Bioinformatics, 39(4), 2023, btad189
https://doi.org/10.1093/bioinformatics/btad189
Advance Access Publication Date: 13 April 2023
Original Paper

OXFORD

Structural bioinformatics

Structure-aware protein self-supervised learning

Can (Sam) Chen [1,2,*], **Jingbo Zhou** [3,*], **Fan Wang**[4], **Xue Liu**[1], **Dejing Dou**[5]

[1]School of Computer Science, McGill University, 845 Rue Sherbrooke O, Montreal, Quebec H3A 0G4, Canada
[2]MILA—Quebec AI Institute, 6666 Rue Saint-Urbain, Montreal, Quebec H2S 3H1, Canada
[3]Baidu Research, Xibeiwang East Road, Haidian District, Beijing 100193, China
[4]Baidu Inc., Xuefu Road East, Nanshan District, Shenzhen 518000, China
[5]BCG X, Level 22, West Tower, Genesis Beijing 8 Xinyuan South Road, Chaoyang District, Beijing 100027, China

*Corresponding author. Baidu Research, Xibeiwang East Road, Haidian District, Beijing 100193, China. E-mail: zhoujingbo@baidu.com (J.Z.); MILA—Quebec AI Institute, 6666 Rue Saint-Urbain, Montreal, Quebec H2S 3H1, Canada. E-mail: an.chen@mila.quebec (C.S.C.)
Associate Editor: Lenore Cowen

Received 16 June 2022; revised 26 December 2022; accepted 12 April 2023

Abstract

Motivation: Protein representation learning methods have shown great potential to many downstream tasks in biological applications. A few recent studies have demonstrated that the self-supervised learning is a promising solution to addressing insufficient labels of proteins, which is a major obstacle to effective protein representation learning. However, existing protein representation learning is usually pretrained on protein sequences without considering the important protein structural information.

Results: In this work, we propose a novel structure-aware protein self-supervised learning method to effectively capture structural information of proteins. In particular, a graph neural network 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 graph neural network model via a novel pseudo bi-level optimization scheme. We conduct experiments on three downstream tasks: the binary classification into membrane/non-membrane proteins, the location classification into 10 cellular compartments, and the enzyme-catalyzed reaction classification into 384 EC numbers, and these experiments verify the effectiveness of our proposed method.

Availability and implementation: The Alphafold2 database is available in https://alphafold.ebi.ac.uk/. The PDB files are available in https://www.rcsb.org/. The downstream tasks are available in https://github.com/phermosilla/IEConv_proteins/tree/master/Datasets. The code of the proposed method is available in https://github.com/GGchen1997/STEPS_Bioinformatics.

1 Introduction

A variety of machine learning-based biological tasks heavily rely on effective protein representation learning, which aims to extract rich sequential and structural information of a protein into a high-dimensional vector. The learned protein representation can be used in many downstream tasks, such as protein function annotation (Gligorijević et al. 2021), enzyme-catalyzed reaction prediction (Hermosilla et al. 2020), and protein classification (Almagro Armenteros et al. 2017). With this topic attracting lots of research attention recently, different neural network architectures are adopted to learn different levels of protein information based on the labeled protein data via supervised learning. 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 (GNNs) (Gligorijević et al. 2021) and convolutional neural networks (Hermosilla et al. 2020) are used to model the structural information. Though these deep learning-based models prove to be effective, one major obstacle to such approach is the lack of labeled data, which is much more severe than the one in computer vision and natural language processing areas since the wet-lab experiment on protein is quite expensive.

Inspired by the remarkable progress of self-supervised learning in other domains, there are a few recent work to perform self-supervised learning for protein from a sequence perspective (Bepler and Berger 2019; Rao et al. 2019, 2020; Vig et al. 2020; Elnaggar et al. 2021; Rives et al. 2021). 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 (Elnaggar et al. 2021; Rives

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.

Though protein structural information determines a wide range of protein properties (Gligorijević et al. 2021), how to incorporate protein structural information into protein self-supervised learning is 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 representation learning. Meanwhile, the number of protein sequences is still orders of magnitude 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 '**ST**rucure-awar**E** **P**rotein **S**elf-supervised Learning' (STEPS) method. This method can not only explicitly incorporate protein structural information into protein representations, but also leverage the existing protein language model to enhance protein representation learning. More specifically, we leverage a GNN to model protein structure and propose two novel self-supervised learning tasks to incorporate the distance information and the angle information into protein representation learning. In particular, the GNN model takes the masked protein structure as input and aims to reconstruct the pairwise residue distance information and the dihedral angle information, 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 the relation between the sequential information and the structural information by maximizing the mutual information between the sequential representation and the structural representation. Then a 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 name 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 STEPS.

In summary, we make the following contributions:

- To the best of our knowledge, we are the first to explicitly incorporate finer 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 article.

Protein sequence. Each protein $S(V, \mathcal{E})$ is a sequence of residues linked by peptide bonds. 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 20 different types of amino acids where each unit is commonly named as residue after being joined by peptide bonds.

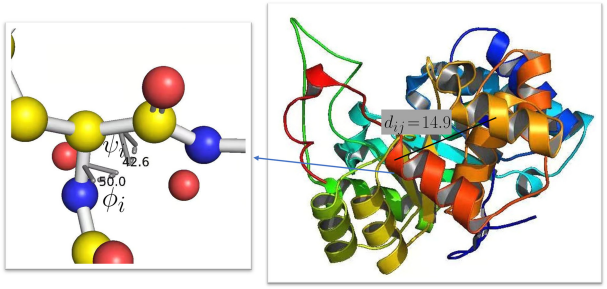

Figure 1. Protein structure.

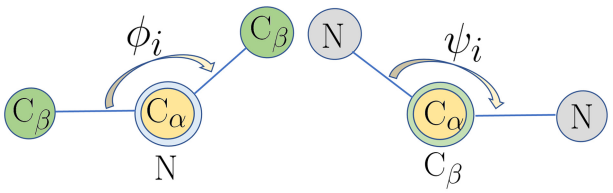

Figure 2. The dihedral angle ϕ_i and ψ_i.

Protein structure. We illustrate an example of protein structure in Fig. 1. As shown in the right part of Fig. 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 C_α 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 Fig. 1, the protein backbone consists of consecutive units of C_α–CO–NH, and the rotation information around C_α provides further structural information of the protein backbone. A simplified illustration is shown in Fig. 2 where the two dihedral angles ϕ_i and ψ_i capture the rotation of the N–C_α bond and the C_α–C_β bond in the residue i, respectively. In the protein structure, the dihedral angles ϕ_i and ψ_i are two important attributes for each residue i.

Protein structure as a graph. We model each protein as a graph G(V, E), where V denotes the set of nodes in the protein graph and each node represents a residue. Each node $v \in$ V has a node feature $\boldsymbol{X_v}$ 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 $\boldsymbol{F_e}$ represents the pairwise residue distance information for $e \in$ 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 Fig. 3.

3.1 Protein modeling

We model a protein structure as a graph and adopt a GNN model (Xu et al. 2018) to encode the pairwise residue distance information and the dihedral angle information. The designed GNN model parameterized by $\boldsymbol{\omega}$ takes as input the protein structural information including node features $\boldsymbol{X}$ and edge features $\boldsymbol{F}$, and outputs the node representations and the graph representation.

Denote the node representation for the ith node in the kth layer of the GNN as $\mathrm{h}_i^{(k)}$. The hidden representation $\mathrm{h}_i^{(k)}$ is then given by

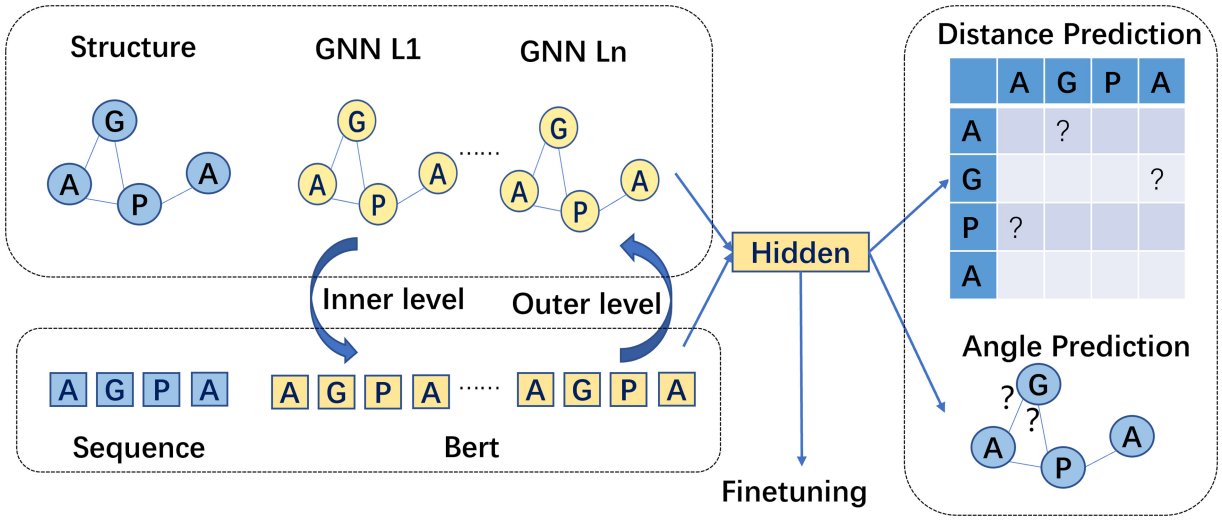

Figure 3. Framework. The GNN model captures protein structural information with two self-supervised tasks: the pairwise distance prediction task and the dihedral angle prediction task. Furthermore, a pseudo bi-level optimization scheme identifies the relation between the protein LM and the GNN model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

$$\mathrm{a}_i^{(k)} = \mathrm{AGGREGATE}^{(k)}(e_{iv}\mathrm{h}_v^{(k-1)} || v \in \mathcal{N}(i)), \tag{1}$$

$$\mathrm{h}_i^{(k)} = \mathrm{COMBINE}^{(k)}(\mathrm{h}_i^{(k-1)}, \mathrm{a}_i^{(k)}), \tag{2}$$

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), where we feed the sum of the $\mathrm{h}_i^{(k-1)}$ and $\mathrm{a}_i^{(k)}$ as the input for the linear layer. We use the mean READOUT function to output the graph representation of the protein as:

$$\mathrm{h}_G = \mathrm{MEAN}^{(K)}(\mathrm{h}_i^{(K)} || i \in V). \tag{3}$$

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. Specifically, we concatenate the node representation from the pretrained language model and the angle vector into a single input and feed this input into the GNN. The edge feature e_{iv} refers to the inverse of the square of pairwise residue distance.

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

$$\mathrm{h}_i = \mathrm{h}_i^s + \mathrm{h}_i^{(K)}, \tag{4}$$

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.

3.2.1 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 $\mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{dis}}}(\cdot)$, 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. 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, $\mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{dis}}}(\cdot)$ can be written as:

$$d'_{ij} = \mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{dis}}}(\mathrm{h}_i - \mathrm{h}_j), \tag{5}$$

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. We parameterize $\mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{dis}}}(\cdot)$ as two fully-connected layers with a ReLU activation in the middle. For a protein, we optimize the Cross Entropy loss among all residue pairs:

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

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

3.2.2 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 the alpha carbon, dihedral angles of residues are of considerable importance since pairwise residue distances alone cannot 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 $\mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{ang}}}(\cdot)$, 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, we have

$$E_k(x) = \exp(-\gamma||x - u_j||^2), \tag{7}$$

where γ determines the kernel shape and $\{u_j\}$ represents the center ranging from −1 to 1. Denote the final masked representation of the residue i as h_i^m and then the dihedral angles of the residue i can be predicted as:

$$\overline{\phi}_i, \overline{\psi}_i = \mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{ang}}}(\mathrm{h}_i^m), \tag{8}$$

where we parameterize $\mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{ang}}}(\cdot)$ as two fully connected layers with a ReLU activation in the middle. The mean squared error loss is adopted:

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

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

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

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

where $\boldsymbol{\theta}$, $\boldsymbol{\omega}$, and $\boldsymbol{\alpha}$ denote the parameters of the protein LM, the GNN model, and the prediction networks, respectively.

3.3 Pseudo bi-level optimization

Yet, directly fusing representations in Equation (4) cannot 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 maximizing the mutual information between the sequential representation and the structural representation. We adopt the Jensen–Shannon MI estimator in Nowozin et al. (2016) to estimate the mutual information. Denote $\boldsymbol{x}$ as a protein sample from $\mathbb{P}$ and $\tilde{\boldsymbol{x}}$ as another protein sample from $\tilde{\mathbb{P}} = \mathbb{P}$, and then, we have:

$$\begin{aligned} I(\boldsymbol{\theta}, \boldsymbol{\omega}) = \mathbb{E}_{\mathbb{P}}[-\mathrm{sp}(-\mathrm{T}_{\boldsymbol{\beta}}(h_{\boldsymbol{\theta}}^s(\boldsymbol{x}), h_{\boldsymbol{\omega}}^{(K)}(\boldsymbol{x}))] - \\ \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[\mathrm{sp}(\mathrm{T}_{\boldsymbol{\beta}}(h_{\boldsymbol{\theta}}^s(\boldsymbol{x}), h_{\boldsymbol{\omega}}^{(K)}(\tilde{\boldsymbol{x}}))], \end{aligned} \tag{11}$$

where $\mathrm{T}_{\boldsymbol{\beta}}$ denotes the discriminator parameterized by $\boldsymbol{\beta}$ and sp is the softplus function. For the details of T_{β}, we feed the positive and negative examples into a three-layered fully connected network with jumping connections and relu activations, and then output the dot-product of the two representations. Then, the relation between the sequential parameters $\boldsymbol{\theta}$ and the structural parameters $\boldsymbol{\omega}$ can be identified by maximizing mutual information:

$$\boldsymbol{\theta}^*(\boldsymbol{\omega}) = \arg\max_{\boldsymbol{\theta}} I(\boldsymbol{\theta}, \boldsymbol{\omega}). \tag{12}$$

This relation captures the correspondence between the sequential representation and the structural representation for a certain protein (Anfinsen 1973). Note that, we do not actually update $\boldsymbol{\theta}$ to $\boldsymbol{\theta}(\boldsymbol{\omega})$ in the end, but only leverage the relation to update the GNN model, which means the final adopted $\boldsymbol{\theta}$ remains the same as that of the protein LM. In this way, the GNN is updated as:

$$\boldsymbol{\omega}' = \arg\min_{\boldsymbol{\omega}} \mathcal{L}(\boldsymbol{\theta}(\boldsymbol{\omega}), \boldsymbol{\omega}, \boldsymbol{\alpha}), \tag{13}$$

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

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

$$\min_{\boldsymbol{\omega}, \boldsymbol{\alpha}} \quad \mathcal{L}(\boldsymbol{\theta}(\boldsymbol{\omega}), \boldsymbol{\omega}, \boldsymbol{\alpha}), \tag{14}$$

$$\text{s.t.} \quad \boldsymbol{\theta}^*(\boldsymbol{\omega}) = \arg\max_{\boldsymbol{\omega}} I(\boldsymbol{\theta}, \boldsymbol{\omega}). \tag{15}$$

Different from the traditional bi-level optimization, we do not update $\boldsymbol{\theta}$ 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 ascent step:

$$\boldsymbol{\theta}(\boldsymbol{\omega}) = \boldsymbol{\theta} + \eta * \frac{\partial I(\boldsymbol{\theta}, \boldsymbol{\omega})}{\partial \boldsymbol{\theta}}. \tag{16}$$

Similarly, the outer level can be solved as:

$$\boldsymbol{\omega}' = \boldsymbol{\omega} - \eta' * \frac{\partial \mathcal{L}(\boldsymbol{\theta}(\boldsymbol{\omega}), \boldsymbol{\omega}, \boldsymbol{\alpha})}{\partial \boldsymbol{\omega}}, \tag{17}$$

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \eta' * \frac{\partial \mathcal{L}(\boldsymbol{\theta}(\boldsymbol{\omega}), \boldsymbol{\omega}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}. \tag{18}$$

4 Experiments

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

4.1 Pretraining settings

Datasets. We sample an independent set from the alphafold protein database (https://alphafold.ebi.ac.uk/) and remove protein sequences, which have more than 25% sequence similarity with the test proteins, forming a size-40 000 pretraining set.

Training details. For the GNN model, we set the dimension of hidden representation as 1280 and the layer number as 2 in 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). For $\mathrm{NN}_{\boldsymbol{\alpha}_{\mathrm{dis}}}(\cdot)$, 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 $1e^{-3}$ and the learning rate for the protein LM as $5e^{-5}$ 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 downstream tasks: the binary classification into membrane/non-membrane proteins, the location classification into 10 cellular compartments (Almagro Armenteros et al. 2017), and the enzyme-catalyzed reaction classification (Hermosilla et al. 2020) into 384 Enzyme Commission numbers. Hereafter, we denote the three tasks as C2, C10, and C384, respectively, for convenience. The train/test sizes of C2, C10, and C384 are 2221/568, 3874/988, and 15 001/2799, respectively. For the binary classification, the membrane/non-membrane protein of the train is 1123/1098 and the membrane/non-membrane protein of the test is 297/271. The performance is evaluated as the mean accuracy (acc) following the setting in 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 a fully connected layer with tanh activation as the head for finetuning. The head takes the mean pooling over residue representations as input and outputs scores.
- DeepFRI (Gligorijević et al. 2021): this method adopts the Graph Convolutional Network (GCN) to predict protein functions by leveraging structural features. It also adopts a pretrained language model to obtain the residue embedding as the input of the GCN model.
- STEPS-w/oLM: we pretrain the GNN model in STEPS without considering protein LMs. We add a layer on GNN for finetuning. The initial node representation for GNN is the concatenation of the one-hot encoding of amino acid and the dihedral angle vectors.

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.

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 five epochs with the Adam optimizer for a fair comparison.

4.3 Result analysis

As shown in Tables 1–3, we report the best results in bold and mark the second best results (excluding STEPS-H) 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 1.39% relative gain in C2, 2.63% relative gain in C10, and 36.68% 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 is 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). We conduct additional experiments where the test set structures are from Alphafold instead of the PDB database on C384, and the results are very similar. The STEPS acc on C384 is 65.38% when using the structures from Alphafold and is 65.74% when using the PDB database. Furthermore, we can observe that Pre-LM and STEPS-w/oLM perform worse than STEPS, which verifies the necessity of structural information and sequential information for protein pretraining. At last, we can observe STEPS-w/oLM performs better than Pre-LM by 19.82% in C2, 7.51% in C10, and 2.57% in C384, which indicates structural information is more important than sequential information for protein representation learning.

4.4 Ablation studies

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

4.4.1 Mutual information

We first remove the mutual information from STEPS, denoted as w/o Mutual, and only use the self-supervised learning losses. As shown in Fig. 4, this removal leads to decreases on all three tasks, which demonstrates the importance of the mutual information.

4.4.2 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 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 Fig. 4, STEPS consistently outperforms STEPS w/o Bi-level in all tasks, which verifies the effectiveness of the proposed pseduo bi-level optimization.

4.4.3 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 Fig. 4, removing either task leads to noticeable

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	77.82	87.68	**89.61**

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	42.51	69.84	**71.05**

Table 3. Experimental results on C384 for comparison.

Method	Blast	IEConv	DeepFRI	Pre-LM	STEPS-w/oLM	STEPS-H	STEPS
Acc (%)	12.83	<u>28.70</u>	15.72	1.32	3.89	50.13	**65.38**

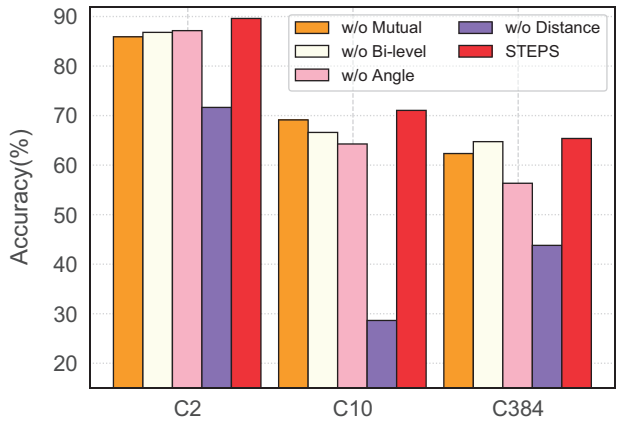

Figure 4. Ablation studies.

performance degradation, which proves the necessity of both self-supervised learning tasks. Moreover, we observe that STEPS w/o Distance results in 15.50% performance decline in C2, 35.63% performance decline in C10, and 12.54% 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. We follow the work (Xia and Ku 2021; Fang et al. 2022) to model the distance prediction as a multi-class classification problem rather than a regression problem. The regression formulation leads to 3.87% performance decline in C2, 1.01% performance decline in C10, and 2.11% performance decline in C384.

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 1D amino acid sequence. For example, Hou et al. (2018) adopt 1D convolutional neural networks to derive hidden representation for classification. Structure-based methods consider the 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; Wang et al. 2023; Zhou et al. 2023). 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. Our distance prediction task and the contact prediction task in Bepler and Berger (2019, 2021) can both incorporate the distance information into the learned protein representation. The difference is that the distance prediction task models the distance as a multi-class classification and this can explicitly consider more protein structural information compared with the binary classification of the contact prediction. Rives et al. (2021) are 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 autoregressive language models and autoencoder models on large datasets, and validate the feasibility of training big language models on proteins. Rao et al. (2020) and 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. (2022) design similar self-supervised learning tasks for molecules while STEPS considers 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. (2022) model molecule from only the structure view. A concurrent work of STEPS (Zhang et al. 2022, 2023) adopts a well-designed GNN for protein pretraining. A future direction may be adopting the pretrained protein models for protein optimization (Chen et al. 2023).

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; Chen et al. 2021, 2022a,b,c) including neural architecture search, instance weighting, initial condition, learning to optimize, data augmentation, etc. In this article, 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 article, to effectively capture protein structural information, we investigate a novel structure-aware self-supervised protein learning approach. Along this line, two novel self-supervised learning tasks on a GNN model are 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 our method STEPS.

Acknowledgements

We would like to thank the editorial team and the publisher of Bioinformatics for granting us a waiver of the open access fee for this publication. We appreciate their support in promoting open access and helping our research reach a wider audience.

Conflict of interest

None declared.

Financial Support: None declared.

References

Almagro Armenteros JJ, Sónderby CK, Sónderby SK *et al.* DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 2017; **33**:4049.

Anfinsen CB. Principles that govern the folding of protein chains. *Science* 1973;**181**:223–30.

Bepler T, Berger B. Learning protein sequence embeddings using information from structure. In: *ICLR*. 2019.

Bepler T, Berger B. Learning the protein language: evolution, structure, and function. *Cell Syst* 2021;**12**:654–69.e3.

Callaway E. Revolutionary cryo-EM is taking over structural biology. *Nature* 2020;**578**:201.

Chen C, Chen X, Ma C *et al.* Gradient-based bi-level optimization for deep learning: a survey. *arXiv, arXiv:2207*.11719. 2022a, preprint: not peer reviewed.

Chen C, Ma C, Chen X *et al.* Unbiased implicit feedback via bi-level optimization. *arXiv, arXiv:2206*.00147. 2022b, preprint: not peer reviewed.

Chen C, Zhang Y, Fu J *et al.* Bidirectional learning for offline infinite-width model-based optimization. In: *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. 2022c.

Chen C, Zhang Y, Liu X *et al.* Bidirectional learning for offline model-based biological sequence design. *arXiv, arXiv:2301*.02931. 2023, preprint: not peer reviewed.

Chen C, Zheng S, Chen X *et al.* Generalized dataweighting via class-level gradient manipulation. In: *NIPS*. 2021.

Cohen M, Potapov V, Schreiber G. Four distances between pairs of amino acids provide a precise description of their interaction. *PLoS Comput Biol* 2009;**5**:e1000470.

Dodge J, Ilharco G, Schwartz R *et al.* Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. *arXiv, arXiv:2002*.06305. 2020, preprint: not peer reviewed.

Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. In: *TPAMI*. 2021.

Fang X, Liu L, Lei J *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nat Mach Intell* 2022;**4**:127–34.

Gligorijević V, Renfrew PD, Kosciolek T *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021; **12**:3168.

Hermosilla P, Schäfer M, Lang M *et al.* Intrinsic-extrinsic convolution and pooling for learning on 3D protein structures. In: *ICLR*. 2020.

Hospedales T, Antoniou A, Micaelli P *et al.* Meta-learning in neural networks: a survey. *arXiv, arXiv:2004.05439*. 2020, preprint: not peer reviewed.

Hou J, Adhikari B, Cheng J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics* 2018;**34**: 1295–303.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Nowozin S, Cseke B, Tomioka R. f-GAN: training generative neural samplers using variational divergence minimization. *Adv Neural Inf Process Syst* 2016;**29**.

Radivojac P, Clark WT, Oron TR *et al.* A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**:221–7.

Rao R, Bhattacharya N, Thomas N *et al.* Evaluating protein transfer learning with TAPE. *Proc. Adv. Neur. Inf. Proc. Syst (NeurIPS)* 2019;**32**:9689–701.

Rao R, Meier J, Sercu T *et al.* Transformer protein language models are unsupervised structure learners. In: *ICLR*. 2020.

Rives A, Meier J, Sercu T *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 2021;**118**:e2016239118.

Somnath VR, Bunne C, Krause A. Multi-scale representation learning on proteins. In: *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

Sønderby SK, Sønderby CK, Nielsen H *et al.* Convolutional LSTM networks for subcellular localization of proteins. In: *AICoB*, 68–80. 2015.

Townshend R, Bedi R, Suriana P *et al.* End-to-end learning on 3D protein structure for interface prediction. In: *Proceedings of the Advances in Neural Information Processing Systems* (NeurIPS). 2019.

Vig J, Madani A, Varshney LR *et al.* BERTology meets biology: interpreting attention in protein language models. In: *ICLR*. 2021.

Wang D, Ye F, Zhou H. On pre-training language model for antibody. In: *ICLR*. 2023.

Wang T, Zhu J-Y, Torralba A *et al.* Dataset distillation. *arXiv, arXiv:1811.10959*. 2018, preprint: not peer reviewed.

Xia T, Ku W-S. Geometric graph representation learning on protein structure prediction. In: *KDD*, 1873–83. 2021.

Xu K, Hu W, Leskovec J *et al.* How powerful are graph neural networks? In: *ICLR*. 2018.

Zhang Z, Xu M, Chenthamarakshan V *et al.* Enhancing protein language models with structure-based encoder and pre-training. *arXiv, arXiv:2303.06275*. 2023, preprint: not peer reviewed.

Zhang Z, Xu M, Jamasb A *et al.* Protein representation learning by geometric structure pretraining. In: *ICLR*. 2023.

Zhou H-Y, Fu Y, Zhang Z *et al.* Protein representation learning via knowledge enhanced primary structure reasoning. In: *ICLR*. 2023.