Light Attention Predicts Protein Location from the Language of Life

Hannes Stärk [*1] **Christian Dallago** [*1 2] **Michael Heinzinger** [1 2] **Burkhard Rost** [1 3 4]

Abstract

Although knowing where a protein functions in a cell is important to characterize biological processes, this information remains unavailable for most known proteins. Machine learning narrows the gap through predictions from expertly chosen input features leveraging evolutionary information that is resource expensive to generate. We showcase using embeddings from protein language models for competitive localization predictions not relying on evolutionary information. Our lightweight deep neural network architecture uses a softmax weighted aggregation mechanism with linear complexity in sequence length referred to as light attention (LA). The method significantly outperformed the state-of-the-art for ten localization classes by about eight percentage points (Q10). The novel models are available as a web-service and as a stand-alone application at `http://embed.protein.properties`.

Introduction. Proteins are the machinery of life involved in all essential biological processes. Knowing where in the cell a protein functions, referred to as its *subcellular localization*, is important for unraveling biological function. The standard tool in molecular biology for inferring localization, namely homology-based inference (HBI), accurately transfers annotations from experimentally annotated to sequence-similar un-annotated proteins. However, HBI is not available or unreliable for most proteins (Goldberg et al., 2014; Mahlich et al., 2018). Machine learning methods perform less well (lower precision) but are available for all proteins (high recall). The best methods use evolutionary information from families of related proteins as input (Goldberg et al., 2012; Almagro Armenteros et al., 2017).

Recently, protein sequence representations (embeddings) have been learned from databases using language models (LMs) (Rives et al., 2019; Alley et al., 2019; Elnaggar et al., 2020) initially used in natural language processing (NLP). Models trained on protein embeddings via transfer learning tend to be outperformed by approaches using evolutionary information (Rao et al., 2019; Heinzinger et al., 2019). For location prediction, embedding-based models (Heinzinger et al., 2019; Elnaggar et al., 2020) remained inferior to the state-of-the-art using evolutionary information, e.g., represented by DeepLoc (Almagro Armenteros et al., 2017). In this work, we leverage protein embeddings to predict cellular location without evolutionary information. We propose a deep neural network architecture using light attention (LA) inspired by previous attention mechanisms.

1. Methods

Data. We use a data set introduced by *DeepLoc* (Almagro Armenteros et al., 2017) for training and testing. The dataset contains 13 858 proteins annotated with experimental evidence for one of ten location classes. 2 768 proteins made up the test set (henceforth called *setDeepLoc*). To rule out that methods had been optimized for the static standard test set (*setDeepLoc*), we create a new independent test set *setHARD*. It contains 490 samples that are more difficult to predict as more stringent redundancy reduction was applied. They also follow a different class distribution than *setDeepLoc*.

Language model protein embeddings. As input to the LA architectures, we extract embeddings from three protein language models: the bidirectional LSTM *SeqVec* (Heinzinger et al., 2019) trained on UniRef50, the encoder-only model *ProtBert* (Elnaggar et al., 2020) trained on BFD (Steinegger & Söding, 2018), and the encoder-only model *ProtT5* (Elnaggar et al., 2020) trained on BFD and fine-tuned on Uniref50. For *SeqVec*, the per-residue embeddings are generated by summing the representations of each layer. For *ProtBert* and *ProtT5*, the per-residue embeddings are extracted from the last hidden layer of the models. Thus we obtain protein embeddings of size $1024 \times L$, where L is the

[*]Equal contribution [1]TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany [2]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Boltzmannstr. 11, 85748 Garching/Munich, Germany [3]Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany [4]TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany. Correspondence to: Christian Dallago <christian.dallago@tum.de>.

The 2021 ICML Workshop on Computational Biology. Copyright 2021 by the author(s).

length of the protein sequence.

Light Attention (LA) architecture. The input to light attention (LA) classifiers are protein embeddings $X \in \mathbb{R}^{1024\times L}$. In the architecture, the input is first transformed by two separate 1D convolutions with filter sizes s parameterized by learned weights $W^{(e)}, W^{(v)} \in \mathbb{R}^{s\times 1024\times d_{out}}$. The convolutions are applied over the length dimension to produce attention coefficients and value features $e, v \in \mathbb{R}^{d_{out}\times L}$. To use the coefficients as attention distribution over all j, we softmax-normalize over protein length. The attention weight $\alpha_{i,j} \in \mathbb{R}$ for the j-th residue and the i-th feature dimension is calculated as:

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum_{l=1}^{L} exp(e_{i,l})} \tag{1}$$

Note that the weight distributions for each feature dimension i are independent, and they can generate different attention patterns. The attention distributions are used to compute weighted sums of the transformed residue embeddings $v_{i,j}$. Thus, we obtain a fixed-size representation $x' \in \mathbb{R}^{d_{out}}$ for the whole protein, independent of its length.

$$x'_i = \sum_{j=1}^{L} \alpha_{i,j} v_{i,j} \tag{2}$$

We concatenate x'_i with the maximum of the values over the length dimension $v^{max} \in \mathbb{R}^{d_{out}}$, meaning $v_i^{max} = \max_{1\leq j\leq L}(v_{i,j})$. This concatenated vector is the input for a two-layer multi-layer perceptron (MLP) $f : \mathbb{R}^{2d_{out}} \mapsto \mathbb{R}^{d_{class}}$ with d_{class} as the number of classes. The softmax over the MLP output represents the class probabilities indexed by c ($\oplus$ denotes concatenation):

$$p(c|x) = softmax_c(f(x' \oplus m)) \tag{3}$$

Baselines. We compare against the SOTA *DeepLoc* and all the methods they used as Baselines. Additionally, we train a two-layer MLP. Instead of per-residue embeddings in $\mathbb{R}^{1024\times L}$, the MLPs use sequence-embeddings in $\mathbb{R}^{1024}$, obtained from averaging over the length dimension. Furthermore, for these representations, we perform annotation transfer (dubbed AT) based on embedding space similarity. Following this approach, proteins in *setDeepLoc* and *setHARD* are annotated by transferring the class of the nearest neighbor in the DeepLoc training set (given by L1 distance). We use the majority classifier as a naive baseline. All evaluation scripts to reproduce results are available at https://github.com/HannesStark/protein-localization.

2. Results and Discussion

Embeddings outperformed evolutionary information. The simple AT approach already outperforms some methods using evolutionary information (Figure 2: *AT**) and the simple MLP trained on *ProtT5* embeddings outperforms SOTA. Methods based on *ProtT5* embeddings consistently yield better results than *ProtBert* and *SeqVec* (**ProtT5* vs **ProtBert*/**SeqVec* in Figure 2). The light attention (LA) architecture consistently achieves better results than other embedding-based approaches, irrespective of the protein LM. Using *ProtBert* embeddings, LA outperforms the SOTA (Almagro Armenteros et al., 2017) by 1 and 2 percentage points on *setHARD* and *setDeepLoc*. More importantly, for both test sets, LA raises the bar for either set by 8 percentage points when using *ProtT5* embeddings.

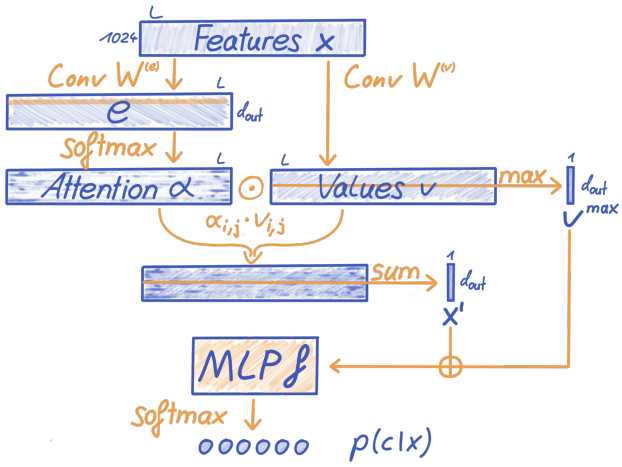

Figure 1. **Sketch of LA solution.** The LA architecture is parameterized by two weight matrices $W^{(e)}, W^{(v)} \in \mathbb{R}^{s\times 1024\times d_{out}}$ and the weights of an MLP $f : \mathbb{R}^{2d_{out}} \mapsto \mathbb{R}^{d_{class}}$.

Light aggregation (LA) mechanism crucial. To probe the effectiveness of LA's aggregation mechanism on *ProtT5* embeddings we run additional tests with obvious baselines and two ablations as detailed in Table 1. Performance deterioration by dropping the softmax or max-pooling aggregation confirms that both aspects are crucial and lead to better performance (the same did not hold for additional mean-, sum-, or min-aggregation). Furthermore, LA is especially apt at extracting information from LM embeddings, while it performs poorly on other protein representations, e.g., one-hot encodings.

Why light attention succeeds. The central challenge for the improvement introduced here is to convert the residue-embeddings (NLP equivalent: word embeddings) from protein language models to meaningful per sequence-embeddings (NLP equivalent: document). A qualitative evaluation of the influence of the attention mechanism (Figure 3) highlights its ability to efficiently aggregate information. Although averaging surpasses evolutionary-information-based methods using simple similarity-based annotation transfer (Figure 2: AT*) and in one instance

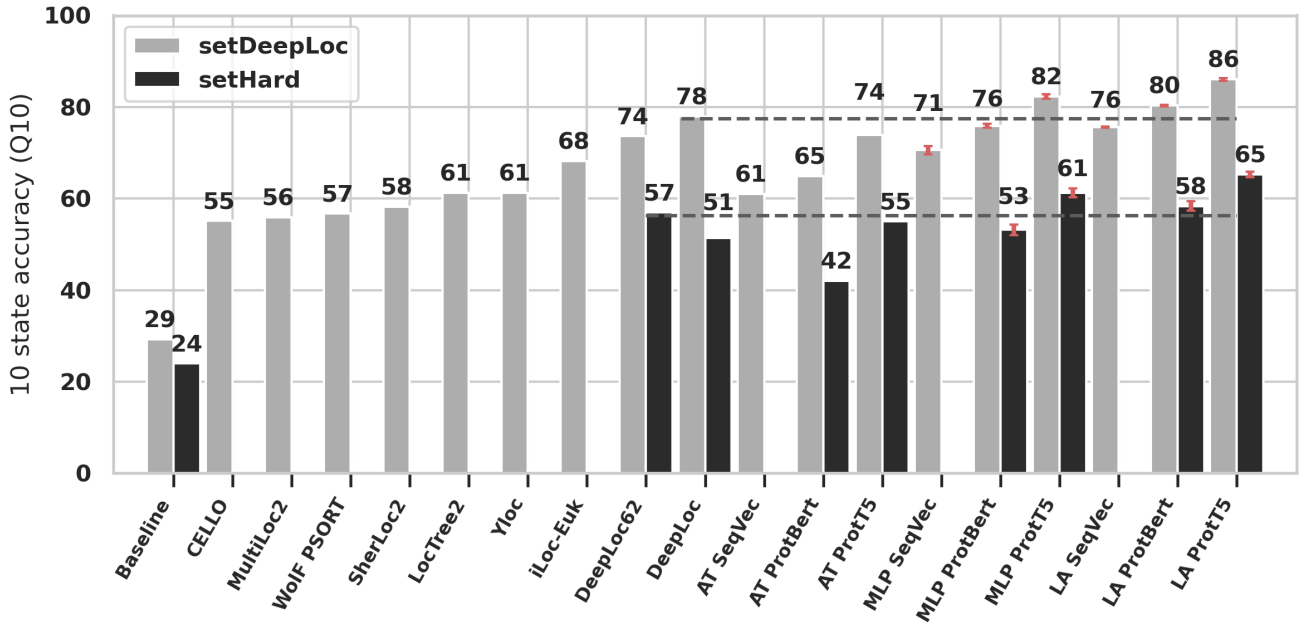

Figure 2. **LA architectures perform best.** Bars give the ten-class accuracy (Q10) for popular location prediction methods on *setDeepLoc* (light-gray bars) and *setHARD* (dark-gray bars). Baseline is the most common class in each set. Horizontal gray dashed lines mark the previous SOTA on either set. Estimates for standard errors are marked in orange for the methods introduced here. *setHARD* results are provided for a subset of methods that yielded the best results on *setDeepLoc* (see *Methods* for detail on the external methods used; tabular data in *Appendix: Additional Results*). Two results stood out: (i) the LA approaches introduced here outperformed the top methods although not using evolutionary information (highest bars), and (ii) the performance estimates differed completely between the two data sets (difference light/dark gray).

even SOTA using a simple feed-forward network (Figure 2: *DeepLoc* vs. *MLP ProtT5*), LA is able to consistently distill more information from embeddings. Most likely, the improvement can be attributed to LA's ability to regulate the immense difference in lengths of proteins (varying from 30 to 30 000 residues) by learning attention distributions over the sequence positions. LA models appear to capture relevant long-range dependencies while retaining the ability to focus on specific sequence regions such as beginning and end, which play a particularly important role in determining protein location for some proteins (Lange et al., 2007).

First win over evolutionary information. Effectively, LA trained on protein LM embeddings from *ProtT5* is at the heart of the first method that clearly appears to outperform the best existing method in a statistically significant manner on two test sets (Figure 2). To the best of our knowledge, this improvement is the first instance ever that embedding-based transfer learning substantially outperforms AI/ML methods using evolutionary information for function prediction. Even if embeddings are extracted from LMs trained on large sequence data originating from evolution, the vast majority of data learned originates from much more generic constraints informative of protein structure and function.

Overfitting through standard data set? For protein subcellular location prediction, the data set of *DeepLoc* has become a standard in the field. Such static standards facilitate method comparisons but can lead to overfitting on that data set. To further probe results, we create a new test set (*setHARD*), which is redundancy-reduced both with respect to itself and all proteins in the *DeepLoc* set. For this set, the 10-state accuracy (Q10) dropped, on average, 22 percentage points with respect to the static standard, but LA remains best by 8 percentage points (Figure 2).

Better and faster than profiles. At inference, the embeddings needed as input for the LA models come with three advantages over the historically most informative evolutionary information, i.e., protein profiles, which were essential for methods such as *DeepLoc* (Almagro Armenteros et al., 2017) to achieve SOTA. Chiefly, embeddings can be obtained in far less time than is needed to generate profiles and require fewer compute resources. Even the lightning-fast MMseqs2 (Steinegger & Söding, 2017), which is not the standard in bioinformatics (other methods 10-100x slower), is 3x slower than *ProtT5*, and 5x slower than *ProtBert*. Moreover, these MMseqs2 stats derive from runs on a machine with $>$ 300GB of RAM and 2x40cores/80threads CPUs, while generating LM embeddings required only a moderate

Table 1. **LA + ProtT5 the winning combination.** Accuracy of baselines and ablations using *ProtT5* embeddings (above the line), one-hot residue encodings or profiles for *setDeepLoc* and *setHARD* for various architectures. LA ProtT5: The proposed light attention architecture. LA - Softmax: replaced softmax aggregation that previously produced x' with averaging of the coefficients e over the length dimension. LA - MaxPool: discarded max-pooled values v^{max} as input to the MLP, aka. only the softmax aggregated features x' were used. Attention from v: attention coefficients e were obtained via a convolution over the values v instead of over the inputs x. DeepLoc LSTM: the architecture of DeepLoc (Almagro Armenteros et al., 2017) was used instead of LA. Conv + AdaPool: a stack of convolutions (kernel-size 3, 9, and 15) followed by adaptive pooling to a length of 5 and an MLP was used instead of LA. LA on OneHot: LA using one-hot encodings of residues in a protein sequence as input. LA on Profiles: LA using evolutionary information in the form of protein profiles (Gribskov et al., 1987) as input.

METHOD	SETDEEPLOC	SETHARD
LA PROTT5	**86.01**± 0.34	**65.21**± 0.61
LA - SOFTMAX	85.30± 0.32	64.72± 0.70
LA - MAXPOOL	84.79± 0.19	63.84± 0.67
ATTENTION FROM v	85.41± 0.27	64.77± 0.93
DEEPLOC LSTM	79.40± 0.88	59.36± 0.84
CONV + ADAPOOL	82.09± 0.92	60.79± 2.01
LA ON ONEHOT	43.53± 1.48	32.57± 2.38
LA ON PROFILES	43.78± 1.25	33.35± 1.82

machine (8 cores, 16GB RAM) equipped with a modern GPU with >7GB of vRAM. Additionally, extracting profiles relies on the use of tools (e.g., MMseqs2) that are sensitive to parameter changes, ultimately an extra complication for users.

Model trainable on consumer hardware. The final LA architecture, made of 18 940 224 parameters, can be trained on an Nvidia GeForce GTX 1060 with 6GB vRAM in 18 hours or on a Quadro RTX 8000 with 48GB vRAM in 2.5 hours.

What can users expect from subcellular location predictions? If the top accuracy for one data set was Q10 $\sim 60\%$ and Q10 $\sim 80\%$ for the other, what can users expect for their next ten queries: six correct or eight, or 6-8? The answer depends on the query: if those proteins are sequence similar to proteins with known location (case: redundant): the answer is eight. Conversely, for new proteins (without homologs of known location), six in ten will be correctly predicted, on average. In turn, this implies that for novel proteins, there seems to be significant room for pushing performance to further heights, possibly by combining *LA ProtBert*/*LA ProtT5* with evolutionary information.

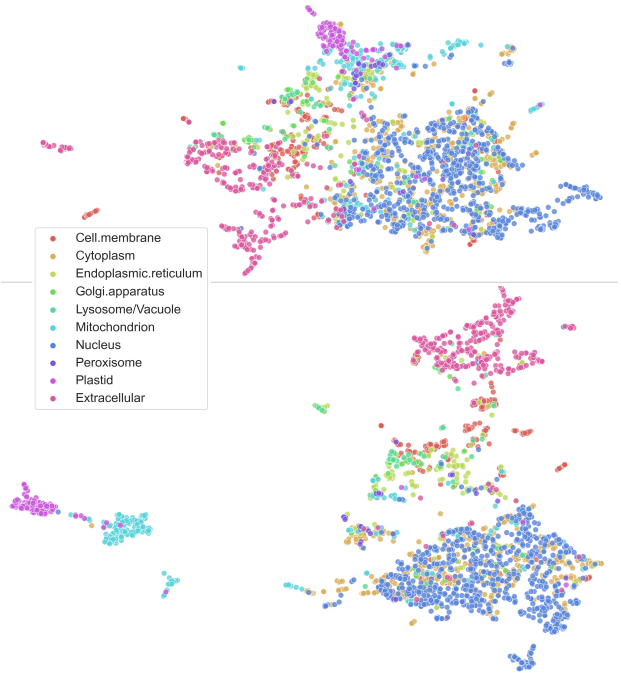

Figure 3. **Qualitative analysis confirms: attention effective.** UMAP (McInnes et al., 2018) projections of per-protein embeddings colored according to subcellular location (*setDeepLoc*). Both plots were created with the same default values of the python *umap-learn* library. Top: ProtT5 embeddings (LA input; x) mean-pooled over protein length (as for MLP/AT input). Bottom: ProtT5 embeddings (LA input; x) weighted according to the attention distribution produced by LA (this is not x' as we sum the input features x and not the values v after the convolution).

3. Conclusion

Conclusion. We presented light attention (LA) operating on language model embeddings of protein sequences. By implicitly assigning a different importance score for each sequence position, the method succeeds in predicting protein subcellular location 8 percentage points more accurately than previous methods. Thus, LA manages to outperform the SOTA without using evolutionary-based inputs, i.e., the single most important input feature for previous methods. This constitutes an important breakthrough: although many methods had come close to the SOTA using embeddings instead of evolutionary information, none had ever overtaken as the methods presented here.

References

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, De-

cember 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0598-1. URL https://www.nature.com/articles/s41592-019-0598-1. Number: 12 Publisher: Nature Publishing Group.

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, November 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx431. URL https://academic.oup.com/bioinformatics/article/33/21/3387/3931857. tex.ids: almagroarmenterosDeepLocPredictionProtein2017a publisher: Oxford Academic.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv*, pp. 2020.07.12.199554, July 2020. doi: 10.1101/2020.07.12.199554. URL https://www.biorxiv.org/content/10.1101/2020.07.12.199554v2. tex.ids: elnaggarProtTransCrackingLanguage2020a publisher: Cold Spring Harbor Laboratory section: New Results.

Goldberg, T., Hamp, T., and Rost, B. LocTree2 predicts localization for all domains of life. *Bioinformatics*, 28 (18):i458–i465, September 2012.

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., Bernhofer, M., Betz, A., Cizmadija, L., Do, K. T., Gerke, J., Greil, R., Joerdens, V., Hastreiter, M., Hembach, K., Herzog, M., Kalemanov, M., Kluge, M., Meier, A., Nasir, H., Neumaier, U., Prade, V., Reeb, J., Sorokoumov, A., Troshani, I., Vorberg, S., Waldraff, S., Zierer, J., Nielsen, H., and Rost, B. LocTree3 prediction of localization. *Nucleic Acids Research*, 42(W1): W350–W355, 2014. ISSN 0305-1048. doi: 10.1093/nar/gku396. URL https://doi.org/10.1093/nar/gku396. _eprint: https://academic.oup.com/nar/article-pdf/42/W1/W350/17423232/gku396.pdf.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, 84(13): 4355–4358, July 1987. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.84.13.4355. URL https://www.pnas.org/content/84/13/4355. tex.ids= gribskovProfileAnalysisDetection1987a publisher: National Academy of Sciences section: Research Article.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723, December 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3220-8. URL https://doi.org/10.1186/s12859-019-3220-8. tex.ids: heinzinger-ModelingAspectsLanguage2019a.

Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E., and Corbett, A. H. Classical Nuclear Localization Signals: Definition, Function, and Interaction with Importin alpha,. *Journal of Biological Chemistry*, 282(8):5101–5105, February 2007. ISSN 0021-9258. doi: 10.1074/jbc.R600026200. URL http://www.sciencedirect.com/science/article/pii/S0021925820688019.

Mahlich, Y., Steinegger, M., Rost, B., and Bromberg, Y. HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics*, 34(13):i304–i312, July 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty262. URL https://doi.org/10.1093/bioinformatics/bty262.

McInnes, L., Healy, J., Saul, N., and Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. Evaluating Protein Transfer Learning with TAPE. *Advances in neural information processing systems*, 32:9689–9701, December 2019. ISSN 1049-5258. URL https://pubmed.ncbi.nlm.nih.gov/33390682.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *bioRxiv*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/early/2019/04/29/622803.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9 (1):2542, June 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04964-5. URL https://doi.org/10.1038/s41467-018-04964-5.