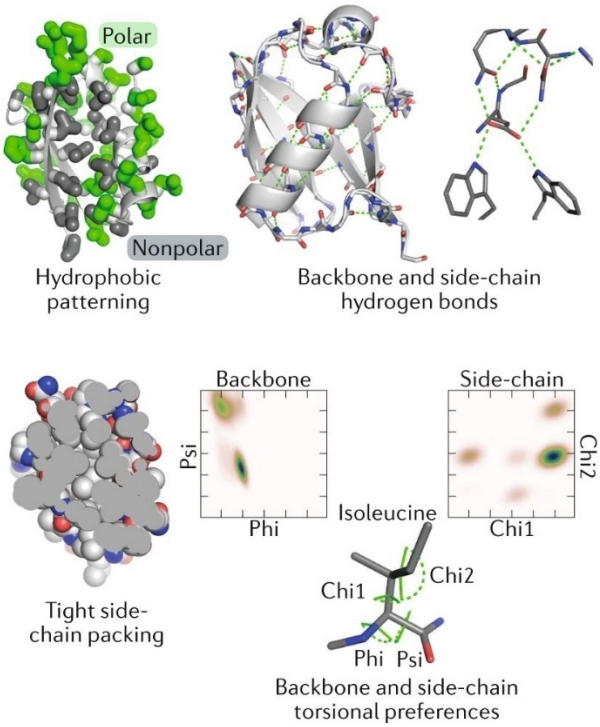


图 1.2 蛋白质能量函数的典型能量项[4]

如图 1.2 所示，蛋白质的能量函数是疏水相互作用、主链和侧链氢键、侧链堆积、主链和侧链扭转偏好等典型能量项的线性组合，蛋白质计算设计的主流策略也即通过对能量函数进行自动优化，得到能量最低的主链结构及氨基酸序列。20 世纪 90 年代，Dahiyat 等[6] 通过经验拟合的能量函数和适当的侧链选择算法首次实现了固定主链的氨基酸序列设计的自动优化方法。在过去近 30 年内，蛋白质计算设计的能量函数和优化算法不断发展，基于 Rosetta 能量函数的软件 Rosetta[7] 已经被广泛应用于蛋白质计算设计领域，在氨基酸序列从头设计任务上结合实验筛选表现出较高的成功率。Xiong 等[8] 建立了 ABACUS 统计能量函数，并通过实验验证了基于 ABACUS 的蛋白质计算设计可以得到远超天然蛋白质的热稳定性的人工蛋白质。能量函数中拟合得到的各能量项主要分为物理能量项和统计能量项两类，例如 Rosetta 使用物理方法对范德华相互作用、静电相互作用、溶剂化作用和氢键相

互作用的势函数进行拟合和线性组合，而 ABACUS 通过对天然蛋白的侧链构象和原子距离等进行统计，把不同的结构特征进行组合，根据玻尔兹曼分布（Boltzmann Distribution）

$$P(r) \propto \exp\left(-\frac{E(r)}{k_B T}\right)$$

从统计热力学的角度由统计得到的概率分布$P(r)$得到能量分布$E(r)$，其中$r$为微观状态的坐标。

随着计算机技术的进步，人工智能（Artificial Intelligence, AI）技术和深度神经网络（Deep Neural Network, DNN）的发展为蛋白质计算设计的能量函数拟合和能量优化过程提供了全新的途径。例如，Huang 等[9]提出的 SCUBA（Side Chain Unspecialized Backbone Arrangement）模型使用神经网络对各能量项进行拟合，高效率地实现了较高的拟合准确度。此外，借助神经网络直接拟合可用于结构优化的梯度场（Gradient Field）也是当下化学信息学领域的热点课题，该方法对于蛋白质的结构优化具有重要的启发意义，具体将在本文 1.3 节进行介绍。

## 1.2 侧链无关的蛋白质主链优化

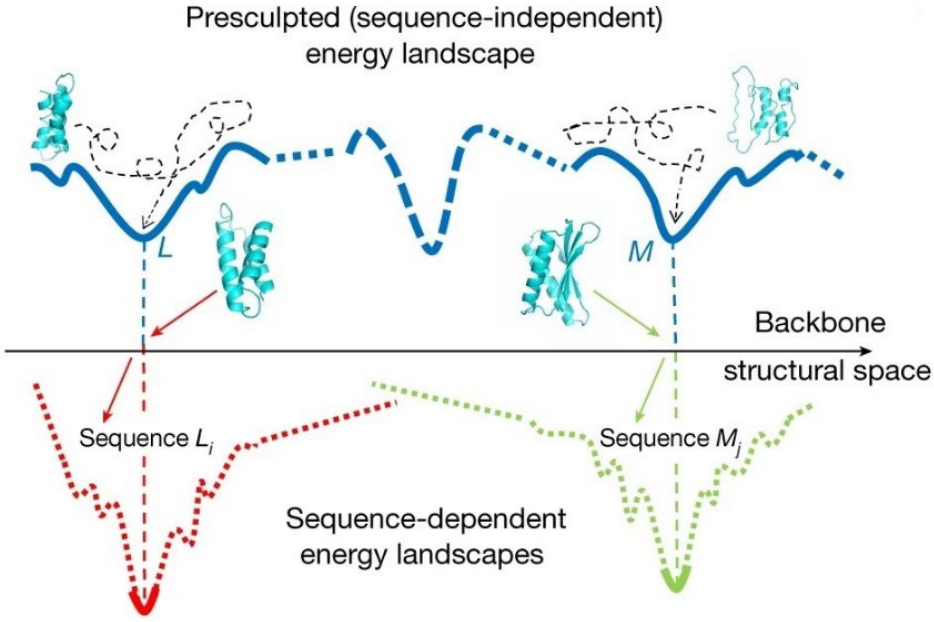


图 1.3 侧链相关与侧链无关的蛋白质能量景观的对应关系[9]

在蛋白质计算设计的过程中，主链结构的优化具有重要的意义。具体地，在蛋白质主链构象空间中，只有占比极小的构象具有相对较高的可设计性（Designability），即存在氨基酸序列能够自发稳定地折叠成所需的蛋白质结构。Hoang 等[10]的研究表明，侧链独立（Side Chain-independent）或侧链类型不敏感（Side Chain Type-insensitive）的分子相互作用塑造了蛋白质的能量景观。具体地，如图 1.3 所示，侧链无关的蛋白质能量景观（蓝色

曲线）的局部极小值L和M所对应的蛋白质主链构象分别精确地再现了侧链相关的蛋白质能量景观（红色和绿色曲线）下不同氨基酸序列折叠过程中能量极小值$L_i$和$M_j$所对应的构象。相对于对氨基酸序列空间和特定序列的构象空间同时进行搜索，对不考虑侧链具体类型的的主链构象空间进行搜索能够大大减小计算成本，并且能够避免在能量极小化过程中落入特定氨基酸序列的能量曲面的局部极小值，有利于得到具有较高可设计性的蛋白质主链结构。

蛋白质的从头设计给定对于所需蛋白质的基本的二级结构的堆积方式也即折叠模式的描述，根据描述搭建蛋白质的初始主链结构，对主链结构进行优化以提高可设计性，再进行后续固定主链的氨基酸序列设计。现有的主链结构设计大部分基于启发式的主链设计方法，使用天然蛋白质的片段库进行片段的拼接组装[11]，随后再使用全原子能量函数对所得结构进行优化。该方法可以基于给定的参数快速生成相对合理的蛋白质骨架，但难以设计较为复杂的主链结构；另一方面，依赖侧链的主链优化过程原理上需要在已知氨基酸序列即侧链信息的前提下进行结构优化，增加了计算的复杂度。当前被广泛应用的蛋白质计算设计软件 Rosetta 在进行主链结构优化时使用的全原子能量函数将序列能量与主链能量相互耦合，因此 Rosetta 实际上采用了假定序列-优化主链-重新设计序列-优化主链的方式进行主链结构优化[12]，上述过程极大地增加了计算成本，并且难以实现主链结构的大范围变动。因此，如果能够设计出具有较高通用性的侧链无关的主链能量函数，实现侧链无关的主链结构优化，能够大大提升蛋白质从头设计的效率。

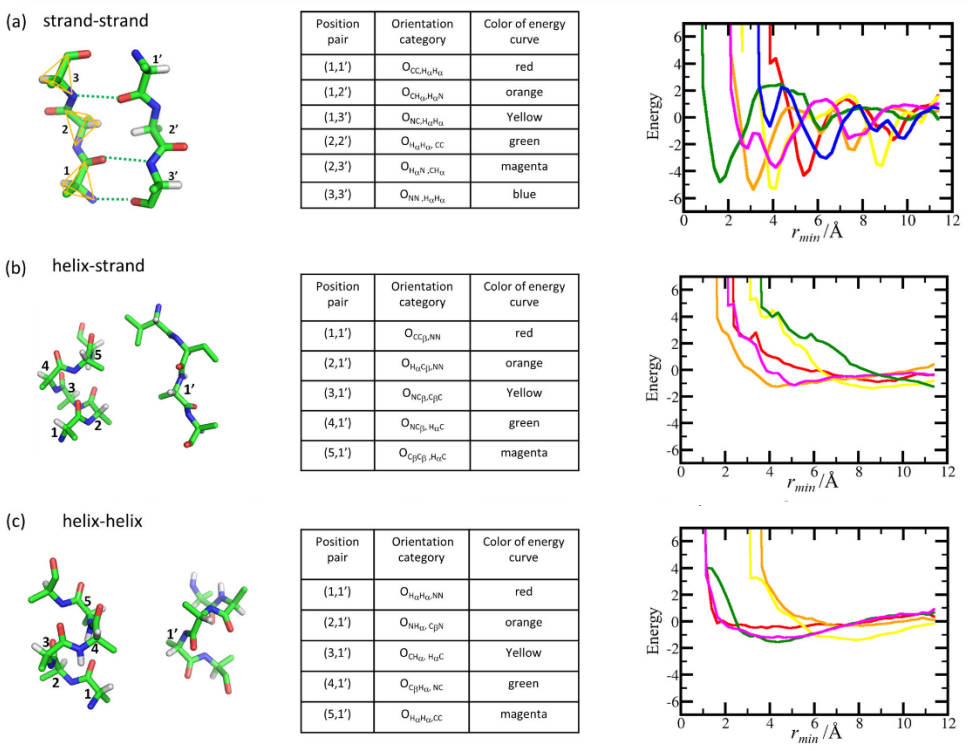


图 1.4 tetraBASE 能量函数的构建。(a) 两条反平行 β 折叠片；(b) 一条 α 螺旋和一条 β 折叠片；(c) 两条反平行 α 螺旋。[13]

2010 年，MacDonald 等 [14] 构建了基于 Cα 原子的能量函数，通过高分辨率蛋白质数据集对能量函数中的各能量项进行参数化，较好地模拟了蛋白质主链局部构象以及主链氢键。该研究表明在侧链无关的情况下，在该能量函数下的一些低能量结构仍能够很好地复现天然蛋白质结构片段，验证了经过优化后能量较低的主链结构具有较高的可设计性。但是该模型重点对于蛋白质局部骨架进行真实建模，而对于蛋白质中长距离相互作用的描述较为粗糙，因此在优化完整的主链结构时相对实际主链结构的偏差较大。2018 年，Chu 等 [13] 发展了基于统计的侧链无关的能量函数 tetraBASE，对于二级结构单元（Secondary Structure Element, SSE）之间的空间堆积进行建模。如图 1.4 所示，该研究将主链 Cα 视为四面体，假设 SSE 之间的相互作用取决于二级结构的类型、原子间距离和主链 Cα 的相对取向，通过对天然蛋白质中 SSE 堆积情况的统计，确定不同 SSE 堆积下不同取向的主链 Cα 之间相互作用的势能曲线，基于得到的能量函数通过蒙特卡洛（Monte Carlo, MC）模拟退火对初始结构中的 SSE 相对位置进行优化，高精度地再现了天然蛋白质中 SSE 的堆积。但是，tetraBASE 能量函数对主链 Cα 的取向和原子间距离进行了离散划分，得到的能量函数不是连续可微的，这对结构优化过程造成了限制。

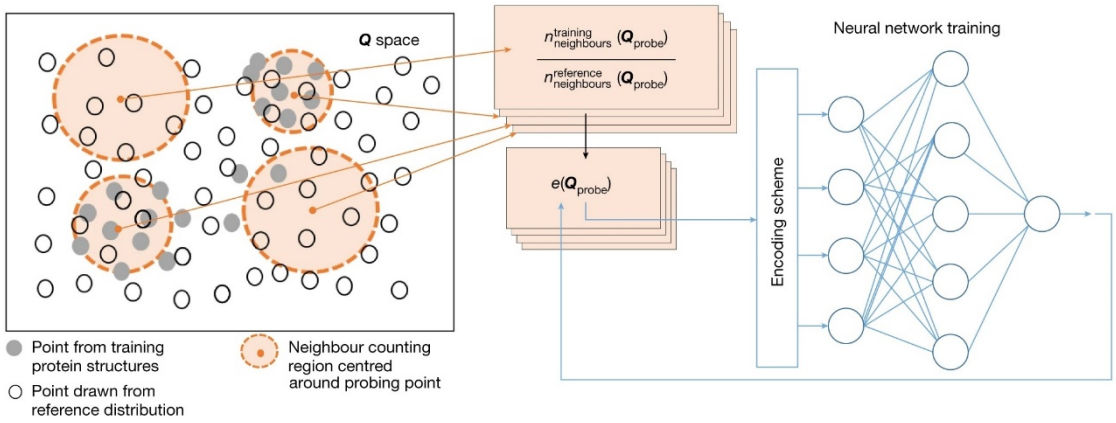


图 1.5 SCUBA 模型中近邻计数-神经网络的原理 [9]

2022 年，Huang 等 [9] 在 tetraBASE 的基础上，使用近邻计数-神经网络（Neighbor Counting-Neural Network, NC-NN）方法拟合侧链无关的能量函数中的各能量项，实现了对柔性主链结构的完整描述，并基于该统计能量函数发展了蛋白质主链结构优化方法 SCUBA，能够通过随机动力学模拟（Stochastic Dynamics, SD）方法对主链结构高效地进行优化。具体地，如图 1.5 所示，考虑结构变量$\boldsymbol{Q}$的高维变量空间，作为训练数据的天然蛋白质结构数据在该空间中以灰色实心圆点示意，而从人为构建的无相互作用下的参考分布中采样获得的结构数据表示为空心圆点；选定空间中特定探测点$\boldsymbol{Q}_{\text{probe}}$进行近邻计数，分别计算近邻的训练数据点和参考数据点的数量$n_{\text{neighbors}}^{\text{observed}}(\boldsymbol{Q}_{\text{probe}})$和$n_{\text{neighbors}}^{\text{reference}}(\boldsymbol{Q}_{\text{probe}})$，从而得到单点统计能量

$$e(\boldsymbol{Q}_{\text{probe}}) = -\ln\left(\frac{n_{\text{neighbors}}^{\text{observed}}(\boldsymbol{Q}_{\text{probe}})}{n_{\text{neighbors}}^{\text{reference}}(\boldsymbol{Q}_{\text{probe}})}\right)$$

将大量根据上述方法得到的单点统计能量输入神经网络，从而对能量函数的不同能量项进行拟合。该研究通过上述方法得到了连续可微的侧链无关的能量函数，能够很好地学习到高维结构数据中的能量相关信息。但是，上述方法本质上仍然构建了传统的基于统计的能量函数，在学习高维空间中的联合分布时对数据量有着较高要求。

## 1.3 基于图神经网络的构象生成

深度学习（Deep Learning, DL）作为机器学习（Machine Learning）的重要分支，通过多层神经元的连接提取数据中隐藏的模式特征，具有强大的学习能力，近年来在机器视觉、自然语言处理等一系列领域取得了巨大的成功。图神经网络（Graph Neural Network, GNN）能够处理以图的形式表示的数据，通过在大量图构成的数据集上进行训练，学习与图的结构相关的信息。

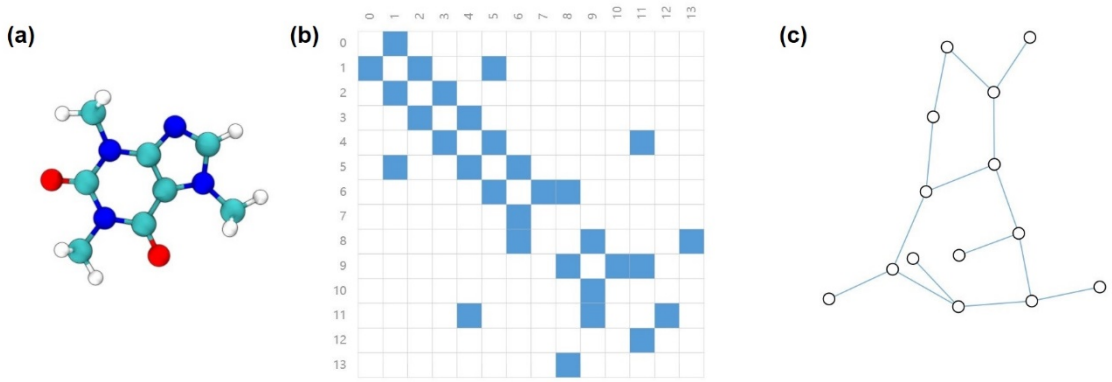


图 1.6 咖啡因分子的三维结构(a)、化学键的邻接矩阵(b)及分子图表示(c)

化学分子的结构可以用分子图（Molecule Graph）的形式表示，分子图能够直观高效地保存和传递化学分子的信息，因此 GNN 在化学信息学中得到了极其广泛的应用，在化学分子性质预测、分子力场拟合、药物分子虚拟筛选等领域都表现出了优异的性能。本文以咖啡因（Caffeine）分子为例，简要介绍与分子图相关的基本概念，以便后文对本课题所建立的模型的表述。如图 1.6 所示，咖啡因分子三维结构中的每个原子i视为无向分子图$G = (V, E)$中的一个节点（Node）$v_i \in V$，连接两个原子$i, j$的化学键视为分子图中连接两个节点的边（Edge）$e_{ij} \in E$，节点间边的连接情况由邻接矩阵（Adjacent Matrix）$\boldsymbol{A}$表示，定义为

$$\boldsymbol{A}_{\text{ij}} = \begin{cases} 1, e_{ij} \in G \\ 0, e_{ij} \notin G \end{cases}$$

分子图$G$的每个节点赋予特征向量，表示节点对应的原子的信息；每条边赋予特征向量，表示边所对应的化学键的信息。事实上，图的表示不仅可以应用于记录化学分子结构信息，

同样可以对蛋白质等大分子的结构信息加以表示，相应的图的构建方法是类似的。

在化学信息学领域，已有研究基于 GNN 实现小分子的构象生成。例如，Mansimov 等[15]使用条件生成图神经网络（Conditional Deep Generative Neural Network）直接在小分子数据集上学习能量函数，从而高效地生成具有较高几何多样性的分子构象；Simm 等[16]和 Xu 等[17]发展了基于变分自编码器（Variational Autoencoder, VAE）和连续流（Continuous Flow）和 GNN 两步生成分子构象的算法，即首先预测原子间距离，再把原子间距离转换成整个分子的三维坐标。但是，对于分子构象生成问题，需要实现分子构象信息在三维空间中的旋转不变性（Roto-translation Equivariance），先前的算法或者未能实现分子构象的旋转不变性，或者为了实现旋转不变性而向模型中引入大量噪声，极大影响了模型的表现。

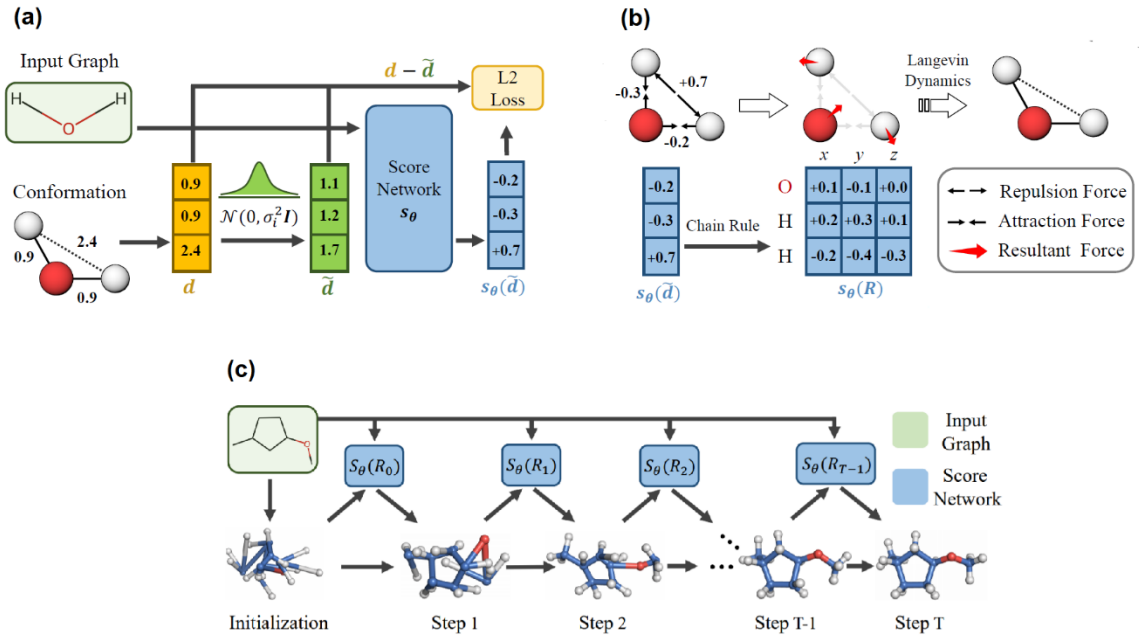


图 1.7 ConfGF 模型的原理[18]。(a) 模型训练过程；(b) 降噪分数匹配过程；(c) 分子构象生成过程。

2021 年，Shi 等[18]基于等变图神经网络（Equivariant Graph Neural Network, EGNN）和降噪分数匹配（Denoising Score Matching）方法提出了分子构象生成模型 ConfGF，一定程度上解决了先前模型的缺陷，对本课题的研究具有关键的启发意义。如图 1.7(a)所示，该模型的核心原理是 Song 等[19,20]提出的降噪分数匹配方法，具体地，EGNN 的输入是分子图和分子构象信息即原子间的距离，将原子间距离使用不同强度的高斯噪声（Gaussian Noise）进行扰动，使用 EGNN 训练一系列由小到大的高斯噪声扰动下作为梯度场的一系列得分函数（score function）$\boldsymbol{s_\theta}$；如图 1.7(b)所示，训练得到的基于原子间相对距离的初始得分函数 $\boldsymbol{s_\theta}(\tilde{\boldsymbol{d}})$经过链式法则（Chain Rule）转换成基于原子三维坐标的得分函数$\boldsymbol{s_\theta}(\boldsymbol{R})$，前者保持了模型训练过程的旋转不变性[21]，后者便于进行实际的坐标更新；如图 1.7(c)所示，在训练得

到一系列得分函数$\boldsymbol{s_\theta(R)}$后，对输入的分子图进行坐标随机初始化输入模型，通过朗之万动力学（Langevin Dynamics）对原子坐标进行更新，经过一系列的降噪得分函数$\boldsymbol{s_\theta(R)}$的梯度更新，最终生成合理的分子构象。该研究在分子构象生成问题上取得了突破性进展，对后续研究具有高度的启发性。

2021 年，Wu 等[22]在上述研究的基础上把相似的模型应用于蛋白质折叠问题上，考虑通过结合 EGNN 与降噪分数匹配方法实现蛋白质的结构优化和从头折叠，获得了较好的结果。该研究把蛋白质视为图，每个残基视为图的一个节点，基于残基之间的共价连接、残基空间位置上的接近和残基在序列中的位置三个标准构建了图的边，训练得到了能够高效实现蛋白质结构优化的模型 EBM-Fold，验证了 EGNN 结合降噪分数匹配方法的架构对于蛋白质结构优化问题的可行性。

## 1.4 课题目的与研究思路

本课题关注作为计算生物学领域重点课题的蛋白质主链结构优化问题，试图提出一种基于 GNN 的新方法，实现侧链无关的蛋白质主链结构优化。具体地，如本文 1.1 节所述，蛋白质计算设计在近几十年来实现了突破性发展，而蛋白质主链结构优化能够提高蛋白质初始结构的可设计性，在蛋白质从头设计中具有关键的意义。但如本文 1.2 节所述，Rosetta 等目前通用的软件在实现蛋白质主链结构优化时使用了与侧链耦合的主链能量函数，大大降低了计算效率和准确性，而目前已有研究使用传统的统计方法拟合侧链无关的主链能量函数，在高维空间中因“维数灾难”（Curse of Dimensionality）而对于数据量要求极高、拟合效率较低。本课题受到本文 1.3 节所述的分子构象生成模型及蛋白质折叠模型的启发，试图构建已在多个问题上被证明行之有效的 EGNN 结合降噪分数匹配的模型架构，在经过处理后的蛋白质数据集上进行训练，从而借助 GNN 的强大学习能力拟合可以用于蛋白质结构优化的梯度场，实现高效的侧链无关的主链结构优化；本课题把训练得到的模型应用于实际的蛋白质主链结构问题上，试图证明该模型改善了初始蛋白质骨架的可设计性，在实际的蛋白质计算设计过程中具有一定程度的实用价值。

# 第二章 材料与方法

## 2.1 数据集的建立

### 2.1.1 蛋白质结构数据的收集

本课题通过 GNN 提取蛋白质结构背后所隐藏的信息，拟合可用于侧链无关的蛋白质主链结构优化的梯度场，因此需要大量高质量的蛋白质结构数据所组成的数据集。本课题使用 Wang 等[23]开发的 PISCES 服务器对整个 PDB（Protein Data Bank）[24]进行筛选，将筛选得到的残基数适中、高分辨率、低序列一致度（Sequence Identity）的 PDB 的子集作为本课题使用的数据集，设定的具体筛选项目和筛选条件如表 2.1 所示。

表 2.1 PISCES 服务器的筛选项目和筛选条件

| 筛选项目 | 筛选条件 |
|---|---|
| 残基数 | 40~1000 |
| 分辨率 | < 2.5 Å |
| 序列一致度（Sequence Identity） | < 30% |

根据上述标准，筛选得到 14606 条蛋白质链，作为初始数据集进行后续处理。

### 2.1.2 蛋白质结构数据的预处理

初始数据集中的蛋白质结构数据需要进行预处理，以便输入 GNN 模型中进行训练。初始结构中的杂原子（Hetero Atoms, HETATM）对后续模型读取和处理 PDB 文件数据造成影响，本课题使用 Rodriguez 等[25]开发的工具 pdb-tools 清除杂原子，并经过处理得到有效的 PDB 文件。此外，初始数据集中部分结构无法在后续训练中被读入，会导致内部模块的错误，例如图 2.1(a)所示的结构末端第 334 位的赖氨酸仅有 N 原子、在 PDB 文件中标注为蓝色，图 2.1(b)所示的蛋白质链中间无序部分的结构未解出，这些蛋白质结构数据均需要在数据集预处理步骤筛去。

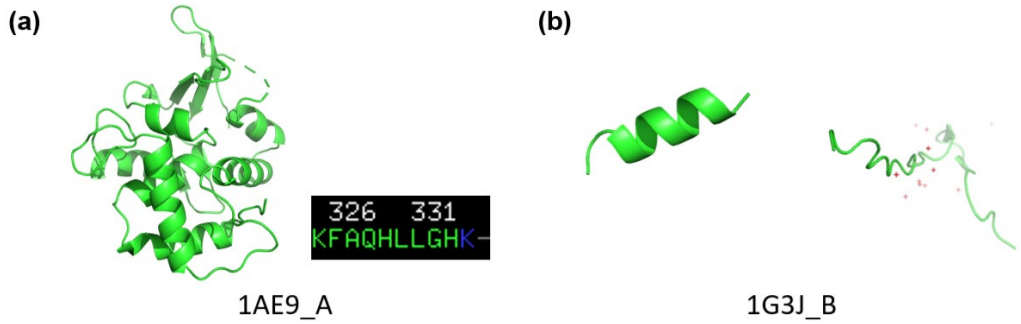


图 2.1 初始数据集中需要筛去的结构示例。(a) 末端残基仅有 N 原子；(b) 中间无序结构未解出。

### 2.1.3 数据集的划分

初始数据集经过前文所述的预处理后得到 9224 条蛋白质结构数据。本课题将数据集按照训练集（Training Set）：验证集（Validation Set）：测试集（Test Set）=18: 1: 1 的比例随机进行划分，得到的数据量分别为 8301、461、462。在后续实验中，训练集用于模型学习蛋白质结构信息，验证集用于监测模型训练的表现、监控模型是否发生过拟合，而测试集用于检验模型在实际情境下的表现，以作为模型优化的依据。

## 2.2 开发环境与硬件

本课题的程序在 Windows 10 的 Visual Studio Code IDE 下进行开发，各项功能主要使用 Python 语言编写，作为模型主体的 GNN 基于成熟的开源 Python 机器学习库 PyTorch 及其附加模块 PyTorch-Geometric（PyG）、PyTorch-Scatter、PyTorch-Sparse、PyTorch-Spline-Conv、PyTorch-Cluster 等实现，结果和蛋白质分子结构的可视化呈现由 Matplotlib 库和 PyMOL 软件实现，模型运行过程中及输出结果所需的统计分析由 NumPy 等常用库实现。具体的开发环境和关键模块信息如表 2.2 所示。此外，本课题使用 Rosetta 程序实现固定骨架的蛋白质计算设计，版本为 Rosetta 2022.11。

表 2.2 开发环境和关键模块信息

| 开发环境或模块 | 版本 |
|---|---|
| Python | 3.7.12 |
| CUDA Toolkit | 10.2.89 |
| DSSP | 2.0.4 |
| Matplotlib | 3.5.1 |
| Numba | 0.55.1 |
| NumPy | 1.21.5 |
| pdb-tools | 2.4.3 |
| PyG | 2.0.3 |
| PyMOL | 2.5.0 |
| PyTorch | 1.10.2 |
| PyTorch-Cluster | 1.5.9 |
| PyTorch-Scatter | 2.0.9 |
| PyTorch-Sparse | 0.6.12 |
| PyTorch-Spline-Conv | 1.2.1 |
| Torchaudio | 0.10.2 |
| Torchvision | 0.11.3 |

本课题中程序的编写与调试在个人 PC 上完成，CPU 为 Intel i7-9750H，内存为 16 GB DDR4，GPU 为 NVIDIA GTX 1660 Ti。模型的训练、评估和应用等过程在北京大学高性能计算平台（PKUHPC）上完成，使用 gpu_2l（CPU: Intel Gold 6132; GPU: NVIDIA Tesla V100）、gpu_4l（CPU: Intel Gold 6132; GPU: NVIDIA Tesla V100）、cn-short（CPU: Intel E5-2670 v2）、cn_nl（CPU: Intel Gold 6132）节点提交任务进行计算。

## 2.3 模型的构建与优化

本课题参考Shi等[18]提出的分子构象生成模型ConfGF和Wu等[22]提出的蛋白质从头折叠模型EBM-Fold的架构，基于四层EGNN+降噪分数匹配构建模型，在前述蛋白质结构数据集上进行训练，从而开发可用于侧链无关的蛋白质主链结构优化的全新模型。在前述研究的基础上，本课题重点关注了模型对蛋白质结构优化问题的适配，隐去训练集中蛋白质结构数据中具体的残基类型信息，使得模型基于二级结构信息学习蛋白质主链结构的优化，从而实现“侧链无关”这一模型关键特性。本课题基于理论分析和蛋白质结构优化问题实际情况，调整了EGNN+降噪分数匹配模型中高斯噪声强度等超参数，并且对降噪分数匹配过程的细节进行了改进。

### 2.3.1 基于蛋白质结构数据的图的构建

以PDB文件形式保存的蛋白质结构数据需要以图的形式被模型读取，用于EGNN的训练。本课题基于蛋白质结构数据构建图的方法简述如下，具体的python代码详见附录A。

**节点** 蛋白质每个残基的Cα原子作为图的一个节点。具体地，节点位置是从PDB文件中读取的Cα的三维坐标，节点类型使用一位数字表示，代表该残基所属的蛋白质二级结构类型；二级结构类型通过DSSP程序[26]进行判断，除α-螺旋和β-折叠外的二级结构被不加区分地归属于一类，具体的二级结构类型、DSSP编码和相应的模型中的编码如表2.3所示。

表 2.3 二级结构类型、DSSP 编码与模型编码的对应关系

| 二级结构类型 | DSSP 编码 | 模型编码 |
|---|---|---|
| α-螺旋 | H | 0 |
| β-折叠 | E | 1 |
| 孤立的 β-桥残基 | B | 2 |
| $3_{10}$-螺旋 | G | 2 |
| π-螺旋 | I | 2 |
| 转角 | T | 2 |
| 弯曲 | S | 2 |
| 无 | - | 2 |

需要说明的是，在模型中被归入同一类的$3_{10}$-螺旋、π-螺旋等与一般的无规卷曲（Coil）具有不同的结构特征，但蛋白质结构数据集中上述结构所占比例极小，如果单独分类，模型训练过程中无法很好地学到上述结构与其他二级结构之间作用的相关信息。为保证模型的泛化性能，本研究不加区别地将上述结构与一般的无规卷曲归入一类进行模型训练。

**边** 蛋白质一对残基之间的相互作用对应图中的一条边，边的长度根据节点位置进行计算，定义为一对残基中Cα原子的距离。模型构建了两种类型的边。第一种类型的边基于蛋白质结构中的氢键相互作用，使用DSSP程序判断主链氢键是否存在，对可能存在的氢键要求键能

$$E_{\text{hbond}} < 0.5\ \text{kJ} \cdot \text{mol}^{-1}$$

并且只在残基所属的二级结构为α-螺旋或β-折叠的情况下，在残基之间建立边的联系，用来捕捉α-螺旋和β-折叠两种二级结构内部和不同二级结构之间的氢键相互作用。第二种类型的边基于序列信息，分别在第$i$个残基和第$i \pm 1$、$i \pm 2$个残基之间建立边的联系，用来捕捉残基的共价连接和在序列位置上的邻近所带来的相互作用。已经被归入基于氢键信息的上述边的连接基于不同类型在模型中被赋予不同编码，具体如表2.4所示。

表 2.4 边的类型与模型编码的对应关系

| 边的类型 | 模型编码 |
|---|---|
| 基于氢键相互作用 | 0 |
| 基于序列信息，残基 $i-i\pm1$ | 1 |
| 基于序列信息，残基 $i-i\pm2$ | 2 |

根据前述方式，基于蛋白质结构数据构建图的节点和边，输入EGNN中进行训练。

2.3.2 模型的原理与基本架构

模型的原理与基本架构如图2.2所示，其中图2.2(a)示出模型的训练过程，图2.2(b)示出利用训练得到的模型实现蛋白质主链结构优化的过程。具体地，在模型训练过程中，对于输入模型的蛋白质结构数据，一方面按照2.3.1所述的方法构建图，另一方面计算存在边的连接的残基的Cα原子之间的欧氏距离，构建距离向量$\boldsymbol{d}$；使用一系列不同强度$\{\sigma_i\}_{i=1}^{L}$的高斯噪声$\boldsymbol{N}(0,\sigma_i^2\boldsymbol{I})$对残基Cα原子的距离向量$\boldsymbol{d}$进行扰动，噪声强度等比分布，得到扰动后的距离向量$\tilde{\boldsymbol{d}}$，即

$$\tilde{\boldsymbol{d}} = \boldsymbol{d} + \boldsymbol{d}_{\text{noise}}$$

其中

$$\boldsymbol{d}_{\text{noise}} \sim \boldsymbol{N}(0,\sigma_i^2\boldsymbol{I})$$

将根据蛋白质结构数据构建的图和扰动后的距离向量$\tilde{\boldsymbol{d}}$输入打分网络（Score Network）$\boldsymbol{s_\theta}$，

使用该网络对梯度$\boldsymbol{s_\theta}(\tilde{\boldsymbol{d}})$进行预测，将模型预测的梯度$\boldsymbol{s_\theta}(\tilde{\boldsymbol{d}})$与实际需要拟合的、由高斯噪声产生的梯度$\boldsymbol{d}-\tilde{\boldsymbol{d}}$作差取模，得到L2损失函数

$$s_\theta(\tilde{d},\sigma)=\frac{1}{2L}\sum_{i=1}^{L}\sigma_i^2 E_{p(d|G)}E_{q_{\sigma_i}(\tilde{d}|d,G)}\left[\left\|\frac{s_\theta(\tilde{d})}{\sigma_i}-\frac{d-\tilde{d}}{\sigma_i^2}\right\|_2^2\right]$$

其中L是噪声级别的数目，$G$为用于训练的蛋白质结构数据所构建的图，$q$为给定的强度为$\sigma_i$的噪声的分布。基于该L2损失函数，通过梯度下降对打分网络$\boldsymbol{s_\theta}$进行训练。在每一个结构数据上进行训练时，高斯噪声的强度σ逐渐由小调大，从而希望提升打分网络$\boldsymbol{s_\theta}$的稳健性（Robustness）。

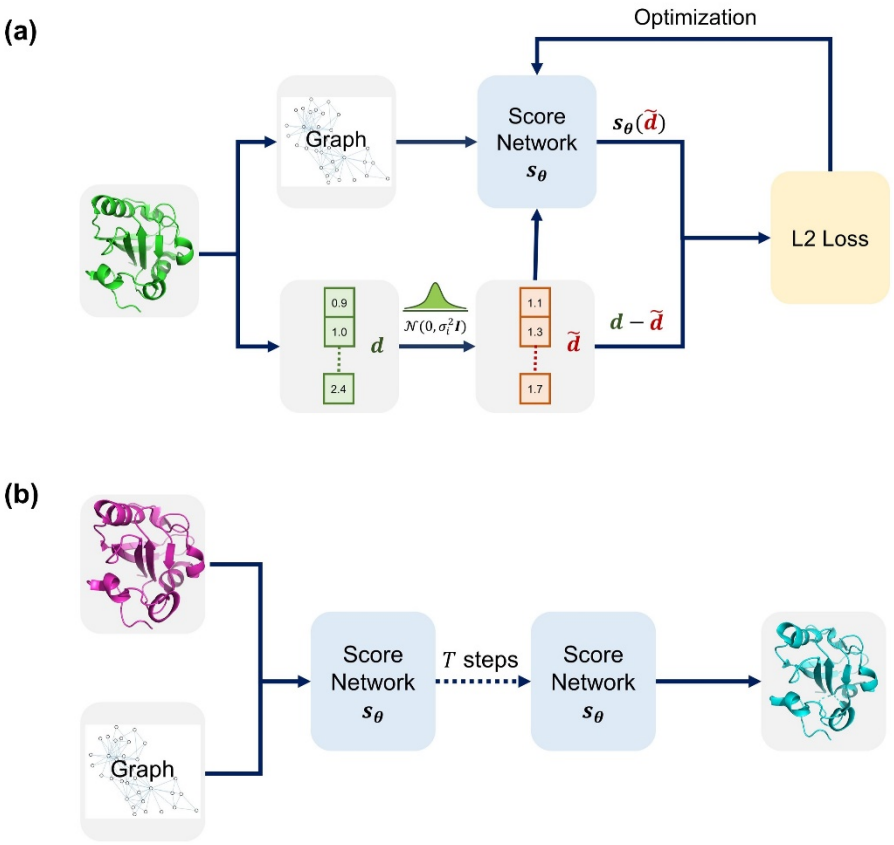


图 2.2 模型的原理与基本架构。(a) 模型训练过程；(b) 蛋白质主链结构优化过程。

在训练得到打分网络$\boldsymbol{s_\theta}$后，使用模拟退火朗之万动力学采样（Annealed Langevin Dynamics Sampling）的方法对蛋白质主链结构进行优化。具体地，如图2.2(b)所示，根据蛋白质已知的结构信息（序列长度、二级结构的类型和堆积方式等）构建图，将构建的图和蛋白质的初始骨架结构输入模型中，在模拟退火过程中逐渐将噪声强度由大调小，通过打分网络$\boldsymbol{s_\theta}$对初始骨架结构进行优化，从而最终得到优化后的蛋白质主链结构，通过上述优化过程提升蛋白质骨架的可设计性。

### 2.3.3 模型参数的确定

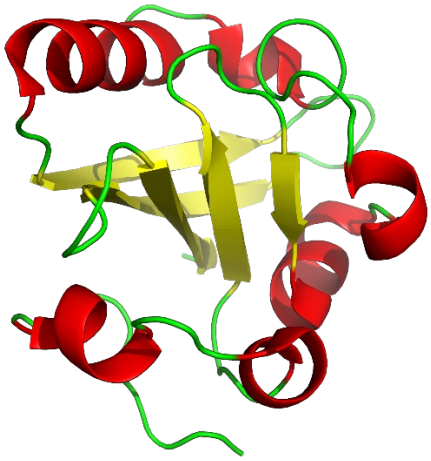

图 2.3 模型调试中选择的代表性蛋白质结构（PDB ID: 1NWZ）

实验中，本课题基于模型在验证集上的表现和模型在实际优化问题上的表现，对模型的超参数进行调整。具体地，考虑到模型调试效率，不失一般性地，本课题在实验调试过程中选择了一个具有代表性的蛋白质结构，即细菌光受体黄色光敏蛋白（Bacterial Light Receptor Photoactive Yellow Protein, PDB ID: 1NWZ）对模型在蛋白质主链结构优化问题上的表现进行验证。如图2.3所示，该结构序列长度适中（125个残基），同时具有多个α-螺旋和β-折叠片，实验中使用高斯噪声对该结构的主链Cα原子坐标进行扰动，使用所得模型对扰动后的结构进行优化，分别计算扰动后结构与真实结构之间以及优化后结构与真实结构之间主链Cα原子坐标的均方根偏差（Root Mean Square Deviation, RMSD），对上述两个RMSD的数值进行比较，从而判断模型的实际表现，作为参数调节的重要参照。

表 2.5 模型的主要超参数

| 参数 | 数值 |
|---|---|
| batch size | 64 |
| anneal power | 2.0 |
| learning rate | 0.001 |
| learning rate decay | 0.95 |
| hidden dimension | 256 |
| $\sigma_{\text{begin}}$ | 1 |
| $\sigma_{\text{end}}$ | 0.001 |
| $n_{\text{noise level}}$ | 50 |

对于EGNN+降噪分数匹配模型，最重要的参数是对初始结构进行扰动的一系列高斯噪声的强度$\{\sigma_i\}_{i=1}^{L}$，也即最大扰动对应的高斯噪声强度$\sigma_{\text{begin}}$和最小扰动对应的高斯噪声强度$\sigma_{\text{end}}$。本课题参考了Song等[27]关于基于分数的生成模型（Score-based Generative Model）如何选择噪声形式以实现最佳训练效果的研究，并且考虑到课题研究的实际背景，也即对蛋白质结构施加的扰动不能过大，否则难以仅根据二级结构信息实现蛋白质的从头折叠。经过实验调试，确定包括高斯噪声强度在内的模型的参数如表2.5所示。

2.3.4 模型性能的研究

基于蛋白质结构数据构建的图同时使用了残基处二级结构信息和氢键相互作用信息，分别作为节点和边的特征加入图中，让模型基于这些特征拟合可用于蛋白质结构优化的梯度场。本课题研究了上述两类信息对模型是否都是必要的，以更好地判断模型是否实际上既学到了不同二级结构所具有的不同特征，又学到了二级结构内部及二级结构之间相互作用的信息。

具体地，在实验过程中，本课题在2.3.1所述的图的构建方式的基础上，分别去除二级结构信息（所有节点记为同一类型，在模型中编码为0）和基于氢键相互作用的边，使用2.3.3中确定的模型参数进行训练，观察训练得到的模型的表现。

2.3.5 模型的改进

本课题在前述的模型构建方法上加以改进，在已有模型的基础上试图进一步改善模型表现。在实验过程中注意到，原有模型在实际的蛋白质主链结构优化问题上表现尚不令人满意，主要原因是未能很好地设计出蛋白质结构中的疏水核心，因此影响了所设计蛋白质骨架的稳定性。本课题因此在图的构建中加入基于侧链相互作用的新的边，希望模型在训练集上学到与蛋白质疏水核心相关的侧链相互作用。

具体地，本课题根据下述两条标准构建基于侧链相互作用的边。第一，使用DSSP程序计算残基的相对溶剂可及性（Relative Solvent Accessibility, RSA）[28]，要求残基的RSA $\leq 0.2$，即侧链被包埋的表面积占比$\geq 80\%$。第二，计算残基Cβ原子之间的距离，要求两个残基的侧链在空间上相互靠近，Cβ原子之间的距离$\leq 6$ Å（不计入甘氨酸）。模型在满足上述标准且尚无其他类型边的连接的一对残基之间建立新的边的连接，在模型中赋予该类边的编码为3，在加入上述边的基础上对模型进行训练和评估。

## 2.4 模型的训练与评估

本课题中各模型的训练在PKUHPC的gpu_2l和gpu_4l节点上进行，每个模型训练1000个epoch，选取在验证集上损失函数取得最小值的epoch所对应的模型进行评估。

模型的评估在测试集上进行，在大小为462的初始测试集中筛选序列长度在100~200个残基之间的蛋白质链，得到168条蛋白质结构数据，作为实际用于模型评估的测试集；使用

pdb-tools程序对这些蛋白质结构所对应的PDB文件进行处理，得到仅保留主链Cα原子的PDB文件；使用$\sigma = 0.2$的高斯噪声对主链Cα原子的三维坐标进行扰动，写入新的PDB文件，生成扰动后的结构；使用模型对扰动后的结构进行优化，分别计算扰动后结构与真实结构之间以及优化后结构与真实结构之间主链Cα原子坐标的RMSD，记为$\mathrm{RMSD}_{\mathrm{perturb}}$和$\mathrm{RMSD}_{\mathrm{opt}}$；若

$$\mathrm{RMSD}_{\mathrm{opt}} < \mathrm{RMSD}_{\mathrm{perturb}}$$

则认为优化成功，计算RMSD的变化

$$\Delta\mathrm{RMSD} = \mathrm{RMSD}_{\mathrm{perturb}} - \mathrm{RMSD}_{\mathrm{opt}}$$

在模型对整个测试集中的蛋白质结构优化完成后，计算模型在整个测试集上的优化成功率、平均的$\mathrm{RMSD}_{\mathrm{perturb}}$、$\mathrm{RMSD}_{\mathrm{opt}}$以及$\Delta$RMSD，作为对模型表现进行评估的标准。

表 2.6 蛋白质主链结构优化过程的参数调试

| steps_lr | steps | $\mathrm{RMSD}_{\mathrm{opt}}$ / Å | $\Delta$RMSD / Å |
|---|---|---|---|
| $2.4\times10^{-10}$ | 1000 | 1.369 | -0.284 |
| $2.4\times10^{-11}$ | 1000 | 1.020 | 0.065 |
| $2.4\times10^{-11}$ | 2000 | **1.004** | **0.081** |
| $2.4\times10^{-11}$ | 3000 | 1.011 | 0.074 |
| $2.4\times10^{-11}$ | 5000 | 1.175 | -0.090 |
| $2.4\times10^{-12}$ | 1000 | 1.073 | 0.012 |
| $2.4\times10^{-12}$ | 5000 | 1.061 | 0.024 |
| $2.4\times10^{-13}$ | 1000 | 1.081 | 0.004 |
| $2.4\times10^{-13}$ | 5000 | 1.116 | -0.031 |

在通过模拟退火朗之万动力学对扰动后的蛋白质结构进行优化的过程中，模型在每个高斯噪声强度下的优化步数（记为steps）和优化步长学习率（记为step_lr）对优化质量具有一定程度的影响。本课题以前述蛋白质结构（PDB ID: 1NWZ）为例，使用$\sigma = 0.1$的高斯噪声对主链Cα原子的三维坐标直接进行扰动，与真实结构进行比较，计算$\mathrm{RMSD}_{\mathrm{perturb}} = 1.085$ Å；通过对扰动后的结构进行优化，计算优化后的$\mathrm{RMSD}_{\mathrm{opt}}$和$\Delta$RMSD，对上述参数进行调试，部分调试结果如表2.6所示。根据实验结果确定模型评估所使用的参数为$\mathrm{steps} = 2000$，$\mathrm{step\_lr} = 2.4 \times 10^{-11}$。

## 2.5 模型的应用

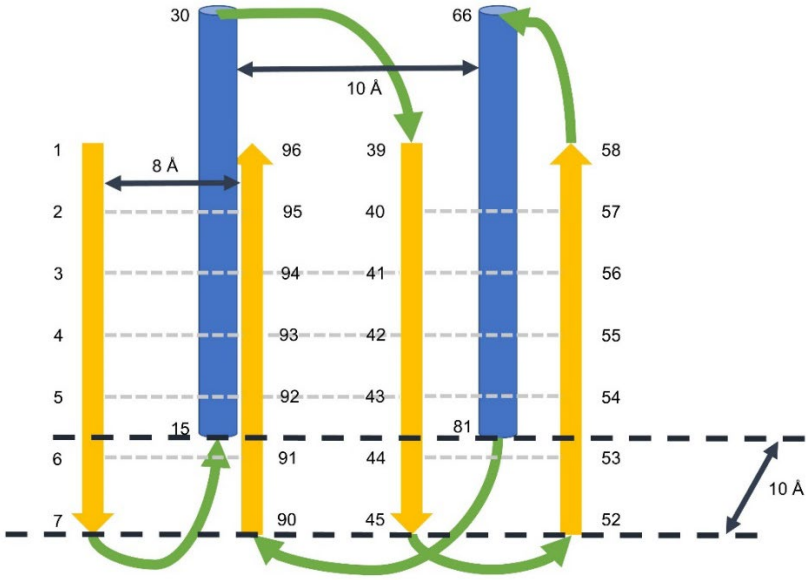


图 2.4 进行主链结构优化的蛋白质初始结构草图

本课题尝试使用2.4中训练得到、经由2.5所述方法改进的模型进行侧链无关的蛋白质主链结构优化和后续的序列设计。具体地，本课题参考Taylor等[29]提出的球状蛋白质结构的“周期表”（Periodic Table），设计了如图2.4所示的目标蛋白质的折叠模式，该结构由2个反平行的α-螺旋（图2.3蓝色圆柱）和4条反平行的β-折叠股（图2.3黄色箭头）组成的β-折叠片构成，图中示出各二级结构的相对位置以及各残基的分布，灰色虚线示出初始结构中定义的β-折叠股间的氢键，作为设计蛋白质结构时具有的先验信息输入模型中。

```
# beta-sheet hydrogen bond information (residue labels start from 1)
2 95
3 94
4 93
5 92
6 91
94 41
93 42
92 43
40 57
41 56
42 55
43 54
44 53


# hydrophobic-core-based neighbor information (residue labels start from 1)
13 89
18 21
18 91
21 76
21 79
28 69
28 72
44 91
60 63
66 69
69 72
76 79
```

图 2.5 模型输入文件 sketch_extra_info.txt 内容

基于图2.3对于蛋白质初始结构的描述，本课题使用Huang等[9]开发的SCUBA程序生成初始的蛋白质骨架。SCUBA程序在生成蛋白质骨架时仅满足二级结构之间的几何位置关系，对骨架结构的合理性未作过多要求，且生成蛋白质骨架的过程具有随机性，依赖于输入程序的随机数种子。本课题随机选择互不重复的1~10000以内的整数作为随机数种子，根据图2.3的描述，使用SCUBA程序生成了10个初始的蛋白质骨架结构。本课题随后使用Rosetta程序的fixbb（Fixed Backbone Design）模块[30]分别对上述10个结构进行固定主链的序列设计，允许全部20种氨基酸在各残基处放置，每个结构生成10个可能的序列（即设定nstruct为10）；对于生成的所有序列，使用Rosetta程序的relax模块[31,32]对全原子的蛋白质结构进行精细优化，对每个序列生成5个可能的优化后结构（即设定nstruct=5），结构优化的循环数（Cycle Number）设定为5。从输出文件score.sc中读取所设计序列经过relax模块精细优化后的得分并取均值，作为初始蛋白质骨架结构质量的评价标准。

本课题随后使用2.5所描述的模型对初始的蛋白质结构进行优化，根据图2.4对蛋白质结构的设计要求，预先指定β-折叠片之间存在氢键相互作用的残基和参与形成疏水核心的、存在侧链相互作用的残基，相关信息写入如图2.5所示的sketch_extra_info.txt文件中，模型在进行主链结构优化时读取sketch_extra_info.txt文件中的描述，从而基于蛋白质初始结构而构建图的节点和边。优化过程所使用的模型参数与2.4一致，即$\mathrm{steps} = 2000$，$\mathrm{step\_lr} = 2.4 \times 10^{-11}$，将优化后的蛋白质主链Cα的三维坐标信息写入新的PDB文件。由于Rosetta fixbb模块需要完整的主链结构，本课题使用PD2 ca2main服务器[33]根据主链Cα坐标重建主链全部原子的坐标。对于经过上述步骤得到的优化后的主链结构，同样使用Rosetta fixbb模块进行固定主链的序列设计，每个结构生成10个可能的序列，使用Rosetta程序的relax模块对设计得到的蛋白质结构进行精细优化，对每个序列生成5个可能的优化后结构，结构优化的循环数设定为5。从输出文件score.sc中读取所设计序列经过relax模块精细优化后的得分并取均值，作为优化后蛋白质骨架结构质量的评价标准。通过比对优化前后序列设计的得分，判断本课题所构建的模型在蛋白质主链结构优化上的效果。

# 第三章 结果与讨论

## 3.1 模型训练与评估结果

### 3.1.1 基础模型的训练与评估

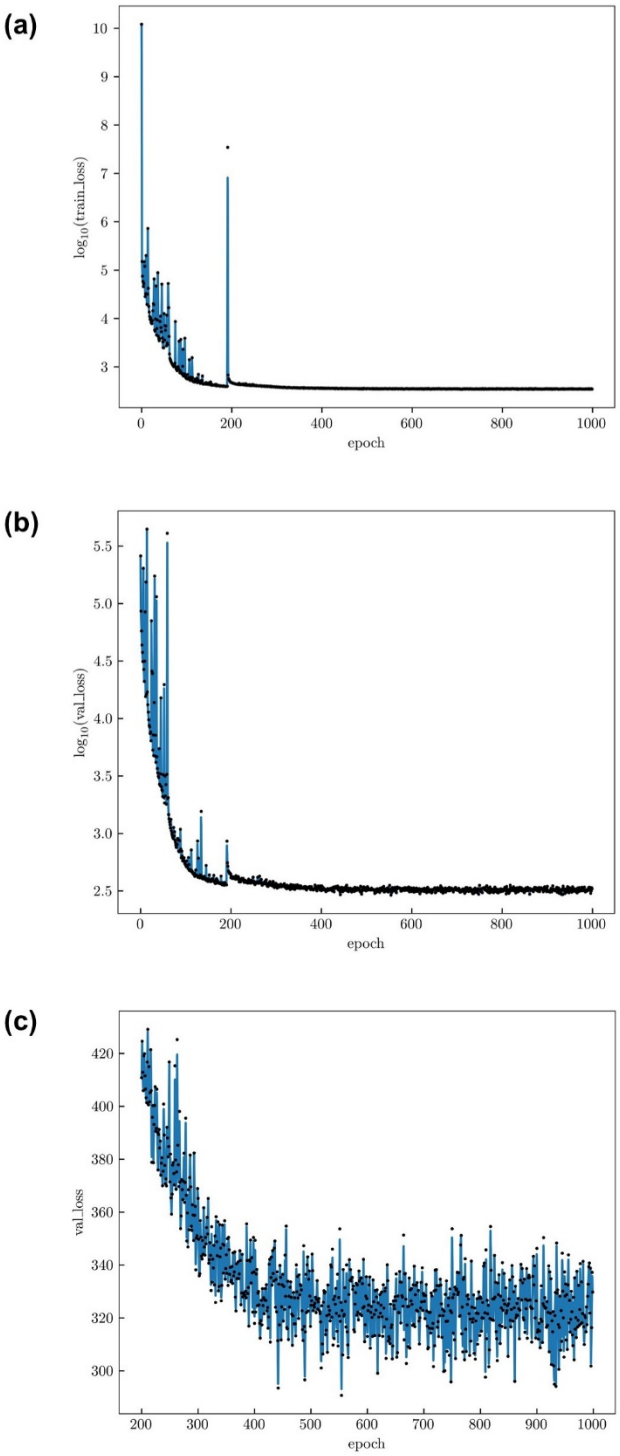


图 3.1 基础模型训练过程的损失曲线。(a) 训练集上的损失曲线，纵坐标为对数坐标；(b) 验证集上的损失曲线，纵坐标为对数坐标；(c) 验证集上 200~1000 epoch 的损失曲线，纵坐标为线性坐标。

在如 2.3 节所述方式构建的基础模型的训练过程中，训练集和验证集上的损失曲线如图 3.1 所示，可见训练过程中损失函数在训练集和验证集上都逐步下降，在约 400 个 epoch 后达到稳定；在验证集上，损失函数在第 554 个 epoch 达到最小值，后续使用第 554 个 epoch 所对应的模型进行评估。

如 2.3.3 节所述，本课题为研究模型是否学到了二级结构信息以及二级结构相关的氢键相互作用的信息，在基础模型的图的构建方式的基础上，分别去除二级结构信息（所有节点记为同一类型，在模型中编码为 0）和基于氢键相互作用的边，使用相同的模型参数进行训练，对模型的表现进行评估。上述三个模型（基础模型、无二级结构信息模型、无氢键相互作用模型）的评估结果如表 3.1 所示。

表 3.1 基础模型及其变种的评估结果

| 模型 | 优化成功率 | $\overline{\mathrm{RMSD}_{\mathrm{perturb}}}$ / Å | $\overline{\mathrm{RMSD}_{\mathrm{opt}}}$ / Å | $\overline{\Delta\mathrm{RMSD}}$ / Å |
|---|---|---|---|---|
| 基础模型 | **100%** | 2.172 ± 0.084 | 1.808 ± 0.086 | **0.363 ± 0.076** |
| 无二级结构信息 | 0% | - | - | - |
| 无氢键相互作用 | 7.4% | 2.158 ± 0.058 | 2.075 ± 0.081 | 0.083 ± 0.058 |

根据表 3.1 可以看出：

一．基础模型在测试集上的蛋白质主链优化成功率为 100%，优化后的结构相比优化前与真实结构间的 RMSD 平均下降了 0.363 Å，证明基础模型能够对经由高斯噪声扰动的蛋白质主链结构起到良好的优化作用，提升了蛋白质主链结构的质量。

直观地，如图 3.2 所示，对于前述代表性的、具有多个 α-螺旋和 β-折叠片的蛋白质结构（PDB ID: 1NWZ），模型的优化显著改善了蛋白质主链结构与真实结构的叠合，对于图中的多处 α-螺旋和 loop 区结构都具有明显的优化。

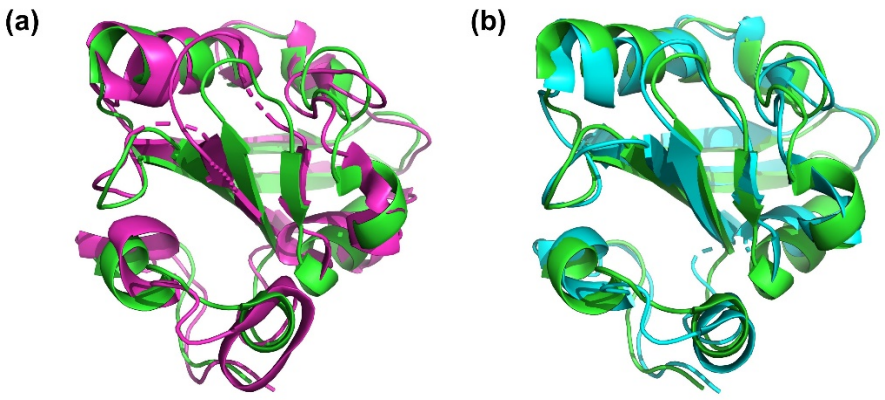


图 3.2 蛋白质（PDB ID: 1NWZ）真实结构（绿色）、扰动后结构（紫色）与优化后结构（蓝色）对比。(a) 真实结构对比扰动后结构；(b) 真实结构对比优化后结构。

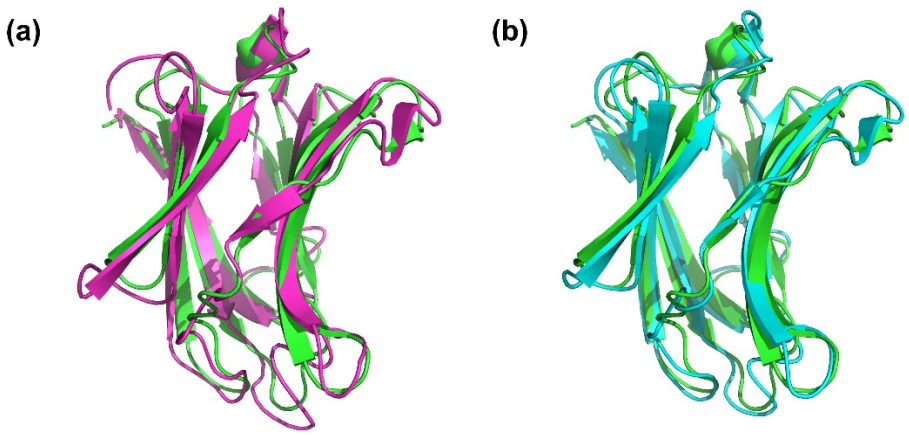


图 3.3 蛋白质（PDB ID: 3ZSJ）真实结构（绿色）、扰动后结构（紫色）与优化后结构（蓝色）对比。(a) 真实结构对比扰动后结构；(b) 真实结构对比优化后结构。

如图 3.3 所示，对于含有多个 β-折叠股的蛋白质结构（PDB ID: 3ZSJ），经过本课题构建的模型的优化，使得蛋白质多个 loop 区的构象明显改善，β-折叠片的位置相对结构优化前更接近真实结构。

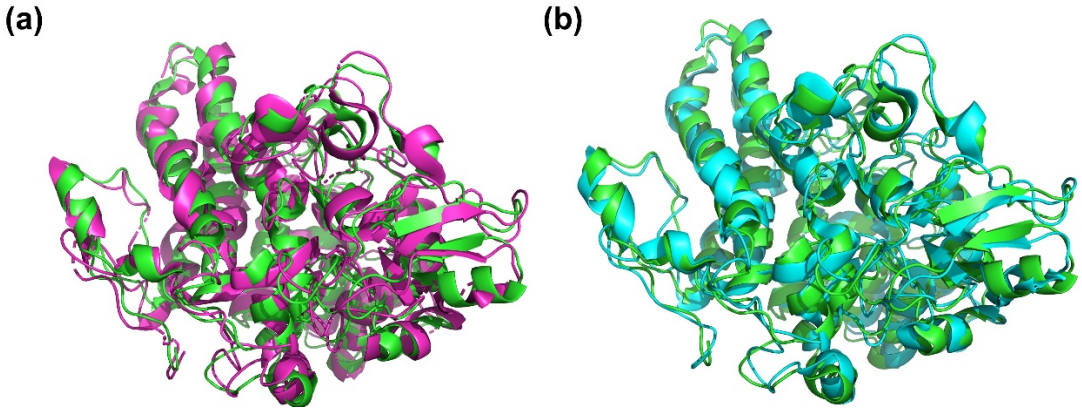


图 3.4 蛋白质（PDB ID: 6GCZ）真实结构（绿色）、扰动后结构（紫色）与优化后结构（蓝色）对比。(a) 真实结构对比扰动后结构；(b) 真实结构对比优化后结构。

如图 3.4 所示，对于序列长度达到 461、包含大量 α-螺旋和少量 β-折叠片的蛋白质结构（PDB ID: 6GCZ），本课题所构建的模型对多个扰动后被破坏的 α-螺旋的结构具有显著的恢复作用，使得优化后结构的 α-螺旋和 loop 区与真实结构很好地叠合，说明本课题构建的模型对较大的蛋白也有着较好的主链结构优化效果。

上述三个实例证明，本课题构建的模型确实学到了与二级结构和氢键等相互作用相关的信息，能够用于侧链无关的主链结构优化。

二．在模型学习的过程中，二级结构信息和氢键相互作用信息都是必要的，而模型也确实学到了上述知识。虽然二级结构信息和氢键相互作用信息部分重叠，但若仅保留其中一者则会导致模型表现大幅下降，以至于无法完成侧链无关的主链结构优化过程。

### 3.1.2 改进模型的训练与评估

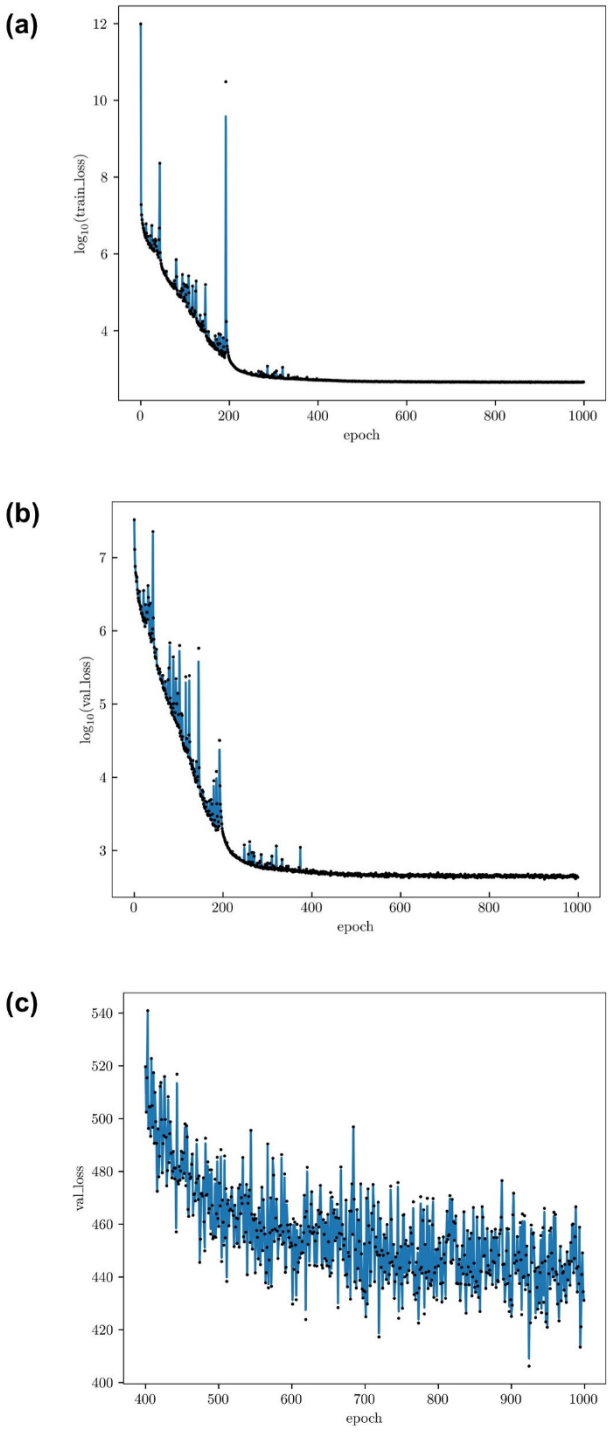


图 3.5 改进模型训练过程的损失曲线。(a) 训练集上的损失曲线，纵坐标为对数坐标；(b) 验证集上的损失曲线，纵坐标为对数坐标；(c) 验证集上 400~1000 epoch 的损失曲线，纵坐标为线性坐标。

如 2.5 所述，本课题对基础模型进行改进，改进模型的训练过程中，训练集和验证集上的损失曲线如图 3.5 所示，可见训练过程中损失函数在训练集和验证集上都逐步下降；

在验证集上，损失函数在第 924 个 epoch 达到最小值，后续使用第 924 个 epoch 所对应的模型进行评估。

表 3.2 基础模型与改进模型的评估结果对比

| 模型 | 优化成功率 | $\overline{\mathrm{RMSD}_{\mathrm{perturb}}}$ / Å | $\overline{\mathrm{RMSD}_{\mathrm{opt}}}$ / Å | $\overline{\Delta\mathrm{RMSD}}$ / Å |
|---|---|---|---|---|
| 基础模型 | **100%** | 2.172 ± 0.084 | 1.808 ± 0.086 | 0.363 ± 0.076 |
| 改进模型 | 92.2% | 2.165 ± 0.071 | 1.762 ± 0.082 | **0.403 ± 0.076** |

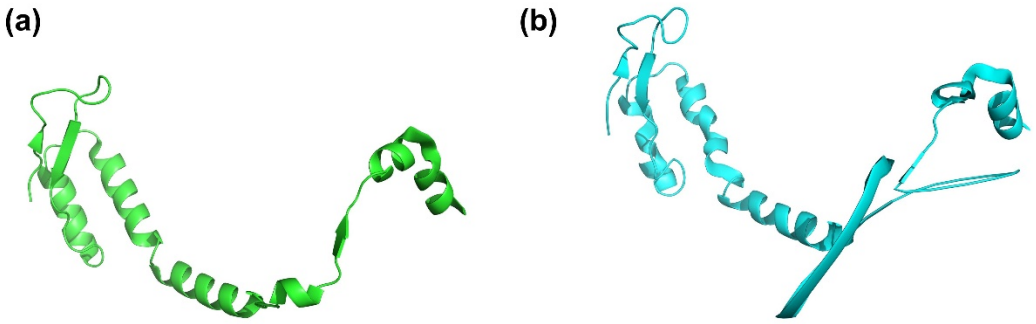


图 3.6 优化失败的蛋白质（PDB ID: 1JYO）真实结构(a)与优化后结构(b)对比

改进模型与基础模型评估结果对比如表 3.2 所示。可见向模型中加入蛋白质疏水核心处的残基侧链相互作用使得主链结构优化过程中平均的 RMSD 下降幅度更大，模型的优化性能得到了一定程度的提升。检视改进模型优化失败的案例，例如图 3.6 所示的优化失败的蛋白质（PDB ID: 1JYO），发现由于模型在计算时抽出整个结构中的一条链进行计算，该链自身无法形成稳定的疏水核心，因此导致优化出现问题，在优化过程中主链部分 Cα 原子被错误优化到明显不合理的位置。后续将对模型作出进一步改进，以避免对自身无法形成稳定结构的单条链进行结构优化。

## 3.2 蛋白质主链结构优化与序列设计结果

如 2.6 所述，将 SCUBA 程序根据描述生成的 10 个初始蛋白质主链结构和这些结构经过优化后得到的蛋白质主链结构通过 Rosetta fixbb 模块进行序列设计，通过 relax 模块进行结构精细优化和打分，结果如表 3.3 所示。除编号为 5 的结构因为 PD2 ca2main 服务器由 Cα 原子坐标构建主链原子坐标过程失败、从而无法完成后续的 Rosetta 序列设计和打分外，其余 9 个结构在 Rosetta 程序固定骨架的序列设计中的平均生成序列打分均有一定程度的下降，证明本课题所构建的模型能够通过侧链无关的主链结构优化一定程度上降低初始骨架结构的能量，从而提升初始主链的可设计性。成功实现优化的 9 个结构与初始结构的叠合对比如图 3.7 所示。

表 3.3 主链初始结构与优化结构的 Rosetta 序列设计结果

| 结构编号 | RMSD / Å | 初始结构平均打分 | 优化结构平均打分 |
|---|---|---|---|
| 1 | 0.802 | -142 ± 3 | -174 ± 7 |
| 2 | 0.766 | -121 ± 9 | -153 ± 13 |
| 3 | 0.857 | -75 ± 12 | -123 ± 10 |
| 4 | 0.739 | -131 ± 8 | -131 ± 9 |
| 5 | 0.792 | -158 ± 20 | 骨架重建失败 |
| 6 | 0.841 | -109 ± 12 | -128 ± 10 |
| 7 | 0.763 | -102 ± 14 | -125 ± 12 |
| 8 | 0.825 | -115 ± 7 | -143 ± 11 |
| 9 | 0.794 | -104 ± 11 | -144 ± 10 |
| 10 | 0.738 | -127 ± 9 | -128 ± 14 |

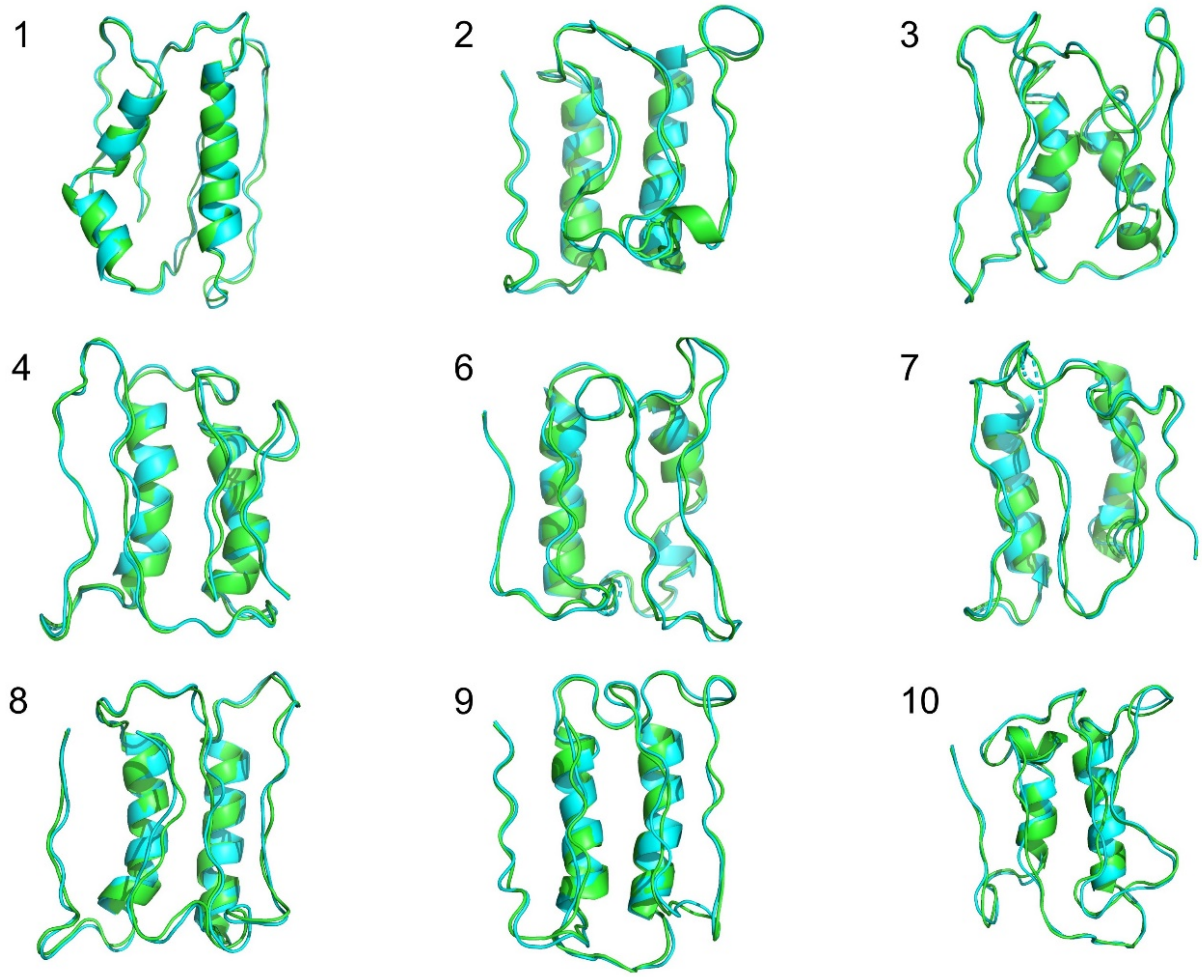


图 3.7 9 个优化成功的主链结构（蓝色）与优化前的主链结构（绿色）对比

需要指出的是，本课题目前得到的模型对蛋白质主链结构的优化仍然具有一定程度的局限性，无法通过模型一步优化为可以用于实际蛋白质计算设计的骨架，设计的序列仍然具有一定程度的不合理，而进一步调整模型参数未能实现模型表现的明显提升。该模型可能的局限性分析如下：在模型训练过程中，使用高斯噪声对训练集中结构的三维坐标进行扰动，希望模型拟合具有降噪效果的梯度场，这样的训练过程本质上主要捕捉了蛋白质主链局部的信息，模型未能很好地学到二级结构之间的相互作用，因此不能实现蛋白质初始主链结构的大幅度变动和优化。

# 第四章 结论与展望

本课题构建了基于等变图神经网络和降噪分数匹配方法的模型，试图实现侧链无关的蛋白质主链结构优化，从而在蛋白质从头计算设计中提升蛋白质初始骨架的可设计性，有助于后续的序列设计等步骤。具体地，本课题将已在分子构象生成等多个领域表现出优异性能的 EGNN+降噪分数匹配模型迁移到蛋白质主链结构优化问题上，针对蛋白质的特性设计了图的表示形式，将残基所属的二级结构信息、残基之间的距离和相互作用等特征嵌入图中，在筛选后的蛋白质结构数据集上对模型进行训练，通过模型对高斯噪声扰动下蛋白质主链结构的优化效果评价模型的表现。实验证明，对于与真实结构偏差 RMSD= 2.172 ± 0.084 Å的扰动后的结构，模型能够以 100%的优化成功率实现 0.363 ± 0.076 Å的 RMSD 下降，从而显著优化经过扰动后质量较差的蛋白质主链结构。本课题研究了二级结构信息和氢键相关的二级结构相互作用信息在模型学习过程中的重要性，证明了模型确实一定程度上学到了上述知识。本课题在基础模型上进行改进，向模型中添加了基于蛋白质疏水核心周围残基侧链相互作用的信息，实验证明对于与真实结构偏差 RMSD= (2.165 ± 0.071) Å的扰动后的结构，模型能够以 92.2%的优化成功率实现 0.403 ± 0.076 Å的 RMSD 下降，相较于基础模型取得了一定程度的改善。本课题将获得的模型应用于实际的蛋白质主链结构优化问题上，使用 Rosetta 程序的 fixbb 模块对优化前后的蛋白质主链结构进行固定骨架的序列设计，使用 relax 模块进行结构的精细优化和打分，在优化成功的骨架结构中，Rosetta 对于设计的序列的打分一定程度地下降，证明了模型能够一定程度上改善初始骨架结构的可设计性。

本课题构建的模型仍具有一定程度的局限性，在进行实际的蛋白质主链结构优化时难以一步直接优化得到合理的可设计骨架，对于骨架结构的优化是有限的。本课题计划接下来对模型架构进行系统改变，考虑在 EGNN 中加入注意力机制等，以更好地捕捉二级结构之间远距离、大范围相互作用的信息，从而更有效地实现对蛋白质初始骨架结构的优化；另一方面，本课题进行模型训练的过程使用高斯噪声对蛋白质主链 Cα 原子的三维坐标进行扰动，模型本质上学到的仍然是蛋白质主链局部的信息，不能很好地学到二级结构之间相互作用的相关知识，因此本课题未来考虑对训练集中的结构施加不同类型的噪声，使得二级结构整体发生平移或转动，从而希望模型学到超越蛋白质局部信息的全局知识。

# 参考文献


[1] M. S. Packer, D. R. Liu, “Methods for the Directed Evolution of Proteins”, *Nat. Rev. Genet.*, **16** (7): 379–394, **2015**

[2] M. G. F. Sun, M.-H. Seo, S. Nim, C. Corbi-Verge, P. M. Kim, “Protein Engineering by Highly Parallel Screening of Computationally Designed Variants”, *Sci. Adv.*, **2** (7): e1600692, **2016**

[3] C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, “The Genetic Control of Tertiary Protein Structure: Studies With Model Systems”, *Cold Spring Harb. Symp. Quant. Biol.*, **28**: 439–449, **1963**

[4] B. Kuhlman, P. Bradley, “Advances in Protein Structure Prediction and Design”, *Nat. Rev. Mol. Cell Biol.*, **20** (11): 681–697, **2019**

[5] P.-S. Huang, S. E. Boyken, D. Baker, “The Coming of Age of de Novo Protein Design”, *Nature*, **537** (7620): 320–327, **2016**

[6] B. I. Dahiyat, C. A. Sarisky, S. L. Mayo, “De Novo Protein Design: Towards Fully Automated Sequence Selection”, *J. Mol. Biol.*, **273** (4): 789–796, **1997**

[7] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan,; K. Kappel et al., “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. ” *J. Chem. Theory Comput.*, **13** (6): 3031–3048, **2017**

[8] P. Xiong, M. Wang, X. Zhou, T. Zhang, J. Zhang, Q. Chen, H. Liu, “Protein Design with a Comprehensive Statistical Energy Function and Boosted by Experimental Selection for Foldability”, *Nat. Commun.*, **5** (1): 5330, **2014**

[9] B. Huang, Y. Xu, X. Hu, Y. Liu, S. Liao, J. Zhang, C. Huang, J. Hong, Q. Chen, H. Liu, “A Backbone-Centred Energy Function of Neural Networks for Protein Design”, *Nature*, **602** (7897): 523–528, **2022**

[10] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, A. Maritan, “Geometry and Symmetry Presculpt the Free-Energy Landscape of Proteins”, *Proc. Natl. Acad. Sci.*, **101** (21): 7960–7964, **2004**

[11] C. O. Mackenzie, G. Grigoryan, “Protein Structural Motifs in Prediction and Design”, *Curr. Opin. Struct. Biol.*, **44**: 161–167, **2017**

[12] 操帆，陈耀晞，“蛋白质计算设计：方法和应用展望”，合成生物学，**2** (1)：15–32，**2020**

[13] H. Chu, H. Liu, “TetraBASE: A Side Chain-Independent Statistical Energy for Designing Realistically Packed Protein Backbones”, *J. Chem. Inf. Model.*, **58** (2): 430–442, **2018**

[14] J. T. MacDonald, K. Maksimiak, M. I. Sadowski, W. R. Taylor, “De Novo Backbone Scaffolds for Protein Design”, *Proteins Struct. Funct. Bioinforma*, **78** (5): 1311–1325, **2010**

[15] E. Mansimov, O. Mahmood, S. Kang, K. Cho, “Molecular Geometry Prediction Using a Deep Generative Graph Neural Network”, *Sci. Rep.*, **9** (1): 20381, **2019**

[16] G. Simm, J. M. Hernandez-Lobato, “A Generative Model for Molecular Distance Geometry”, In *Proceedings of the 37th International Conference on Machine Learning*; PMLR, pp 8949–8958, **2020**

[17] M. Xu, S. Luo, Y. Bengio, J. Peng, J. Tang, “Learning Neural Generative Dynamics for Molecular Conformation Generation”, arXiv March 30, **2021**

[18] C. Shi, S. Luo, M. Xu, J. Tang, “Learning Gradient Fields for Molecular Conformation Generation”, In *Proceedings of the 38th International Conference on Machine Learning*; PMLR, pp 9558–9568, **2021**

[19] Y. Song, S. Ermon, “Generative Modeling by Estimating Gradients of the Data Distribution”, In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Vol. 32., **2019**

[20] Y. Song, S. Garg, J. Shi, S. Ermon, “Sliced Score Matching: A Scalable Approach to Density and Score Estimation”, In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*; PMLR, pp

574–584, **2020**

[21] J. Han, Y. Rong, T. Xu, W. Huang, “Geometrically Equivariant Graph Neural Networks: A Survey”, arXiv February 21, **2022**

[22] J. Wu, T. Shen, H. Lan, Y. Bian, J. Huang, “SE(3)-Equivariant Energy-Based Models for End-to-End Protein Folding”; preprint; Bioinformatics, **2021**

[23] G. Wang, R. L. Dunbrack, “PISCES: A Protein Sequence Culling Server”, *Bioinformatics*, **19** (12): 1589–1591, **2003**

[24] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, “The Protein Data Bank”, *Nucleic Acids Res.*, **28** (1): 235–242, **2000**

[25] J. P. G. L. M. Rodrigues, J. M. C. Teixeira, M. Trellet, A. M. J. J. Bonvin, “Pdb-Tools: A Swiss Army Knife for Molecular Structures”, *F1000Research*, **7**: 1961, **2018**

[26] W. Kabsch, C. Sander, “Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features”, *Biopolymers*, **22** (12): 2577–2637, **1983**

[27] Y. Song, S. Ermon, “Improved Techniques for Training Score-Based Generative Models”, arXiv October 23, **2020**

[28] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, C. O. Wilke, “Maximum Allowed Solvent Accessibilites of Residues in Proteins”, *PLoS ONE*, **8** (11): e80635, **2013**

[29] W. R. Taylor, “A ‘Periodic Table’ for Protein Structures”, *Nature*, **416** (6881): 657–660, **2002**

[30] A. Leaver-Fay, B. Kuhlman, J. Snoeyink, “An Adaptive Dynamic Programming Algorithm For The Side Chain Placement Problem”, In *Biocomputing 2005*; WORLD SCIENTIFIC: Hawaii, USA; pp 16–27, **2004**

[31] P. Conway, M. D. Tyka, F. DiMaio, D. E. Konerding, D. Baker, “Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling”, *Protein Sci. Publ. Protein Soc.*, **23** (1): 47–55, **2014**

[32] L. G. Nivón, R. Moretti, D. Baker, “A Pareto-Optimal Refinement Method for Protein Design Scaffolds”, *PLoS ONE*, **8** (4): e59004, **2013**

[33] B. L. Moore, L. A. Kelley, J. Barber, J. W. Murray, J. T. MacDonald, “High–Quality Protein Backbone Reconstruction from Alpha Carbons Using Gaussian Mixture Models”, *J. Comput. Chem.* **34** (22): 1881–1889, **2013**

# 附录 A

附录 A.1：基于蛋白质结构数据构建图的 python 代码

```
import os
from tqdm import tqdm
import warnings
import pickle
from shutil import copy

import numpy as np

import torch
from torch_geometric.data import Data


from Bio.PDB import PDBParser, PDBIO, DSSP, NeighborSearch
from Bio.PDB.Selection import unfold_entities
from Bio.PDB.Polypeptide import is_aa


def set_working_dir(dir):
    """
    set default working directory
    """
    os.chdir(
        dir
    )


def pdb_to_data(pdb_file, hbond_threshold = -0.5, rsa_threshold = 0.2,
CB_dist_threshold = 6):
    """
    Covert a pdb file to a pyg object that can be fed into GNN
    """
    p = PDBParser(PERMISSIVE=1)
    model = p.get_structure(pdb_file[:4], pdb_file)[0]
    try:
        chain = model[pdb_file[5]]
    except:
        chain = model['A']

    dssp = DSSP(model, pdb_file)
```

```
dssp_keys = list(dssp.keys())

chain_len = 0
pos_list = []
ss_list = []
CB_item_list = []


for res in chain.get_residues():
    # exclude situations like 1OJH_A.pdb ('H_MSE', 25, ' ')
    if is_aa(res):
        chain_len += 1

        # 1. obtain secondary structure type
        ss_type = dssp[dssp_keys[chain_len-1]][2]

        if ss_type == 'H':
            ss_list.append(0)
        elif ss_type == 'E':
            ss_list.append(1)
        else:
            ss_list.append(2)


        # 2. obtain C-alpha and C-beta position
        atom_CA = res['CA']
        pos_list.append(list(atom_CA.coord))

        try:
            atom_CB = res['CB']
            CB_item_list.append(atom_CB)
        except:
            CB_item_list.append('')


# avoid residue missing in the middle of the sequence (see 1G3J_B.pdb)
if chain_len != dssp_keys[-1][1][1] - dssp_keys[0][1][1] + 1:
    raise IndexError('residue index mislabeled!')


# 3. construct edges
edge_list = []
edge_type = []
```

```
    # i) hydrogen-bond-based neighbors
    for i in range(0, chain_len):
        for col in [6, 8, 10, 12]:
            hbond_id = int(dssp[dssp_keys[i]][col])
            hbond_energy = float(dssp[dssp_keys[i]][col+1])

            # only preserve alpha-alpha, alpha-beta, beta-beta
            if (hbond_energy <= hbond_threshold) and (hbond_id != 0) and ([i,
i+hbond_id] not in edge_list) and ((i+hbond_id) in range(0, chain_len)) and
(dssp[dssp_keys[i]][2] in ['H', 'E']) and (dssp[dssp_keys[i+hbond_id]][2] in
['H', 'E']):
                edge_list.append([i, i+hbond_id])
                edge_list.append([i+hbond_id, i])
                edge_type += 2 * [0]
                # print([i, i+hbond_id])


    # ii) sequence-based neighbors
    for i in range(0, chain_len):
        for j in range(i+1, chain_len):
            if ((j-i) == 1) and ([i, j] not in edge_list):
                edge_list.append([i, j])
                edge_list.append([j, i])
                edge_type += 2 * [1]
            elif ((j-i) == 2) and ([i, j] not in edge_list):
                edge_list.append([i, j])
                edge_list.append([j, i])
                edge_type += 2 * [2]


    # iii) hydrophobic-core-based neighbors
    for i in range(0, chain_len):
        rsa_i = dssp[dssp_keys[i]][3]
        # print(rsa_i)
        if rsa_i <= rsa_threshold:
            for j in range(i+1, chain_len):
                rsa_j = dssp[dssp_keys[j]][3]
                # print(rsa_j)
                if rsa_j <= rsa_threshold:
                    # print([i, j])
                    try:
                        CB_dist = CB_item_list[i] - CB_item_list[j]
                        # print(CB_dist)
```

```
                    if (CB_dist <= CB_dist_threshold) and ([i, j] not in
edge_list):
                        edge_list.append([i, j])
                        edge_list.append([j, i])
                        edge_type += 2 * [3]
                        print(f'{i+1} {j+1}')
                except:
                    pass


    node_feature = torch.tensor(ss_list, dtype=torch.long)
    pos = torch.tensor(pos_list, dtype=torch.float32)
    edge_index = torch.tensor(edge_list, dtype=torch.long).t().contiguous()
    edge_type = torch.tensor(edge_type)
    graph_label = pdb_file[:-4]

    data = Data(x=node_feature, edge_index=edge_index, edge_type=edge_type,
pos=pos, y=graph_label)

    return data
```

# 致谢

“我们在窗口拥抱，人们从街上张望：
是让他们知道的时候了！
是石头要开花的时候了，
时间动荡有颗跳动的心，
是过去成为此刻的时候了。”

坐在荧荧的电脑屏幕前敲下毕业论文的致谢，大学四年生活的分分秒秒仿佛在我眼前全部复苏。正如本雅明所谓的，曾经与当下在一闪现中聚合成一个星丛，无数过往的瞬间相互交叠，从不同的时空位置向我投来柔和的光芒。

感谢我的导师来鲁华教授，大二那年能够加入您的课题组、加入分子设计实验室大概是我大学四年间最幸运的事情。在过去的两年多里，您对于科研工作的热忱、对我的关怀和教诲时常让我深深地感动，我的课题推进和毕业论文写作也得到了您的极大帮助。感谢您将我带进了化学信息学和计算生物学的大门，让我在交叉学科领域确定了自己的志趣所向，坚定地走上科研工作的道路。

感谢对我的课题进行了细致指导的张长胜老师，难以忘记与您在办公室中的定期讨论，您的帮助使得我在课题研究和毕业论文写作的过程中少走了许多弯路，在我最终完成课题研究的过程中起到了重要的作用。感谢本课题组的裴剑锋老师、王世伟师兄、李亦博师兄、林康杰师兄、黄志贤师兄、刘佳乐师兄、王凡灏师兄等组内成员的帮助，各位老师、师兄与师姐在组会上的讨论和建议让我受益良多。感谢同组的李隽仁、王应泽、钟书辰、于中天同学，与大家的讨论让我学到很多，我会怀念本科阶段与大家相处的快乐时光。

感谢我的父母，你们的支持和理解让我在大学四年间受益匪浅，感谢你们多年以来对我的抚养和教育。感谢我的舍友王崇斌、叶开和翁培壹同学，与你们一同度过的宿舍生活总是充满了亮色与笑声，希望我们都会收拾行囊、奔向各自明亮的前程。感谢我的好友韩易，多年以后当我们因不同的科研课题而通宵时，或许会想起一起下楼去全家买夜宵、从 AlphaFold 2 聊到 djent、前卫金属或当代艺术的那些夏夜。感谢化院辩论队的同学们，与大家一起打比赛的经历是我最快乐的记忆之一，和大家的讨论让我一次次重新充满活力，“我曾经垂垂老矣，而现在却风华正茂”。感谢艺双小组的各位同学，选择艺术双学位让我领略了全新的学科范式、让长久以来的梦想得以实现，而与大家一同选课、做 pre、吃吃喝喝的经历共同组成了我大学生活中难忘的独特记忆。

特别感谢我的女朋友汤一可，与你一同度过的时光充满了安宁与喜悦，让我从未如此真切地感到生命中的一切艰难都无足畏惧。言语总是拙于表达爱意本身，但每当我想起你、想起与你共度的这些时光，总是如此真诚地、不由自主地泛起微笑，憧憬着与你一同坚定而坦诚地走向未来、走向远方，与你一起延展着生命的宽度。“当我跨过沉沦的一切、向着

永恒开战的时候，你是我的军旗。”

感谢大学四年间所有课程的老师。特别感谢教授高等数学课程的谭小江老师、教授普通物理课程的穆良柱老师、教授量子力学课程的叶林晖老师、教授中级物理化学和量子化学的蒋鸿老师、教授物理化学的吴凯老师、教授数值方法课程的周铁老师和教授计算神经科学课程的吴思老师，你们悉心教授的课程让我受益匪浅，为我打开了全新的世界，对我的科研工作助益良多。感谢毕明辉老师、戴锦华老师、陈斯一老师、Matteo Ravasio 老师、唐宏峰老师、朱晓阳老师和李洋老师，你们让我一窥古典乐、电影、哲学、艺术理论或人类学的堂奥，满足了我对于兼容并包、富有人文关怀的北京大学的全部想象。感谢中国电影资料馆，四年里我曾一次次往返于小西天与北京大学之间，在荧荧的银幕前与所有观众一同分享了无数难忘的观影体验。感谢塔可夫斯基、基耶斯洛夫斯基、安哲罗普洛斯和阿巴斯，感谢 Pink Floyd、Dream Theater、Joy Division 和 King Crimson，感谢博尔赫斯、略萨、胡安·鲁尔福和托卡尔丘克。政治使人们分离，而艺术把人联合起来，建造起近神的巴别塔。

本课题只是在计算生物学领域作出了一点微不足道的贡献，未来的科研道路或许不会平坦。但是我在内心深处始终相信，深度学习的发展和“AI for Science”的理念将为生命科学的研究范式带来前所未有的革命，而我乐于成为这场革命中的一员，“看见风暴而激动如大海”。最后，请允许我以希尔伯特的墓志铭作为本篇致谢的结尾：

“Wir müssen wissen. Wir werden wissen.”

（“我们必须知道。我们终将知道。”）

# 北京大学学位论文原创性声明和使用授权说明

## 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，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。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。本声明的法律结果由本人承担。

论文作者签名：王宇哲

日期： 2022 年 6 月 2 日

## 学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：王宇哲 导师签名：张长胜

日期： 2022 年 6 月 2 日