Protein Representation Learning by Geometric Structure Pretraining

通过几何结构预训练进行蛋白质表示学习

Zuobai Zhang 1 2 Minghao Xu 1 2 Arian Jamasb 3 Vijil Chenthamarakshan 4

Aurelie Lozano´ 4 Payel Das 4 Jian Tang 1 5 6

# Abstract

Learning effective protein representations is critical in a variety of tasks in biology such as predicting protein function or structure. Existing approaches usually pretrain protein language models on a large number of unlabeled amino acid sequences and then finetune the models with some labeled data in downstream tasks. Despite the effectiveness of sequence-based approaches, the power of pretraining on smaller numbers of known protein structures has not been explored for protein property prediction, though protein structures are known to be determinants of protein function. In this paper, we propose to pretrain protein representations according to their 3D structures. We first present a simple yet effective encoder to learn protein geometry features. We pretrain the protein graph encoder by leveraging multiview contrastive learning and different selfprediction tasks. Experimental results on both function prediction and fold classification tasks show that our proposed pretraining methods outperform or are on par with the state-of-the-art sequence-based methods using much less data. All codes and models will be published upon acceptance.

学习有效的蛋白质表征在生物学的各种任务中至关重要，例如预测蛋白质功能或结构。现有方法通常在大量未标记的氨基酸序列上预先训练蛋白质语言模型，然后在下游任务中用一些标记数据微调模型。尽管基于序列的方法有效，但尚未探索对较少数量的已知蛋白质结构进行预训练的力量，以用于蛋白质特性预测，尽管已知蛋白质结构是蛋白质功能的决定因素。在本文中，我们建议根据蛋白质的3D结构对其进行预训练。我们首先提出一个简单而有效的编码器来学习蛋白质几何特征。我们通过利用多视图对比学习和不同的自我预测任务来预训练蛋白质图编码器。在函数预测和折叠分类任务上的实验结果表明，我们提出的预训练甲基苯丙胺优于或与使用更少数据的基于序列的最新方法相当。所有代码和模型将在接受后公布。

蛋白质是由氨基酸以“脱水缩合”的方式组成的多肽链经过盘曲折叠形成的具有一定[空间结构](https://baike.so.com/doc/5364411-5600015.html)的物质

# 1. Introduction

Proteins are workhorses of the cell and are implicated in a broad range of applications ranging from therapeutics to material. They consist of a linear chain of amino acids (residues) folding into a specific 3D conformation. Due to the advent of low cost sequencing technologies, in recent years a massive volume of protein sequences has been newly discovered. As the function annotation of a new protein sequence remains costly and time-consuming, accurate and efficient in silico protein function annotation methods are needed to bridge the existing sequence-function gap.

arXiv:2203.06125v2 [cs.LG] 14 Mar 2022

蛋白质是细胞的主力，涉及从治疗到材料的广泛应用。它们由折叠成特定3D构象的氨基酸（残基）的线性链组成。由于低成本测序技术的出现，近年来新发现了大量的蛋白质序列。由于新蛋白质序列的功能注释仍然昂贵且耗时，因此需要准确和高效的计算机蛋白功能注释方法来弥合现有的序列 - 功能差距。

As a large number of protein functions are governed by their folded structures, several data-driven approaches rely on learning representations of the protein structures, which then can be used for a variety of tasks such as protein design (Ingraham et al., 2019; Strokach et al., 2020; Cao et al., 2021; Jing et al., 2021), structure classification (Hermosilla et al., 2021), model quality assessment (Baldassarre et al., 2021; Derevyanko et al., 2018), and function prediction (Gligorijevic et al.´ , 2021). Unfortunately, due to the challenge of experimental protein structure determination, the number of reported protein structures is orders of magnitude lower than the size of datasets in other machine learning application domains. For example, 182K structures exist in Protein Data Bank (PDB) (Berman et al., 2000) (*vs* 47M protein sequences in Pfam (Mistry et al., 2021)) and *vs* 10M annotated images in ImageNet (Russakovsky et al., 2015).

由于大量蛋白质功能由其折叠结构控制，因此几种数据驱动的方法依赖于学习蛋白质结构的表示，然后可用于各种任务，例如蛋白质设计，结构分类，模型质量评估，和功能预测。不幸的是，由于实验蛋白质结构测定的挑战，报告的蛋白质结构的数量比其他机器学习应用领域的数据集大小低几个数量级。例如，蛋白质数据库（PDB）中存在182K结构与Pfam中的47M蛋白质序列（Mistry等人，2021）相比和图像网络中的10M注释图像（Russakovsky等人，2015）。

As a remedy to this problem, recent works have leveraged the large volume of unlabeled protein sequences data to learn an effective representation of known proteins (Bepler & Berger, 2019; 2021; Rives et al., 2021; Elnaggar et al., 2021). A number of studies have employed pretraining protein encoders on millions of sequences via self-supervised learning (Hadsell et al., 2006; Devlin et al., 2018; Chen et al., 2020). However, these methods neither directly capture nor leverage the available protein structural information that is known to be the determinants of protein function.

作为解决这个问题的补救措施，最近的工作利用大量未标记的蛋白质序列数据来学习已知蛋白质的有效表示（Bepler &Berger，2019;里夫斯等人，2021;埃尔纳加等人，2021）。许多研究通过自我监督学习在数百万个序列上使用预训练蛋白质编码器。然而，这些方法既不直接捕获也不利用已知的蛋白质功能决定因素的可用蛋白质结构信息。

To better utilize this critical structural information, several structure-based protein encoders (Hermosilla et al., 2021; Hermosilla & Ropinski, 2022) are proposed. However, due to the scarcity of protein structures, these encoders are often specifically designed for individual tasks, and it is unclear how well they generalize to other tasks. Little attempts have been made until recently to develop universal protein encoders that exploit 3D structures (Hermosilla & Ropinski, 2022) due to the above-mentioned reasons of scarcity and sparsity of protein structures. Thanks to the recent advances in highly accurate deep learning-based protein structure prediction methods (Baek et al., 2021; Jumper et al., 2021), it is now possible to efficiently predict the structure of a large number of protein sequences with reasonable confidence.

为了更好地利用这些关键的结构信息，几种基于结构的蛋白质编码器（Hermosilla等人，2021;赫莫西拉和罗平斯基，2022年）被提出。然而，由于蛋白质结构的稀缺性，这些编码器通常是专门为单个任务设计的，目前尚不清楚它们如何推广到其他任务。直到最近，由于上述蛋白质结构的稀缺性和稀疏性的原因，很少有人尝试开发利用3D结构的通用蛋白质编码器（Hermosilla &Ropinski，2022）。由于基于高精度深度学习的蛋白质结构预测方法的最新进展（Baek等人，2021;Jumper等人，2021），现在可以合理地自信地有效地预测大量蛋白质序列的结构。

Motivated by this development, we develop a universal protein encoder pretrained on the largest possible number of protein structures (both experimental and predicted) that is able to generalize to a variety of property prediction tasks. We propose a simple yet effective structure-based encoder called GeomEtry-Aware Relational Graph Neural Network (GearNet), which encodes spatial information by adding different types of sequential or structural edges and then performs relational message passing on protein residue graphs. Inspired by the recent geometry-based encoders for small molecules (Klicpera et al., 2020), we propose an edge message passing mechanism to enhance the protein structure encoder.

受这一发展的激励，我们开发了一种通用蛋白质编码器，该编码器对最大数量的蛋白质结构（实验和预测）进行了预训练，能够推广到各种属性预测任务。我们提出了一种简单而有效的基于结构的编码器，称为GeomEtry-Aware关系图神经网络（GearNet），它通过添加不同类型的顺序或结构边缘来编码空间信息，然后在蛋白质残基图上执行关系消息传递。受最近基于几何的小分子编码器（Klicpera等人，2020）的启发，我们提出了一种边缘消息传递机制来增强蛋白质结构编码器。

We further introduce different geometric pretraining methods for learning the protein structure encoder by following popular self-supervised learning frameworks such as contrastive learning and self-prediction. In the contrastive learning scenario, we aim to maximize the similarity between the learned representations of different augmented views from the same protein, while minimizing the similarity between those from different proteins. For self-prediction, the model performs masked prediction of different geometric or biochemical properties, such as residue types, Euclidean distances, angles and dihedral angles, during training.

我们通过遵循流行的自我监督学习框架（如对比学习和自我预测）进一步介绍了用于学习蛋白质结构编码器的不同几何预训练方法。在对比学习场景中，我们的目标是最大化来自同一蛋白质的不同增强视图的学习表示之间的相似性，同时最小化来自不同蛋白质的视图之间的相似性。对于自我预测，该模型在训练期间对不同的几何或生化性质（例如残基类型，欧氏距离，角度和二面角）执行掩蔽预测。

Extensive experiments on several existing benchmarks, including Enzyme Commission number prediction (Gligorijevic et al.´ , 2021), Gene Ontology term prediction (Gligorijevic et al.´ , 2021), fold classification (Hou et al., 2018) and reaction classification (Hermosilla et al., 2021) verify our GearNet augmented with edge message passing can consistently and significantly outperform existing protein encoders on most tasks in a supervised setting. Further, by employing the proposed pretraining methods, our model trained on less than a million samples achieves state-of-the-art results on different tasks, even better than the sequence-based encoders pretrained on million- or billion-scale datasets.

对几个现有基准的广泛实验，包括酶委员会数量预测（Gligorijevic等人，2021年），基因本体学术语预测（格利戈列维奇等人，2021年），折叠分类（侯等人，2018年）和反应分类（Hermosilla等人，2021年）验证了我们的GearNet增强边缘消息传递可以在监督设置中的大多数任务上一致且显着优于现有的蛋白质编码器。此外，通过采用所提出的预训练方法，我们在不到一百万个样本上训练的模型在不同的任务上实现了最先进的结果，甚至优于在百万或十亿级数据集上预训练的基于序列的编码器。

# 2. Related Work

Previous works seek to learn protein representations based on different modalities of proteins, including amino acid sequences (Rao et al., 2019; Elnaggar et al., 2021; Rives et al., 2021), multiple sequence alignments (MSAs) (Rao et al., 2021; Biswas et al., 2021; Meier et al., 2021) and protein structures (Hermosilla et al., 2021; Gligorijevic et al.´ , 2021; Somnath et al., 2021). These works share the common goal of learning informative protein representations that can benefit various downstream applications, like predicting protein function (Gligorijevic et al.´ , 2021; Rives et al., 2021) and protein-protein interaction (Wang et al., 2019), as well as designing protein sequences (Biswas et al., 2021).

以前的工作试图根据蛋白质的不同模式（包括氨基酸序列）学习蛋白质表示（Rao等人，2019;埃尔纳加等人，2021;里夫斯等人，2021），多序列比对（MSA）（Rao等人，2021;比斯瓦斯等人，2021;迈尔等人，2021）和蛋白质结构（埃莫西拉等人，2021;格里戈里耶维奇等人'，2021;索姆纳特等人，2021）。这些工作的共同目标是学习信息丰富的蛋白质表示，这些蛋白质表示可以有益于各种下游应用，例如预测蛋白质功能（Gligorijevic等人，2021;Rives等人，2021）和蛋白质 - 蛋白质相互作用（Wang等人，2019），以及设计蛋白质序列（Biswas等人，2021）。

2.1. Sequence-Based Methods

Sequence-based protein representation learning is mainly inspired by the methods of modeling natural language sequences. Recent methods aim to capture the biochemical and co-evolutionary knowledge underlying a large-scale protein sequence corpus by self-supervised pretraining, and such knowledge is then transferred to specific downstream tasks by finetuning. Typical pretraining objectives explored in existing methods include next amino acid prediction (Alley et al., 2019; Elnaggar et al., 2021), masked language modeling (MLM) (Rao et al., 2019; Elnaggar et al., 2021; Rives et al., 2021), pairwise MLM (He et al., 2021) and contrastive predictive coding (CPC) (Lu et al., 2020). Compared to sequence-based approaches that learn in the whole sequence space, MSA-based methods (Rao et al., 2021; Biswas et al., 2021; Meier et al., 2021) leverage the sequences within a protein family to capture the conserved and variable regions of homologous sequences, which imply specific structures and functions of the protein family.

2.1. 基于序列的方法 基于序列的蛋白质表示学习主要受到自然语言序列建模方法的启发。最近的方法旨在通过自我监督的预训练来捕获大规模蛋白质序列语料库背后的生化和共进化知识，然后通过微调将这些知识转移到特定的下游任务中。现有方法中探索的典型预训练目标包括下一个氨基酸预测（Alley等人，2019;埃尔纳加等人，2021），蒙面语言建模（MLM）（Rao等人，2019;埃尔纳加等人，2021;Rives等人，2021年），成对MLM（他等人，2021年）和对比预测编码（CPC）（Lu等人，2020年）。与在整个序列空间中学习的基于序列的方法相比，基于MSA的方法（Rao等人，2021;比斯瓦斯等人，2021;Meier等人，2021）利用蛋白质家族中的序列来捕获同源序列的保守和可变区域，这意味着蛋白质家族特殊的架构和功能。

2.2. Structure-Based Methods基于结构的方法

Although sequence-based methods pretrained on large-scale databases perform well, structure-based methods should be, in principle, a better solution to learning an informative protein representation, as the function of a protein is determined by its structure. This line of works seeks to encode spatial information in protein structures by 3D CNNs (Derevyanko et al., 2018) or graph neural networks (GNNs) (Gligorijevic et al.´ , 2021; Baldassarre et al., 2021; Jing et al., 2021). Among these methods, IEConv (Hermosilla et al., 2021) tries to fit the inductive bias of protein structure modeling, which introduced a graph convolution layer incorporating intrinsic and extrinsic distances between nodes. Another potential direction is to extract features from protein surfaces (Gainza et al., 2020; Sverrisson et al., 2021; Dai & Bailey-Kellogg, 2021). Somnath et al. (2021) combined the advantages of both worlds and proposed a parameterefficient multi-scale model. Besides, there are also works that enhance pretrained sequence-based models by incorporating structural information in the pretraining stage (Bepler & Berger, 2021) or finetuning stage (Wang et al., 2021).

尽管在大规模数据库上预先训练的基于序列的方法表现良好，但原则上，基于结构的方法应该是学习信息性蛋白质表示的更好解决方案，因为蛋白质的功能由其结构决定。这一系列工作试图通过3D CNN（Derevyanko等人，2018）或图神经网络（GNN）（格里戈里耶维奇等人）在蛋白质结构中编码空间信息，2021;巴尔达萨尔等人，2021;荆等人，2021）。在这些方法中，IEConv（Hermosilla等人，2021）试图拟合蛋白质结构建模的归纳偏差，其引入了一个包含节点之间内在和外在距离的图卷积层。另一个潜在的方向是从蛋白质表面中提取特征（Gainza等人，2020;斯维里森等人， 2021;戴和贝利-凯洛格，2021）。Somnath等人（2021）结合了两个世界的优点，提出了一种参数效率高的多尺度模型，此外，还有一些作品通过在预训练阶段（Bepler &Berger，2021）或微调阶段（Wang等人，2021）中结合结构信息来增强预训练的基于序列的模型。

Despite progress in the design of structure-based encoders, there are few works focusing on structure-based pretraining for proteins. To the best of our knowledge, the only attempt is a concurrent work (Hermosilla & Ropinski, 2022), which applies a contrastive learning method on an encoder (Hermosilla et al., 2021) relying on cumbersome convolutional and pooling layers. Compared with these existing works, our proposed encoder is conceptually simpler and more effective on many different tasks, thanks to the proposed relational graph convolutional layer and edge message passing layer, which are able to efficiently capture the sequential and structural information. Furthermore, we introduce five structure-based pretraining methods within the contrastive learning and self-prediction frameworks, which can serve as a solid starting point for enabling self-supervised learning on protein structures.

尽管基于结构的编码器在设计方面取得了进展，但很少有工作专注于基于结构的蛋白质预训练。据我们所知，唯一的尝试是并发工作（Hermosilla &Ropinski，2022），它依靠繁琐的卷积和池化层在编码器上应用对比学习方法（Hermosilla等人，2021）。与这些现有工作相比，我们提出的编码器在概念上更简单，在许多不同的任务上更有效，这要归功于所提出的关系图卷积层和边缘消息传递层，它们能够有效地捕获顺序和结构信息。此外，我们在对比学习和自我预测框架中引入了五种基于结构的预训练方法，可以作为实现蛋白质结构自我监督学习的坚实起点。

# 3. Structure-Based Protein Encoder

Existing protein encoders are either designed for specific tasks or cumbersome for pretraining due to the dependency on computationally expensive convolutions. In contrast, here we propose a simple yet effective protein structure encoder, named *GeomEtry-Aware Relational Graph Neural Network (GearNet缩写，即几何关系感知图神经网络)*. We utilize *edge message passing* to enhance the effectiveness of GearNet, which is novel in the field of protein structure modeling. In contrast, previous works (Hermosilla et al., 2021; Somnath et al., 2021) only consider message passing among residues or atoms.

现有的蛋白质编码器要么是为特定任务而设计的，要么是由于依赖于计算成本高昂的卷积而难以进行预训练。相比之下，在这里我们提出了一个简单而有效的蛋白质结构编码器，名为GeomEtry感知关系图神经网络（GearNet）。我们利用边缘消息传递来增强GearNet的有效性，这在蛋白质结构建模领域是新颖的。相比之下，以前的作品（埃莫西拉等人，2021;Somnath等人，2021）只考虑在残基或原子之间传递信息。

3.1. Geometry-Aware Relational Graph Neural Network

Given protein structures, our model aims to learn representations encoding their spatial and chemical information. These representations should be invariant under translations and rotations in 3D space. To achieve this requirement, we first construct our protein graph based on the spatial features invariant under these transformations.

3.1. 几何感知关系图神经网络 给定蛋白质结构，我们的模型旨在学习编码其空间和化学信息的表示。这些表示在 3D 空间中的平移和旋转下应该是不变的。为了满足这一要求，我们首先基于这些变换下不变的空间特征构建蛋白质图。

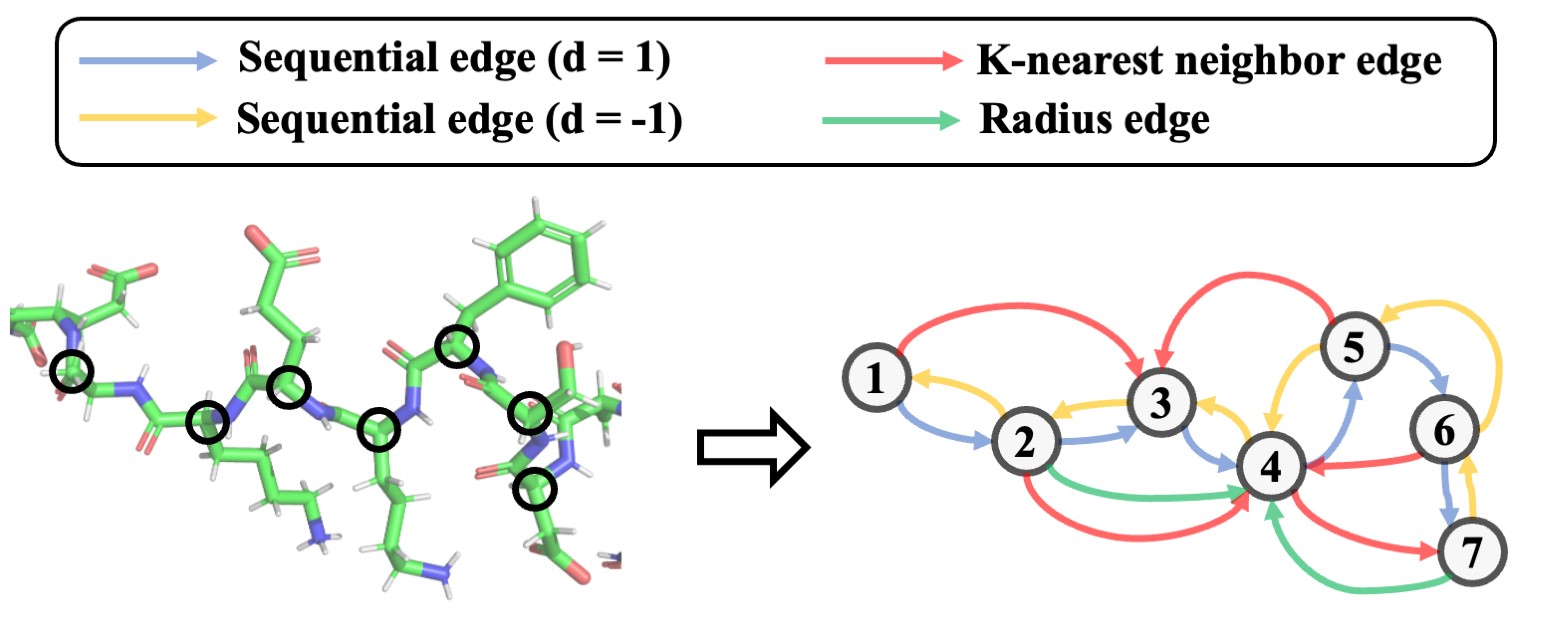


Figure 1: Protein residue graph construction. Sequential, radius and knn edges are added into the graph and treated as different types. Some edges are omitted to save space.

Protein graph construction. We represent the structure of a protein as a relational graph G = (V*,*E*,*R), where V and E denotes the set of nodes and edges respectively, and R is the set of edge types. We use (*i,j,r*) to denote the edge from node *i* to node *j* with type *r*. We use *n* and *m* to denote the number of nodes and edges, respectively. In this work, each node in the protein graph represents the alpha carbon of a residue with the 3D coordinates of all nodes *x* ∈ R*n*×3.

图1：蛋白质残基图结构。连续边、半径边和K-近邻边被添加到图形中，并被视为不同的类型。省略某些边以节省空间。蛋白质图结构。我们将蛋白质的结构表示为关系图G = （V，E，R），其中V和E分别表示节点和边缘的集合，R是边缘类型的集合。我们使用 （i，j，r） 来表示从节点 i 到节点 j 的边，类型为 r。我们分别使用 n 和 m 来表示节点和边的数量。在这项工作中，蛋白质图中的每个节点都表示残基的α碳，其所有节点的3D坐标x∈Rn×3。（n行3列的矩阵）

Then, we add three different types of edges into our graphs: sequential edges, radius edges and K-nearest neighbor edges. Among these, sequential edges will be further divided into different edge types according to the relative position *d* between two end nodes. These edge types reflect different geometric properties, which all together yield a comprehensive featurization of proteins. The graph construction process is shown in Fig. 1 and more details can be found in Appendix C.1.

然后，我们在图形中添加三种不同类型的边：连续边、半径边和 K-最近邻边。其中，顺序边将根据两个端点节点之间的相对位置d进一步划分为不同的边类型。这些边缘类型反映了不同的几何特性，它们共同产生了蛋白质的全面特征。图形构建过程如图1所示，更多细节可以在附录C.1中找到。

Node and edge features. Most previous structure-based encoders designed for biological molecules used many chemical and spatial features, some of which are difficult to obtain or time-consuming to calculate. In contrast, we only use the one-hot encoding of residue types with one additional dimension for unknown types as node features, denoted as *f* ∈ {0*,*1}*n*×21, which is enough to learn good representation as shown in our experiments.

节点和边缘要素。以前为生物分子设计的大多数基于结构的编码器都使用了许多化学和空间特征，其中一些特征难以获得或计算起来很耗时。相比之下，我们只使用残基类型的独热编码，对于未知类型，有一个额外的维度作为节点特征，表示为f ∈{0，1}n×21，这足以学习我们的实验中所示的良好表示。

The feature *f*(*i,j,r*) for an edge (*i,j,r*) is the concatenation of the node features of two end nodes, the one-hot encoding of the edge type, and the sequential and spatial distances between them:

*f*(*i,j,r*) = Cat(*fi,fj,*onehot(*r*)*,*|*i* − *j*|*,*k*xi* − *xj*k2)*,*

(1)

where Cat(·) denotes the concatenation operation.

边 （i，j，r） 的特征 f（i，j，r） 是两个终端节点的节点特征的串联、边类型的[独热编码](https://www.cnblogs.com/zongfa/p/9305657.html)以及它们之间的顺序和空间距离：f（i，j，r） = Cat（fi，fj，onehot（r），|i − j|，kxi − xjk2），（1） 其中 Cat（·） 表示[串联](https://blog.csdn.net/weixin_45770896/article/details/121246660)操作。

Relational graph convolutional layer. Upon the protein graphs defined above, we utilize a GNN to derive per-residue and whole-protein representations. One simple example of GNNs is the GCN (Kipf & Welling, 2017), where messages are computed by multiplying node features with a convolutional kernel matrix shared among all edges. To increase the capacity of GCN in protein structure modeling, IEConv (Hermosilla et al., 2021) proposed to apply a learnable kernel function on edge features. In this way, *m*

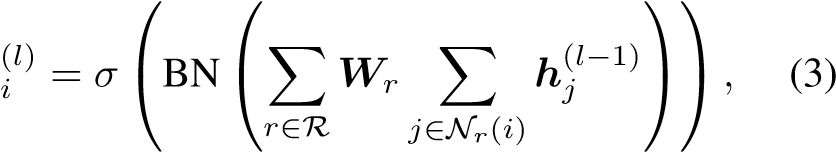
different kernel matrices can be applied on different edges, which achieves good performance but induces high memory costs. To balance model capacity and memory cost, we use a relational graph convolutional neural network (Schlichtkrull et al., 2018) to learn graph representations, where a convolutional kernel matrix is shared within each edge type and there are |R| different kernel matrices in total.

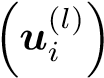
关系图卷积层。在上面定义的蛋白质图上，我们利用GNN来推导每个残基和全蛋白质表示。GNN的一个简单例子是GCN(图卷积神经网络)（Kipf和Welling，2017），其中消息的计算方法是将节点特征与所有边缘共享的卷积核矩阵相乘。为了提高GCN在蛋白质结构建模中的能力，IEConv（Hermosilla等人，2021）建议在边缘特征上应用可学习的核函数。通过这种方式，可以在不同的边缘上应用m个不同的内核矩阵，从而实现良好的性能，但会导致高内存成本。为了平衡模型容量和内存成本，我们使用关系图卷积神经网络（Schlichtkrull等人，2018）来学习图表示，其中卷积核矩阵在每个边缘类型中共享，并且存在|R|总共不同的内核矩阵。

Formally, the relational graph convolutional layer used in our model is defined as

形式上，我们模型中使用的关系图卷积层定义为

*h**,* (2)

*u*

*h*Dropout*.* (4)

Specifically, we first apply a linear transformation FC(·) and a batch norm layer BN(·) on node features to get the initial representation for each node. Then, given the input node representation *h* for node *i* at the *l*-th layer, we compute updated node representation  by aggregating features from neighboring nodes N*r*(*i*), where N*r*(*i*) = {*j* ∈ V|(*j,i,r*) ∈ E} denotes the neighborhood of node *i* with respect to the edge type *r* and *Wr* denotes the learnable convolutional kernel matrix for edge type *r*. Here we use a ReLU function as the activation *σ*(·). Finally, we apply the dropout on the update  and add a residual connection from the previous layer.

具体来说，我们首先在节点特征上应用线性变换FC（·）和批处理范数层BN（·）来获得每个节点的初始表示。然后，给定第 l 层节点 i 的输入节点表示 h/，我们通过聚合来自相邻节点 Nr（i） 的特征来计算更新的节点表示 /，其中 Nr（i） = {j ∈ V|（j，i，r） ∈ E} 表示节点 i 相对于边缘类型 r 的邻域，而 Wr 表示边缘类型 r 的可学习卷积核矩阵。这里我们使用 ReLU 函数作为激活σ（·）。最后，我们在更新/上应用掉线，并添加来自前一层的残余连接。

3.2. Edge Message Passing Layer

As in the literature of molecular representation learning, many geometric encoders show benefits from explicitly modeling interactions between edges. For example, DimeNet (Klicpera et al., 2020) uses a 2D spherical FourierBessel basis function to represent angles between two edges and pass messages between edges. Inspired by this mechanism, we propose a variant of our structure-based encoder enhanced with an edge message passing layer.

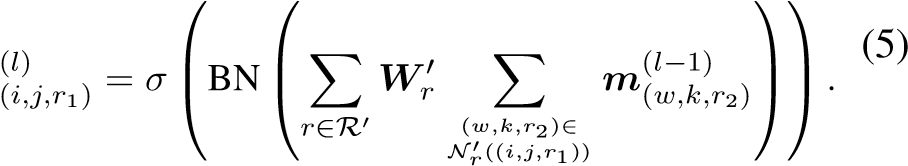
We first construct a relational graph G0 = (V0*,*E0*,*R0) among edges, which is also known as *line graph* in the literature (Harary & Norman, 1960). Each node in the graph G0 corresponds to an edge in the original graph. G0 links edge (*i,j,r*1) in the original graph to edge (*w,k,r*2) if and only if *j* = *w* and *i* 6= *k*. The type of this edge is determined by the angles between (*i,j,r*1) and (*w,k,r*2). We discretize the range [0*,π*] into 8 bins and use the index of the bin as the edge type.

3.2. 边缘消息传递层 正如在分子表示学习的文献中，许多几何编码器都显示出显式建模边缘之间相互作用的好处。例如，DimeNet（Klicpera等人，2020）使用2D球面傅里叶贝塞尔基函数来表示两条边之间的角度并在边之间传递消息。受此机制的启发，我们提出了一种基于结构的编码器的变体，该编码器通过边缘消息传递层进行了增强。我们首先在边缘之间构造一个关系图G0 = （V0，E0，R0），这在文献中也称为折线图（Harary &Norman，1960）。图形 G0 中的每个节点对应于原始图形中的一条边。G0 将原始图形中的边 （i，j，r1） 链接到边 （w，k，r2），当且仅当 j = w 和 i 6= k。此边的类型由 （i，j，r1） 和 （w，k，r2） 之间的角度决定。我们将范围 [0，π] 离散化为 8 个条柱，并使用条柱的索引作为边缘类型。

Then, we apply a similar relational graph convolutional network on the graph G0 to obtain the message function for each edge. Formally, the edge message passing layer is defined as

然后，我们在图G0上应用类似的关系图卷积网络，以获得每个边缘的消息函数。形式上，边缘消息传递层定义为

*m**f*(*i,j,r*1)*,*

*m*

(*l*)

Here we use *m*(*i,j,r*1) to denote the message function for edge (*i,j,r*1) in the *l*-th layer. Similar as Eq. (3), the message function for edge (*i,j,r*1) will be updated by aggregating features from its neighbors N*r*0((*i,j,r*1)), where N*r*0((*i,j,r*1)) = {(*w,k,r*2) ∈ V0|((*w,k,r*2)*,*(*i,j,r*1)*,r*) ∈ E0} denotes the set of incoming edges of (*i,j,r*1) with relation type *r* in graph G0.

Finally, we replace the aggregation function Eq. (3) in the original graph with the following one:

这里我们使用 m（i，j，r1） 来表示第 l 层中边缘 （i，j，r1） 的消息函数。与等式 （3） 类似，边缘 （i，j，r1） 的消息函数将通过聚合来自其邻居 Nr0（（i，j，r1）） 的特征来更新，其中 Nr0（（i，j，r1）） = {（w，k，r2） ∈ V0|（（w，k，r2），（i，j，r1），r） ∈ E0} 表示图 G0 中关系类型为 r 的 （i，j，r1） 的传入边的集合。最后，我们将原始图中的聚合函数 Eq. （3） 替换为以下函数：



This variant of GearNet with edge message passing mechanism is referred as GearNet-Edge in the sequel.

Notably, though most components in our model are adapted from encoders designed for small molecules, the idea of using a relational graph to model spatial information within protein structures is novel. In addition, to the best of our knowledge, this is the first work that explores edge message passing for protein representation learning.

这种具有边缘消息传递机制的齿轮网变体在续集中被称为齿轮网边缘。值得注意的是，尽管我们模型中的大多数组件都是由为小分子设计的编码器改编的，但使用关系图来模拟蛋白质结构中的空间信息的想法是新颖的。此外，据我们所知，这是第一部探索蛋白质表示学习的边缘消息传递的工作。

# 4. Geometric Pretraining Methods

As a large amount of unlabeled protein structures exist, which are potentially helpful to protein representation learning, we study how to leverage these unlabeled structures. We follow two popular self-supervised learning frameworks: contrastive learning and self-prediction, and propose five different pretraining strategies for structure-based encoders

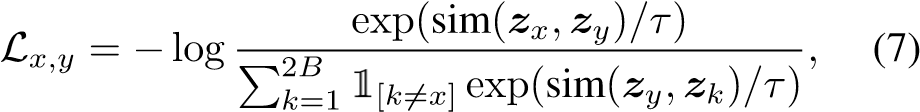
几何预训练方法 由于存在大量未标记的蛋白质结构，这些结构可能对蛋白质表示学习有帮助，因此我们研究了如何利用这些未标记的结构。我们遵循两种流行的自我监督学习框架：对比学习和自我预测，并为基于结构的编码器提出了五种不同的[预训练](https://blog.csdn.net/weixin_48192256/article/details/120937144)策略。

4.1. Multiview Contrastive Learning多视图对比学习

Inspired by recent contrastive learning methods (Oord et al., 2018; Chen et al., 2020; He et al., 2020), our framework learns representations by maximizing the similarity between representations of different augmented views of the same protein while minimizing the agreement between views of different proteins. The high-level idea is illustrated in Fig. 2.

More specifically, given a protein graph G, we first sample two different views G*x* and G*y* via a stochastic augmentation module. We then compute the graph representations *hx* and *hy* of the two views using our structure-based encoder. Following SimCLR (Chen et al., 2020), a two-layer projection head is further applied to map the representations to lower-dimensional space, denoted as *zx* and *zy*. Finally, a contrastive loss function is defined by distinguishing views from the same or different proteins using their similarities. For a positive pair *x* and *y*, we treat views from other proteins in the same mini-batch as negative pairs. Mathematically, the loss function for a positive pair of views *x* and *y* can be written as:

受到最近对比学习方法的启发，我们的框架通过最大化同一蛋白质的不同增强视图的表示之间的相似性来学习表示，同时最小化不同蛋白质视图之间的一致性。高层次的想法如图2所示。更具体地说，给定一个蛋白质图G，我们首先通过随机增强模块对两个不同的视图Gx和Gy进行采样。然后，我们使用基于结构的编码器计算两个视图的图形表示 hx 和 hy。在SimCLR（Chen等人，2020）之后，进一步应用双层投影头将表示映射到低维空间，表示为zx和zy。最后，对比损失函数的定义是通过使用它们的相似性从相同或不同的蛋白质中区分观点。对于正对 x 和 y，我们将来自同一小批次中其他蛋白质的观点视为负对。在数学上，正对视图 x 和 y 的损失函数可以写为：



where *B* denotes the batch size, 1[*k*6=*x*] ∈ {0*,*1} is an indicator function that is equal to 1 if and only if *k* 6= *x*. And the similarity function sim(*u,v*) is defined by the cosine similarity between *u* and *v*.

其中 B 表示批大小，1[k6=x] ∈ {0，1} 是一个指示符函数，当且仅当 k 6= x 时等于 1。相似性函数 sim（u，v） 由 u 和 v 之间的余弦相似性定义。

Data augmentation. As suggested in Chen et al. (2020), diverse data augmentation schemes play an important role in contrastive learning. In this work, we propose a principled way to generate diverse views of a protein.

We first randomly choose a *cropping* function to make the protein graph smaller so that large-size batches can be used for pretraining. Here we consider two different cropping functions:

数据增强。正如Chen等人（2020）所建议的那样，不同的数据增强方案在对比学习中起着重要作用。在这项工作中，我们提出了一种有原则的方式来产生蛋白质的不同观点。我们首先随机选择一个裁剪函数，使蛋白质图更小，以便大批量可用于预训练。在这里，我们考虑两种不同的裁剪函数：

|  |
| --- |
| Figure 2: Demonstration of geometric pretraining methods. For multiview contrastive learning, we aim to align representations of different views from the same protein while minimizing the similarity between those from different proteins. For self-prediction methods, we construct four masked prediction objectives by inferring masked geometric or biochemical |

quantities with learned representations.

几何预训练方法的演示。对于多视图对比学习，我们的目标是对齐来自同一蛋白质的不同观点的表示，同时最大限度地减少来自不同蛋白质的观点之间的相似性。对于自我预测方法，我们通过用学习的表示推断被掩盖的几何或生化量来构建四个被屏蔽的预测目标。

* Subsequence: randomly sample a consecutive segment of protein sequences and take the corresponding subgraph from the protein residue graph.
* Subspace: randomly sample a residue as the center and select all residues within a Euclidean ball with a predefined radius.

Then, we randomly choose one of the following two transformations with equal probability to apply on the cropped protein graphs.

* Identity: no transformation.
* Random edge masking: randomly mask edges with a fixed mask rate equal to 0*.*15.

子序列：随机采样蛋白质序列的连续片段，并从蛋白质残基图中获取相应的子图。

子空间：随机采样一个残基作为中心，并选择欧几里得球内具有预定义半径的所有残基。

然后，我们随机选择以下两个转换中的一个，具有相等的概率应用于裁剪后的蛋白质图。

身份：无转换。

随机边缘遮罩：随机遮罩固定遮罩速率等于 0.15 的边缘。

4.2. Self-Prediction Methods自我预测方法

Another line of research is based on the recent progress of self-prediction methods in natural language processing (Devlin et al., 2018; Brown et al., 2020). Given a protein, our objective is to predict one part of the protein given the remainder of the structure. Here, we propose four self-supervised tasks based on geometric or biochemical properties.

Residue Type Prediction. Our first method is based on the masked language modeling objective, which has been widely used in pretraining large-scale protein language models (Bepler & Berger, 2021). For each protein, we randomly mask the types (*i.e.*, node features) of some residues and then let the structure-based encoders to predict these masked residue types. For a masked node *i*, the learning objective is defined as:

另一条研究线是基于自然语言处理中自我预测方法的最新进展（Devlin等人，2018;布朗等人，2020）。给定一种蛋白质，我们的目标是在给定结构的其余部分的情况下预测蛋白质的一部分。在这里，我们提出了四个基于几何或生化特性的自我监督任务。

残差类型预测。我们的第一种方法基于掩码语言建模目标，该目标已广泛用于预训练大规模蛋白质语言模型（Bepler &Berger，2021）。对于每种蛋白质，我们随机掩蔽一些残基的类型（即节点特征），然后让基于结构的编码器预测这些被掩蔽的残基类型。对于屏蔽节点 i，学习目标定义为：

L*i* = CE(*f*residue*,*

where *f*residue(·) is an MLP classification head and CE denotes the cross entropy loss. This method is also known as Attribute Masking in the literature of pretraining on small molecules (Hu et al., 2019).

其中，弗雷西杜（·）是MLP分类头，CE表示交叉熵损失。这种方法在小分子预训练的文献中也称为属性掩蔽（Hu等人，2019）。

Distance Prediction. In order to learn local spatial structures, we use our learned representations to predict the Euclidean distance between two nodes connected in the protein graph. First, we randomly select a fixed number of edges from the original graph. Then, these edges will be removed when feeding the graph into the encoder. For a selected edge, the representations of its two end nodes will be concatenated to predict the distance between them. More concretely, the loss function for an edge (*i,j,r*) will be defined as:

距离预测。为了学习局部空间结构，我们使用学习的表示来预测蛋白质图中连接的两个节点之间的<欧几里得距离>。首先，我们从原始图形中随机选择固定数量的边。然后，将图形输入编码器时，这些边将被删除。对于所选边，将连接其两个端点的表示以预测它们之间的距离。更具体地说，边 （i，j，r） 的损失函数将定义为：

L(*i,j,r*) = (*f*dist(*h*0*i,h*0*j*) − k*xi* − *xj*k2)2*,*

where *f*dist(·) is an MLP prediction head and *h*denotes the representations of node *i* and *j* after masking selected edges.

其中 fdist（·） 是 MLP 预测头，h/ 表示在屏蔽所选边后节点 i 和 j 的表示。

Angle Prediction. Besides distances, angles between edges are also important features that reflect the relative position between residues. Similarly, we can define a masked geometric loss by 1) randomly selecting two adjacent edges, 2) removing these edges from the graph, 3) using the three end nodes of the pair of edges sharing a single end node to predict the angles between them. Here we discretized the angles by cutting the range [0*,π*] into 8 bins. The objective for a selected pair of edges (*i,j,r*1) and (*j,k,r*2) is to predict which bin the angle between them will belong to:

角度预测。除了距离之外，边之间的角度也是反映残留物之间相对位置的重要特征。类似地，我们可以通过以下方式定义一个屏蔽的几何损失：1）随机选择两条相邻的边，2）从图形中删除这些边，3）使用共享单个端节点的一对边的三个端点节点来预测它们之间的角度。在这里，我们通过将范围 [0，π] 切成 8 个箱来离散化角度。所选边对 （i，j，r1） 和 （j，k，r2） 的目标是预测它们之间的角度将属于哪个条柱：

*,*

where *f*angle(·) is an MLP classification head and bin(·) is used to discretize the angle.

其中扇形（·）是MLP分类头，bin（·）用于离散化角度。

Dihedral Prediction. As shown in Klicpera et al. (2021), the dihedral angles between three edges can provide important clues about the relative directional information. Therefore, we can also construct a masked geometric objective by predicting the dihedral angles between three consecutive edges (*i,j,r*1)*,*(*j,k,r*2) and (*k,t,r*3):

二面体预测。如Klicpera等人（2021）所示，三条边之间的二面角可以提供关于相对方向信息的重要线索。因此，我们还可以通过预测三个连续边 （i，j，r1）、（j，k，r2） 和 （k，t，r3） 之间的二面角来构造一个屏蔽的几何目标：

*,*

where *f*(·) is an MLP classification head and bin(·) is used to discretize the dihedral angles.

The framework of self-prediction methods is depicted in Fig. 2. In the sequel, we will treat the tasks above as four different pretraining methods.

其中 f（·） 是 MLP 分类头，bin（·） 用于离散二面角。自我预测方法的框架如图2所示。在续集中，我们将上述任务视为四种不同的预训练方法。

# 5. Experiments实验

In this section, we first introduce our experimental setup for pretraining and then evaluate our models on four standard downstream tasks including Enzyme Commission number prediction, Gene Ontology term prediction, fold classification and reaction classification. More experiments on protein engineering tasks can be found in Appendix G.

在本节中，我们首先介绍用于预训练的实验设置，然后评估四个标准下游任务的模型，包括酶委员会数预测，基因本体项预测，折叠分类和反应分类。有关蛋白质工程任务的更多实验可以在附录G中找到。

5.1. Pretraining Setup预训练建立

We use the AlphaFold protein structure database (Jumper et al., 2021; Varadi et al., 2021) for pretraining. This database contains the protein structures predicted by the AlphaFold2 model, and we employ both 365K proteomewide predictions and 440K Swiss-Prot (Consortium, 2021) predictions in our experiments. In Appendix F, we further report the results of pretraining on different datasets.

For pretraining, the model is trained on the AlphaFold protein database for 50 epochs. All these models are trained on 4 Tesla A100 GPUs. See Appendix E.3 for more details.

我们使用阿尔法折叠蛋白质结构数据库，用于预训练。该数据库包含AlphaFold2模型预测的蛋白质结构，我们在实验中采用了365K蛋白质组范围的预测和440K Swiss-Prot（联盟，2021）预测。在附录F中，我们进一步报告了对不同数据集的预训练结果。对于预训练，模型在AlphaFold蛋白质数据库上训练50个epoch。所有这些型号都在4个特斯拉A100 GPU上训练。详情见附录E.3。

5.2. Downstream Tasks 1 & 2: EC and GO Prediction下游任务 1 & 2：EC 和 GO 预测

We first adopt two tasks proposed in Gligorijevic et al.´

(2021) for downstream evaluation. Enzyme Commission (EC) number prediction seeks to predict the EC numbers of different proteins, which describe their catalysis of biochemical reactions. The EC numbers are selected from the third and fourth levels of the EC tree, forming 538 binary classification tasks. Gene Ontology (GO) term prediction aims to predict whether a protein belongs to some GO terms. These terms classify proteins into hierarchically related functional classes organized into three ontologies: molecular function (MF), biological process (BP) and cellular component (CC).

We follow the split method in Gligorijevic et al.´ (2021) to ensure that the test set only contains PDB chains with sequence identity no more than 95% to the training set.

我们首先采用格里戈里耶维奇等人（2021）中提出的两项任务进行下游评估。酶委员会（EC）数预测旨在预测不同蛋白质的EC数，其描述它们对生化反应的催化作用。EC 编号是从 EC 树的第三和第四级中选择的，形成 538 个二元分类任务。基因本体（GO）项预测旨在预测蛋白质是否属于某些GO项。这些术语将蛋白质分为分层相关的功能类别，分为三个本体：分子功能（MF），生物过程（BP）和细胞组分（CC）。我们遵循Gligorijevic等人（2021）中的拆分方法，以确保测试集仅包含序列标识不超过训练集95%的PDB链。

Baselines. Following Wang et al. (2021), we compare our encoders with many existing protein representation learning methods, including four sequence-based encoders (CNN (Shanehsazzadeh et al., 2020), ResNet (Rao et al., 2019), LSTM (Rao et al., 2019) and Transformer (Rao et al., 2019)), four structure-based encoders (GAT (Velickoviˇ c´ et al., 2018), GVP (Jing et al., 2021), DeepFRI (Gligorijevic et al.´ , 2021) and New IEConv (Hermosilla & Ropinski, 2022)). We also include three models pretrained on largescale sequence datasets (ProtBERT-BFD (Elnaggar et al.,

2021), ESM-1b (Rives et al., 2021) and LM-GVP (Wang et al., 2021)).

基线。继Wang等人（2021）之后，我们将编码器与许多现有的蛋白质表示学习方法进行了比较，包括四种基于序列的编码器（CNN（Shanehsazzadeh等人，2020年），ResNet（饶等人，2019年），LSTM（饶等人，2019年）和变压器（饶等人，2019年）），四种基于结构的编码器（GAT（Velickoviˇ c' 等人，2018年），GVP（荆等人，2021年），DeepFRI（格里戈里耶维奇等人，2019年），四种基于结构的编码器（Velickoviˇ c' 等人，2018年），GVP（荆等人，2021年），DeepFRI（格里戈里耶维奇等人。' 2021）和新IEConv（赫莫西拉和罗宾斯基，2022））。我们还包括三个在大规模序列数据集上预训练的模型（ProtBERT-BFD（埃尔纳加等人，2021年），ESM-1b（里夫斯等人，2021年）和LM-GVP（王等人，2021年））。

Training and evaluation. For comparison, we train GearNet and GearNet-Edge from scratch and also include the results of GearNet-Edge pretrained with five proposed methods. The models are trained for 200 epochs. We evaluate the performance with two metrics commonly used in the CAFA challenges (Radivojac et al., 2013): (1) proteincentric maximum F-score Fmax; (2) pair-centric area under precision-recall curve AUPRpair (See Appendix E.2 for details). Models are selected based on the Fmax on validation sets.

培训和评估。为了进行比较，我们从头开始训练齿轮网和齿轮网边缘，还包括使用五种建议的方法预先训练的齿轮网边缘的结果。这些模型经过 200 个 epoch 的训练。我们用中央美术学院挑战赛中常用的两个指标来评估表现（Radivojac等人，2013）：（1）以蛋白质为中心的最大F得分Fmax;（2）精确召回率曲线AUPPair下的对中心区域（详见附录E.2）。根据验证集上的 Fmax 选择模型

Results. Table 1 summarizes the results on EC and GO prediction. For encoders without pretraining, we find that the vanilla GearNet can already obtain competitive results with other baselines. After adding the edge message passing mechanism, our method GearNet-Edge significantly outperforms other baselines on EC, GO-BP and GO-MF and achieves competitive results on GO-CC. With pretraining , all five proposed variants show large improvements over models trained from scratch. After pretraining with the best among these methods, *i.e.*, Multiview Contrast, our model can achieve state-of-the-art results on EC and GO-BP and competitive results on GO-MF against pretrained sequencebased encoders. It should be noted that our models are pretrained on a dataset with fewer than one million structures, whereas all three language model based baselines are pretrained on million- or billion-scale sequence databases.

结果。表 1 总结了 EC 和 GO 预测的结果。对于没有预训练的编码器，我们发现vanilla GearNet已经可以获得与其他基线相比具有竞争力的结果。在添加边缘消息传递机制后，我们的方法 GearNet-Edge 在 EC、GO-BP 和 GO-MF 上的表现显著优于其他基线，并在 GO-CC 上取得了具有竞争力的结果。通过预训练，所有五个提议的变体都比从头开始训练的模型有很大的改进。在使用这些方法中最好的方法（即多视图对比度）进行预训练后，我们的模型可以在EC和GO-BP上获得最先进的结果，并在GO-MF上获得与预训练的基于序列的编码器的竞争结果。应该注意的是，我们的模型是在结构少于一百万的数据集上进行预训练的，而所有三个基于语言模型的基线都是在百万或十亿级序列数据库上进行预训练的。

5.3. Downstream Tasks 3 & 4: Fold and Reaction下游任务 3 & 4：折叠和反应分类

Classification

We further include two tasks used in Hermosilla et al. (2021). Fold classification is first proposed in Hou et al. (2018),

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Results on EC and GO prediction. [\*] means the results are taken from Wang et al. (2021).  我们进一步包括埃莫西拉等人（2021）中使用的两个任务。折叠分类首先在Hou等人（2018）中提出，表1：EC和GO预测的结果。[\*] 表示结果取自 Wang 等人 （2021）   |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Category | Method | EC | | | GO-BP | | | GO-MF | | | GO-CC | | | AUPRpair Fmax | | | AUPRpair | Fmax | AUPRpair | | Fmax | AUPRpair | | Fmax | |  |  | Without Pretraining | | |  |  |  | |  |  | |  | | Sequence-based | CNN (Shanehsazzadeh et al., 2020)  ResNet (Rao et al., 2019) LSTM (Rao et al., 2019) | | 0.540 0.137  0.032 | 0.545 0.187  0.082 | 0.165 0.166  0.130 | 0.244 0.280  0.248 | 0.380 0.281  0.100 | | 0.354 0.267  0.166 | 0.261 0.266  0.150 | | 0.387  0.403  0.320 | |  | Transformer (Rao et al., 2019) | | 0.187 | 0.219 | 0.135 | 0.257 | 0.172 | | 0.240 | 0.170 | | 0.380 | | Structure-based | GAT (Velickoviˇ c et al.´ , 2018)  GVP (Jing et al., 2021)  DeepFRI (Gligorijevic et al.´ , 2021) | | 0.320  0.482  0.547 | 0.368  0.489  0.631 | 0.171\*  0.224\*  0.282 | 0.284\*  0.326\*  0.399 | 0.329\*  0.458\*  0.462 | | 0.317\*  0.426\*  0.465 | 0.249\*  0.278\*  0.363 | | 0.385\*  0.420\*  0.460 | |  | New IEConv (Hermosilla & Ropinski, 2022) | | 0.775 | 0.735 | 0.273 | 0.374 | 0.572 | | 0.544 | 0.316 | | 0.444 | | Ours | GearNet  GearNet-Edge | | 0.751 0.872 | 0.730 0.810 | 0.211  0.251 | 0.356 0.403 | 0.490 0.570 | | 0.503 0.580 | 0.276  0.303 | | 0.414  0.450 | |  |  | | With Pretraining | |  |  |  | |  |  | |  | | Sequence Pretrained | ESM-1b (Rives et al., 2021)  ProtBERT-BFD (Elnaggar et al., 2021) | | 0.889  0.859 | 0.864  0.838 | 0.343  0.188\* | 0.470  0.279\* | 0.639  0.464\* | | 0.657  0.456\* | 0.384  0.234\* | | 0.488  0.408\* | |  | LM-GVP (Wang et al., 2021) | | 0.710 | 0.664 | 0.302\* | 0.417\* | 0.580\* | | 0.545\* | 0.423\* | | 0.527\* | | Structure Pretrained | GearNet-Edge (Multiview Contrast)  GearNet-Edge (Residue Type Prediction)  GearNet-Edge (Distance Prediction) | | 0.892 0.870  0.863 | 0.874 0.834  0.839 | 0.292  0.267  0.274 | 0.490  0.430  0.448 | 0.596 0.583  0.586 | | 0.650 0.604  0.616 | 0.336 0.311  0.327 | | 0.486  0.465  0.464 | |  | GearNet-Edge (Angle Prediction) | | 0.880 | 0.853 | 0.291 | 0.458 | 0.603 | | 0.625 | 0.331 | | 0.473 | |  | GearNet-Edge (Dihedral Prediction) | | 0.881 | 0.859 | 0.304 | 0.458 | 0.603 | | 0.626 | 0.338 | | 0.465 |   Table 2: Accuracy (%) on fold and reaction classification. [\*] denotes results taken from Hermosilla et al. (2021) and Hermosilla & Ropinski (2022). For pretraining, we select the model with the best performance when training from scratch.  表2：折叠和反应分类的精度（%）。[\*] 表示从赫莫西拉等人（2021年）和赫莫西拉·罗平斯基（2022年）获得的结果。对于预训练，我们选择从头开始训练时性能最佳的模型   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Category | Method | Fold Classification | | | Reaction | | Fold Super. Fam. Avg. | | | Acc. | |  | Without Pretraining |  |  |  |  | | Sequence-based | CNN (Shanehsazzadeh et al., 2020)  ResNet (Rao et al., 2019)  LSTM (Rao et al., 2019) | 11.3 10.1  6.41 | 13.4 7.21  4.33 | 53.4 26.0 23.5 13.6  18.1 9.61 | 51.7 24.1  11.0 | |  | Transformer (Rao et al., 2019) | 9.22 | 8.81 | 40.4 19.4 | 26.6 |   GCN (Kipf & Welling, 2017) 16.8\* 21.3\* 82.8\* 40.3\* 67.3\*  GVP (Jing et al., 2021) 16.0 22.5 83.8 40.7 65.5  3DCNN MQA (Derevyanko et al., 2018) 31.6\* 45.4\* 92.5\* 56.5\* 72.2\*   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Structure-based | GraphQA (Baldassarre et al., 2021) | 23.7\* | 32.5\* | 84.4\* | 46.9\* | 60.8\* | |  | DeepFRI (Gligorijevic et al.´ , 2021) | 15.3\* | 20.6\* | 73.2\* | 36.4\* | 63.3\* | |  | IEConv (Hermosilla et al., 2021) | 45.0\* | 69.7\* | 98.9\* | 71.2\* | 87.2\* | |  | New IEConv (Hermosilla & Ropinski, 2022) | 47.6\* | 70.2\* | 99.2\* | 72.3\* | 87.2\* | | Ours | GearNet  GearNet-IEConv  GearNet-Edge | 28.4 42.3  44.0 | 42.6 64.1  66.7 | 95.3 99.1  99.1 | 55.4 68.5  69.9 | 79.4 83.7  86.6 | |  | GearNet-Edge-IEConv | 48.3 | 70.3 | 99.5 | 72.7 | 85.3 | |  | With Pretraining |  |  |  |  |  | | Sequence Pretrained | ESM-1b (Rives et al., 2021)  ProtBERT-BFD (Elnaggar et al., 2021) | 26.8  26.6\* | 60.1  55.8\* | 97.8  97.6\* | 61.5  60.0\* | 83.1  72.2\* | | Structure Pretrained | New IEConv (Contrast Learning)  GearNet-Edge(-IEConv) (Multiview Contrast)  GearNet-Edge(-IEConv) (Residue Type Prediction) GearNet-Edge(-IEConv) (Distance Prediction) | 50.3\*  54.1 48.8  50.9 | 80.6\*  80.5 71.0  73.5 | 99.7\*  99.9 99.4  99.4 | 76.9\*  78.1 73.0  74.6 | 87.6\*  87.5 86.6  87.5 | |  | GearNet-Edge(-IEConv) (Angle Prediction) | 56.5 | 76.3 | 99.6 | 77.4 | 86.8 | |  | GearNet-Edge(-IEConv) (Dihedral Prediction) | 51.8 | 77.8 | 99.6 | 75.9 | 87.0 | |

with the goal to predict the fold class label given a protein. Reaction classification aims to predict the enzyme-catalyzed reaction class of a protein, in which all four levels of the EC number are employed to depict the reaction class. Although the latter task is essentially the same with EC prediction, we include it to make a fair comparison with the baselines

in Hermosilla et al. (2021).

目标是预测给定蛋白质的折叠类标签。反应分类旨在预测蛋白质的酶催化反应类，其中EC编号的所有四个水平都用于描述反应类。尽管后一项任务与EC预测基本相同，但我们将其包括在内，以便与Hermosilla等人（2021）中的基线进行公平的比较。

Dataset splits. For fold classification, Hou et al. (2018) provides three different test sets: *Fold*, in which proteins from the same superfamily are unseen during training; *Superfamily*, in which proteins from the same family are Table 3: Ablation study results on EC prediction.

数据集拆分。对于折叠分类，Hou等人（2018）提供了三种不同的测试集：Fold，其中来自同一超科的蛋白质在训练过程中是看不见的;超家族，其中来自同一家族的蛋白质是表3：EC预测的消融研究结果。

|  |  |  |
| --- | --- | --- |
| Method | AUPRpair | Fmax |
| GearNet-Edge | 0.872 | 0.810 |
| - w/o relational convolution | 0.778 | 0.752 |
| - w/o edge message passing | 0.751 | 0.730 |
| GearNet-Edge (Multiview Contrast) | 0.892 | 0.874 |
| - subsequence + identity | 0.874 | 0.866 |
| - subspace + identity | 0.894 | 0.872 |
| - subsequence + random edge masking | 0.885 | 0.869 |
| - subspace + random edge masking | 0.887 | 0.876 |
| GearNet-Edge (Dihedral Prediction) | 0.881 | 0.859 |
| - w/ random sampling | 0.841 | 0.821 |

not present during training; and *Family*, in which proteins from the same family are present during training. For reaction classification, we adopt dataset splits proposed in Hermosilla et al. (2021), where proteins have less than 50% sequence similarity in-between splits.

训练期间不在场;和家族，其中来自同一家族的蛋白质在训练期间存在。对于反应分类，我们采用了Hermosilla等人（2021）中提出的数据集分裂，其中蛋白质在分裂之间的序列相似性小于50%。

Training and evaluation. We find that the IEConv layer is important for predicting fold labels (but not for GO and EC prediction tasks as there was no performance improvement observed), so we enhance our model by incorporating this as an additional layer for fold prediction task. The models will be referred as GearNet-IEConv and GearNetEdge-IEConv, respectively. All models are trained for 300 epochs on the downstream datasets. Details of these models can be found in Appendix C.2. We pretrain GearNet-Edge and GearNet-Edge-IEConv with our proposed five methods. The performance is measured with the mean accuracy on test sets and validation sets are used for model selection.

培训和评估。我们发现IEConv层对于预测折叠标签很重要（但对于GO和EC预测任务并不重要，因为没有观察到性能改善），因此我们通过将其作为折叠预测任务的附加层来增强我们的模型。这些模型将分别称为齿轮网-IEConv和齿轮网-IEConv。所有模型都在下游数据集上针对 300 个 epoch 进行训练。这些模型的详细信息可在附录 C.2 中找到。我们用我们提出的五种方法对齿轮网-边缘和齿轮-边缘-IEConv进行预训练。使用测试集的平均精度来测量性能，并使用验证集进行模型选择。

Baselines. To verify the effectiveness of our approaches, we compare them with the baselines provided in Hermosilla et al. (2021). Moreover, we add more baselines including four sequence-based encoders (CNN (Shanehsazzadeh et al., 2020), ResNet (Rao et al., 2019), LSTM (Rao et al., 2019) and Transformer (Rao et al., 2019)) and the SOTA method New IEConv (Hermosilla & Ropinski, 2022).

基线。为了验证我们方法的有效性，我们将其与Hermosilla等人（2021）中提供的基线进行了比较。此外，我们添加了更多基线，包括四个基于序列的编码器（CNN（谢内萨扎德等人，2020年），ResNet（饶等人，2019年），LSTM（饶等人，2019年）和变形金刚（饶等人，2019年））和SOTA方法新IEConv（埃莫西拉和罗宾斯基，2022年）。

Results. Table 2 summarizes the results on fold and reaction classification. As observed in the table, our model GearNet-Edge-IEConv can achieve better results than the SOTA methods on fold classification tasks. Both the edge message passing and IEConv layers contribute a lot to the final performance. With pretraining, the model can significantly benefit from all the five proposed methods. After pretraining with the Multiview Contrast method, the model can achieve SOTA results on both tasks, which proves the effectiveness of our pretraining strategies.

结果。表2总结了折叠和反应分类的结果。如表中所示，我们的模型齿轮网-边缘-IEConv可以在折叠分类任务上获得比SOTA方法更好的结果。边缘消息传递和 IEConv 层都对最终性能有很大贡献。通过预训练，该模型可以从所有五种建议的方法中获益匪浅。使用多视图对比方法进行预训练后，该模型可以在两个任务上获得SOTA结果，这证明了我们预训练策略的有效性。

5.4. Ablation Studies医学研究

To analyze the contribution of different components in our proposed methods, we perform ablation studies on the EC prediction task. The results are shown in Table 3.

为了分析不同组分在我们提出的方法中的贡献，我们对EC预测任务进行了消融研究。结果示于表3中。

Relational graph convolutional layers. To show the effects of relational convolutional layers, we replace it with graph convolutional layers that share a kernel matrix among all edges. As reported in the table, results can be significantly improved by using relational convolution, which suggests the importance of treating edges as different types.

关系图卷积层。为了显示关系卷积层的效果，我们将其替换为在所有边缘之间共享核矩阵的图卷积层。如表中所述，通过使用关系卷积可以显著改善结果，这表明将边视为不同类型的重要性。

Edge message passing layers. We also compare the results of GearNet with and without edge message passing layers, the results of which are shown in Tables 1, 2, 3. It can be observed that the performance consistently increases after performing edge message passing. This demonstrates the effectiveness of our proposed mechanism.

边缘消息传递层。我们还比较了具有和不具有边缘消息传递层的 GearNet 的结果，其结果如表 1、2、3 所示。可以观察到，在执行边缘消息传递后，性能会持续提高。这证明了我们提议的机制的有效性。

Different augmentations in Multiview Contrast. We investigate the contribution of each augmentation operation proposed in the Multiview Contrast method. Instead of randomly sampling cropping functions and transformations, we pretrain our model with four deterministic combinations of augmentations, respectively. As shown in Table 3, all the four combinations can yield good results, which suggests that arbitrary combinations of the proposed cropping and transformation schemes can yield informative partial views of proteins.

多视图对比度中的不同增强功能。我们研究了多视图对比方法中提出的每个增强操作的贡献。我们没有随机采样裁剪函数和变换，而是分别使用四种确定性的增强组合来预训练模型。如表3所示，所有四种组合都可以产生良好的结果，这表明所提出的裁剪和转化方案的任意组合可以产生蛋白质的信息性部分视图

Sampling schemes in Self-Prediction methods. Different sampling schemes may lead to different results for self-prediction methods. We study the effects of sampling schemes using Dihedral Prediction as an example. Instead of sampling dihedral angles formed by three consecutive edges, we try to predict the dihedrals formed by four randomly sampled nodes. We observe that this change of sampling schemes will make the self-prediction task more difficult to solve, which even brings negative effects after pretraining.

自预测方法中的抽样方案。不同的抽样方案可能导致自我预测方法的不同结果。我们以二面体预测为例研究了抽样方案的效果。我们不是对由三条连续边形成的二面角进行采样，而是尝试预测由四个随机采样节点形成的二面体角。我们观察到，抽样方案的这种变化会使自预测任务更难解决，甚至在预训练后带来负面影响。

# 6. Conclusion

In this work, we propose a simple yet effective structurebased encoder for protein representation learning, which performs relational message passing on protein residue graphs. A novel edge message passing mechanism is introduced to explicitly model interactions between edges, which show consistent improvements. Moreover, five self-supervised pretraining methods are proposed following two standard frameworks: contrastive learning and self-prediction methods.

在这项工作中，我们提出了一种简单而有效的基于结构的编码器，用于蛋白质表示学习，该编码器在蛋白质残基图上执行关系消息传递。引入了一种新颖的边缘消息传递机制来显式模拟边缘之间的交互，这显示了一致的改进。此外，还提出了五种自我监督的预训练方法，遵循两个标准框架：对比学习和自我预测方法。Comprehensive experiments over multiple benchmark tasks verify that our model outperforms previous state-ofthe-art baselines under both from scratch and pretrained settings.

对多个基准测试任务的全面实验证明，我们的模型在从头开始和预先训练的设置下都优于以前最先进的基线。

We believe that our work is an important step to adopt selfsupervised learning methods on protein structure understanding. As the AlphaFold Protein Structure Database is planned to cover over 100 million proteins in UniRef90, it would be possible to train huge and more advanced structurebased models on larger datasets in the future. Another promising direction is to apply our proposed methods on more tasks, *e.g.*, protein-protein interaction modeling and protein-guided ligand molecule design, which underpins many important biological processes and applications. More broader impacts will be discussed in Appendix B.

我们相信，我们的工作是在蛋白质结构理解上采用自我监督学习方法的重要一步。由于AlphaFold蛋白质结构数据库计划在UniRef90中覆盖超过1亿个蛋白质，因此将来有可能在更大的数据集上训练巨大且更先进的基于结构的模型。另一个有希望的方向是将我们提出的方法应用于更多任务，例如蛋白质 - 蛋白质相互作用建模和蛋白质引导的配体分子设计，这支持了许多重要的生物过程和应用。附录B将讨论更广泛的影响。

# Acknowledgments鸣谢

This project is supported by AIHN IBM-MILA partnership program, the Natural Sciences and Engineering Research

Council (NSERC) Discovery Grant, the Canada CIFAR

AI Chair Program, collaboration grants between Microsoft Research and Mila, Samsung Electronics Co., Ltd., Amazon Faculty Research Award, Tencent AI Lab Rhino-Bird Gift Fund, a NRC Collaborative R&D Project (AI4D-CORE-06) as well as the IVADO Fundamental Research Project grant PRF-2019-3583139727.

# A. More Related Work

A.1. Structure-based Encoders for Biological Molecules用于生物分子的基于结构的编码器

Following the early efforts (Behler & Parrinello, 2007; Bartok et al.´ , 2010; 2013; Chmiela et al., 2017) of building machine learning systems for molecules by hand-crafted atomic features, recent works exploited end-to-end message passing neural networks (MPNNs) (Gilmer et al., 2017) to encode the structures of small molecules and macromolecules like proteins. Specifically, existing methods employed node/atom message passing (Gilmer et al., 2017; Schutt et al.¨ , 2017a;b), edge/bond message passing (Jørgensen et al., 2018; Chen et al., 2019) and directional information (Klicpera et al., 2020; Liu et al., 2021; Klicpera et al., 2021) to encode 2D or 3D molecular graphs.

在早期的努力之后（贝勒和帕里内洛，2007;巴托克等人'，2010;2013;Chmiela等人，2017）通过手工制作的原子特征为分子构建机器学习系统，最近的工作利用端到端消息传递神经网络（MPNN）（Gilmer等人，2017）来编码小分子和大分子的结构，如蛋白质。具体来说，现有方法采用节点/原子消息传递（Gilmer等人，2017;̈ ， 2017a;b）， 边缘/债券消息传递 （约根森等人， 2018;陈等人，2019）和方向信息（克里ICPera等人，2020;刘等， 2021;Klicpera等人，2021）编码2D或3D分子图。

Compared to small molecules, structural representations of proteins are more diverse, including residue-level graphs, atom-level graphs and protein surfaces. There are some recent models designed to encode residue-level graphs (Hermosilla et al., 2021; Hermosilla & Ropinski, 2022) and protein surfaces (Gainza et al., 2020; Sverrisson et al., 2021), and they achieved impressive results on various tasks. However, these models are either not expressive enough to capture edge interactions or too complicated for representation learning.

与小分子相比，蛋白质的结构表示更加多样化，包括残基水平图，原子水平图和蛋白质表面。最近有一些模型设计用于编码残留水平图（Hermosilla等人，2021;赫莫西拉和罗平斯基，2022）和蛋白质表面（Gainza等人，2020;Sverrisson等人，2021），他们在各种任务上取得了令人印象深刻的成果。但是，这些模型要么没有足够的表现力来捕获边缘交互，要么对于表示学习来说太复杂了。

A.2. Pretraining Graph Neural Networks预训练图神经网络

Our work is also related to the recent efforts of pretraining graph neural networks (GNNs), which sought to learn graph representations in a self-supervised fashion. In this domain, various self-supervised pretext tasks, like edge prediction (Kipf & Welling, 2016; Hamilton et al., 2017), context prediction (Hu et al., 2019; Rong et al., 2020), node/edge attribute reconstruction (Hu et al., 2019) and contrastive learning (Hassani & Khasahmadi, 2020; Qiu et al., 2020; You et al., 2020; Xu et al., 2021), are designed to acquire knowledge from unlabeled graphs. In this work, we focus on learning representations of residue-level graphs of proteins in a self-supervised way. To attain this goal, we design several novel protein-specific pretraining methods to learn the proposed structure-based GNN encoder.

我们的工作也与最近预训练图神经网络（GNN）的努力有关，该网络试图以自我监督的方式学习图表示。在这个领域，各种自我监督的前置任务，如边缘预测（Kipf &Welling，2016;汉密尔顿等人，2017），上下文预测（Hu等人，2019;Rong等人，2020），节点/边缘属性重建（Hu等人，2019）和对比学习（哈萨尼和哈萨哈马迪，2020;邱等， 2020;你等人，2020;Xu等人，2021），旨在从未标记的图形中获取知识。在这项工作中，我们专注于以自我监督的方式学习蛋白质残基水平图的表示。为了实现这一目标，我们设计了几种新型的蛋白质特异性预训练方法来学习所提出的基于结构的GNN编码器。

# B. Broader Impact广泛的影响

This research project focuses on learning effective protein representations via pretraining with a large number of unlabeled protein structures. Compared to the conventional sequence-based pretraining methods, our approach is able to leverage structural information and thus provide better representations. This merit enables more in-depth analysis of protein research and can potentially benefit many real-world applications, like protein function prediction and sequence design.

However, it cannot be denied that some harmful activities could be augmented by powerful pretrained models, *e.g.*, designing harmful drugs. We expect future studies will mitigate these issues.

该研究项目的重点是通过对大量未标记的蛋白质结构进行预训练来学习有效的蛋白质表示。与传统的基于序列的预训练方法相比，我们的方法能够利用结构信息，从而提供更好的表示。这一优点可以对蛋白质研究进行更深入的分析，并可能使许多实际应用受益，如蛋白质功能预测和序列设计。然而，不可否认的是，一些有害的活动可以通过强大的预训练模型来增强，例如设计有害药物。我们预计未来的研究将缓解这些问题。

# C. More Details of GearNet

In this section, we describe more details about the implementation of our GearNet. The whole pipeline of our structurebased encoder is depicted in Fig. 3.

在本节中，我们将介绍有关 GearNet 实现的更多详细信息。基于结构的编码器的整个管道如图 3 所示。

C.1. Protein Graph Construction

For graph construction, we use three different ways to add edges:

1. Sequential edges. The *i*-th residue and the *j*-th residue will be linked by an edge if the sequential distance between them is below a predefined threshold *d*seq, i.e., |*j* − *i*| *< d*seq. The type of each sequential edge is determined by their relative position *d* = *j* − *i* in the sequence. Hence, there are 2*d*seq − 1 types of sequential edges.
2. Radius edges. Following previous works, we also add edges between two nodes *i* and *j* when the Euclidean distance between them is smaller than a threshold *d*radius.
3. K-nearest neighbor edges. Since the scales of spatial coordinates may vary among different proteins, a node will be also connected to its k-nearest neighbors based on the Euclidean distance. In this way, the density of spatial edges are guaranteed to be comparable among different protein graphs.

C.1. 蛋白质图结构

对于图构建，我们使用三种不同的方法来添加边缘：

顺序边。如果第 i 个残差和第 j 个残差之间的顺序距离低于预定义的阈值 dseq，即 |j − i|<每个连续边的类型由它们在序列中的相对位置 d = j − i 确定。因此，有 2dseq − 1 种类型的顺序边。

半径边。在前面的工作之后，当两个节点 i 和 j 之间的欧几里得距离小于阈值 dradius 时，我们还在它们之间添加边。

K-最近邻边。由于空间坐标的尺度在不同蛋白质之间可能有所不同，因此节点也将根据欧几里得距离连接到其k-最近邻。通过这种方式，可以保证空间边缘的密度在不同的蛋白质图之间具有可比性。

Since we are not interested in spatial edges between residues close together in the sequence, we further add a filter to the latter two kinds of edges. Specifically, for an edge connecting the *i*-th residue and *j*-th residue, it will be removed if the sequential distance between them is lower than a long range interaction cutoff *d*long, *i.e.*, |*i* − *j*| *< d*long.

In this paper, we set the sequential distance threshold *d*seq =

3, the radius *d*radius = 10*.*0A˚ , the number of neighbors *k* = 10 and the long range interaction cutoff *d*long = 5. By regarding radius edges and KNN edges as two separate edge types, there will be totally 2*d*seq + 1 = 7 different types of edges.

由于我们对序列中紧密相连的残差之间的空间边缘不感兴趣，因此我们进一步向后两种边缘添加过滤器。具体来说，对于连接第i个残基和第j个残基的边缘，如果它们之间的顺序距离低于长程相互作用截止长度，即|i − j|< *d*long。在本文中，我们设置了序列距离阈值 dseq = 3，半径德拉迪乌斯 = 10.0A°，邻居数 k = 10，长距离相互作用截止值 dlong = 5。通过将半径边和 KNN 边视为两种不同的边类型，将完全存在 2dseq + 1 = 7 种不同类型的边。

C.2. Enhance GearNet with IEConv Layers使用 IEConv 层增强齿轮网

|  |
| --- |
| 在我们的实验中，我们发现IEConv层对于预测折叠标签非常有用，尽管它们在函数预测任务上的性能相对较差。因此，我们通过添加简化的IEConv层作为  Figure 3: The pipeline for GearNet and GearNet-edge. First, we construct a relational protein residue graph with sequential, radius and knn edges (some edges are omitted in the figure to save space). Then, a relational graph convolutional layer is applied. Similar message passing layers can be applied on the edge graph to improve the model capacity. This figure shows the update iteration for node 4 and edge (4*,*7*,*red), respectively.  图 3：齿轮网和齿轮网边缘的管道。首先，我们构建一个具有顺序，半径和kn边缘的关系蛋白质残基图（图中省略了一些边缘以节省空间）。然后，应用关系图卷积层。可以在边缘图上应用类似的消息传递层，以提高模型容量。此图分别显示了节点 4 和 Edge（4、7、红色）的更新迭代。 |

In our experiments, we find that IEConv layers are very useful for predicting fold labels in spite of their relatively poor performance on function prediction tasks. Therefore, we enhance our models by adding a simplified IEConv layer as an

additional layer, which achieve better results than the original IEConv. In this section, we describe how to simplify the IEConv layer and how to combine it with our model.

在我们的实验中，我们发现IEConv层对于预测折叠标签非常有用，尽管它们在函数预测任务上的性能相对较差。因此，我们通过添加简化的IEConv层作为附加层来增强我们的模型，从而获得比原始IEConv更好的结果。在本节中，我们将介绍如何简化 IEConv 层以及如何将其与我们的模型结合使用。

Simplify the IEConv layer. The original IEConv layer relies on the computation of intrinsic and extrinsic distances between two nodes, which are computationally expensive. Hence, we follow the modifications proposed in Hermosilla

& Ropinski (2022), which show improvements as reported in their experiments. Although these modifications are not proposed by us, we still briefly describe the model for completeness.

简化集成电路层。原始的IEConv层依赖于计算两个节点之间的内在距离和外在距离，这在计算上是昂贵的。因此，我们遵循赫莫西拉和罗宾斯基（2022）中提出的修改，这些修改显示了实验中报告的改进。虽然这些修改不是我们提出的，但我们仍然简要描述模型的完整性

In the IEConv layer, we keep the edges in our graph G and use *h*˜*i*(*l*) to denote the hidden representation for node *i* in the *l*-th layer. The update equation for node *i* is defined as:

在IEConv层中，我们保留图G中的边，并使用h ̃i（l）来表示第l层中节点i的隐藏表示。节点 i 的更新公式定义为：

*h*˜(*il*) = X *ko*(*f*(G*,i,j*)) · *h*(*jl*−1)*,* (8)

*j*∈N(*i*)

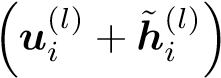
where N(*i*) is the set of neighbors of *i*, *f*(G*,i,j*) is the edge feature between *i* and *j* and *ko*(·) is an MLP mapping the feature to a kernel matrix. Instead of intrinsic and extrinsic distances in the original IEConv layer, we follow New IEConv, which adopts three relative positional features proposed in Ingraham et al. (2019) and further augments them with additional input functions.

We aim to apply this layer on our constructed protein residue graph instead of the radius graph in the original paper. Therefore, we simply remove the dynamically changed receptive fields, pooling layer and smoothing tricks in our setting.

其中 N（i） 是 i 的邻居的集合，f（G，i，j） 是 i 和 j 之间的边缘特征，ko（·） 是将特征映射到核矩阵的 MLP。我们遵循New IEConv，它采用了Ingraham等人（2019）中提出的三个相对位置特征，并通过额外的输入函数进一步增强了它们，而不是原始IEConv层中的内在距离和外在距离。我们的目标是将这一层应用于我们构建的蛋白质残基图，而不是原始论文中的半径图。因此，我们只需删除设置中动态变化的接收字段，池化层和平滑技巧。

Combine IEConv with GearNet. Our model is very flexible to incorporate other message passing layers. To incorporate IEConv layers, we just use our graph and hidden representations as input and replace the update equation Eq. 4 with

将IEConv与齿轮网相结合。我们的模型非常灵活地合并其他消息传递层。要合并IEConv层，我们只需使用我们的图形和隐藏表示作为输入，并将更新方程方程4替换为

 Dropout*.* (9)

# D. Augmentations in Multiview Contrast多视图对比度中的增强

For multiview contrastive learning methods, we propose two cropping functions and two transformation functions to generate different views of proteins. Here we illustrate these augmentation functions in Fig. 4.

对于多视图对比学习方法，我们提出了两个裁剪函数和两个转换函数来生成蛋白质的不同视图。在这里，我们在图 4 中说明了这些增强函数。

# E. Experimental Details实验细节

E.1. Dataset Statistics数据集统计

Dataset statistics of our four downstream tasks are summarized in Table 4. More details are introduced as follows.

表 4 总结了我们四个下游任务的数据集统计数据。更多细节介绍如下。

|  |
| --- |
| Figure 4: Illustration of four different augmentation functions. First, we randomly apply one of the two cropping functions shown in the figure. For subsequence, we randomly sample a consecutive segment of the protein (2-7 in this case) and take the corresponding subgraph. For subspace, we first sample a center residue (4 in this case) and then sample all residues  图 4：四种不同增强函数的图示。首先，我们随机应用图中所示的两个裁剪函数之一。对于亚序列，我们随机采样蛋白质的连续片段（在本例中为2-7）并获取相应的子图。对于子空间，我们首先对中心残基（在本例中为4个）进行采样，然后对所有残差进行取样  within a distance threshold *r*. Then, a random transformation function will be applied on the output subgraph in the last step. For identity, we directly return the graph without transformation while for random edge masking, we randomly remove a  在距离阈值 r 内。然后，在最后一步中，将在输出子图上应用随机变换函数。对于恒等式，我们直接返回图形而不进行转换，而对于随机边缘掩码，我们随机删除 |

Enzyme Commission and Gene Ontology. Following DeepFRI (Gligorijevic et al.´ , 2021), the EC numbers are selected from the third and fourth levels of the EC tree, forming 538 binary classification tasks, while the GO terms with at least 50 and no more than 5000 training samples are selected. The non-redundant sets are partitioned into fixed ratio of edges from the graph.

酶委员会和基因本体论。在DeepFRI（Gligorijevic等人，2021）之后，从EC树的第三和第四级中选择EC编号，形成538个二元分类任务，而GO术语至少选择50个和不超过5000个训练样本。非冗余集从图中划分为固定比例的边。

Table 4: Dataset statistics for downstream tasks.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | # Train | # Proteins  # Validation | # Test |
| Enzyme Commission | 15,550 | 1,729 | 1,919 |
| Gene Ontology | 29,898 | 3,322 | 3,415 |
| Fold Classification - *Fold* | 12,312 | 736 | 718 |
| Fold Classification - *Superfamily* | 12,312 | 736 | 1,254 |
| Fold Classification - *Family* | 12,312 | 736 | 1,272 |
| Reaction Classification | 29,215 | 2,562 | 5,651 |

training, validation and test set according to the sequence identity. We retrieve all protein chains from PDB used the code provided in their codebase and remove those with obsolete pdb ids, so the statistics will be slightly different from the number reported in the original paper.

根据序列标识进行训练、验证和测试集。我们使用其代码库中提供的代码从PDB中检索所有蛋白质链，并删除具有过时pdb ID的蛋白质链，因此统计数据将与原始论文中报告的数字略有不同。

Fold Classification. We directly use the dataset processed in (Hermosilla et al., 2021), which consolidated 16,712 proteins with 1,195 different folds from the SCOPe 1.75 database (Murzin et al., 1995).

Reaction Classification. The dataset comprises 37,428 proteins categorized into 384 reaction classes. The split methods are described in Hermosilla et al. (2021), where they cluster protein chains via sequence similarities and ensure that protein chains belonging to the same cluster are in the same set.

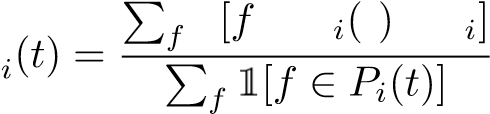
折叠分类。我们直接使用处理在（Hermosilla等人，2021）中的数据集，该数据集整合了16，712种蛋白质，其中有1，195个不同的折叠，来自SCOPe 1.75数据库（Murzin等人，1995）。反应分类。该数据集包含37，428种蛋白质，分为384个反应类。Hermosilla等人（2021）描述了分离方法，其中它们通过序列相似性对蛋白质链进行聚类，并确保属于同一簇的蛋白质链位于同一组中。

E.2. Evaluation Metrics评估指标

Now we introduce the details of evaluation metrics for EC and GO prediction. These two tasks aim to answer the question: whether a protein has a particular function, which can be seen as multiple binary classification tasks.

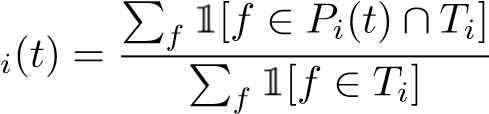
The first metric, protein-centric maximum F-score Fmax, is defined by first calculating the precision and recall for each protein and then take the average score over all proteins. More specifically, for a given target protein *i* and some decision threshold *t* ∈ [0*,*1], the precision and recall are computed as:

现在，我们介绍 EC 和 GO 预测的评估指标的详细信息。这两个任务旨在回答这个问题：蛋白质是否具有特定的功能，这可以看作是多个二元分类任务。第一个指标，即以蛋白质为中心的最大F分数Fmax，是通过首先计算每种蛋白质的精度和召回率，然后取所有蛋白质的平均分数来定义的。更具体地说，对于给定的目标蛋白 i 和某个决策阈值 t ∈ [0，1]，精度和召回率计算如下：

1 ∈ *P t* ∩ *T*

precision*,* (10)

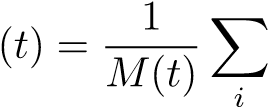
and

recall*,* (11)

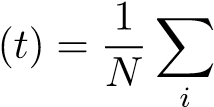
where *f* is a function term in the ontology, *Ti* is a set of experimentally determined function terms for protein *i*, *Pi*(*t*) denotes the set of predicted terms for protein *i* with scores greater than or equal to *t* and 1[·] ∈ {0*,*1} is an indicator function that is equal to 1 iff the condition is true.

Then, the average precision and recall over all proteins at threshold *t* is defined as:

其中 f 是本体中的函数项，Ti 是蛋白质 i 的一组实验确定的函数项，Pi（t） 表示分数大于或等于 t 的蛋白质 i 的预测项的集合，并且 1[·] ∈ {0，1} 是条件为真则等于 1 的指示函数。然后，阈值t处所有蛋白质的平均精度和召回率定义为：

precisionprecision*i*(*t*)*,* (12)

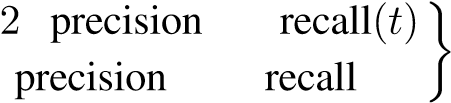
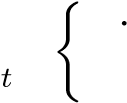
and

recallrecall*i*(*t*)*,* (13)

where we use *N* to denote the number of proteins and *M*(*t*) to denote the number of proteins on which at least one prediction was made above threshold *t*, *i.e.*, |*Pi*(*t*)| *>* 0.

Combining these two measures, the maximum F-score is defined as the maximum value of F-measure over all thresholds. That is,

其中，我们使用 N 表示蛋白质的数量，使用 M（t） 表示至少一个预测高于阈值 t 的蛋白质数量，即|圆周率（t）|> 0.结合这两个度量值，最大 F 得分定义为所有阈值上 F 度量值的最大值。那是

(*t*) ·

Fmax = max  *.* (14)

(*t*) + (*t*)

The second metric, pair-centric area under precision-recall curve AUPRpair, is defined as the average precision scores for all protein-function pairs, which is exactly the micro average precision score for multiple binary classification.

第二个指标，即精度召回曲线AUPRpair下的以对为中心的区域，定义为所有蛋白质 - 功能对的平均精度分数，这恰恰是多个二元分类的微观平均精度分数。

E.3. Implementation Details执行细节

In this subsection, we describe implementation details of all baselines and our methods. For all models, the outputs will be fed into a three-layer MLP to make final prediction. The dimension of hidden layers in the MLP is equal to the dimension of model outputs.

在本小节中，我们将介绍所有基线和方法的实现细节。对于所有模型，输出将被馈送到三层MLP中以进行最终预测。MLP 中隐藏层的维度等于模型输出的维度。

CNN (Shanehsazzadeh et al., 2020). Following the finding in Shanehsazzadeh et al. (2020), we directly employ a shallow convolutional neural network (CNN) to encode protein sequences. Specifically, 2 convolutional layers with 1024 hidden dimensions and kernel size 5 constitute this baseline model.

ResNet (Rao et al., 2019). We also adopt a deep CNN model, *i.e.*, the ResNet for protein sequences proposed by Rao et al. (2019), in our benchmark. This model is with 12 residual blocks and 512 hidden dimensions, and it uses the GELU (Hendrycks & Gimpel, 2016) activation function.

LSTM (Rao et al., 2019). The bidirectional LSTM model proposed by (Rao et al., 2019) is another baseline for protein sequence encoding. It is composed of three bidirectional LSTM layers with 640 hidden dimensions.

卷积神经网络（谢内萨扎德等人，2020）。在Shanehsazzadeh等人（2020）的发现之后，我们直接使用浅卷积神经网络（CNN）来编码蛋白质序列。具体而言，具有 1024 个隐藏维度和内核大小 5 的 2 个卷积层构成了此基线模型。瑞斯网（饶等人，2019）。我们还采用了深度CNN模型，即Rao等人（2019）在我们的基准中提出的蛋白质序列的ResNet。该模型具有12个残差块和512个隐藏维度，并使用GELU（亨德利克和吉姆佩尔，2016）激活功能。伦敦钢铁公司（饶毅等人，2019）。（Rao等人，2019）提出的双向LSTM模型是蛋白质序列编码的另一个基线。它由三个双向LSTM层组成，具有640个隐藏维度。

Transformer (Rao et al., 2019). The self-attention-based

Transformer encoder (Vaswani et al., 2017) is a strong model in natural language processing (NLP), Rao et al. (2019) adapts this model into the field of protein sequence modeling. We also adopt it as one of our baselines. This model has a comparable size with BERT-Small (Devlin et al., 2018), which contains 4 Transformer blocks with 512 hidden dimensions and 8 attention heads, and it is activated by GELU (Hendrycks & Gimpel, 2016).

转化（饶等人，2019）。基于自我注意力的转化编码器（Vaswani等人，2017）是自然语言处理（NLP）的强大模型，Rao等人（2019）将此模型应用于蛋白质序列建模领域。我们还将其作为我们的基准之一。该模型的尺寸与BERT-Small相当（Devlin等人，2018），其中包含4个变压器块，具有512个隐藏尺寸和8个注意头，并由GELU激活（亨德利克斯&Gimpel，2016）。

ESM-1b (Rives et al., 2021). Besides the from-scratch sequence encoders above, we also compare with two state-ofthe-art pretrained protein language models. ESM-1b (Rives et al., 2021) is a huge Transformer encoder model whose size is larger than BERT-Large (Devlin et al., 2018), and it is pretrained on 24 million protein sequences from UniRef50 (Suzek et al., 2007) by masked language modeling (MLM) (Devlin et al., 2018). In our evaluation, we finetune the ESM-1b model with the learning rate that is one-tenth of that of the MLP prediction head.

ESM-1b（里夫斯等人，2021）。除了上面的从头开始的序列编码器之外，我们还与两种最先进的预训练蛋白质语言模型进行了比较。ESM-1b（Rives等人，2021）是一个巨大的变压器编码器模型，其尺寸大于BERT-Large（Devlin等人，2018），并且通过掩蔽语言建模（MLM）（Devlin等人，2018）对UniRef50（Suzek等人，2007）的2400万个蛋白质序列进行了预训练。在我们的评估中，我们微调了ESM-1b模型，其学习速率是MLP预测头的十分之一。

ProtBERT-BFD (Elnaggar et al., 2021). The other protein language model evaluated in our benchmark is ProtBERT-BFD (Elnaggar et al., 2021) whose size also excesses BERT-Large (Devlin et al., 2018). This model is pretrained on 2.1 billion protein sequences from BFD (Steinegger & Soding¨ , 2018) by MLM (Devlin et al., 2018). The evaluation of ProtBERT-BFD uses the same learning rate configuration as ESM-1b.

普罗特伯特-BFD（埃尔纳加尔等人，2021）。在我们的基准测试中评估的另一种蛋白质语言模型是ProtBERT-BFD（埃尔纳加等人，2021），其大小也超过了BERT-大（Devlin等人，2018）。该模型由MLM（Devlin等人，2018）在BFD（斯坦尼格和索丁，2018）的21亿个蛋白质序列上进行预训练。对普罗伯特-BFD 的评估使用与 ESM-1b 相同的学习速率配置。

GVP (Jing et al., 2021). The GVP model (Jing et al.,

2021) is a decent protein structure encoder. It iteratively updates the scalar and vector representations of a protein, and these representations possess the merit of invariance and equivariance. In our benchmark, we evaluate this baseline method following the official source code. In specific, 3 GVP layers with 32 feature dimensions (20 scalar and 4 vector channels) constitute the GVP model.

总决赛（荆等人，2021）。GVP模型（Jing等人，2021）是一个不错的蛋白质结构编码器。它迭代地更新蛋白质的标量和向量表示，这些表示具有不变性和等变性的优点。在我们的基准测试中，我们按照官方源代码评估此基线方法。具体而言，具有 32 个特征维度（20 个标量和 4 个矢量通道）的 3 个 GVP 层构成了 GVP 模型。

LM-GVP (Wang et al., 2021). To further enhance the effectiveness of GVP (Jing et al., 2021), Wang et al. (2021) proposed to prepend a protein language model, *i.e.* ProtBERT (Elnaggar et al., 2021), before GVP to additionally utilize protein sequence representations. We also adopt this hybrid model as one of our baselines, and its implementation follows the official source code.

LM-GVP（王等人，2021）。为了进一步增强GVP的有效性（Jing等人，2021），Wang等人（2021）提出在GVP之前预先附加蛋白质语言模型，即ProtBERT（Elnaggar等人，2021），以额外利用蛋白质序列表示。我们还采用这种混合模型作为我们的基准之一，其实现遵循官方源代码。

DeepFRI (Gligorijevic et al.´ , 2021). We also evaluate DeepFRI (Gligorijevic et al.´ , 2021) in our benchmark, which is a popular structure-based encoder for protein function prediction. DeepFRI employs an LSTM model to extract residue features and further constructs a residue graph to propagate messages among residues, in which a 3-layer graph convolutional network (GCN) (Kipf & Welling, 2017) is used for message passing. We directly utilize the official model checkpoint for baseline evaluation.

深弗里（格里戈里耶维奇等人'，2021）。我们还在我们的基准测试中评估了DeepFRI（格里戈里耶维奇等人，2021），这是一种用于蛋白质功能预测的流行的基于结构的编码器。DeepFRI采用LSTM模型来提取残基特征，并进一步构建残差图以在残差之间传播消息，其中3层图卷积网络（GCN）（Kipf和Welling，2017）用于消息传递。我们直接利用官方模型检查点进行基线评估。

Table 5: Hyperparameter configurations of our model on different datasets. The batch size reported in the table referrs to the batch size on each GPU. All the hyperparameters are chosen by the performance on the validation set.

表 5：不同数据集上模型的超参数配置。表中报告的批大小是指每个 GPU 上的批大小。所有超参数都由验证集的性能选择。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Hyperparameter | | EC | GO | Fold | Reaction |
| GNN | #layer hidden dim. | 6  512 | 6  512 | 6  512 | 6  512 |
|  | dropout | 0.1 | 0.1 | 0.2 | 0.2 |
| Learning | optimizer learning rate weight decay | AdamW  1e-4  0 | AdamW  1e-4  0 | SGD  1e-3  5e-4 | SGD  1e-3  5e-4 |
|  | batch size | 2 | 2 | 2 | 2 |
|  | # epoch | 200 | 200 | 300 | 300 |

New IEConv (Hermosilla & Ropinski, 2022). Since the code for New IEConv has not been made public when the paper is written, we reproduce the method according to the description in the paper and achieve similar results on Fold and Reaction classification tasks. Then, we evaluate the method on EC and GO prediction tasks with the default hyperparameters reported in the original paper and follow the standard training procedure on these two tasks.

新IEConv（赫莫西拉和罗宾斯基，2022年）。由于新IEConv的代码在撰写论文时尚未公开，因此我们根据论文中的描述重现了该方法，并在折叠和反应分类任务上取得了类似的结果。然后，我们使用原始论文中报告的默认超参数评估EC和GO预测任务的方法，并遵循这两个任务的标准训练过程。

Our methods. For pretraining, we use Adam optimizer with learning rate 0.001 and train a model for 50 epochs. Then, the pretrained model will be finetuned on downstream datasets.

For Multiview Contrast, we set the cropping length of subsequence operation as 50, the radius of subspace operation as 15A˚ , the mask rate of random edge masking operation as 0.15. The temperature *τ* in the InfoNCE loss function is set as 0.07. When pretraining GearNet-Edge and GearNetEdge-IEConv, we use 96 and 24 as batch sizes, respectively.

我们的方法。对于预训练，我们使用学习速率为 0.001 的 Adam 优化器，并训练 50 个 epoch 的模型。然后，将在下游数据集上微调预训练的模型。对于多视图对比度，我们将子序列操作的裁剪长度设置为50，子空间操作的半径设置为15A°，随机边缘遮罩操作的遮罩率为0.15。InfoNCE 损耗函数中的温度 τ 设置为 0.07。在预训练齿轮网边缘和齿轮网边缘IEConv时，我们分别使用96和24作为批量大小。

For Distance Prediction, we set the number of sampled residue pairs as 256. And the batch size will be set as 128 and 32 for GearNet-Edge and GearNet-Edge-IEConv, respectively. For Residue Type, Angle and Dihedral Prediction, we set the number of sampled residues, residue triplets and residue quadrants as 512. And the batch size will be set as 96 and 32 for GearNet-Edge and GearNet-Edge-IEConv, respectively.

对于距离预测，我们将采样残差对的数量设置为 256。齿轮网-边缘和齿轮网络-边缘-IEConv 的批大小将分别设置为 128 和 32。对于残基类型、角度和二面体预测，我们将采样残基、残基三重态和残基象限的数量设置为 512。齿轮网边缘和齿轮网边缘IEConv的批量大小将分别设置为96和32。

For downstream evaluation, the hidden representations in each layer of GearNet will be concatenated for the final prediction. Table 5 lists the hyperparameter configurations for different downstream tasks. For the four tasks, we use the same optimizer and number of epochs as in the original papers to make fair comparison. And for EC and GO prediction, we use ReduceLROnPlateau scheduler with factor 0*.*6 and patience 5, while we use StepLR scheduler with step size 50 and gamma 0*.*5 for fold and reaction classification.

对于下游评估，GearNet 每层中的隐藏表示将被连接起来以进行最终预测。表 5 列出了不同下游任务的超参数配置。对于这四个任务，我们使用与原始论文中相同的优化器和 epoch 数来进行公平的比较。对于 EC 和 GO 预测，我们使用系数为 0.6 且耐心为 5 的 ReduceLROnPlateau 调度程序，而我们使用步长为 50 且伽玛值为 0.5 的 StepLR 调度程序进行折叠和反应分类。

# F. Pretraining on Different Datasets在不同数据集上的预训练

We use the AlphaFold protein structure database as our pretraining database, as it contains the largest number of protein structures and is planned to cover over 100 million proteins in the future. However, the structures in this database are not experimentally determined but predicted by AlphaFold2. Therefore, it is interesting to see the performance of our methods when pretraining on different datasets.

我们使用AlphaFold蛋白质结构数据库作为我们的预训练数据库，因为它包含最多数量的蛋白质结构，并计划在未来覆盖超过1亿个蛋白质。然而，该数据库中的结构不是通过实验确定的，而是由AlphaFold2预测的。因此，在对不同数据集进行预训练时，看到我们的方法的性能是很有趣的。

To study the effects of the choice of pretraining dataset, we build another dataset using structures extracted from Protein Data Bank (PDB) (Berman et al., 2000). Specifically, we extract 123,505 experimentally-determined protein structures from PDB whose resolutions are between 0.0 and 2.5 angstroms, and we further extract 305,265 chains from these proteins to construct the final pretraining dataset.

为了研究选择预训练数据集的效果，我们使用从蛋白质数据库（PDB）中提取的结构构建另一个数据集（Berman等人，2000）。具体来说，我们从PDB中提取123，505个实验确定的蛋白质结构，其分辨率在0.0到2.5埃之间，我们进一步从这些蛋白质中提取305，265个链来构建最终的预训练数据集。

Next, we pretrain our five methods on AlphaFold Database v1 (proteome-wide structure predictions), AlphaFold Database v2 (Swiss-Prot structure predictions) and Protein

Data Bank and then evaluate the pretrained models on the EC prediction task. The results are reported in Table 6. As can be seen in the table, our methods can achieve comparable performance on different pretraining datasets. Consequently, our methods are robust to the choice of pretraining datasets.

接下来，我们在 AlphaFold 数据库 v1（蛋白质组范围的结构预测）、AlphaFold 数据库 v2（瑞士-Prot 结构预测）和蛋白质数据库上预训练我们的五种方法，然后在 EC 预测任务上评估预训练的模型。结果报告在表6中。从表中可以看出，我们的方法可以在不同的预训练数据集上实现可比的性能。因此，我们的方法对于预训练数据集的选择是稳健的。

# G. Results on Protein Engineering Tasks蛋白质工程任务的结果

Besides four standard tasks considered in Section 5, another important kind of downstream tasks is related to protein engineering, which is heavily relied on mutations on protein sequences. These tasks aim to predict the ability of a protein to perform a desired function, termed protein fitness. Good models are expected to have sufficient precision to distinguish between closely-related protein sequences upon mutations. In this section, we further evaluate our model on four protein engineering related tasks.

除了第5节中考虑的四种标准任务外，另一种重要的下游任务与蛋白质工程有关，蛋白质工程严重依赖于蛋白质序列上的突变。这些任务旨在预测蛋白质执行所需功能的能力，称为蛋白质适应性。好的模型应该具有足够的精度，以区分突变时密切相关的蛋白质序列。在本节中，我们进一步评估了四个蛋白质工程相关任务的模型。

G.1. Setup

We choose two protein engineering tasks from Rao et al. (2019) and two landscape prediction tasks from Dallago et al. (2021), all of which are standard benchmarks to evaluate protein language models. The statistics of four datasets are shown in Table 7 and we describe each task as follows.

Fluoresence landscape prediction (Sarkisyan et al., 2016) This task aims to predict the log-fluorescence intensity of mutants of the parent green fluorescent protein (GFP). The training set consists of single, double and triple mutants, while the test set includes variants with four or more mutations.

我们从Rao等人（2019）和Dallago等人（2021）中选择了两个蛋白质工程任务，所有这些都是评估蛋白质语言模型的标准基准。表 7 显示了四个数据集的统计信息，我们对每个任务的描述如下。荧光景观预测（Sarkisyan等人，2016）该任务旨在预测母体绿色荧光蛋白（GFP）突变体的对数荧光强度。训练集由单一、双和三重突变体组成，而测试集包括具有四个或更多突变的变体。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 6: Results of GearNet-Edge pretrained on different pretraining datasets with different methods. Models are evaluated on the EC prediction task.  表 6：使用不同方法在不同的预训练数据集上进行预训练的 GearNet-Edge 的结果。在 EC 预测任务上评估模型  Multivew Contrast Residue Type Prediction Distance Prediction Angle Prediction Dihedral Prediction   |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  | AUPRpair | Fmax | AUPRpair | Fmax | AUPRpair | Fmax | AUPRpair | Fmax | AUPRpair | Fmax | | AlphaFold Database (v1 + v2) | 804,872 | 0.892 | 0.874 | 0.870 | 0.834 | 0.863 | 0.839 | 0.880 | 0.853 | 0.881 | 0.859 | | AlphaFold Database (v1) | 365,198 | 0.890 | 0.874 | 0.869 | 0.842 | 0.871 | 0.843 | 0.879 | 0.854 | 0.877 | 0.852 | | AlphaFold Database (v2) | 439,674 | 0.890 | 0.874 | 0.868 | 0.838 | 0.868 | 0.846 | 0.881 | 0.853 | 0.883 | 0.861 | | Protein Data Bank | 305,265 | 0.881 | 0.859 | 0.870 | 0.841 | 0.865 | 0.847 | 0.880 | 0.857 | 0.886 | 0.858 |   Dataset # Proteins |

Table 7: Dataset statistics for protein engineering tasks.

表 7：蛋白质工程任务的数据集统计数据

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | # Train | # Proteins  # Validation | # Test |
| Fluorescence | 21,446 | 5,362 | 27,217 |
| Stability | 53,614 | 2,512 | 12,851 |
| GB1 | 4,580 | 509 | 3,644 |
| Thermostability | 5,149 | 643 | 1,366 |

Stability landscape prediction (Rocklin et al., 2017) This is a regression task that maps each input protein to a value measuring the most extreme circumstances in which the protein can maintain its fold above a concentration threshold. To test the generalization ability from a broad set of relevant sequences to local neighborhood of a few sequences, the training set includes proteins from four rounds of experimental design, whereas the test set only contains 1-hop neighbors of top candidate proteins.

稳定性景观预测（Rocklin等人，2017）这是一个回归任务，它将每个输入蛋白质映射到一个值，该值测量蛋白质可以将其折叠保持在浓度阈值以上的最极端情况。为了测试从广泛的相关序列到几个序列的局部邻域的泛化能力，训练集包括来自四轮实验设计的蛋白质，而测试集仅包含顶级候选蛋白质的1-hop邻居

GB1 (Wu et al., 2016) This task uses the GB1 landscape to test the model’s ability to predict the effects of interactions between mutations, termed epistasis. We adopt the low-vs-high split proposed in Dallago et al. (2021), where sequences with fitness value equal to or below wild type are used to train, while sequences with fitness value above wild type are used to test.

Thermostability (Jarzab et al., 2020) We use the screening landscape curated from Dallago et al. (2021) to measure the model’s ability to predict thermostability of proteins. This landscape includes both global and local variation instead of only mutants of a single protein. Similarly, we use the low-vs-high split proposed in Dallago et al. (2021).

All these four tasks are evaluated via Spearman’s *ρ* (rank correlation coefficient).

GB1（Wu等人，2016）该任务使用GB1景观来测试模型预测突变之间相互作用的影响的能力，称为上位。我们采用Dallago等人（2021）中提出的低与高分割，其中使用适应度值等于或低于野生类型的序列进行训练，而使用适应度值高于野生类型的序列进行测试。热稳定性（Jarzab等人，2020）我们使用Dallago等人（2021）策划的筛选景观来测量模型预测蛋白质热稳定性的能力。这种景观包括全球和局部变异，而不仅仅是单一蛋白质的突变体。同样，我们使用Dallago等人（2021）中提出的低与高分割。所有这四个任务都通过斯皮尔曼ρ（秩相关系数）进行评估。

G.2. Implementation Details

Protein structure generation. Since all these tasks are originally designed for evaluating sequence-based encoders, the experimentally-determined structures for proteins in these datasets are not available. To solve this issue, we use AlphaFold2 to generate structures for all datasets except Table 8: Hyperparameter configurations of our model on protein engineering datasets. The batch size reported in the table referrs to the batch size on each GPU. All the hyperparameters are chosen by the performance on the validation set.

蛋白质结构生成。由于所有这些任务最初都是为评估基于序列的编码器而设计的，因此这些数据集中通过实验确定的蛋白质结构不可用。为了解决这个问题，我们使用AlphaFold2为所有数据集生成结构，除了表8：蛋白质工程数据集上模型的超参数配置。表中报告的批大小是指每个 GPU 上的批大小。所有超参数都由验证集的性能选择。

Hyperparameter Fluores Stability GB1 Thermo

#layer 6 6 6 6

GNN hidden dim. 512 512 512 512

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | dropout | 0.1 | 0.1 | 0.1 | 0.1 |
| Learning | optimizer learning rate weight decay | AdamW  1e-4  0 | AdamW  1e-4  0 | AdamW  1e-4  0 | AdamW  1e-4  0 |
|  | batch size | 8 | 32 | 8 | 2 |
|  | # epoch | 200 | 200 | 200 | 200 |

GB1. For GB1 dataset, we only generate the structure of the parent protein. Because the differences between mutant structures are almost negligible on the residue level, we directly use the residue graph constructed from the parent protein for all mutants and replace node and edge features with corresponding residue types after mutations. We scale up the number of predictions that can be performed by AlphaFold2 by using the fast homology search of MMSeqs2 (Steinegger & Soding¨ , 2017) and running AlphaFold2 in batch mode using ColabFold (Mirdita et al., 2021). Five predictions were made for each sequence and prediction with the highest pLDDT score was chosen. The number of recycles were set to 3 and relaxation using amber force fields was not used. For those proteins the structures of which AlphaFold2 fails to generate, we only add sequential edges in the protein residue graph based on sequential information.

千兆字节1.对于GB1数据集，我们只生成母体蛋白的结构。由于突变体结构之间的差异在残基水平上几乎可以忽略不计，因此我们直接使用从母体蛋白构建的所有突变体的残基图，并在突变后用相应的残基类型替换节点和边缘特征。通过使用MMSeqs2的快速同源搜索（斯坦尼格和索丁，2017），并使用ColabFold以批处理模式运行AlphaFold2，我们扩大了AlphaFold2可以执行的预测数量（Mirdita等人，2021）。对每个序列进行了五次预测，并选择了pLDDT得分最高的预测。回收数量设置为3，不使用琥珀力场松弛。对于那些AlphaFold2无法生成的结构的蛋白质，我们只根据顺序信息在蛋白质残基图中添加顺序边缘。

Training details. We use the same pretraining and downstream setting as in Section E.3. Hyperparameters for each dataset are described in Table 8.

培训详细信息。我们使用与 E.3 节中相同的预训练和下游设置。表 8 中描述了每个数据集的超参数。

G.3. Results

The results on four protein engineering tasks are reported in Table 9. First, our model achieves the best result on all tasks among models without pretraining. Moreover, it obtains comparable or even better performance than pretrained sequence encoders on Fluoresence and Stability tasks. This can be understood because our model is a generalization of

表9中报告了四个蛋白质工程任务的结果。首先，我们的模型无需预训练即可在模型中的所有任务上获得最佳结果。此外，它在荧光和稳定性任务方面的性能与预先训练的序列编码器相当甚至更好。这可以理解，因为我们的模型是

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 9: Spearman’s *ρ* (rank correlation coefficient) on four protein engineering tasks. [\*] means the results are taken from Wang et al. (2021), while [†] means the results are taken from Dallago et al. (2021).  表9：四个蛋白质工程任务的斯皮尔曼ρ（秩相关系数）。[\*] 表示结果取自 Wang 等人 （2021），而 [†] 表示结果取自 Dallago 等人 （2021）。   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Category | Method | Fluores | Stability | GB1 | Thermo | |  | Without Pretraining |  |  |  |  | | Sequence-based | CNN (Shanehsazzadeh et al., 2020)  ResNet (Rao et al., 2019) LSTM (Rao et al., 2019) | 0.656 0.369  0.124 | 0.717 0.478  0.477 | 0.51†  0.294  0.552 | 0.49†  0.412  0.142 | |  | Transformer (Rao et al., 2019) | 0.522 | 0.645 | 0.001 | OOM | | Structure-based | GAT (Velickoviˇ c et al.´ , 2018)  GVP (Jing et al., 2021) | 0.390\*  0.545\* | 0.565\*  0.680\* | -  - | -  - | |  | New IEConv (Hermosilla & Ropinski, 2022) | 0.635 | 0.529 | 0.205 | OOM | | Ours | GearNet  GearNet-Edge | 0.682 0.677 | 0.719 0.740 | 0.546  0.545 | 0.632 0.654 | |  | With Pretraining |  |  |  |  |   ESM-1b (Rives et al., 2021) 0.682 0.734 0.59† 0.76†  Sequence Pretrained ProtBERT-BFD (Elnaggar et al., 2021) 0.677\* 0.734\* - -  LM-GVP (Wang et al., 2021) 0.679\* 0.733\* - - |

CNN enhanced with structural information encoded in the protein graph and thus at least as good as CNN. Surprisingly, the improvements of edge message passing aren’t significant on these four tasks. One potential reason is that the structures of protein mutants are undistinguishable on the residue level and some structures are missing in the dataset.

As for pretraining, structure-based methods show significant improvements on Stability and GB1 datasets, which achieve the state-of-the-art performance. However, no positive effects of pretraining are shown on the other two tasks. Besides the reason mentioned above, it may be also because sequential information plays a more important role in these tasks.

Overall, our methods achieve the state-of-the-art performance on three of four datasets. Nevertheless, it still needs further exploration and ablation studies on these mutationbased datasets, which should be considered in future work.

CNN通过蛋白质图中编码的结构信息增强，因此至少与CNN一样好。令人惊讶的是，边缘消息传递的改进在这四项任务中并不重要。一个潜在的原因是蛋白质突变体的结构在残基水平上无法区分，并且数据集中缺少一些结构。至于预训练，基于结构的方法在稳定性和GB1数据集上显示出显着的改进，从而实现了最先进的性能。但是，预训练对其他两项任务没有显示出积极的影响。除了上面提到的原因之外，也可能是因为顺序信息在这些任务中起着更重要的作用。总体而言，我们的方法在四个数据集中的三个数据集上实现了最先进的性能。尽管如此，它仍然需要对这些基于突变的数据集进行进一步的探索和消融研究，这应该在未来的工作中加以考虑。

# H. Latent Space Visualization

For qualitatively evaluating the quality of the protein embeddings learned by our pretraining method, we visualize the latent space of the GearNet-Edge model pretrained by Multiview Contrast. Specifically, we utilize the pretrained model to extract the embeddings of all the proteins in AlphaFold Database v1, and these embeddings are mapped to the two-dimensional space by UMAP (McInnes et al., 2018) for visualization. Following Akdel et al. (2021), we highlight the 20 most common superfamilies within the database by different colors. The visualization results are shown in Fig. 5. It can be observed that our pretrained model tends to group the proteins from the same superfamily together and divide the ones from different superfamilies apart. In particular, it succeeds in clearly separating three superfamilies, *i.e.*, Protein kinase superfamily, Cytochrome P450 family and TRAFAC class myosin-kinesin ATPase superfamily. Such a decent capability of discriminating protein superfamilies, to some degree, interprets our model’s superior performance on Fold Classification.

为了定性评估通过我们的预训练方法学习的蛋白质包埋的质量，我们可视化了通过多视图对比预训练的GearNet-Edge模型的潜在空间。具体来说，我们利用预训练模型提取AlphaFold数据库v1中所有蛋白质的嵌入，这些嵌入被UMAP映射到二维空间（McInnes等人，2018）进行可视化。继Akdel等人（2021）之后，我们通过不同的颜色突出显示了数据库中最常见的20个超级家族。可视化结果如图5所示。可以观察到，我们的预训练模型倾向于将来自同一超科的蛋白质分组在一起，并将来自不同超科的蛋白质分开。特别是，它成功地清楚地分离出三个超家族，即蛋白激酶超家族，细胞色素P450家族和TRAFAC类肌球蛋白 - 驱动蛋白ATP酶超家族。如此体面的蛋白质超家族鉴别能力，如此...

# I. Residue-Level Explanation

Protein functions are often reflected by specific regions on the 3D protein structures. For example, the binding ability of a protein to a ligand is highly related to the binding interface between them. Hence, to better interpret our prediction, we apply Integrated Gradients (IG) (Sundararajan et al., 2017), a model-agnostic attribution method, on our model to obtain residue-level interpretation. Specifically, we first select two molecular functions, ATP binding (GO:0005524) and Heme binding (GO:0020037), from GO terms that are related to ligand binding. For each functional term, we pick one protein and feed it into the best model trained on the GO-MF dataset. Then, we use IG to generate the feature attribution scores for each protein. The method will integrate the gradient along a straight-line path between a baseline input and the original input. Here the original input and baseline input are the node feature *f* and a zero vector, respectively. The final attribution score for each protein will be obtained by summing over the feature dimension. The normalized score distribution over all residues are visualized in Figure 6. As can be seen, our model is able to identify the active sites around the ligand, which are likely to be responsible for binding. Note that these attributions are directly generated from our model without any supervision, which suggests the decent interpretability of our model.

蛋白质功能通常由3D蛋白质结构上的特定区域反映。例如，蛋白质与配体的结合能力与它们之间的结合界面高度相关。因此，为了更好地解释我们的预测，我们在模型上应用了综合梯度（IG）（Sundararajan等人，2017），这是一种与模型无关的归因方法，以获得残差水平解释。具体来说，我们首先从与配体结合相关的GO术语中选择两个分子功能，ATP结合（GO：0005524）和血红素结合（GO：0020037）。对于每个功能项，我们选择一种蛋白质并将其输入到GO-MF数据集上训练的最佳模型中。然后，我们使用IG来生成每种蛋白质的特征归因分数。该方法将沿基线输入和原始输入之间的直线路径积分梯度。此处的原始输入和基线输入分别是节点特征 f 和零向量。每种蛋白质的最终归因分数将通过对特征维度求和来获得。所有残差的归一化分数分布如图 6 所示。可以看出，我们的模型能够识别配体周围的活性位点，这些位点可能负责结合。请注意，这些归因是直接从我们的模型生成的，没有任何监督，这表明我们的模型具有良好的可解释性。

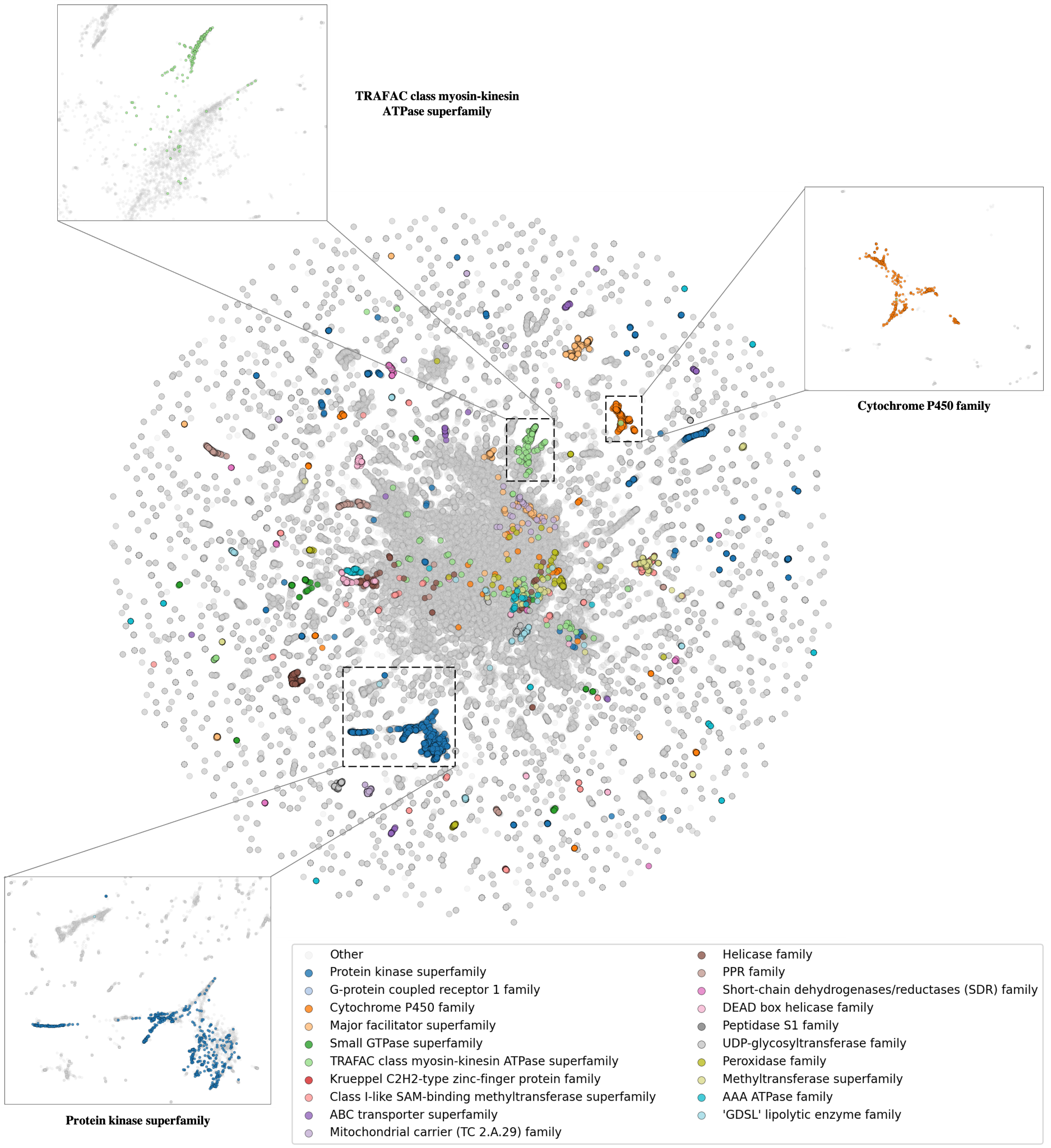


Figure 5: Latent space visualization of GearNet-Edge (Multiview Contrast) on AlphaFold Database v1.

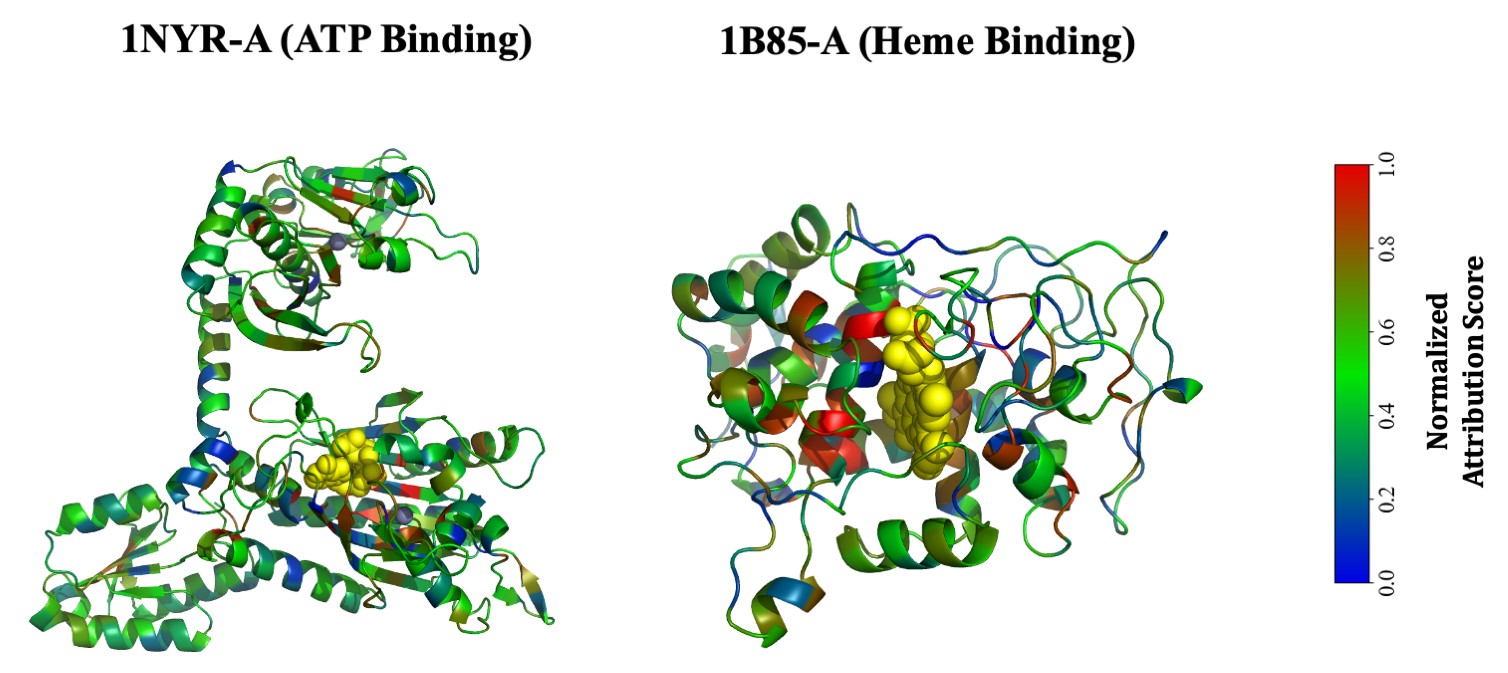
Figure 6: Identification of active sites on proteins responsible for binding based on attribution scores. Two proteins binding to specific targets are selected for illustration (1NYR-A for ATP binding and 1B85-A for Heme binding). For these two complexes, ligands are shown in yellow spheres while the residues of the receptors are colored based on attribution scores. Residues with higher attribution scores are colored in red while those with lower scores are colored in blue.

图6：根据归因评分鉴定负责结合的蛋白质上的活性位点。选择与特定靶标结合的两种蛋白质进行说明（1NYR-A用于ATP结合，1B85-A用于血红素结合）。对于这两种复合物，配体以黄色球体显示，而受体的残基则根据归因评分进行着色。归因分数较高的残渣以红色显示，而得分较低的残差以蓝色显示