STRUCTURE-INFORMED PROTEIN LANGUAGE MODEL

Zuobai Zhang[1,2], Jiarui Lu[1,2], Vijil Chenthamarakshan[3],
Aurélie Lozano[3], Payel Das[3], Jian Tang[1,4,5]
Mila - Québec AI Institute[1], Université de Montréal[2]
IBM Research[3], HEC Montréal[4], CIFAR AI Chair[5]
{zuobai.zhang, jiarui.lu}@mila.quebec,
{aclozano,ecvijil,daspa}@us.ibm.com, jian.tang@hec.ca

ABSTRACT

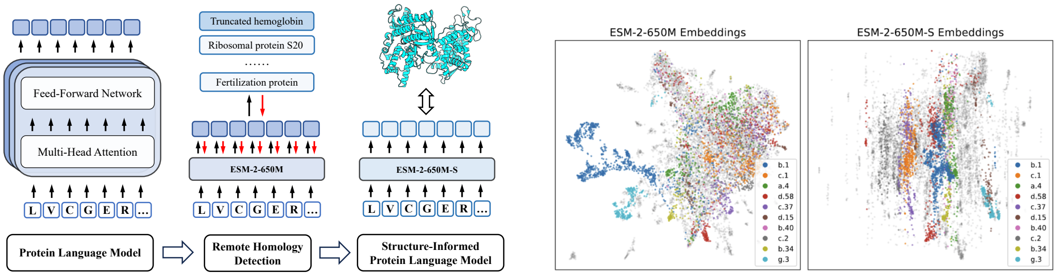
Protein language models are a powerful tool for learning protein representations through pre-training on vast protein sequence datasets. However, traditional protein language models lack explicit structural supervision, despite its relevance to protein function. To address this issue, we introduce the integration of remote homology detection to distill structural information into protein language models without requiring explicit protein structures as input. We evaluate the impact of this structure-informed training on downstream protein function prediction tasks. Experimental results reveal consistent improvements in function annotation accuracy for EC number and GO term prediction. Performance on mutant datasets, however, varies based on the relationship between targeted properties and protein structures. This underscores the importance of considering this relationship when applying structure-aware training to protein function prediction tasks. Code and model weights are available at https://github.com/DeepGraphLearning/esm-s.

1 INTRODUCTION

Proteins play a fundamental role in biological processes, and a deeper understanding of them can pave the way for groundbreaking advancements in medical, pharmaceutical, and genetic research. A cutting-edge technology that has emerged to represent proteins is the Protein Language Model (PLM) (Rao et al., 2019a). Inspired by Natural Language Processing (NLP) methodologies, PLMs have demonstrated remarkable performance in capturing long-range residue correlations - also known as co-evolution - through self-supervised training on vast repositories of protein residue sequences (Rao et al., 2021). Prominent PLMs like ESM (Rives et al., 2021b; Lin et al., 2023) have shown the ability to implicitly capture evolutionary and structural information and demonstrate outstanding performance across various tasks related to protein structures and functions.

However, vanilla PLMs face a significant limitation: their absence of explicit supervision based on protein structure information, despite its critical relevance to protein function. To address this limitation, recent studies have developed models that combine large-scale pre-training on protein sequences with the integration of structural information as input (Zhang et al., 2023a; Su et al., 2023). While these models have demonstrated impressive performance in function prediction, their reliance on protein structures as input introduces an additional computational burden for structure prediction and limits their application to proteins with indistinct structures, such as mutant data. Therefore, a key question remains: how to distill structural knowledge into protein language models without requiring explicit structure as input, and how will this impact downstream function prediction tasks?

In this study, we investigate the use of remote homology detection tasks as a means of incorporating structural information into protein language models (Chen et al., 2018). This task aims to identify proteins with similar structures but low sequence similarity, thus complementing the training of protein language models. We train ESM-2 models (Lin et al., 2023) on this task and obtain *structure-informed protein language models*. To assess the impact of structurally training, we evaluate our models on downstream function prediction tasks taken from Gligorijević et al. (2021), Xu et al. (2022), and Dallago et al. (2021). We find that incorporating structural information leads to consistent improvement in function annotation tasks like Enzyme Commission (EC) number and Gene Ontology



(A) Training Process.

(B) UMAP Embeddings on SCOPe.

Figure 1: Illustration of Training Procedure and Embeddings for Structure-Informed Protein Language Models. (A) Protein language models like ESM-2-650M are enhanced with structural information through training on remote homology detection tasks. This process results in the structure-informed model, ESM-2-650M-S, whose embeddings represent more structural characteristics. (B) We present UMAP embeddings of both ESM-2-650M and ESM-2-650M-S on the SCOPe dataset. After targeted training, ESM-2-650M-S embeddings show improved separability for different protein folds.

(GO) term prediction. However, improvement on mutant data highly depends on the targeted property's relationship to protein structures. This highlights the importance of considering the relationship between protein structures and targeted properties when applying structure-informed training to protein function prediction tasks. We hope this study encourages further exploration of structural knowledge in protein language models, leading to better protein representation learning.

Related work. Protein language models view protein sequences as the language of life and employ masked language modeling loss for pre-training transformer-based models (Rao et al., 2019b; Elnaggar et al., 2021; Rives et al., 2021a; Elnaggar et al., 2023). Structure-based representation learning has also shown promise in incorporating structural information for better representation learning methods (Jing et al., 2021; Zhang et al., 2023b; Chen et al., 2022; Zhang et al., 2023c). To combine the advantages of both approaches, recent studies have designed architectures that take both sequences and structures as input (Zhang et al., 2023a; Su et al., 2023). However, these methods require protein structures as input, limiting their application to datasets without protein structures. In this study, we propose the use of remote homology detection to inject structural information into protein language models. While this concept has been utilized in previous works (Bepler & Berger, 2018; Hamamsy et al., 2022), their focus is on structural alignment rather than representation learning. Here, we investigate the impact of incorporating structural information on learning protein representations by evaluating our models on downstream function prediction tasks.

2 Method

Proteins. Proteins are made up of amino acids, also known as residues, that form chains via peptide bonds. There are 20 standard types of residues, and their varied combinations lead to the vast variety of proteins in nature. The specific arrangement of these residues is crucial in determining the three-dimensional coordinates of every atom in the protein, thus shaping what is known as protein structure. Protein structures are known to be a direct determinant of protein functions. In this study, we focus on learning representations based on protein sequences. These sequences are denoted as $\mathcal{R} = [r_1, r_2, \cdots, r_n]$, where each $r_i \in \{1, ..., 20\}$ corresponds to the type of the i-th residue.

Protein Language Models. To effectively encode protein sequences, recent research has treated them as the "language of life", employing methods from large pre-trained language models. This approach aims to capture evolutionary patterns across billions of protein sequences using self-supervised learning. A notable instance of this is the transformer-based protein language model ESM (Rives et al., 2021a; Lin et al., 2023). This model takes residue type sequences as input, integrating several self-attention layers and feed-forward networks to model dependencies among residues. These models are pre-trained with a masked language modeling (MLM) loss, which involves predicting the type of a masked residue based on its surrounding context. An additional linear head uses the final-layer representations for this prediction. The loss function for each sequence is defined as:

$$\mathcal{L}_{MLM} = -\mathbb{E}_M[\textstyle\sum_{i \in M} \log p(r_i | r_{/M})]. \tag{1}$$

Here, a randomly chosen set of indices M is used for masking, replacing the true token at each index i with a mask token. The model's objective is to minimize the negative log likelihood of the correct residue r_i, using the masked sequence $r_{/M}$ as context. By leveraging vast amounts of unlabeled data, these models have set new benchmarks in a variety of protein-related tasks (Lin et al., 2023).

Injecting Structural Information via Protein Remote Homology Detection. Protein language models are known to reflect sequence similarity, thereby facilitating the identification of protein homology – the shared ancestry in the evolutionary history of life (Rives et al., 2021a). The challenge in this field, however, goes beyond simple homology detection. Over the course of natural evolution, protein structures and functions tend to be more conserved than their sequences (Pál et al., 2006; Liu et al., 2014). This means proteins with similar structures and functions might exhibit low sequence identities. Therefore, in the realm of protein homology detection, it is relatively straightforward to identify homologs with high sequence identity, but far more challenging to detect those with low sequence identity. This specific task of identifying homologous proteins that share structural and functional similarities but differ significantly in sequence is termed *protein remote homology detection*. The ability to detect remote homologous proteins is crucial in various fields, including proteomics (Kim et al., 2014) and biomedical sciences (Standley et al., 2008).

While Rives et al. (2021a) has shown that masked language modeling pre-training enables ESM representations to capture remote homology information, these protein language models do not inherently process protein structures as input, nor are they trained with any specific structural loss. To explicitly incorporate structural information into these models, we now take the step to fine-tune the ESM model specifically for the task of protein remote homology detection.

We employ the remote homology detection dataset from Hou et al. (2018), which is derived from SCOPe 1.75 (Murzin et al., 1995). This dataset is composed of genetically distinct domain sequence subsets that share less than 95% identity. It includes a total of 12,312 proteins, categorized into 1,195 distinct folds, where proteins within the same fold exhibit similar structure patterns. To adapt our protein language model for this task, we fine-tune it to take protein sequences as input and attach an MLP head for predicting the fold class label of each protein. Formally, our objective is to train a protein language model with parameters ϕ through training on a protein database $\mathcal{R}_D$ with corresponding fold labels c_D. The optimization involves maximizing the log likelihood:

$$\max_\phi \ \log p_\phi(c_D|\mathcal{R}_D) = \textstyle\sum_{n \in D} \sum_c [c_n = c] \log p_\phi(c_n = c|\mathcal{R}_n). \tag{2}$$

Through this training process, we aim to enhance the model's ability to generate similar representations for proteins that belong to the same fold, thereby injecting structural information. The high-level idea and effect of structural information injection is shown in Fig. 1. In practice, limited by the computation resources, we only consider ESM-2-{8,35,150,650}M for illustration and exclude ESM-2-3B and ESM-2-15B. The models are trained on the dataset for 50 epochs using the Adam optimizer with a batch size of 8. To preserve the pre-trained representations, we set the learning rate for ESM to 1e-5, while the learning rate for the prediction head is set to 1e-4. Models after training on remote homology detection are denoted with a suffix "-S".

3 Experiment

In this section, we aim to evaluate the impact of incorporating structural information by comparing protein language models with structure-informed models across various protein function prediction tasks. Our experiments are run upon ESM-2-{8,35,150,650}M, which serves as our protein sequence feature extractor. We explore two approaches to leverage these features: (1) feeding them into a 2-layer Multilayer Perceptron (MLP) predictor (see Sec. 3.1) and (2) utilizing them as a similarity metric to retrieve proteins with similar characteristics (see Sec. 3.2). Predictor-based methods aim to determine whether the representations are easily distinguishable for different functions, while retriever-based methods explore whether structural similarity aids in determining protein functions.

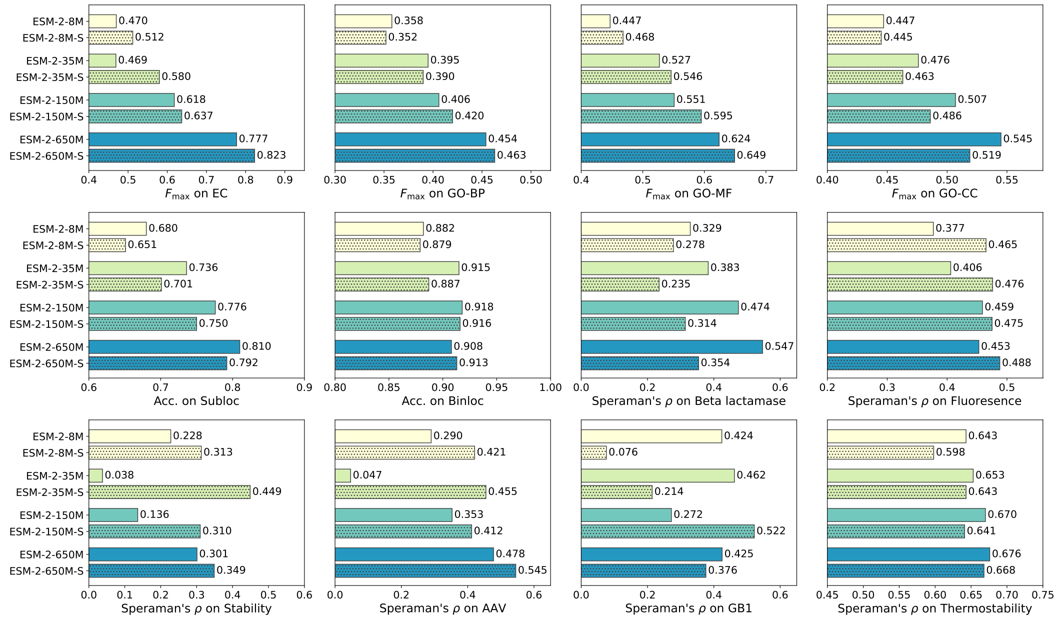


Figure 2: Results on function prediction tasks with various sizes of ESM-2 models as feature extractors. Structure-informed models are denoted with suffixes "-S" and highlighted with dots.

3.1 Evaluation with Predictor-based Methods

Setup. We evaluate the methods using three different categories of function prediction tasks. The first category is function annotation tasks in Gligorijević et al. (2021), *i.e.*, Enzyme Commission (EC) prediction and Gene Ontology (GO) prediction. EC prediction aims to determine whether a protein can catalyze a biochemical reaction, while GO prediction aims to identify a protein's involvement in specific molecular functions (MF), biological processes (BP), and cellular components (CC). We split them based on sequence identity cutoff, with the test sets containing sequences that have no more than 95% similarity to the training set. The evaluation of performance is based on the protein-centric maximum F-score, denoted as $\mathrm{F}_{\max}$ (Radivojac et al., 2013).

In addition to function annotation tasks, we also include protein localization prediction tasks, which are related to the in vivo functionality of proteins. For this evaluation, we utilized two datasets from Almagro Armenteros et al. (2017): Subcellular localization and Binary localization prediction. These tasks aim to predict the cellular location of natural proteins and are measured by accuracy.

Next, we select a subset of mutation-based tasks from Xu et al. (2022) and Dallago et al. (2021). These include Beta-lactamase activity (Gray et al., 2018), Fluorescence (Sarkisyan et al., 2016), Stability (Rocklin *et al.*, 2017), AAV fitness (Bryant et al., 2021), GB1 fitness (Wu et al., 2016) and Thermostability (Jarzab et al., 2020). The datasets are split based on the number of mutations and measured by Spearman's correlation. Please refer to the original papers for more details.

Results. We freeze the protein language model encoders and feed their outputs into a two-layer MLP head for prediction. The MLP head are trained for 100 epochs on each task. The results are reported in Fig. 2. From the figure, it can be observed that the improvement or decline brought by structurally training are consistent across different sizes of protein language models. We can gain insights from the results based on different types of tasks. Firstly, structure-informed ESMs consistently outperform vanilla ESMs in function annotation tasks such as EC, GO-BP, and GO-CC. This can be attributed to the fact that protein structures directly determine their functions, such as catalysis. However, for tasks related to cellular location like GO-CC, Subloc, and Binloc, it can be observed that the models perform worse after incorporating structure information. This is likely because protein structures have little influence on where proteins perform their functions. As for the remaining tasks based on protein mutants, whether structure-informed models provide benefits depends on whether sequence-based evolutionary information plays a more crucial role than structural information in determining the functions. In summary, it is important to consider the relationship between protein structures and targeted properties when applying structure-informed training for function prediction.

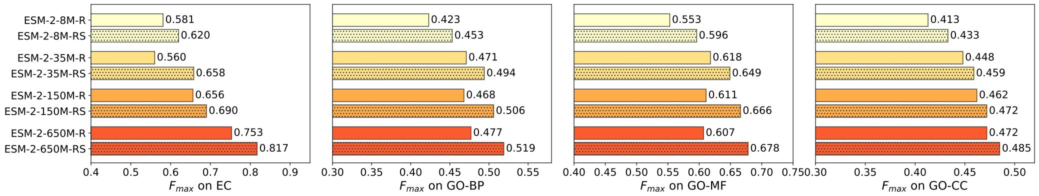


Figure 3: F_{max} on function annotation with various sizes of ESM-2 models as retrievers with suffixes "-R". Structure-informed retrievers are denoted with suffixes "-RS" and highlighted with dots.

3.2 Evaluation with Retriever-based Methods

Besides predictor-based methods, another approach for protein function annotation is to annotate function labels based on labels from similar proteins. In this subsection, we evaluate the capability of ESM and structure-informed ESM for measure protein similarity for function annotation.

We first focus on the EC and GO prediction tasks previously discussed. To annotate an unseen protein within the test set, we calculate the cosine similarity between the language model representations of proteins and select the top-5 proteins with the highest similarity scores. Subsequently, we determine the probability of each label by averaging the retrieved proteins' labels, weighted by their similarity scores. The results are presented in Fig. 3. Notably, when utilizing structure-informed training as retrievers, we observe consistent improvement across all tasks and model sizes. This finding further emphasizes the impact of protein structures in determining their functions.

In addition, we also explore studies that test EC number annotation under more realistic and challenging settings (Yu et al., 2023; Sanderson et al., 2021). We utilize the Swiss-Prot dataset collected in Yu et al. (2023) with 227,363 protein sequences, as the retrieval dataset. We then test various retriever-based methods on two independent test sets. The first, an enzyme sequence dataset, includes 392 sequences that span 177 different EC numbers. These sequences were released after April 2022, which reflects a real-world scenario where the functions of the query sequences are still unknown. The second test set, known as Price-149, is a benchmark dataset curated by Sanderson et al. (2021). It consists of experimentally validated findings from the study by Price et al. (2018). This dataset includes sequences that were previously mislabeled or inconsistently annotated in automated systems, making it a challenging benchmark for evaluation. To establish baselines, we consider four EC number prediction tools: CLEAN (Yu et al., 2023), ProteInfer (Sanderson et al., 2021), ECPred (Dalkiran et al., 2018) and DeepEC (Ryu et al., 2019). The results of these tools are directly taken from the CLEAN paper by (Yu et al., 2023). For comparison, we test the performance of two neural retrievers introduced in our paper : ESM-2-650M-R and ESM-2-650M-RS.

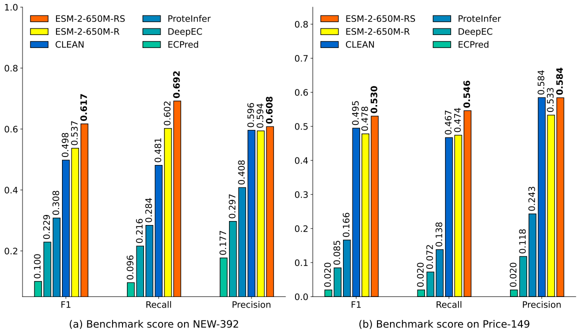


Figure 4: Results of EC annotation on NEW-392 and Price-149 test sets. Two proposed retrievers are in warm colors, whereas other baselines are in cold colors.

The results are plotted in Fig. 4. Both our retrievers surpass the performance of CLEAN on the NEW-392 test set in F1 score, despite not undergoing any supervised training on the training set, a process that CLEAN underwent. This underscores the potency of protein language models. Furthermore, the strategy of integrating structural insights into ESM proves to be effective for EC number prediction, with observable enhancements on both test sets. Specifically, on the more challenging Price-149 set, while CLEAN slightly outperforms ESM-2-650M, it falls short against ESM-2-650M when structural information is incorporated. This reaffirms the significance of structural similarity in function similarity assessments. To conclude, structure-informed training continues to demonstrate potential in practical function annotation scenarios, emphasizing the critical role of modeling structural similarities between proteins.

4 Conclusion

In this study, we investigate the integration of remote homology detection tasks for infusing structural information into protein language models. Our experimental findings indicate that incorporating structural information leads to consistent enhancement in function annotation accuracy. Nonetheless, it remains crucial to consider the connection between protein structures and targeted properties when applying structure-informed training. We envision that this study paves the way for further exploration of structural distillation techniques to enhance protein language models.

Acknowledgments

This project is supported by AIHN IBM-MILA partnership program, the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, the Canada CIFAR AI Chair Program, collaboration grants between Microsoft Research and Mila, Samsung Electronics Co., Ltd., Amazon Faculty Research Award, Tencent AI Lab Rhino-Bird Gift Fund, a NRC Collaborative R&D Project (AI4D-CORE-06) as well as the IVADO Fundamental Research Project grant PRF-2019-3583139727.

References

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. In *International Conference on Learning Representations*, 2018.

Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

Can (Sam) Chen, Jingbo Zhou, Fan Wang, Xue Liu, and Dejing Dou. Structure-aware protein self-supervised learning. *Bioinformatics*, 39, 2022.

Junjie Chen, Mingyue Guo, Xiaolong Wang, and Bin Liu. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Briefings in bioinformatics*, 19(2):231–244, 2018.

Alperen Dalkiran, Ahmet Sureyya Rifaioglu, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan. Ecpred: a tool for the prediction of the enzymatic functions of protein sequences based on the ec nomenclature. *BMC bioinformatics*, 19(1):1–13, 2018.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 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 Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *bioRxiv*, pp. 2023–01, 2023.

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.

Vanessa E Gray, Ronald J Hause, Jens Luebeck, Jay Shendure, and Douglas M Fowler. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Systems*, 6(1):116–124, 2018.

Tymor Hamamsy, James T Morton, Daniel Berenberg, Nicholas Carriero, Vladimir Gligorijevic, Robert Blackwell, Charlie EM Strauss, Julia Koehler Leman, Kyunghyun Cho, and Richard Bonneau. Tm-vec: template modeling vectors for fast homology detection and alignment. *bioRxiv*, pp. 2022–07, 2022.

Jie Hou, Badri Adhikari, and Jianlin Cheng. Deepsf: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 2018.

Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, et al. Meltome atlas—thermal proteome stability across the tree of life. *Nature methods*, 17(5):495–503, 2020.

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`.

Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Bin Liu, Deyuan Zhang, Ruifeng Xu, Jinghao Xu, Xiaolong Wang, Qingcai Chen, Qiwen Dong, and Kuo-Chen Chou. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics*, 30(4):472–479, 2014.

Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.

Csaba Pál, Balázs Papp, and Martin J Lercher. An integrated view of protein evolution. *Nature reviews genetics*, 7(5):337–348, 2006.

Morgan N Price, Kelly M Wetmore, R Jordan Waters, Mark Callaghan, Jayashree Ray, Hualan Liu, Jennifer V Kuehl, Ryan A Melnyk, Jacob S Lamson, Yumi Suh, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706):503–509, 2018.

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, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019a.

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*, 2019b.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 18–24 Jul 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021a.

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):e2016239118, 2021b.

Gabriel J Rocklin *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, 116(28):13996–14001, 2019.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. Proteinfer: deep networks for protein functional inference. *Biorxiv*, pp. 2021–09, 2021.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Daron M Standley, Akira R Kinjo, Kengo Kinoshita, and Haruki Nakamura. Protein structure databases with new web services for structural biology and biomedical research. *Briefings in bioinformatics*, 9(4):276–285, 2008.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023.

Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.

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. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures. *arXiv preprint arXiv:2303.06275*, 2023a.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2023b.

Zuobai Zhang, Minghao Xu, Aurelie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Pre-training protein encoder via siamese sequence-structure diffusion trajectory prediction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.