# Structure-aware Protein Self-supervised Learning

**Can Chen**[1] , **Jingbo Zhou**[2] , **Fan Wang**[3] , **Xue Liu**[4] and **Dejing Dou**[2]
[1]McGill University [2]Baidu Research [3]Baidu Inc.
can.chen@mail.mcgill.ca, {zhoujingbo,wangfan04, doudejing}@baidu.com, xueliu@cs.mcgill.ca



## Abstract

Protein representation learning methods have shown great potential to yield useful representation for many downstream tasks, especially on protein classification. Moreover, a few recent studies have shown great promise in addressing insufficient labels of proteins with self-supervised learning methods. However, existing protein language models are usually pretrained on protein sequences without considering the important protein structural information. To this end, we propose a novel structure-aware protein self-supervised learning method to effectively capture structural information of proteins. In particular, a well-designed graph neural network (GNN) model is pretrained to preserve the protein structural information with self-supervised tasks from a pairwise residue distance perspective and a dihedral angle perspective, respectively. Furthermore, we propose to leverage the available protein language model pretrained on protein sequences to enhance the self-supervised learning. Specifically, we identify the relation between the sequential information in the protein language model and the structural information in the specially designed GNN model via a novel pseudo bi-level optimization scheme. Experiments on several supervised downstream tasks verify the effectiveness of our proposed method. Our code will be released upon acceptance.

## 1 Introduction

Protein classification is one of the most important tasks in biological applications [Elnaggar *et al.*, 2021]. For example, many drug discovery studies [Yang *et al.*, 2019] rely on the protein classification to understand the biological activities, which brings great benefit to efficiently select and/or generate proper drugs for human diseases. Protein classification can also help to understand the role of proteins in disease pathobiology. Therefore, many efforts have been devoted to accurately classifying the proteins in past decades. Yet, the performance of feature engineering methods is still not good enough in real-life applications.

Recently protein representation learning, which exhibits promising performance over feature engineering methods, has attracted lots of research attention. Different neural network architectures are adopted to learn different levels of protein information, for example, LSTMs [Sønderby *et al.*, 2015] are used to model the sequential information (i.e. primary structure of protein), and variants of graph neural networks [Gligorijević *et al.*, 2021] and convolutional neural networks [Hermosilla *et al.*, 2020] are used to model the structure information. Though these deep learning-based models prove to be effective, one major obstacle of protein classification is the lack of annotated protein data, which is much more severe than the one in computer vision and natural language processing areas since the wet-lab experiment of a protein annotation is quite expensive.

Inspired by the remarkable progress of self-supervised learning, there are a few recent work to perform self-supervised learning for protein from a sequence perspective [Bepler and Berger, 2019; Rives *et al.*, 2021; Rao *et al.*, 2019; Elnaggar *et al.*, 2021; Rao *et al.*, 2020; Vig *et al.*, 2020]. These sequence-based pretraining methods treat every protein as a sequence of amino acids and use autoregressive or autoencoder methods to obtain the protein representation. Although previous studies [Rives *et al.*, 2021; Elnaggar *et al.*, 2021] found such sequential pretrained protein language models can understand protein structures to some extent, these studies have not explicitly considered modeling structural information of proteins. Protein structural information determines a wide range of protein properties [Gligorijević *et al.*, 2021], but is not explicitly utilized on self-supervised learning for proteins.

Although these protein representation learning methods achieve comparable performance in various tasks compared with feature engineering methods, incorporating protein structural information into self-supervised learning is still overlooked. With the development of structural biology including cryo-EM [Callaway, 2020] and Alphafold2 [Jumper *et al.*, 2021], the availability of reliable protein structures is increasing in recent years. Thus, it is desirable to devise a new mechanism to explicitly incorporate protein structural information into self-supervised learning to boost the performance of protein classification. Meanwhile, the number of protein sequences is still much larger than the number of proteins with reliable protein structures. Therefore, learning protein

representation solely based on the limited number of structural protein data may not be able to show superior performance compared with existing protein language models.

To this end, we propose a novel ***STrucure-awarE Protein Self-supervised Learning*** (**STEPS**) method. This method can not only explicitly incorporate protein structural information into protein modeling, but also leverage the existing protein language model to enhance protein representation learning. More specifically, we leverage a well-designed graph neural network (GNN) to model protein structure and propose two novel self-supervised learning tasks to incorporate the distance information and the angle information into protein modeling. In particular, the GNN model takes the masked protein structure as input and aims to reconstruct the pairwise residue distance and the dihedral angle respectively.

Furthermore, we propose to leverage the available sequential protein language model pretrained on protein sequences (named as protein LM for short) to empower the GNN model via a pseudo bi-level optimization scheme. This optimization scheme aims to effectively transfer the knowledge of the protein LM to the GNN model. The insight is that we identify a relation between the protein LM and the GNN model and the relation is defined by the constraint between protein sequence and protein structure. Then the bi-level optimization scheme is devised to exploit the sequential information in the protein LM by leveraging its relation with the structural information in the GNN model. We named this optimization process as *pseudo bi-level* optimization because we update the GNN model in the outer level, but finally keep the parameters of the protein LM fixed in the inner level to avoid distorting the protein LM. Experiments on several downstream tasks verify the effectiveness of our method.

In summary, we make the following contributions:

- To the best of our knowledge, we are the first to explicitly incorporate protein structural information into self-supervised learning. Two novel self-supervised tasks are proposed to capture the pairwise residue distance information and the dihedral angle information respectively.
- We adopt a pseudo bi-level optimization scheme to exploit the sequential information in the protein LM.
- We conduct various supervised downstream tasks to verify the effectiveness of STEPS.

## 2 Preliminaries

In this section, we first introduce preliminary concepts and some basic notations used in this paper.

**Protein sequence.** Each protein $\mathrm{S}(\mathrm{V}, \mathcal{E})$ is a sequence of residues linked by the peptide bond. Here $\mathcal{V}$ represents the set of $L$ residues in the sequence and $\mathcal{E} \subseteq V \times V$ describes the $L-1$ peptide bonds. The protein sequence is mainly composed of twenty different types of amino acids where each unit of the protein sequence is commonly named residue after being joined by peptide bonds.

**Protein structure.** We illustrate an example of protein structure in Figure 1. As shown in the right part of Figure 1, pairwise residue distances provide important structural information of the protein. We compute the pairwise residue distance $d_{ij}$ between residue $i$ and residue $j$ as the distance between the corresponding $\mathrm{C}_\alpha$ atoms on the protein backbone. Because of the free rotation of the chemical bonds around alpha carbon, distance information alone cannot fully determine protein backbone structure, which necessitates the dihedral angle information. As shown in the left part of Figure 1, the protein backbone consists of consecutive units of $\mathrm{C}_\alpha$–CO–NH, and the rotation information around $\mathrm{C}_\alpha$ provides further structural information of the protein backbone. A simplified illustration is shown in Figure 2 where the two dihedral angles $\phi_i$ and $\psi_i$ capture the rotation of the N–$\mathrm{C}_\alpha$ bond and the $\mathrm{C}_\alpha$–$\mathrm{C}_\beta$ bond in the residue $i$ respectively. In the protein structure, the dihedral angles $\phi_i$ and $\psi_i$ can be considered as two important attributes for each residue $i$.

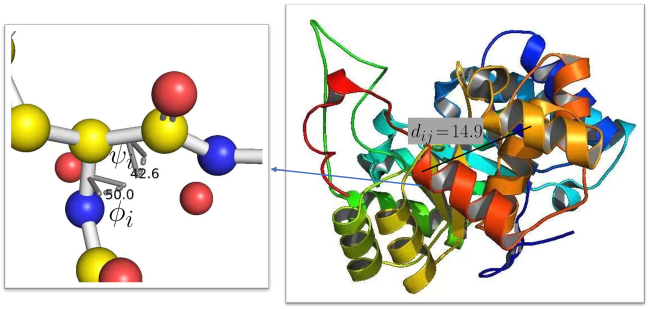


Figure 1: Protein structure.

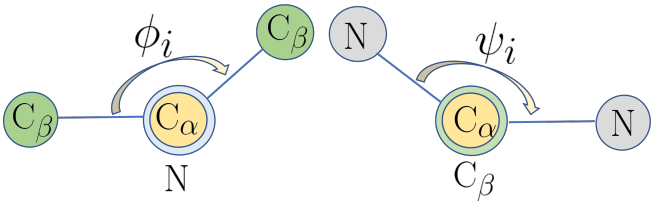


Figure 2: The dihedral angle $\phi_i$ and $\psi_i$.

**Protein structure as a graph.** We model each protein as a graph $\mathrm{G}(\mathrm{V}, \mathrm{E})$ where V denotes the set of nodes in the protein graph where each node represents a residue. Each node $v \in \mathrm{V}$ has a node feature $X_v$ including the initial residue embedding and the dihedral angle information. There is an edge $e$ between two nodes in the graph G if the pairwise residue distance is smaller than a threshold. E represents the set of edges in the protein graph, and $F_e$ represents the pairwise residue distance information for $e \in \mathrm{E}$.

## 3 The STEPS Framework

In this section, we first introduce a protein modeling method using GNN. Second, we present how to use two novel self-supervised tasks to pretrain the GNN model. Finally, we introduce the pseudo bi-level optimization scheme. The overall framework is shown in Figure 3.

### 3.1 Protein Modeling

We model protein structure as a graph and adopt GNN [Xu *et al.*, 2018] to encode the pairwise residue distance information and the dihedral angle information. The designed GNN

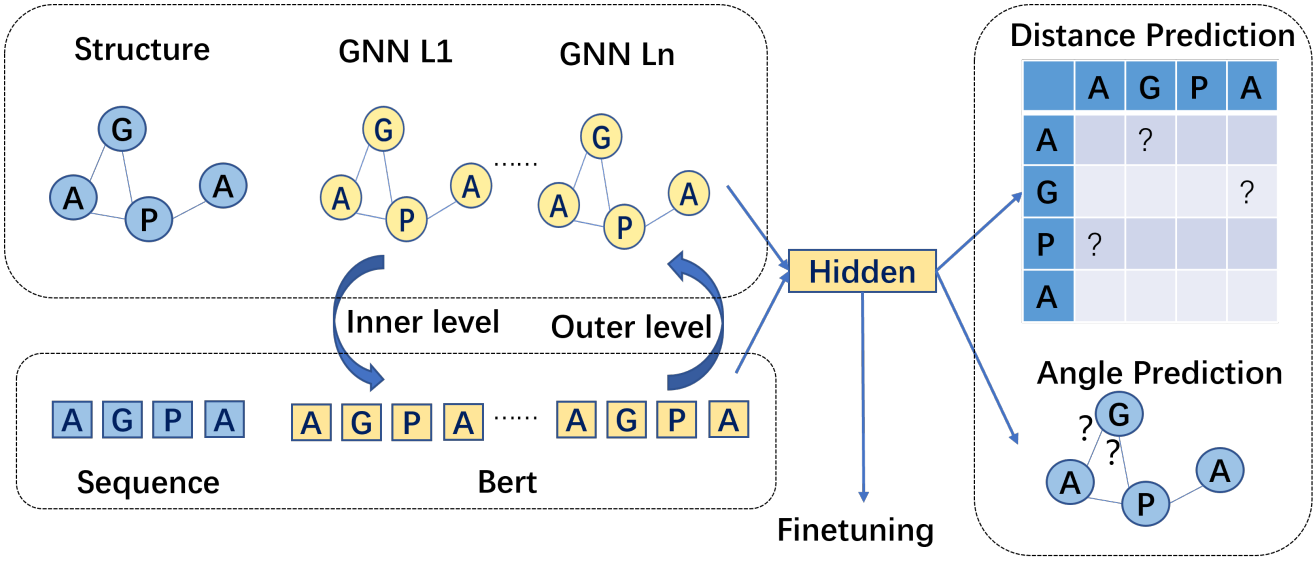


Figure 3: Framework. The GNN model captures protein structural information with two self-supervised tasks: the pairwise distance prediction task and the dihedral angle prediction task. Furthermore, a pseudo bi-level optimization scheme identifies the relation between the protein LM and the GNN model, which enhances the self-supervised learning.

model takes as input the protein structural information including node features $X$ and edge features $F$, and outputs the node representations and the graph representation.

Denote the node representation for the $i_{th}$ node in the $k_{th}$ layer of the GNN as $\mathrm{h_i^{(k)}}$. The hidden representation $\mathrm{h_i^{(k)}}$ is given by

$$\mathrm{a_i^{(k)} = AGGREGATE^{(k)}(e_{iv} h_v^{(k-1)} | v \in \mathcal{N}(i))} \quad (1)$$

$$\mathrm{h_i^{(k)} = COMBINE^{(k)}(\{h_i^{(k-1)}, a_i^{(k)}\})} \quad (2)$$

where $\mathcal{N}(i)$ denotes the neighbors of node $i$ and $e_{iv}$ denotes the feature of the edge between $i$ and $v$. $\mathrm{AGGREGATE^{(k)}}$ is the sum function and $\mathrm{COMBINE^{(k)}}$ is a linear layer for feature transformation following [Xu *et al.*, 2018]. We use the mean READOUT function to output the graph representation of the protein as:

$$\mathrm{h_G = MEAN^{(K)}(h_i^{(K)} | i \in V)} \quad (3)$$

Note that $h^0$ refers to the initial node features X which mainly include the dihedral angle information and the pretrained node embeddings which serve as initialization. The edge feature $e_{iv}$ refers to the inverse of the square of the pairwise residue distance.

To further incorporate the sequential information of a protein, we extract protein sequence representation $h_i^s$ from the protein LM, and fuse sequential representation and structural representation for the residue $i$ as:

$$h_i = h_i^s + h_i^{(K)} \quad (4)$$

where $\mathrm{h_i^{(K)}}$ refers to the final layer hidden representation from the designed GNN model.

### 3.2 Self-supervised learning tasks

We propose two self-supervised learning tasks to explicitly incorporate the distance information and the angle information into protein modeling. The distance prediction task preserves the pairwise residue distance information and the angle prediction task preserves the dihedral angle information. In this way, the GNN model yields protein representation which well captures the overall protein structural information.

**Distance prediction task**

The pairwise residue distance determines the overall shape of a protein backbone and thus determines the function of a protein to a large extent. To this end, we introduce a distance prediction task to encode the pairwise residue distance information into the GNN model.

More specifically, we develop a distance prediction network $\mathrm{NN_{dis}}(\cdot)$ which takes the vector difference between the node hidden representations of residue $i$ and residue $j$ as input, and aims to predict the pairwise residue distance between $i$ and $j$. The intuition for this operation is that the interactions of residues play an important role in determining the diverse functions of protein [Cohen *et al.*, 2009]. Therefore, the residues nearby in the protein backbone should have similar representations for protein classification. Besides, the numerical scale is quite different in the distance matrix even for the same protein. Therefore, it is more effective to formulate this distance prediction task as a multi-class classification problem instead of a regression problem. We divide the distance into $T$ uniform bins and every bin corresponds to a certain class. In this way, $\mathrm{NN_{dis}}(\cdot)$ can be written as:

$$\mathrm{d'_{ij} = NN_{dis}(h_i - h_j)} \quad (5)$$

where $d'_{ij} \in \mathrm{R^T}$ represents the predicted pairwise residue distance distribution between the residue $i$ and the residue $j$

over T classes. For a certain protein, we optimize the Cross Entropy loss among all residue pairs:

$$l_{dis} = \frac{1}{\|V\|^2} \sum_{i,j} -\text{label}(\text{d}_{\text{ij}}) \log(\text{d}_{\text{ij}}^{'}). \tag{6}$$

where $\text{label}(\text{d}_{\text{ij}})$ returns the ground truth one-hot label corresponding to the distance $d_{ij}$.

**Angle prediction task**

We further propose an angle prediction task for incorporating the dihedral angle information into the GNN model. The angle prediction task aims to predict the dihedral angles of every residue. Due to the free rotation of the chemical bonds around alpha carbon, dihedral angles of residues are of considerable importance since pairwise residue distance information alone can not determine the protein backbone structure. Note that the dihedral angles are the attributes of each residue of a protein (instead of between two or more residues).

In particular, we propose an angle prediction network $\text{NN}_{\text{ang}}(\cdot)$ which takes the angle-masked residue representation as input and aims to reconstruct the masked angles. For a certain protein, we randomly mask the feature of $15\%$ of residues, and feed the masked protein to the GNN model, which derives the hidden representation of masked residues. Note that the dihedral angles are continuous features and we first normalize the angles into [-1, 1]. After that, we adopt the Radial Basis Function to extend the scalar angle information into an angle feature vector, which serves as input to the GNN model. More specifically,

$$E_k(x) = \exp(-\gamma \|\text{x} - \text{u}_\text{k}\|^2) \tag{7}$$

where $\gamma$ determines the kernel shape and $\{u_k\}$ represents the center ranging from -1 to 1. Denote the final masked representation of the residue $i$ as $\text{h}_\text{i}^\text{m}$ and then the dihedral angles of the residue $i$ can be predicted as,

$$\bar{\phi}_i, \bar{\psi}_i = \text{NN}_{\text{ang}}(\text{h}_\text{i}^\text{m}) \tag{8}$$

The Mean Squared Error loss is adopted:

$$l_{angle} = \sum_{i \in \mathcal{M}} (\phi_i - \bar{\phi}_i)^2 + (\psi_i - \bar{\psi}_i)^2 \tag{9}$$

where $\mathcal{M}$ denotes the set of masked residues.

To sum up, the loss function for the two self-supervised learning can be written as:

$$\mathcal{L}(\theta, \omega) = l_{dis} + l_{angle} \tag{10}$$

where $\theta$ denote the parameters of the protein LM and $\omega$ denote the parameters of the GNN model including the prediction networks.

### 3.3 Pseudo Bi-level Optimization

Yet, directly fusing representations in Eq. (4) can not capture the relation between the sequential information in the protein LM and the structural information in the GNN model. We propose to identify the relation between the protein LM and the GNN model by minimizing the self-supervised loss which reconstructs the protein structure. This can be written as:

$$\theta^*(\omega) = \arg\min_{\theta} \mathcal{L}(\theta, \omega) \tag{11}$$

This relation captures the correspondance between the sequential representation and the structural representation for a certain protein [Anfinsen, 1973]. This relation is defined by the constraint between protein sequence and protein structure where the protein sequence determines the protein structure [Anfinsen, 1973]. Minimizing Eq. (14) means to adjust $\theta$ to reconstruct the protein structure for a given GNN model parameterized by $\omega$. This implies the relation between $\theta$ and $\omega$ is determined by the common protein structure. Note that we do not actually update $\theta$ to $\theta(\omega)$ in the end, but only leverage the relation to update the GNN model, which means the final adopted $\theta$ remains the same as that of the protein LM. In this way, the GNN is updated as:

$$\omega^{'} = \arg\min_{\omega} \mathcal{L}(\theta(\omega), \omega) \tag{12}$$

which could better exploit the sequential information in the protein LM.

This can be formulated as a bi-level optimization problem:

$$\min_{\alpha} \quad L(\theta(\omega), \omega) \tag{13}$$

$$\text{s.t.} \quad \theta^*(\omega) = \arg\min_{\omega} L(\theta, \omega) \tag{14}$$

where Eq.(13) defines the outer level task and Eq.(14) defines the inner level task. Different from the traditional bi-level optimization, we do not update $\theta$ in the end similar to [Wang *et al.*, 2018] so we name this scheme as pseudo bi-level optimization. The inner level can be solved approximately by a gradient descent step:

$$\theta(\omega) = \theta - \eta * \frac{\partial \mathcal{L}(\theta, \omega)}{\partial \theta}. \tag{15}$$

Similarly, the outer level can be solved as:

$$\omega^{'} = \omega - \eta^{'} * \frac{\partial \mathcal{L}(\theta(\omega), \omega)}{\partial \omega}. \tag{16}$$

## 4 Experiments

In this section, we first introduce the pretraining settings including the datasets and the training details. Second, we evaluate STEPS on three protein classification datasets and compare STEPS with existing SOTA methods. At last, we conduct ablation studies to verify the effectiveness of different components in STEPS.

### 4.1 Pretraining settings

**Datasets.** We merge two datasets from the Deeploc dataset [Almagro Armenteros *et al.*, 2017] and the Enzyme dataset [Hermosilla *et al.*, 2020], and perform pretraining on the merged dataset. For the Deeploc dataset, we acquire the available protein structures from the alphafold protein database [1]. For the Enzyme dataset, we use PDB files in the Protein Data Bank [Berman *et al.*, 2000] to obtain protein structures. Besides, we exclude proteins with a sequence length of more than 400 residues, which results in a total of around 40,000 protein structures for pretraining.

**Training details.** For the GNN model, we set the dimension of hidden representation as 1280 and the layer number as 2 in

[1] https://alphafold.ebi.ac.uk/

our experiments. The threshold to determine whether there is an edge between two residues is set as 7 Å which is consistent with previous study [Xia and Ku, 2021]. We parameterize both $\mathrm{NN}_{\mathrm{dis}}(\cdot)$ and $\mathrm{NN}_{\mathrm{ang}}(\cdot)$ as two fully-connected layers with a ReLU activation in the middle. For $\mathrm{NN}_{\mathrm{dis}}(\cdot)$, we set $T$=30 and apply softmax to the output logits. Besides, we adopt the available protein BERT model in [Elnaggar *et al.*, 2021] as the pretrained protein language model. We use the cosine learning rate decay schedule for a total of 10 epochs for pretraining. We set the learning rate for the GNN model as $0.001$ and the learning rate for the protein LM as $0.0001$ in the pseudo bi-level optimization scheme. The Adam optimizer is adopted to update the GNN parameters with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

## 4.2 Finetuning

**Downstream tasks.** We finetune the pretrained model on three protein classification tasks: the binary classification into membrane/non-membrane proteins [Almagro Armenteros *et al.*, 2017], the location classification into 10 cellular compartments [Almagro Armenteros *et al.*, 2017], and the enzyme-catalyzed reaction classification into 384 Enzyme Commission (EC) numbers [Hermosilla *et al.*, 2020]. Hereafter we denote the three tasks as C2, C10, and C384 respectively for convenience. The performance is evaluated as the mean accuracy (acc) following [Hermosilla *et al.*, 2020].

**Baselines.** We compare STEPS with two groups of baselines: methods without and with pretraining. Methods without pretraining include:

- Blast [Radivojac *et al.*, 2013]: a sequence in the test set receives labels from all labeled sequences in the training set and the prediction is obtained as the highest one. Similar to [Gligorijević *et al.*, 2021], we remove all training sequences with an E-value threshold $1e$-3 to prevent label transfer from homologous sequences.
- IEConv [Hermosilla *et al.*, 2020]: it introduces a novel convolution operator and hierarchical pooling operators to model different particularities for a protein.

Methods with pretraining are:

- Pre-LM [Elnaggar *et al.*, 2021]: it adopts the protein BERT model pretrained on Uniref100 and adds an fully-connected layer with tanh activation as the head for finetuning. The head takes the mean pooling over residue representations as input and outputs classification scores.
- DeepFRI [Gligorijević *et al.*, 2021]: this method adopts the Graph Convolutional Network (GCN) to predict protein functions by leveraging structure features. It also adopts a pretrained language model to obtain the residue embedding as the input of GCN.
- STEPS-w/oLM: we pretrain the GNN model in STEPS without considering the protein LM. Similar to Pre-LM, we add a layer on the GNN model for finetuning.

For the proposed STEPS, we finetune both the GNN model and the linear head. Besides we use STEPS-H to denote the STEPS with only finetuning the linear head.

Table 1: Experimental Results on C2 for Comparison.

| Method | Blast | IEConv | DeepFRI | Pre-LM | STEPS-w/oLM | STEPS-H | STEPS |
|---|---|---|---|---|---|---|---|
| Acc(%) | 65.14 | 62.15 | <u>88.38</u> | 58.00 | 76.94 | 87.32 | **89.26** |

Table 2: Experimental Results on C10 for Comparison.

| Method | Blast | IEConv | DeepFRI | Pre-LM | STEPS-w/oLM | STEPS-H | STEPS |
|---|---|---|---|---|---|---|---|
| Acc(%) | 31.78 | 30.99 | <u>69.23</u> | 35.00 | 43.52 | 69.94 | **70.64** |

**Training details.** For all methods and all datasets, we adopt a cosine learning rate decay with an initial learning rate 1e-4 and train the models for 5 epochs with the Adam optimizer for a fair comparison.

## 4.3 Result Analysis

As shown in Table 1, Table 2 and Table 3, we report the best results in bold and mark the second best results among two groups of baselines by underlines. First, we can observe that STEPS has consistent gains over all comparison methods in the three downstream tasks. More specifically, compared with the second best results, STEPS achieves $0.88\%$ performance gain in C2, $1.41\%$ performance gain in C10 and $39.87\%$ performance gain in C384, which proves the effectiveness of STEPS. Note that STEPS performs better than STEPS-H, which means further finetuning the GNN model on a specific task yields better representation. It worth noting that STEPS significantly outperforms its baselines in C384. A potential reason is that protein structure primarily determines the specific binding sites of an enzyme. Therefore, as the first method to incorporate the structural information into protein pretraining, STEPS performs much better than other methods on the enzyme-catalyzed reaction classification task (i.e. C384). Furthermore, we can observe that Pre-LM and STEPS-w/oLM perform worse than STEPS, which verifies the necessity of sequential information and structural information for protein pretraining. At last, we can observe STEPS-w/oLM performs better than Pre-LM by $18.94\%$ in C2, $8.52\%$ in C10 and $5.58\%$ in C384, which indicates structural information is more important than sequential information for protein classification.

## 4.4 Ablation Studies

In this section, we conduct ablation studies to verify the effectiveness of different components in STEPS.

**Pseduo bi-level optimization**

To verify the effectiveness of the pseudo bi-level optimization, we remove this part from STEPS and only optimize the GNN model, which is denoted as w/o Bi-level. Besides, we consider alternate optimization and joint optimization between the protein LM and the GNN model. We find alternate optimization and joint optimization do not perform well and

Table 3: Experimental Results on C384 for Comparison.

| Method | Blast | IEConv | DeepFRI | Pre-LM | STEPS-w/oLM | STEPS-H | STEPS |
|---|---|---|---|---|---|---|---|
| Acc(%) | 12.83 | <u>26.98</u> | 13.79 | 1.32 | 6.90 | 51.23 | **66.85** |

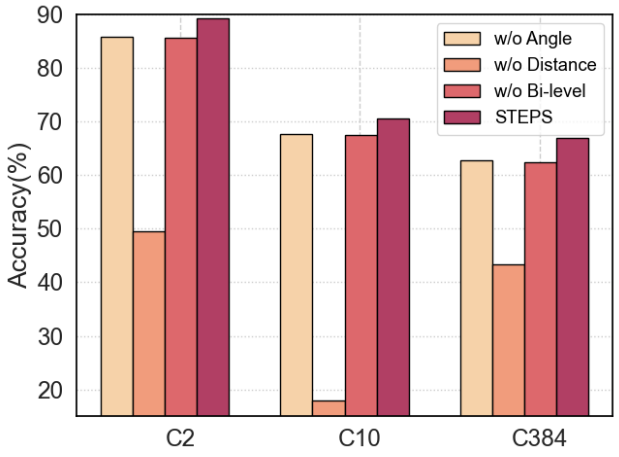


Figure 4: Ablation studies.

are even much worse than without these optimizations, so we do not report these results here. The reason may be the protein LM collapse during updating [Dodge *et al.*, 2020]. As shown in Figure 4, STEPS consistently outperforms STEPS w/o Bi-level in all tasks, which verifies the effectiveness of the proposed pseduo bi-level optimization.

**Self-supervised learning tasks**

We then demonstrate the effectiveness of the two self-supervised learning tasks: the pairwise residue distance prediction task and the dihedral angle prediction task. We remove the angle prediction task from STEPS and denote it as w/o Angle. Similarly, we remove the distance prediction task from STEPS and denote it as w/o Distance. As shown in Figure 4, removing either task leads to noticeable performance degradation, which proves the necessity of both self-supervised learning tasks. Moreover, we observe that STEPS w/o Distance results in $36.27\%$ performance decline in C2, $49.70\%$ performance decline in C10 and $19.44\%$ performance decline in C384 compared with STEPS w/o Angle. This phenomenon indicates that the pairwise residue distance information plays a more important role than the dihedral angle information in protein modeling.

# 5 Related Work

## 5.1 Protein Representation Learning

Protein representation learning methods are mainly classified into two categories: sequence-based methods and structure-based methods. Sequence-based methods model a protein via its one-dimensional amino acid sequence. For example, [Hou *et al.*, 2018] adopt one-dimensional convolutional neural networks to derive hidden representation for classification. Structure-based methods consider the three-dimensional (3D) structure of proteins. For example, [Townshend *et al.*, 2019] leverage 3D convolutional neural networks for protein quality assessment and protein contact prediction. [Hermosilla *et al.*, 2020] propose novel convolutional operators and pooling operators to model the primary, secondary, and tertiary structure effectively, which demonstrates strong performance on protein function prediction tasks. [Gligorijević *et al.*, 2021] leverage the LSTM model to encode the protein sequence and the GCN model to encode the protein tertiary structure for function prediction. [Somnath *et al.*, 2021] connect protein surface to structure modeling and sequence modeling where the learned representation achieves good performance on several downstream tasks.

## 5.2 Protein Pretraining

There are a few studies to perform pretraining on protein sequences. [Bepler and Berger, 2019] propose to train an LSTM on protein sequences, which could implicitly incorporate structural information from the global structural similarity between proteins and the contact maps for individual proteins, while STEPS uses novel self-supervised tasks to explicitly model protein structure. [Rives *et al.*, 2021] is the first to model protein sequences with self-attention, and the learned representation of the pretrained language model contains the protein information of structure and function. [Elnaggar *et al.*, 2021] try to train auto-regressive language models and auto-encoder models on large datasets, and validate the feasibility of training big language models on proteins. [Rao *et al.*, 2020; Vig *et al.*, 2020] study the transformer attention maps from the unsupervised learned language model and uncover the relationship between the attention map and the protein contact map. [Fang *et al.*, 2021] design similar self-supervised learning tasks for molecules while STEPS consideres different information including the pairwise residue distance information and the dihedral angle information for protein modeling. Besides, STEPS models protein from both sequence and structure views while [Fang *et al.*, 2021] model molecule from only the structure view.

## 5.3 Bi-level Optimization

Bi-level optimization is a special kind of optimization problem where one level of problem is embedded in the other level. Bi-level optimization has been widely used in the deep learning community due the hierarchy problem structure in many applications [Hospedales *et al.*, 2020] including neural architecture search, instance weighting, initial condition, learning to optimize, data augmentation, etc. In this paper, similar to [Wang *et al.*, 2018], we develop a pseudo bi-level optimization scheme to identify the relation between the sequential information in the protein LM and the structural information in the GNN model, which can help exploit the sequential information in the protein LM.

# 6 Conclusion

In this paper, to effectively capture protein structural information, we investigate a novel structure-aware self-supervised learning approach. Along this line, two novel self-supervised learning tasks on a GNN model is adopted to capture the pairwise residue distance information and the dihedral angle information, respectively. Also, to leverage the pretrained sequential protein language model to further improve the representation learning, we propose a pseudo bi-level optimization scheme to transfer the knowledge of the protein LM to the GNN model. Finally, the experimental results on several benchmarks for protein classification show the effectiveness and the generalizability of STEPS.

## References


[Almagro Armenteros *et al.*, 2017] José Juan Almagro Armenteros, Casper Kaae Sónderby, Sóren Kaae Sónderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 2017.

[Anfinsen, 1973] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 1973.

[Bepler and Berger, 2019] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *ICLR*, 2019.

[Berman *et al.*, 2000] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 2000.

[Callaway, 2020] Ewen Callaway. Revolutionary cryo-em is taking over structural biology. *Nature*, 2020.

[Cohen *et al.*, 2009] Mati Cohen, Vladimir Potapov, and Gideon Schreiber. Four distances between pairs of amino acids provide a precise description of their interaction. *PLoS computational biology*, 2009.

[Dodge *et al.*, 2020] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[Elnaggar *et al.*, 2021] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *TPAMI*, 2021.

[Fang *et al.*, 2021] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Chemrl-gem: Geometry enhanced molecular representation learning for property prediction. *arXiv preprint arXiv:2106.06130*, 2021.

[Gligorijević *et al.*, 2021] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 2021.

[Hermosilla *et al.*, 2020] Pedro Hermosilla, Marco Schäfer, Matej Lang, Gloria Fackelmann, Pere-Pau Vázquez, Barbora Kozlikova, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. In *ICLR*, 2020.

[Hospedales *et al.*, 2020] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[Hou *et al.*, 2018] Jie Hou, Badri Adhikari, and Jianlin Cheng. Deepsf: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 2018.

[Jumper *et al.*, 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, pages 1–11, 2021.

[Radivojac *et al.*, 2013] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 2013.

[Rao *et al.*, 2019] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *NIPS*, 2019.

[Rao *et al.*, 2020] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *ICLR*, 2020.

[Rives *et al.*, 2021] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2021.

[Somnath *et al.*, 2021] Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-scale representation learning on proteins. *NeurIPS*, 2021.

[Sønderby *et al.*, 2015] Søren Kaae Sønderby, Casper Kaae Sønderby, Henrik Nielsen, and Ole Winther. Convolutional lstm networks for subcellular localization of proteins. In *AICoB*, pages 68–80, 2015.

[Townshend *et al.*, 2019] Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *NeurIPS*, 2019.

[Vig *et al.*, 2020] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

[Wang *et al.*, 2018] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[Xia and Ku, 2021] Tian Xia and Wei-Shinn Ku. Geometric graph representation learning on protein structure prediction. In *Proc. of KDD*, pages 1873–1883, 2021.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2018.

[Yang *et al.*, 2019] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 2019.