Geometric Self-Supervised Pretraining on 3D Protein Structures using Subgraphs

Michail Chatzianastasis [1] **George Dasoulas** [2] **Michalis Vazirgiannis** [1 3]

Abstract

Protein representation learning aims to learn informative protein embeddings capable of addressing crucial biological questions, such as protein function prediction. Although sequence-based transformer models have shown promising results by leveraging the vast amount of protein sequence data in a self-supervised way, there is still a gap in applying these methods to 3D protein structures. In this work, we propose a pre-training scheme going beyond trivial masking methods leveraging 3D and hierarchical structures of proteins. We propose a novel self-supervised method to pretrain 3D graph neural networks on 3D protein structures, by predicting the distances between local geometric centroids of protein subgraphs and the global geometric centroid of the protein. The motivation for this method is twofold. First, the relative spatial arrangements and geometric relationships among different regions of a protein are crucial for its function. Moreover, proteins are often organized in a hierarchical manner, where smaller substructures, such as secondary structure elements, assemble into larger domains. By considering subgraphs and their relationships to the global protein structure, the model can learn to reason about these hierarchical levels of organization. We experimentally show that our proposed pertaining strategy leads to significant improvements in the performance of 3D GNNs in various protein classification tasks.

[1]École Polytechnique, Institut Polytechnique de Paris, Palaiseau, France [2]Harvard University, Cambridge, MA, USA [3]Mohamed bin Zayed University of Artificial Intelligence, UAE. Correspondence to: Michail Chatzianastasis <mixalisx97@gmail.com>.

AI for Science Workshop at International Conference on Machine Learning, Vienna, Austria.

1. Introduction

Proteins are fundamental biological macromolecules, responsible for a variety of functions within living organisms, ranging from catalyzing metabolic reactions, DNA replication, and signal transduction, to providing structural support in cells and tissues (Conrado et al., 2008; Whitford, 2013; Tye, 1999). Accurately predicting protein function is a cornerstone in molecular biology, with extensive applications in drug design, drug discovery and disease modeling (Rezaei et al., 2020). However, the complexity and variability of proteins pose significant challenges for computational prediction models (Radivojac et al., 2013; Schauperl & Denny, 2022). The functionality of a protein is affected by its three-dimensional structure, often dictating its interactions with other molecules (Ivanisenko et al., 2005). The 3D structure of proteins provides critical knowledge that is often much harder to derive from their 1D amino acid sequences alone. Therefore, understanding and predicting protein function based purely on sequence data can be challenging without considering the 3D structural modality (Gligorijević et al., 2021; Ingraham et al., 2019).

In recent years, the advent of 3D graph neural networks (GNNs) has introduced a big potential for protein representation learning. These models utilize the graph structure of proteins, where nodes represent atoms or residues, and edges represent the bonds or spatial relationships between them. GNNs are particularly good at processing the non-Euclidean data represented by 3D protein structures, enabling them to learn complex patterns that dictate protein functionality (Zhang et al., 2022; Abdine et al., 2024).

Despite these advancements, a significant limitation remains in the field: the *absence of a unified approach to effectively leverage unlabeled 3D structures for pretraining deep learning models*. Most current methods depend heavily on labeled data, which is scarce and expensive to produce. In contrast with transformer models, which have effectively used element masking as a pretraining strategy and achieved significant success in various fields (Vaswani et al., 2017), graph models still lack a definitive, universally accepted pretraining approach (Sun et al., 2022). Particularly for 3D structures, graph-based models face challenges in leveraging the extensive, unlabeled data available, while also struggling

to manage computational demands efficiently. Most prominent approaches mask node attributes or edges and then try to predict them (Hu et al., 2020). However, they do not take into account the hierarchical structure of proteins and the important substructures that affect their function.

Our approach tackles these challenges by introducing a novel pretraining strategy for 3D Graph Neural Networks (GNNs), capitalizing on the geometric properties of protein structures. Specifically, we predict the Euclidean distances between the geometric centers of various protein subgraphs and the protein's overall geometric center. This method offers several advantages. First, by utilizing subgraph representations, the model can accurately learn and capture hierarchical patterns within the 3D structure. Second, it captures the relative distances between subgraphs, a valuable feature as some tasks require focusing on surface nodes, while others may need attention on more central nodes. This flexibility increases the model's ability to handle different types of protein-related tasks effectively.

We evaluate our approach, using ProNet (Wang et al., 2023), a state-of-the-art 3D GNN model as the base architecture. We pretrain ProNet in a large amount of 3D structures from AlphaFold database (Varadi et al., 2022), and we demonstrate increased performance in protein classification tasks. Our pretraining strategy is designed to be general and adaptable, as it can be used with any 3D GNN architecture. We believe our approach will lead the way and inspire more geometric self-supervised methods on 3D protein structures.

2. Methods

2.1. 3D Graph Neural Networks

Notation. A 3D graph representing a protein is formally denoted as $G = (V, E, P)$, where V represents the set of nodes, E denotes the edges, and P denotes the spatial coordinates of each node in the graph. In this work, we represent each amino acid as a node, using the position $\boldsymbol{p} \in \mathbb{R}^3$ of the C_α atom as the position of the amino acid. We construct the edges between nodes using a threshold radius t, such that each pair of nodes is connected if their distance is smaller than the threshold t. We encode the aminoacid types as node features and the sequential distances as edge features. We denote as $\boldsymbol{h}_u^l$ the node features of node u at layer l, and $\boldsymbol{e}_{uv}$ the edge feature vector for the edge uv. We denote as N the total number of nodes and $\mathcal{N}_i$ the set of neighbors of node i.

GNN model. In our study, we utilize Graph Neural Networks (GNNs) that are specifically adapted for analyzing (3D) protein structures. 3D GNNs are deployed to process these graphs, utilizing layers of graph convolutions that aggregate information from a node's local neighborhood to capture global structural features. The convolutions in 3D GNNs are designed to take into account the Euclidean distances and other spatial relationships. We use ProNet (Wang et al., 2023) as the base model for our experiments, a recent 3D GNN model that achieves state-of-the-art performance in protein classification tasks. In each layer of ProNet, the representations of the nodes are updated using the following equation:

$$\boldsymbol{h}_i^{l+1} = f_1\left(\boldsymbol{h}_i^l, \sum_{j \in \mathcal{N}_i} f_2\left(\mathbf{v}_j^l, \mathbf{e}_{ji}, \mathcal{F}\left(d_{ji}, \theta_{ji}, \phi_{ji}, \tau_{ji}\right)\right)\right), \tag{1}$$

where f_1 and f_2 functions are parameterized using neural networks and $\mathcal{F}$ is a geometric transformation at the amino acid level. Here $(d_{ji}, \theta_{ji}, \phi_{ji})$ is the spherical coordinate of node j in the local coordinate system of node i to determine the relative position of j, and τ_{ji} is the rotation angle of edge ji. The final protein representation $\boldsymbol{h}_G$ is computed by applying a sum pooling layer in the node representations from the last layer L:

$$\boldsymbol{h}_G = \sum_{i=1}^{N} \boldsymbol{h}_i^L \tag{2}$$

2.2. Geometric Self-Supervised Pretraining

Pretraining plays a crucial role in enhancing the performance of deep neural networks, particularly in domains where labeled data is scarce or expensive to obtain. In this work, we leverage the large amount of available unlabeled 3D protein structures. Specifically, we train the model to predict the distance between the centroid of a subgraph S and the geometric centroid of the entire protein G. The self-supervised objective is to minimize the difference between the predicted and actual Euclidean distances. An overview of the proposed pipeline is illustrated in Figure 1.

Subgraph Computation. While our approach is compatible with any subgraph selection method, for our implementation, we chose 2-hop ego networks centered around 10% of the amino acids in each protein. Therefore, for each protein G, we obtain a set of different subgraphs $\mathcal{S}_\mathcal{G}$, where each subgraph corresponds to a 2-hop ego network.

Firstly, we compute the geometric centroid of the protein and the subgraphs. The geometric centroid c_G of the protein is calculated by averaging the coordinates of all aminoacids

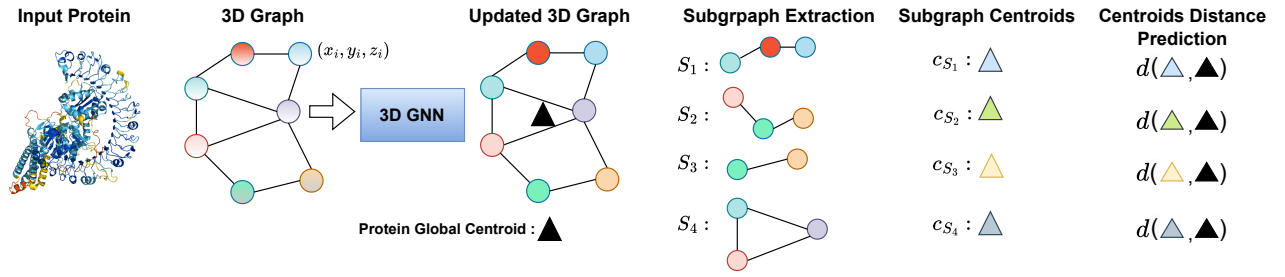

Figure 1. Visualization of the Geometric Centroid Pretraining Strategy for Protein Graph Neural Networks. This diagram illustrates the methodology employed to predict the Euclidean distances between the centroids of various subgraphs (c_S) and the overall protein centroid (c_G).

in the protein:

$$\begin{aligned} \boldsymbol{c}_G &= \frac{1}{|V|} \sum_{i \in V} \boldsymbol{p}_i \\ \boldsymbol{c}_G &= \left(\frac{1}{|V|} \sum_{i \in V} x_i, \frac{1}{N} \sum_{i \in V} y_i, \frac{1}{|V|} \sum_{i \in V} z_i \right) \end{aligned} \tag{3}$$

where (x_i, y_i, z_i) are the coordinates of each node i. Similarly, the centroid c_S for each subgraph $S \in \mathcal{S}_G$ is calculated by averaging the coordinates of the nodes within the subgraph:

$$\boldsymbol{c}_S = \left(\frac{1}{|S|} \sum_{j \in S} x_j, \frac{1}{|S|} \sum_{j \in S} y_j, \frac{1}{|S|} \sum_{j \in S} z_j \right), \tag{4}$$

where $|S|$ is the number of nodes in subgraph S. Then the label $y_{S,G}$ is computed by taking the Euclidean distance between the centroid of the protein and the centroid of subgraph S

$$y_{S,G} = d(\mathbf{c}_S, \mathbf{c}_G) = \|\mathbf{c}_S - \mathbf{c}_G\| \tag{5}$$

Distance Prediction. We calculate the embedding for a subgraph S by aggregating the node representations within this subgraph:

$$\boldsymbol{h}_S = \sum_{i \in S} \boldsymbol{h}_i^L \tag{6}$$

This summation operation merges the features of the nodes in the subgraph from the final layer L of ProNet into a unified vector that represents the entire subgraph. The predicted distance, denoted as $y_{\hat{S},G}$, is derived from the embeddings $\mathbf{h}_G$ and $\mathbf{h}_S$, using a parameterized function f:

$$\hat{y}_{S,G} = f(\mathbf{h}_S \| \mathbf{h}_G). \tag{7}$$

In our experiments, we use a two-layer multilayer perceptron (MLP) to parameterize the function f. The loss function $\mathcal{L}$ is then defined as the mean squared error (MSE) between the actual and predicted distances across all proteins and their respective subgraphs:

$$\mathcal{L}_{\text{pretraining}} = \frac{1}{N} \sum_{G \in \mathcal{D}} \sum_{S \in \mathcal{S}_G} \left(y_{S,G} - \hat{y}_{S,G} \right)^2, \tag{8}$$

where $\mathcal{D}$ is the collection of training protein graphs.

Motivation. The motivation behind our proposed approach is to overcome the limitations of existing pretraining methods for protein representation learning, which often rely on simplistic masking strategies that fail to capture the complex 3D structural patterns crucial for understanding protein function. Our method leverages the geometric and hierarchical properties of protein structures to pretrain 3D GNNs. Specifically, we predict the Euclidean distances between the geometric centers of protein subgraphs and the global geometric center. This strategy is motivated by the importance of spatial arrangements and hierarchical organization within proteins. The relative spatial relationships among protein regions are important for function, and proteins often exhibit hierarchical structures where smaller elements assemble into larger domains. By incorporating subgraph representations and focusing on their distances from the global centroid, our model captures these patterns.

3. Experiments and results

Pretraining Dataset For the pertaining, we used up to 434K proteins from the AlphaFold Database. This decision was driven by the database's extensive collection computationally predicted protein structures, which cover a wide range of known proteins across numerous species. The AlphaFold Database is renowned for its accuracy and the detailed resolution of its protein models, which closely approximate experimental structures.

Fold Classification. Protein fold classification is essential for understanding the relationships between protein struc-

ture and function. We followed the dataset and experimental protocols from (Wang et al., 2023). The dataset encompasses a total of 16,712 proteins categorized into 1,195 different folds. Our evaluation spans three distinct test sets: Fold, Superfamily, and Family. For the Fold Dataset, we used the same dataset as in previous studies (Hermosilla et al., 2020; Wang et al., 2023). To assess the model's ability to generalize, three test sets are used: Fold, where proteins from the same superfamily are not seen during training; Superfamily, where proteins from the same family are excluded from training; and Family, where proteins from the same family are included in the training data. Among these, the Fold test set presents the highest challenge due to its significant divergence from the training set's conditions. For this task, the dataset is divided into 12,312 proteins for training, 736 for validation, and additional subsets for testing: 718 proteins for the Fold test, 1,254 for Superfamily, and 1,272 for Family.

React Classification. An Enzyme Commission (EC) number is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. Each protein in the dataset is associated with an EC number, with annotations for these numbers obtained from the SIFTS database (Dana et al., 2019). The dataset encompasses a total of 37,428 proteins representing 384 distinct EC numbers. We utilized a dataset comprised of 3D protein structures sourced from the Protein Data Bank (PDB) (Berman et al., 2000). Following the experimental setup of (Wang et al., 2023), 29,215 proteins were used for training, 2,562 for validation, and 5,651 for testing. Every EC number is represented across all three dataset splits. Proteins with more than 50% similarity were grouped together in the same split. This setup aids in evaluating the model's ability to generalize across different protein structures.

Setup. We use ProNet as the base architecture. Following (Wang et al., 2023), we also apply Gaussian noise to the input data and to the hidden states to improve the robustness of the model, and we denote this model as ProNet Augmented. We use the best hyperparameters from (Wang et al., 2023) and we pretrain the models for 10 epochs. Then we fine-tune the models for the classification tasks, initialized with the pretrained weights.

Results. We report the results in Table 1. We observe that our pretrained models outperform the baselines in both the base architecture and the augmented one. Additionally, we conduct an ablation study by varying the number of pretraining samples from 33K to 434K. The results indicate that the size of the pretraining dataset significantly affects downstream task performance, with larger datasets leading to higher accuracy.

Table 1. Accuracy (%) on fold and reaction classification tasks.

Method	Pretraining Size	React	Fold			
			Fold	Sup.	Fam.	Avg.
ProNet	-	81.04	44.29	58.37	96.23	66.30
ProNet Pretrained	33K	81.05	46.80	63.56	97.64	69.33
ProNet Pretrained	100K	81.07	49.44	62.84	**97.56**	69.95
ProNet Pretrained	434K	**81.51**	**49.72**	**64.51**	**97.56**	**70.60**
ProNet Augmented	-	84.23	51.50	66.75	98.19	72.15
ProNet Augmented Pretrained	434K	**84.50**	**52.23**	**68.42**	**98.27**	**72.97**

4. Conclusion and Future work

In this work, we proposed a new self-supervised learning method to learn accurate protein representations from 3D structures. By capitalizing on the extensive collection of 3D protein structures available, we pre-trained a 3D GNN model to predict the distance between the geometric centroid of the entire protein and various subgraphs within the protein. We experimentally show that our pretraining strategy leads to improved performance in downstream classification tasks, such as protein fold and reaction classification. For future work, we aim to study the impact of different subgraph selection methods and experimentally test our method on more 3D GNNs.

References

Abdine, H., Chatzianastasis, M., Bouyioukos, C., and Vazirgiannis, M. Prot2text: Multimodal protein's function generation with gnns and transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10): 10757–10765, Mar. 2024. doi: 10.1609/aaai.v38i10.28948. URL https://ojs.aaai.org/index.php/AAAI/article/view/28948.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Conrado, R. J., Varner, J. D., and DeLisa, M. P. Engineering the spatial organization of metabolic enzymes: mimicking nature's synergy. *Current opinion in biotechnology*, 19(5):492–499, 2008.

Dana, J. M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., and Velankar, S. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, 47(D1):D482–D489, 2019.

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., et al. Structure-based protein

function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Hermosilla, P., Schäfer, M., Lang, M., Fackelmann, G., Vázquez, P. P., Kozlíková, B., Krone, M., Ritschel, T., and Ropinski, T. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv preprint arXiv:2007.06252*, 2020.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks, 2020.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

Ivanisenko, V. A., Pintus, S. S., Grigorovich, D. A., and Kolchanov, N. A. Pdbsite: a database of the 3d structure of protein functional sites. *Nucleic Acids Research*, 33 (suppl_1):D183–D187, 2005.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.

Rezaei, M. A., Li, Y., Wu, D., Li, X., and Li, C. Deep learning in drug design: protein-ligand binding affinity prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(1):407–417, 2020.

Schauperl, M. and Denny, R. A. Ai-based protein structure prediction in drug discovery: impacts and challenges. *Journal of Chemical Information and Modeling*, 62(13): 3142–3156, 2022.

Sun, R., Dai, H., and Yu, A. W. Does gnn pretraining help molecular representation? *Advances in Neural Information Processing Systems*, 35:12096–12109, 2022.

Tye, B. K. Mcm proteins in dna replication. *Annual review of biochemistry*, 68(1):649–686, 1999.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, L., Liu, H., Liu, Y., Kurtin, J., and Ji, S. Learning hierarchical protein representations via complete 3d graph networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9X-hgLDLYkQ.

Whitford, D. *Proteins: structure and function*. John Wiley & Sons, 2013.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.