ENHANCING PROTEIN LANGUAGE MODEL WITH STRUCTURE-BASED ENCODER AND PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (PLMs) pre-trained on large-scale protein sequence corpora have achieved impressive performance on various downstream protein understanding tasks. Despite the ability to implicitly capture inter-residue contact information, transformer-based PLMs cannot encode protein structures explicitly for better structure-aware protein representations. Besides, the power of pre-training on available protein structures has not been explored for improving these PLMs, though structures are important to determine functions. To tackle these limitations, in this work, we enhance the PLM with structure-based encoder and pre-training. We first explore feasible model architectures to combine the advantages of a state-of-the-art PLM (*i.e.*, ESM-1b) and a state-of-the-art protein structure encoder (*i.e.*, GearNet). We empirically verify the **ESM-GearNet** that connects two encoders in a series way as the most effective combination model. To further improve the effectiveness of ESM-GearNet, we pre-train it on massive unlabeled protein structures with contrastive learning, which aligns representations of co-occurring subsequences so as to capture their biological correlation. Extensive experiments on EC and GO protein function prediction benchmarks demonstrate the superiority of ESM-GearNet over previous PLMs and structure encoders, and clear performance gains are further achieved by structure-based pre-training upon ESM-GearNet. The source code will be made public upon acceptance.

1 INTRODUCTION

Proteins are functional macromolecules in the cell, governing diverse biological processes and driving life itself. Machine learning methods have shown great promise in predicting protein structures (Jumper et al., 2021; Baek et al., 2021) and understanding protein functions (Gligorijević et al., 2021; Meier et al., 2021). Among these methods, protein language models (PLMs) (Elnaggar et al., 2021; Lu et al., 2020; Rives et al., 2021; Lin et al., 2022) excel at acquiring informative protein representations from large-scale protein sequence corpora and further boost protein structure and function prediction (Lin et al., 2022; Xu et al., 2022b).

Pre-trained with masked language modeling losses, existing PLMs can well capture co-evolutionary information and implicitly capture inter-residue contact information (Rives et al., 2021). However, since they do not explicitly take protein structures as input, it is questionable whether they can capture detailed protein structural characteristics. Given the importance of protein structures on determining functions, there have been some works exploring along this direction by enhancing PLMs with protein structure encoders, *e.g.* DeepFRI (Gligorijević et al., 2021) and LM-GVP (Wang et al., 2022). Nevertheless, their performance are not as good as PLMs and structure-based methods (Zhang et al., 2022b), probably due to the lack of good structure encoders. Besides, these methods only focus on a limited number of labeled structures, while ignoring abundant unlabeled structures available in PDB (Berman et al., 2000) or AlphaFold DB (Varadi et al., 2021).

In this work, we study the problem of how to enhance PLMs with structure-based encoders and pre-training methods. First, to incorporate structural information into PLMs, we combine a state-of-the-art PLM (*i.e.*, ESM-1b) and a state-of-the-art protein structure encoder (*i.e.*, GearNet) into a holistic architecture. We investigate three different model architectures that fuse the two encoders in a *parallel*, *series* or *cross* manner. Based on empirically evaluation, we identify the **ESM-GearNet** with **series fusion architecture** as the most effective sequence-structure hybrid encoder. Based on

Table 1: Comparison of different protein encoders with and without sequence or structure pre-training. MC and DP is the abbreviation of Multiview Contrast and Dihedral Prediction, respectively. Our proposed ESM-GearNet is a sequence-structure multimodal encoder and is pre-trained with unlabeled structures, which is the only model that benefits from both sequence and structure pre-training.

	Method	Sequence Encoder	Structure Encoder	Sequence Pre-Training	Structure Pre-Training
①	CNN (Rao et al., 2019) Transformer (Rao et al., 2019)	✓			
②	GVP (Jing et al., 2021) GearNet (Zhang et al., 2022b)		✓		
③	ESM-1b (Rives et al., 2021) ProtBert (Elnaggar et al., 2021)	✓		✓	
④	DeepFRI (Gligorijević et al., 2021) LM-GVP (Wang et al., 2022) **ESM-GearNet (Ours)**	✓	✓	✓	
⑤	GearNet (MC) (Zhang et al., 2022b) GearNet (DP) (Zhang et al., 2022b)		✓		✓
⑥	**ESM-GearNet (MC) (Ours)**	✓	✓	✓	✓

ESM-GearNet, we perform structure-based pre-training to further improve its effectiveness. To be specific, we adapt the Multiview Contrast algorithm proposed by Zhang et al. (2022b) to our pre-training situation. By maximizing the representation similarity between correlated protein subunits and minimizing that of uncorrelated ones, ESM-GearNet can capture the biological correlation between co-occurring sequence and structure motifs. Compared with previous protein representation learning methods, our method is the only one that considers sequences and structures and benefits both from sequence- and structure-based pre-training, as shown in Table 1.

We run experiments on EC and GO protein function prediction benchmarks. The benchmark results verify the superiority of the proposed ESM-GearNet over vanilla PLMs, various protein structure encoders and existing structure encoder enhanced PLMs. After applying structure-based pre-training, the pre-trained ESM-GearNet achieves new state-of-the-art on both benchmarks. These results illustrate the great promise of structure encoder enhanced PLMs and structure-based pre-training upon such models.

2 Related Work

Protein Language Models (PLMs). Regarding protein sequences as the language of life, PLMs (Elnaggar et al., 2021; Lu et al., 2020; Rives et al., 2021; Lin et al., 2022; Zhang et al., 2022a) aim to learn effective protein representations from large-scale protein sequence corpora. This representation learning is typically performed in a self-supervised fashion via masked language modeling (MLM) (Elnaggar et al., 2021; Rives et al., 2021; Lin et al., 2022), pairwise MLM (He et al., 2021), contrastive learning (Lu et al., 2020), *etc.* PLMs have shown impressive performance on predicting protein structures (Lin et al., 2022) and functionality (Rao et al., 2019; Xu et al., 2022b). Transformer-based PLMs are also verified to be able to implicitly capture inter-residue contact information via their intermediate attention maps (Vig et al., 2020). However, these existing PLMs cannot explicitly encode protein structures, which are actually determinants of diverse protein functions. In this work, we seek to overcome this limitation by enhancing a PLM with a protein structure encoder so as to capture detailed protein structural characteristics.

Protein Structure Encoders. Diverse types of protein structure encoders have been devised to capture different granularities of protein structures, including residue-level structures (Gligorijević et al., 2021; Zhang et al., 2022b; Xu et al., 2022a), atom-level structures (Jing et al., 2021; Hermosilla et al., 2021) and protein surfaces (Gainza et al., 2020; Sverrisson et al., 2021). These structure encoders have boosted protein function understanding (Gligorijević et al., 2021; Zhang et al., 2022b) and protein design (Jing et al., 2021; Hsu et al., 2022). In this work, we aim at injecting residue-level structural information into the protein representations learned by a PLM, where we resort to a SOTA residue-level encoder GearNet (Zhang et al., 2022b) to achieve this goal.

Protein Structure Pre-training. Various self-supervised learning algorithms are designed to learn informative protein structure representations, including contrastive learning (Zhang et al., 2022b; Hermosilla & Ropinski, 2022), self-prediction (Zhang et al., 2022b; Chen et al., 2022) and denoising score matching (Guo et al., 2022; Wu et al., 2022). Structurally pre-trained models outperform PLMs on function prediction tasks (Zhang et al., 2022a; Hermosilla & Ropinski, 2022), given the principle that protein structures are the determinants of their functions. Inspired by this fact, we perform structure-based pre-training to improve the effectiveness of the proposed sequence-structure hybrid encoder ESM-GearNet.

3 METHODS

3.1 BACKGROUND

Proteins. Proteins are macromolecules consisting of a series of residues, *a.k.a.* amino acids, arranged in one or more chains. Although there are only 20 different types of standard residues, their exponential combination makes up the vast variety of proteins in nature. Protein structures are the 3D coordinates of all atoms and mainly determined by the arrangement of these residues. Each residue contains an amino group, a carboxylic acid group, and a side chain group that determines its type, which are all attached to a central carbon atom, called the alpha carbon. In this work, we simply keep the coordinates of alpha carbon atoms for representing the backbone structure of each protein.

Pre-Training with Protein Sequences. Treating protein sequences as the language of life, recent works draw inspirations from large pre-trained language models to learn the evolutionary information from billions of protein sequences via self-supervised learning. Prominent protein language models (PLM) include transformer-based ProtTrans (Elnaggar et al., 2021) and ESM (Rives et al., 2021). These models are pre-trained with masked language modeling (MLM) loss by predicting the type of a masked residue given the surrounding context. By fully utilizing massive unlabeled data, these models have achieved state-of-the-art performance on various protein understanding tasks (Lin et al., 2022; Elnaggar et al., 2023).

Pre-Training with Protein Structures. The success of AlphaFold2 (Jumper et al., 2021) enables accurate protein structure prediction and incentivizes a series of works on pre-training with protein structures (Zhang et al., 2022b; Chen et al., 2022; Zhang et al., 2023), as structures are a direct determinant of protein functions. Given a protein, Zhang et al. (2022b) constructs a residue graph for encoding its structure, the edges of which are classified into different types based on the sequential distance, spatial distance and knn neighbors. Several structure-based pre-training methods have been proposed via self-prediction (Zhang et al., 2022b), multiview contrast (Chen et al., 2022) and denoising objective (Zhang et al., 2023). Pre-trained on a much smaller dataset, structure-based methods have achieved competitive or even better results than sequence-based methods on function prediction tasks (Zhang et al., 2022b).

3.2 ENHANCING PROTEIN LANGUAGE MODEL WITH PROTEIN STRUCTURES

Although protein language models are able to implicitly capture structural contact information (Rives et al., 2021), incorporating detailed structures explicitly can be an effective way to model spatial interactions between residues. Therefore, in this subsection, we propose to enhance protein language model with a protein structure encoder. We choose ESM-1b (Rives et al., 2021) as our baseline backbone model, which takes protein sequences as input and outputs representations for each residue. For structure encoders, following (Zhang et al., 2022b), we construct a residue graph based on protein structures and feed one-hot features of residue types into GearNet as node features. Then, we consider three different fusion methods to combine sequence and structure encoders as shown in Figure 1:

1. *Parallel*: we concatenate the output of ESM-1b and GearNet as the final representations.
2. *Series*: we replace the node features of GearNet with the output of ESM-1b and use the output of GearNet as final representations.
3. *Cross*: we concatenate the output of ESM-1b and GearNet and then feed them into a transformer to perform cross-attention between modalities. The output of the transformer will be used as final representations.

Finally, the final representation will be used for residue-level or protein-level prediction.

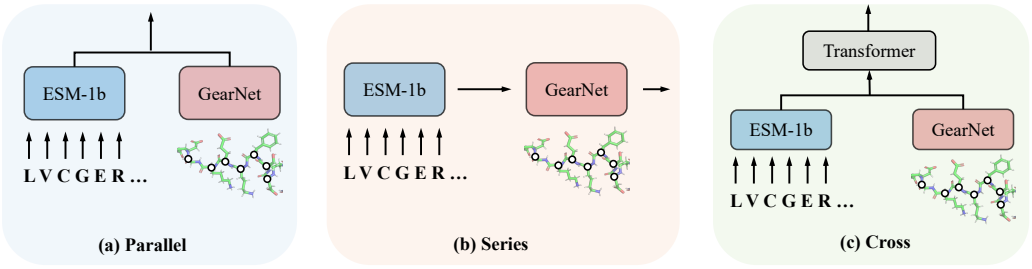

Figure 1: Three different ways to fuse protein sequence and structure encoder. (a) *Parallel*. The outputs of ESM-1b and GearNet are concatenated. (b) *Series*. The output of ESM-1b is fed into GearNet. (c) *Cross*. The outputs of ESM-1b and GearNet are fused via a cross-modal transformer.

In practice, to avoid changing sequence representations dramatically, we use a smaller learning rate for ESM-1b than GearNet. We set the ratio of two learning rates as 0.1 in our experiments and find that this trick is important for the generalization of ESM-GearNet. Also, as shown in Section 4.3, we find that the series connection is the most simple and effective method. Hence, we use this method for fusion between sequences and structures and refer this multi-modal encoder as ESM-GearNet.

3.3 Pre-Training ESM-GearNet with Unlabeled Protein Structures

The current encoder ESM-GearNet is able to extract information learned from massive unlabeled protein sequences and then can be fine-tuned on downstream tasks with labeled protein structures and sequences. Now we move a further step to fully utilize available unlabeled protein structures for pre-training. We select the representative pre-training method in Zhang et al. (2022b): *Multiview Contrast*. We use subsequence operation to extract correlated protein substructures and then apply a random edge masking operation to add noises. Here we discard the subspace operation in the original paper, because PLMs only take consecutive sequences as input. We align their representations from ESM-GearNet in the latent space with an InfoNCE loss. During pre-training, we fix the ESM and only tune the GearNet encoder, which can keep sequence representations from being destroyed.

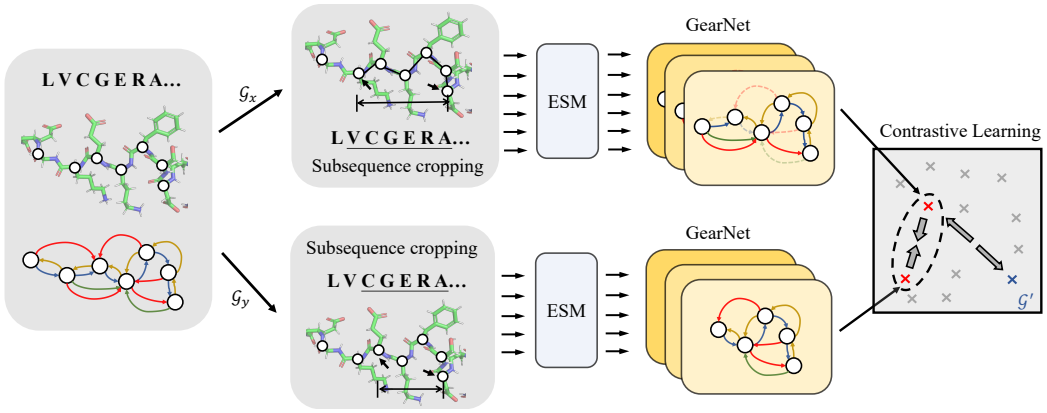

Figure 2: High-level illustration of ESM-GearNet pre-trained with Multiview Contrast. For each protein, we randomly sample subsequences $\mathcal{G}_x$ and $\mathcal{G}_y$ and randomly mask some edges to add noises. Encoding with ESM-GearNet, their representations are aligned in the latent space while minimizing its similarity with a negative sample $\mathcal{G}'$.

In practice, we find pre-training with *Multiview Contrast* can achieve significant improvements over ESM-GearNet. This can be explained as follows. Compared with methods that only apply structural losses for pre-training, *Multiview Contrast* can consider sequence and structure dependency more comprehensively. By aligning representations from subsequences extracted from the same protein, *Multiview Contrast* captures the sequence co-occurance and structure motif dependency, which fully

Table 2: F_{max} and AUPR on EC and GO prediction. ①, ② include sequence and structure encoders without pre-training. ③ ④ include baselines built based on protein language models. ⑤ ⑥ include encoders with structure-based pre-training. The best results in each group are marked in bold.

	Method	EC		GO-BP		GO-MF		GO-CC	
		F_{max}	AUPR	F_{max}	AUPR	F_{max}	AUPR	F_{max}	AUPR
①	CNN	0.545	0.526	0.244	0.159	0.354	0.351	0.287	0.204
	Transformer	0.238	0.218	0.264	0.156	0.211	0.177	0.405	0.210
②	GVP	0.489	0.482	0.326	**0.224**	0.426	0.458	**0.420**	**0.279**
	GearNet	**0.730**	**0.751**	**0.356**	0.211	**0.503**	**0.490**	0.414	0.276
③	ESM-1b	0.864	0.889	0.452	0.332	0.657	0.639	0.477	0.324
	ProtBERT-BFD	0.838	0.859	0.279	0.188	0.456	0.464	0.408	0.234
④	DeepFRI	0.631	0.547	0.399	0.282	0.465	0.462	0.460	0.363
	LM-GVP	0.664	0.710	0.417	**0.302**	0.545	0.580	**0.527**	**0.423**
	ESM-GearNet	**0.883**	**0.890**	**0.491**	**0.301**	**0.677**	**0.632**	0.501	0.345
⑤	GearNet-Edge (Multiview Contrast)	0.874	0.892	0.490	0.292	0.654	0.596	0.488	0.336
	GearNet-Edge (Dihedral Prediction)	0.859	0.881	0.458	0.304	0.626	0.603	0.465	0.338
⑥	**ESM-GearNet (Multiview Contrast)**	**0.894**	**0.907**	**0.516**	**0.301**	**0.684**	**0.621**	**0.506**	**0.359**

utilizes the representations from ESM-1b and GearNet and thus is beneficial for function prediction. We illustrates the idea of *Multiview Contrast* in Figure 2.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed enhancement on two protein function prediction tasks: Enzyme Commission number prediction, Gene Ontology term prediction.

4.1 EXPERIMENTAL SETUP

Pre-training datasets. We follow (Zhang et al., 2022b) to use the AlphaFold protein structure database v1 and v2 (Varadi et al., 2021) for pre-training, which contains 365K proteome-wide predictions and 440K Swiss-Prot (Consortium, 2021) predictions from AlphaFold2.

Downstream datasets. We adopt two function prediction tasks proposed in Gligorijević et al. (2021) for downstream evaluation. *Enzyme Commission (EC) number prediction* predicts the EC numbers of proteins, characterizing biochemical reactions they catalyze. *Gene Ontology (GO) term prediction* seeks to classify proteins into hierarchically related functional classes organized into three ontologies: molecular function (MF), biological process (BP) and cellular component (CC). Following Gligorijević et al. (2021), the dataset are split according to sequence identity cutoff to ensure the test set only contains PDB chains with sequence identity no more than 95% to the training set. We report the protein-centric maximum F-score F_{max} and pair-centric area under precision-recall curve AUPR, which are commonly used in the CAFA challenges (Radivojac et al., 2013).

Baselines. We select representative baselines for each category listed in Table 1. We use CNN (Shanehsazzadeh et al., 2020) and Transformer (Rao et al., 2019) for sequence-based encoders ① and GVP (Jing et al., 2021) and GearNet (Zhang et al., 2022b) for structure-basde encoders ②. For protein language models ③, we choose ESM-1b (Rives et al., 2021) and ProtBERT-BFD (Elnaggar et al., 2021) as baselines. For models combining pre-trained PLM with structural information ④, we run DeepFRI (Gligorijević et al., 2021) and LM-GVP (Wang et al., 2022) as baselines besides our ESM-GearNet. For encoders pre-trained with structure-based methods ⑤, we use the state-of-the-art model GearNet-Edge with two strong pre-training methods (Multiview Contrast and Dihedral Prediction) as baselines. As shown in Table 1, our method ESM-GearNet pre-trained with protein structures are the only model enjoying both sequence and structure-based pre-training ⑥. These methods are categorized into three groups according to their pre-training schemes: ①② w/o pre-training, ③④ w/ sequence pre-training, ⑤⑥ w/ sequence and structure pre-training.

Training. We pre-train the models for 50 epochs on AlphaFold Database and for 200 epochs on EC and GO prediction. We use the default hyperparameter configuration in (Zhang et al., 2022b) for pre-training and fine-tuning. For Multiview Contrast, we use the cropping length of subsequence as

50, the mask rate of random edge masking as 0.15. The temperature in InfoNCE loss is set as 0.07. The model is pre-trained with batch size 256 and learning rate 2e-4. For downstream evaluation, we use batch size 2 and learning rate 1e-4 with ReduceLROnPlateau scheduler with factor 0.6 and patience 5. All these models are implemented with TorchDrug library (Zhu et al., 2022) and trained on 4 Tesla A100 GPUs.

4.2 Results

The results on EC and GO prediction are reported in Table 2, showing results of all baselines in six categories. The effectiveness of our methods can be demonstrated by the following observations.

First, comparing results in the first (①②) and second group (③④), we can observe the significant benefits from sequence-based pre-training. Then, comparing ③ and ④, we find that our proposed ESM-GearNet significantly improves the language model ESM-1b by incorporating structural information, while the other two (DeepFRI and LM-GVP) fail to beat PLM baselines. Finally, after pre-training with Multiview Contrast, the performance of ESM-GearNet increases by a large margin and achieve SOTA results in terms of F_{max}, better than encoders with only structure-based pre-training in ⑤.

4.3 Results of Different Fusion Schemes in ESM-GearNet

As discussed in Section 3.2, we try three different ways to fuse representations from sequence and structure encoders: parallel, series and cross. We conduct an experiment on EC to test their performance, the results of which are reported in Table 3. Surprisingly, we find that direct concatenating outputs of sequence and structure encoders does not perform well, even worse than the protein language model ESM-1b itself. This is probably because the sequence encoder is much larger than the structure encoder and have already been pre-trained on a large corpus. Learning with an unpre-trained structure encoder simultaneously may destroy the learned representations. This problem can be addressed by introducing a cross-modality transformer as in the cross fusion scheme. Since the performance of cross fusion is similar to that of series fusion, we choose to use the simpler series fusion in our study, which introduces no additional learnable parameters besides those of two encoders.

Method	**F_{max}**	**AUPR**
ESM-1b	0.864	0.889
ESM-GearNet		
- w/ parallel fusion	0.733	0.759
- w/ series fusion	**0.883**	0.890
- w/ cross fusion	0.880	**0.893**

Table 3: Results of different fusion schemes on EC.

5 Conclusions and Future Work

In this work, we propose to enhance protein language models with structure-based encoders and pre-training. We investigate three different ways to fuse sequence and structure encoders and find that series connection is the most effective way. Moreover, we pre-train the sequence-structure hybrid encoder with contrastive learning on massive unlabeled structures in AlphaFold DB. As the first method enjoying both sequence- and structure-based pre-training, ESM-GearNet (Multiview Contrast) achieves the state-of-the-art performance on two function prediction tasks. Future directions include exploring the usage of the proposed models on more downstream applications, including protein engineering tasks and protein-protein interactions.

References

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

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*, 28(1): 235–242, 2000.

Can Chen, Jingbo Zhou, Fan Wang, Xue Liu, and Dejing Dou. Structure-aware protein self-supervised learning. *arXiv preprint arXiv:2204.04213*, 2022.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Fouad Omar Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *bioRxiv*, 2023.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

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*, 12 (1):1–14, 2021.

Yuzhi Guo, Jiaxiang Wu, Hehuan Ma, and Junzhou Huang. Self-supervised pre-training for protein embeddings using tertiary structures. In *AAAI*, 2022.

Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021.

Pedro Hermosilla and Timo Ropinski. Contrastive representation learning for 3d protein structures. *arXiv preprint arXiv:2205.15675*, 2022.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *International Conference on Learning Representations*, 2021.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022.

Bowen Jing, Stephan Eismann, Pratham N. Soni, and Ron O. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=1YLJDvSx6J4`.

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*, 596(7873):583–589, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan M Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, 2020.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.

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*, 10(3):221–227, 2013.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019.

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. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Amir Shanehsazzadeh, David Belanger, and David Dohan. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443*, 2020.

Freyr Sverrisson, Jean Feydy, Bruno E Correia, and Michael M Bronstein. Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15272–15281, 2021.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 2021.

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.

Zichen Wang, Steven A Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O Salawu, Colby J Wise, Sri Priya Ponnapalli, et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific reports*, 12(1):1–12, 2022.

Fang Wu, Qiang Zhang, Dragomir Radev, Yuyang Wang, Xurui Jin, Yinghui Jiang, Zhangming Niu, and Stan Z Li. Pre-training of deep protein models with molecular dynamics simulations for drug binding. *arXiv preprint arXiv:2204.08663*, 2022.

Minghao Xu, Yuanfan Guo, Yi Xu, Jian Tang, Xinlei Chen, and Yuandong Tian. Eurnet: Efficient multi-range relational modeling of spatial multi-relational data. *arXiv preprint arXiv:2211.12941*, 2022a.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. *arXiv preprint arXiv:2206.02096*, 2022b.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022a.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022b.

Zuobai Zhang, Minghao Xu, Aurélie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Physics-inspired protein encoder pre-training via siamese sequence-structure diffusion trajectory prediction. *arXiv preprint arXiv:2301.12068*, 2023.

Zhaocheng Zhu, Chence Shi, Zuobai Zhang, Shengchao Liu, Minghao Xu, Xinyu Yuan, Yangtian Zhang, Junkun Chen, Huiyu Cai, Jiarui Lu, et al. Torchdrug: A powerful and flexible machine learning platform for drug discovery. *arXiv preprint arXiv:2202.08320*, 2022.