SCOP: A Sequence-Structure Contrast-Aware Framework for Protein Function Prediction (Appendix)

Runze Ma[§*], Chengxin He[†], Huiru Zheng[‡], Lei Duan[†*]
[§] Shanghai Jiaotong University, Shanghai, China
[†] Sichuan University, Chengdu, China
[‡] Ulster University, Belfast, United Kingdom
* Corresponding authors. Email: runze.ma@outlook.com, leiduan@scu.edu.cn

I. Related Work

A. Protein Pre-training Models

With the widespread application of pre-trained models in fields such as natural language processing and computer vision, researchers also try to introduce pre-training techniques into the field of protein representation learning. Most protein pre-training models rely on a sizable corpus of protein sequences, using large language models (LLMs) and specific pre-training tasks including contrastive predictive coding (CPC) [1], masked language modeling (MLM) [2]–[4], pairwise MLM [5] and next amino acid prediction [3] to learn a comprehensive protein representation. With the increase of highly accurate protein structure data available, several structure-based pre-training methods have emerged as well. Guo *et al.* [6] proposed a pretrained framework for learning structure representations from tertiary structures of proteins; Hermosilla *et al.* [7] applied graph contrastive learning during the pre-training phase; Zhang *et al.* [8] and Chen *et al.* [9] proposed several self-prediction tasks, like bond distance prediction and bond angle prediction, to learn protein representations.

II. Experiments

A. Benchmark Datasets

We utilize four datasets for comparison, i.e., Enzyme Commission (EC) [10], Gene Ontology Molecular Function (GO-MF), Gene Ontology Cellular Component (GO-CC) and Gene Ontology Biological Process (GO-BP) [8], all of which are used for protein function prediction. The details of four benchmark datasets are described below:

- *EC*: A dataset about the protein's enzyme commission (EC) numbers, describing their catalysis of biochemical reactions.
- *GO-MF*: A dataset regarding the molecular function (MF) terms of proteins. The MF term, which corresponds to activities carried out by individual gene products (such as proteins or RNA), is one of the three ontologies of GO annotations.
- *GO-CC*: A dataset about the cellular component (CC) terms of proteins. The term CC refers to the location of a gene product relative to cellular components or structures when it performs its biochemical function (e.g., mitochondrion). The CC term is one of the three domains of GO as well.
- *GO-BP*: A dataset related to the biological process (BP) terms of protein prediction. The BP term is one of the three domains of Gene Ontology (GO) annotations, representing a particular organic process achieved through gene programming, such as DNA repair and signal transduction.

B. Evaluation Metrics

We evaluate SCOP by the area under the precision-recall curve (AUPR) and protein centric maximum F-score (F_{max}) [10]. AUPR summarizes the precision-recall curve by calculating the weighted precision at each threshold [**?**], and it is effective for imbalanced classification. AUPR can be formalized as follows:

$$\text{AUPR} = \sum_{n=1}^{N} \left(R_n - R_{n-1}\right) P_n \tag{1}$$

where R_n and P_n are recall and precision of the n-th threshold, and N is the overall number of thresholds. F_{max} is used to evaluate the accuracy for multi-label classification. In order to calculate F_{max}, the precision and recall of each protein are firstly calculated and then averaged across all proteins. Given a target protein i and decision threshold $\lambda \in [0, 1]$, the precision and recall for this protein is defined as:

$$\text{precision}_i(\lambda) = \frac{\sum_j^J \left(\left(p_i^j \geq \lambda\right) \cap b_i^j\right)}{\sum_j^J \left(p_i^j \geq \lambda\right)} \tag{2}$$

$$\text{recall}_i(\lambda) = \frac{\sum_j^J \left(p_i^j \geq \lambda\right)}{\sum_j^J b_i^j} \tag{3}$$

where p_i^j refers to the predicted probability of the i-th protein belonging to the j-th category, $b_i^j \in \{0, 1\}$ represents the corresponding ground-truth label, and J is the total number of categories. The average precision and recall of all proteins can be computed according to the following formula:

$$\text{precision}(\lambda) = \frac{\sum_i^N \text{precision}_i(\lambda)}{\sum_i^N \left(\left(\sum_j^J \left(p_i^j \geq \lambda \right) \right) \geq 1 \right)} \tag{4}$$

$$\text{recall}(\lambda) = \frac{1}{N} \sum_i^N \text{recall}_i(\lambda) \tag{5}$$

where N is the total count of proteins. F_{max} can be defined as:

$$\text{F}_{\max} = \max_{\lambda \in [0,1]} \left\{ \frac{2 \times \text{precision}(\lambda) \times \text{recall}(\lambda)}{\text{precision}(\lambda) + \text{recall}(\lambda)} \right\} \tag{6}$$

C. Implementation Details

We utilize TorchDrug [11] to transform the original PDB file of a protein into the protein topological graph and protein spatial graph. In order to generate initial features for residues, chemical bonds, and bond angles, we partially adopt the approach proposed by Zhang *et al*. [8]. We leverage one-hot encoding to generate the initial residue feature of a protein as a 21-dimensional tensor, representing twenty standard amino acids and one additional unknown amino acid. To represent the initial features of a chemical bond, we concatenate the original features of the two residues which form this bond with a tensor representing the bond length, resulting in a 62-dimensional tensor. The bond length is converted into a 20-dimensional tensor using a radial basis function (RBF). Similarly, the bond angle between two chemical bond is also converted into a 20-dimensional tensor using another RBF. For a protein sequence, each amino acid symbol is embedded as a 16-dimensional dense tensor. These features will be used to train structure-based and sequence-based encoders.

Given the substantial influence of the pre-training batch size on the model's performance, we select the largest batch size that the available GPU memory (24GB) could support, which is 256, and complete the pre-training. The batch size for the benchmark datasets varies depending on the specific dataset. The temperature coefficient τ_s and τ_m are initialized to 0.07. The hyper-parameter α is obtained by a grid search strategy ranging from 0 to 1 with an interval 0.1. The learning rate for pre-training and benchmark datasets is 1e-4, using AdamW optimizer.

Due to the lengthy computational time required for pre-training, we determine the numbers of layers for the sequence and structure-based encoder in advance on benchmark datasets, which are set to 2 and 6, respectively. And it allows us to achieve optimal results within limited computational resources. We train SCOP from scratch for 50 epochs.

References

[1] A. X. Lu, H. Zhang, M. Ghassemi, et al., "Self-Supervised Contrastive Learning of Protein Representations by Mutual Information Maximization," bioRxiv, 2020.

[2] A. Rives,J. Meier, T. Sercu, et al., "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proceedings of the National Academy of Sciences, vol. 118, no. 15, pp. e2016239118, 2021.

[3] A. Elnaggar, M. Heinzinger, C. Dallago, et al., "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing," bioRxiv, 2020.

[4] R. Rao, N. Bhattacharya, N. Thomas, et al., "Evaluating Protein Transfer Learning with TAPE," in NeurIPS, 2019.

[5] L. He, S. Zhang, L. Wu, et al, "Pre-training Co-evolutionary Protein Representation via A Pairwise Masked Language Model," arXiv, vol. abs/2110.15527, 2021.

[6] Y. Guo, Y. Guo, J. Wu, et al., "Self-Supervised Pre-training for Protein Embeddings Using Tertiary Structures," in AAAI, vol. 36, no. 6, pp. 6801–6809, 2022.

[7] P. Hermosilla and T. Ropinski, "Contrastive Representation Learning for 3D Protein Structures," arXiv, vol. abs/2205.15675, 2022.

[8] Z. Zhang, M. Xu, A. Jamasb, et al., "Protein Representation Learning by Geometric Structure Pretraining," in ICLR, 2023.

[9] C. Chen, J. Zhou, F. Wang, et al., "Structure-aware Protein Self-supervised Learning," arXiv, vol. abs/2204.04213, 2022.

[10] V. Gligorijević, P. D. Renfrew, T. Kosciolek, et al., "Structure-based Protein Function Prediction Using Graph Convolutional Networks," Nature Communications, vol. 12, no. 1, pp. 3168, 2021.

[11] Z. Zhu, C. Shi, Z. Zhang, et al., "Torchdrug: A Powerful and Flexible Machine Learning Platform for Drug Discovery," arXiv, vol. abs/2202.08320, 2022.