Evaluating task specific finetuning for protein language models

Robert Schmirler[*1 2 3] **Michael Heinzinger**[2 3] **Burkhard Rost**[2 4 5]

Abstract

Prediction methods inputting embeddings from protein Language Models (pLMs) have reached or even surpassed state-of-the-art (SOTA) performance on many protein prediction tasks. In natural language processing (NLP) fine-tuning Language Models has become the *de facto* standard. In contrast, most protein-prediction tasks do not backpropagate to the pLM. Here, we compared the use of pre-trained embeddings to fine-tuning three SOTA pLMs (ESM2, ProtT5, Ankh) on eight different tasks. Two results stood out: (1) task-specific supervised fine-tunig mostly increased downstream prediction performance. (2) Parameter-efficient fine-tuning could reach similar improvements consuming substantially fewer resources. These findings suggest task-specific fine-tuning as a generic improvement of pLM-based prediction methods. To help kick-off such an advance, we provided easy-to-use notebooks for parameter efficient fine-tuning of ProtT5 for per-protein (pooling) and per-residue prediction tasks at `https://github.com/agemagician/ProtTrans/tree/master/Fine-Tuning`.

1 Introduction

Transformer [1] based language models (LMs) changed Natural Language Processing (NLP). Large language models (LLMs) tend to perform at or above average human level performance. For the newest generation such as GPT4 [2] and PaLM2 [3] have affected fields from computer vision [4] to time series forecasting [5] and biologic language models [6, 7, 8].

In computational biology, LLMs have learned to extract meaningful features from raw, unlabeled sequence data. The values of the last hidden layers from these pre-trained protein language models (pLMs) [9, 10, 11, 12], dubbed the embeddings, are used as exclusive input to subsequent state-of-the-art (SOTA) prediction methods which tend to reach or outperform methods using evolutionary information, or hand-crafted features, (prediction of secondary structure [9], disorder [13], stability [14, 15], solubility [16], subcellular location [17], the identification of para- [18], epitopes [19] and signal peptides [20], and many other tasks [21][22, 23, 24]). Embedding-based predictions seem particularly successful when experimental data are very limited [25].

These methods leverage the rich internal sequence representation of pLMs. The input is often so informative that small prediction models (thousands rather than millions of free parameters) suffice, although more complex architectures have been explored [17]. Some work further specialized general-purpose pLMs, e.g., to optimize a pLM to specific protein families [26] or to enrich the embeddings with structural information [27]. Specialist models have also been trained for specific proteins such as antibodies [18, 28].

[*]Correspondence to: Robert Schmirler <robert.schmirler@abbvie.com> [1] Innovation Center, BTS IR LU, AbbVie Deutschland GmbH & Co. KG, Ludwigshafen, Germany [2] Faculty of Informatics, TUM (Technical University of Munich), Munich, Germany [3] TUM School of Computation, Information and Technology (CIT), TUM Graduate School, Garching, Germany [4] Institute for Advanced Study (TUM-IAS), TUM, Garching, Germany [5] TUM School of Life Sciences Weihenstephan (WZW), TUM, Freising, Germany

3rd NeurIPS Workshop on Efficient Natural Language and Speech (ENLSP 2023).

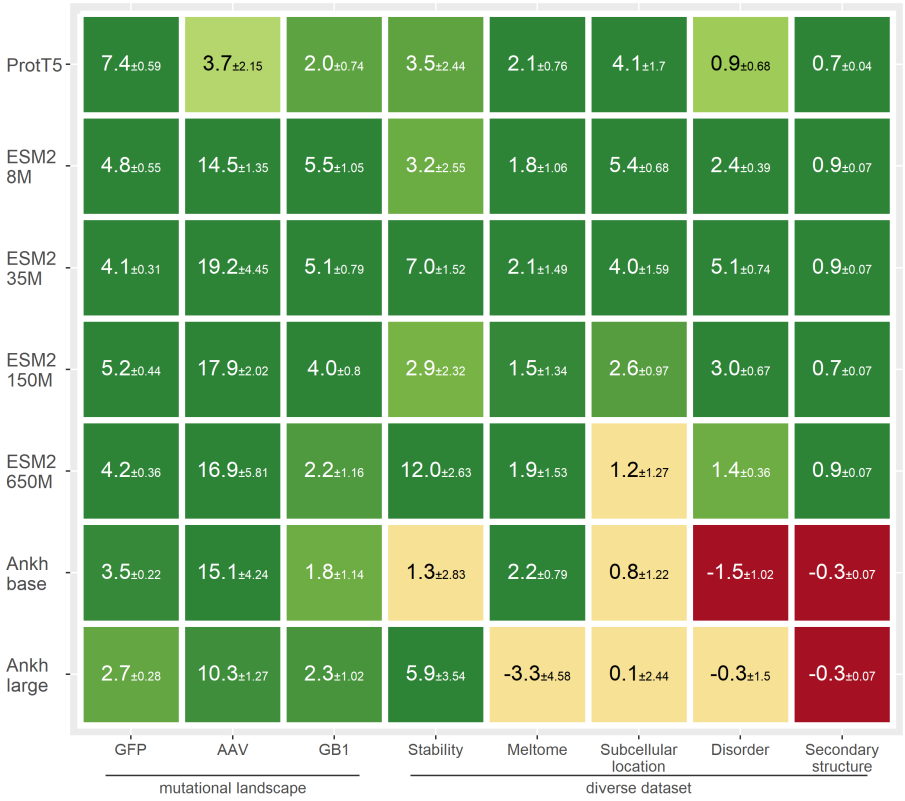

Figure 1: **Comparison of models and tasks.** Values are percentage differences between the fine-tuned and pre-trained models (Eqn.1) for each of the measures employed (Spearman ranking correlation for GFP, AAV, GB1, stability, meltome and disorder; accuracy for 10-class, per-protein sub-cellular location and 3-class per-residue secondary structure). Each tile compares the fine-tuning of one specific model against the pre-trained embedding-based solution for one specific prediction task. Green tiles mark statistically significant (exceeding 1.96 standard errors) increases in performance (finetuning over pretrained embeddings), yellow tiles mark statistically insignificant changes (0 lies within the error margins) and for red tiles supervised fine-tuning significantly decreased performance. Error estimates (±percentage values) represent the 95% confidence intervals (CI) from multiple runs of the same experiment (same model, identical data split) using different random seeds.

In contrast, here we evaluated the impact of task specific supervised fine-tuning on downstream prediction performance. We added a simple network as *prediction head* to the pLM encoder and trained the prediction model (pLM encoder and prediction head) in supervised fashion. We compared to predictions using pre-trained embeddings (training the prediction head using pre-computed embeddings). For the larger T5 based models we applied Low Rank Adaptation (LoRA) [29], a form of Parameter Efficient Fine-Tuning (PEFT) [30]. Freezing most of the model and updating only a small fraction of weights during training, leads to much lower hardware requirements and also prevents catastrophic forgetting [31, 32]. This will become especially relevant as pLMs [12] grow larger.

For pLMs, the advantages of fine-tuning remain less explored compared to NLP [33, 34, 30], although for some prediction tasks supervised, task-specific fine-tuning led to improvements [35, 36, 37]. Here, we assessed diverse prediction tasks from eight previously established benchmark datasets and three model architectures (pLMs) over a wide range of model sizes, aiming to draw a more general conclusion.

2 Results and Discussion

Top-level comparison of models and tasks. We trained 455 individual prediction methods (175 for fine-tuning, 280 using frozen embeddings from pre-trained pLMs) comprising seven models

(Supporting Online Material, SOM, Table S2), each trained on eight different datasets (SOM 1.1, SOM Table S1). Each model-task combination was repeated several times with different seeds for random initialization. We measured model performance on the test set due to some issues with the validation splits (details SOM 3). For each prediction task PT, we compared the performance between fine-tuning and pre-training as follows:

$$\Delta(PT) = performance(PT)_{finetuned} - performance(PT)_{pretrained} \quad (1)$$

For the ProtT5 (ProtT5-XL-U50 [9]) and all four tested ESM2 models [11], supervised fine-tuning significantly increased performance for all combinations of pLMs and tasks (Figure 1, SOM Table S5-S10), with one exception (ESM2-650M / sub-cellular location). Both Ankh [10] models deviated from this result, gaining significantly by fine-tuning only for the mutational landscape data sets (GFP, AAV and GB1: green in Figure 1).

Fine-tuning boosts per-residue disorder prediction. SETH [13] reached the level of MSA-based SOTA methods, such as ODiNPred [38] in the prediction of per-residue protein disorder as described by CheZOD scores [38]. SETH uses a 2-layer CNN trained on top of static embeddings extracted from a frozen pLM. Keeping those hyper-parameters and adding LoRA fine-tuning (dubbed SETH LoRA), improved performance by 2.2 percentage points (from Spearman 0.72 to 0.736, SOM Figure S1). The fine-tuned much smaller 150M parameter ESM2 model bested (average Spearman 0.742) all solutions compared (SOM Figure S1), including its larger counterparts, i.e., ESM2 with 650M parameters (SOM Table S8).

Compared to SETH LoRA, where only 2.5 million out of the 1.2 billion ProtT5 parameters are unfrozen during finetuning, the ESM2 150M updates a lot more parameters. However both finetuing approaches reach the same performance, i.e., show no significant difference (SOM Figure S1). While our results (Figure 1) show improvement from finetuning for most models and all tasks, selection of the absolute best model for a specific task remains less clear. For now we recommend testing different pLMs.

Insignificant gain for secondary structure prediction. For per-residue, three-class secondary structure prediction (helix, strand, other), fine-tuning improved only slightly (SOM Figure S2; up to 1.2 percentage points for both datasets, CASP12[39] and NEW364[9]). We confirmed this for both the original general purpose ProtT5 [9], and for the bilingual, structure-tuned ProstT5 [27] model, whose embeddings were enriched using Foldseek [40] 3Di structure information. We conclude that secondary structure gets already well captured in the model embeddings during unsupervised pretraining.

Sub-cellular location from pooled pLM embedding. A Light Attention (LA) based prediction method using per-residue pLM embeddings can surpass networks that use only average-pooled embeddings, i.e., embeddings derived from averaging over all residue-level embeddings for sub-cellular location prediction [17]. In brief, LA learned weighting for averaging over individual per-residue embeddings for an entire protein, while average pooling weights each residue equally. This might be particularly beneficial for predicting location where the dominant signal originates from the first N residues (with N typically around 50, e.g., for signal peptides) [41, 20]. Toward this end, LoRA fine-tuning combined the best of both worlds: the simplicity of having only a simple one-layer feed-forward neural network put on top of average-pooled last hidden states, paired with the learned, weighted averaging of LA. This allowed LoRA fine-tuning to reach or even surpass the performance of LA (Table 1).

Table 1: Q10 sub-cellular localization prediction * Values are taken from previous work [17]. For ProtT5 FFN LoRA five models were trained, initialized with different random states, and the mean value and standard deviation from these training runs are shown here. The performance metric is 10 class accuracy on the "set_HARD" test dataset [17].

Model	set_HARD
ProtT5 FFN*	61.3 ± 1.0
ProtT5 LA*	65.2 ± 0.6
ProtT5 FFN LoRA	$\mathbf{66.2 \pm 0.6}$

Finetuning improves models in various ways. We showed that finetuning can improve performance of pLM based predictors, although the extent of this improvement varies by task. We assume that finetuning unlocks additional degrees of freedom within the model. It indirectly allows a learned weighted pooling of the last hidden layer which has previously shown to be advantageous for some tasks [17]. Also, the last hidden layer is specialized on the unsupervised pretraining objective, which might not be optimal for all downstream tasks [42]. Finetuning allows the model to adapt accordingly. Lastly, the models might simply extract additional information directly from the task specific datasets. Further research is needed to determine which effects are prevalent.

3 Methods

Data sets. We subdivided the data sets according to two aspects, prediction task level and sequence diversity (SOM Table S1). The prediction task level indicates whether a prediction is either done for each residue in a protein individually (e.g., secondary structure and disorder), or for each protein (e.g., sub-cellular location). Sequence diversity distinguishes between data sets containing a diversity of proteins from those with variations of one (or several) proteins (e.g., mutational landscapes, GFP, AAV and GB1 all generated in deep mutational scanning experiments [43]). All data sets are described in detail in SOM 1.1.

Performance measurements. For simplicity, we confined results to using the measures for performance introduced by others who had introduced the data sets that we used (SOM 1.1). All regression tasks were evaluated using Spearman rank correlations, i.e., all three mutational landscapes (GFP, AAV, GB1), both stability related datasets (Stability and Meltome), as well as, the per-residue regression of Disorder. For the classification tasks, accuracy was defined as a 10-class accuracy for sub-cellular location, and a 3-class accuracy for secondary structure.

Model training. All pre-trained models used in this work are shown elsewhere (SOM Table S2). To initialize these models, we used the checkpoints available on Huggingface. For embedding-based predictions, we first took the pre-trained models and generated embeddings for each data set. Afterwards we trained a single layer neural network as model to predict each task inputting only the fixed embeddings from ESM, ProtT5, and Ankh [9, 10, 11]. For fine-tuning, we took the models and put the same simple prediction network on top of the last hidden layer. We then trained the encoder weights of the pLMs and the prediction head simultaneously (details SOM 1).

Acknowledgments and Disclosure of Funding

RS is an employee of AbbVie. The design, study conduct, and financial support for this research were provided by AbbVie. AbbVie participated in the interpretation of data, review, and approval of the publication. All computational resources for this work were provided by AbbVie. MH and BR were supported by the Bavarian Ministry of Education through funding to the TUM, by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium für Bildung und Forschung), and by a grant from Deutsche Forschungsgemeinschaft (DFG-GZ: RO1320/4-1). No competing interests have to be disclosed.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, and Zhifeng Chen. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, page 10012–10022, 2021.

[5] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI conference on artificial intelligence*, 35:11106–11115, 2021.

[6] Meredith V Trotter, Cuong Q Nguyen, Stephen Young, Rob T Woodruff, and Kim M Branson. Epigenomic language models powered by cerebras. *arXiv preprint arXiv:2112.07571*, 2021.

[7] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, and Guillaume Richard. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679.

[8] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 10 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z.

[9] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv*, 1 2021. doi: 10.1101/2020.07.12.199554.

[10] 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. *arXiv preprint arXiv:2301.06568*, 2023.

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

[12] Bo Chen, Xingyi Cheng, Yangli-ao Geng, Shen Li, Xin Zeng, Boyan Wang, Jing Gong, Chiming Liu, Aohan Zeng, and Yuxiao Dong. xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein. *bioRxiv*, 2023. doi: 10.1101/2023.07.05.547496.

[13] Dagmar Ilzhöfer, Michael Heinzinger, and Burkhard Rost. Seth predicts nuances of residue disorder from protein embeddings. *Frontiers in Bioinformatics*, 2, 2022. ISSN 2673-7647.

[14] 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*, 32:9689–9701, 2019.

[15] Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, 2021. doi: 10.1101/2021.11.09.467890.

[16] Jiangyan Feng, Min Jiang, James Shih, and Qing Chai. Antibody apparent solubility prediction from sequence by transfer learning. *Iscience*, 25(10), 2022.

[17] Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021. doi: 10.1093/bioadv/vbab035.

[18] Jinwoo Leem, Laura S. Mitchell, James HR Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7), 2022.

[19] Joakim Nøddeskov Clifford, Magnus Haraldson Høie, Sebastian Deleuran, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. Bepipred-3.0: Improved b-cell epitope prediction using protein language models. *Protein Science*, 31(12):e4497, 2022.

[20] Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology*, 40(7):1023–1025, 2022.

[21] Peter Mørch Groth, Richard Michael, Jesper Salomon, Pengfei Tian, and Wouter Boomsma. Flop: Tasks for fitness landscapes of protein wildtypes. *bioRxiv*, 2023. doi: 10.1101/2023.06.21.545880.

[22] Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647, 10 2022. ISSN 1432-1203. doi: 10.1007/s00439-021-02411-y.

[23] Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022.

[24] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N. Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *International Conference on Machine Learning*, page 16990–17017, 2022.

[25] Maria Littmann, Michael Heinzinger, Christian Dallago, Konstantin Weissenow, and Burkhard Rost. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports*, 11(1):23916, 2021.

[26] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z. Sun, and Richard Socher. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, page 1–8, 2023.

[27] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Martin Steinegger, and Burkhard Rost. Prostt5: Bilingual language model for protein sequence and structure. *bioRxiv*, 2023. doi: 10.1101/2023.07.23.550085.

[28] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[29] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[30] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, and Weize Chen. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[31] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135, 1999.

[32] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[33] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339, 2018. doi: 10.18653/v1/P18-1031.

[34] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, and Hongseok Namkoong. Robust fine-tuning of zero-shot models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 7959–7971, 2022.

[35] Vineet Thumuluri, Hannah-Marie Martiny, Jose J. Almagro Armenteros, Jesper Salomon, Henrik Nielsen, and Alexander Rosenberg Johansen. Netsolp: predicting protein solubility in escherichia coli using language models. *Bioinformatics*, 38(4):941–946, 2022.

[36] Danqing Wang, Y. E. Fei, and Hao Zhou. On pre-training language model for antibody. *The Eleventh International Conference on Learning Representations*, 2022.

[37] Alexandru Dumitrescu, Emmi Jokinen, Anja Paatero, Juho Kellosalo, Ville O. Paavilainen, and Harri Lähdesmäki. Tsignal: a transformer model for signal peptide prediction. *Bioinformatics*, 39(Supplement_1): i347–i356, 2023.

[38] Rupashree Dass, Frans AA Mulder, and Jakob Toudahl Nielsen. Odinpred: comprehensive prediction of protein order and disorder. *Scientific Reports*, 10(1):14780, 2020.

[39] Luciano A. Abriata, Giorgio E. Tamò, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Matteo Dal Peraro. Assessment of hard target modeling in casp12 reveals an emerging role of alignment-based contact prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 86:97–112, 2018.

[40] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, page 1–4, 2023.

[41] 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. doi: 10.1093/bioinformatics/btx548.

[42] Francesca-Zhoufan Li, Ava P Amini, Kevin K Yang, and Alex X Lu. Pretrained protein language model transfer learning: is the final layer representation what we want. *Proceedings of the Machine Learning for Structural Biology Workshop, NeurIPS 2022*, 2022.

[43] Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.