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

Summary: 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 expert-designed input features leveraging information from multiple sequence alignments (MSAs) that is resource expensive to generate. Here, we showcased using embeddings from protein language models (pLMs) for competitive localization prediction without MSAs. Our lightweight deep neural network architecture used 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 (SOTA) for ten localization classes by about eight percentage points (Q10). So far, this might be the highest improvement of *just embeddings* over MSAs. Our new test set highlighted the limits of standard static data sets: while inviting new models, they might not suffice to claim improvements over the SOTA.
Availability: Online predictions are available at http://embed.protein.properties. Predictions for the human proteome are available at https://zenodo.org/record/5047020. Code is provided at https://github.com/HannesStark/protein-localization.

[*]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>.

1. Introduction

Prediction bridges gap between proteins with and without location annotations. Proteins are the machinery of life involved in all essential biological processes (*Appendix: Biological Background*). Knowing where in the cell a protein functions, *natively*, i.e. its *subcellular location* or *cellular compartment* (for brevity, abbreviated by *location*), is important to unravel biological function (Nair & Rost, 2005; Yu et al., 2006). Experimental determination of protein function is complex, costly, and selection biased Ching et al., 2018. In contrast, protein sequences continue to explode (Consortium, 2021). This increases the sequence-annotation gap between proteins for which only the sequence is known and those with experimental function annotations . Computational methods have been bridging this gap (Rost et al., 2003), e.g. by predicting protein location (Goldberg et al., 2012; 2014; Almagro Armenteros et al., 2017; Savojardo et al., 2018). The standard tool in molecular biology, namely homology-based inference (HBI), accurately transfers annotations from experimentally annotated to sequence-similar un-annotated proteins. However, HBI is either unavailable 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 as computed from families of related proteins identified in multiple sequence alignments (MSAs) as input (Nair & Rost, 2005; Goldberg et al., 2012; Almagro Armenteros et al., 2017). Although the marriage of evolutionary information and machine learning has influenced computational biology for decades (Rost & Sander, 1993), due to database growth, MSAs have become costly.

Protein Language Models (pLMs) better represent sequences. Recently, protein sequence representations (embeddings) have been learned from databases (Steinegger & Söding, 2018; Consortium, 2021) using language models (LMs) (Bepler & Berger, 2019; Alley et al., 2019; Heinzinger et al., 2019; Rives et al., 2021; Elnaggar et al., 2021) initially used in natural language processing (NLP) (Peters et al., 2018; Devlin et al., 2019; Raffel et al., 2020). Models trained on protein embeddings via transfer learning tend to be outperformed by approaches using MSAs (Rao et al., 2019; Heinzinger et al., 2019). However, embedding-based solutions can outshine HBI (Littmann et al., 2021)

and advanced protein structure prediction methods (Bhattacharya et al., 2020; Rao et al., 2020; Weißenow et al., 2021). Yet, for location prediction, embedding-based models (Heinzinger et al., 2019; Elnaggar et al., 2021; Littmann et al., 2021) remained inferior to the state-of-the-art (SOTA) using MSAs, such as DeepLoc (Almagro Armenteros et al., 2017).

In this work, we leveraged protein embeddings to predict cellular location without MSAs. We proposed a deep neural network architecture using light attention (LA) inspired by previous attention mechanisms (Bahdanau et al., 2015).

2. Related Work

The best previous predictions of location prediction combined HBI, MSAs, and machine learning, often building prior expert-knowledge into the models. For instance, *LocTree2* (Goldberg et al., 2012) implemented profile-kernel (Support Vector Machines (SVMs) (Cortes & Vapnik, 1995) which identified k-mers conserved in evolution and put them into a hierarchy of models inspired by cellular sorting pathways. *BUSCA* (Savojardo et al., 2018) combined three compartment-specific SVMs based on MSAs (Pierleoni et al., 2006; Savojardo et al., 2017). *DeepLoc* (Almagro Armenteros et al., 2017) used convolutions followed by a bidirectional long short-term memory (LSTM) module (Hochreiter & Schmidhuber, 1997) employing the Bahdanau-Attention (Bahdanau et al., 2015). Using the BLOSUM62 substitution metric (Henikoff & Henikoff, 1992) for fast and MSAs for slower, refined predictions, DeepLoc rose to become the SOTA. Embedding-based methods (Heinzinger et al., 2019) have not yet consistently outperformed this SOTA, although *ProtTrans* (Elnaggar et al., 2021), based on very large data sets, came close.

3. Methods

3.1. Data

Standard setDeepLoc. Following previous work (Heinzinger et al., 2019; Elnaggar et al., 2021), we began with a data set introduced by *DeepLoc* (Almagro Armenteros et al., 2017) for training (13 858 proteins) and testing (2 768 proteins). All proteins have experimental evidence for one of ten location classes (nucleus, cytoplasm, extracellular space, mitochondrion, cell membrane, Endoplasmatic Reticulum, plastid, Golgi apparatus, lysosome/vacuole, peroxisome). The 2 768 proteins making up the test set (dubbed *setDeepLoc*), had been redundancy reduced to the training set (but not to themselves), and thus share $\leq 30\%$ PIDE (pairwise sequence identity) and E-values $\leq 10^{-6}$ to any sequence in training. To avoid overestimations by tuning hyper-parameters, we split the DeepLoc training set into: training-only (9 503 proteins) and validation sets (1 158 proteins; $\leq 30\%$ PIDE; *Appendix: Datasets*).

Novel *setHARD*. To catch over-fitting on a static standard data set, we created a new independent test set from *SwissProt* (Consortium, 2021). Applying the same filters as *DeepLoc* (only eukaryotes; all proteins ≥ 40 residues; no fragments; only experimental annotations) gave 5 947 proteins. Using *MMseqs2* (Steinegger & Söding, 2017), we removed all proteins from the new set with $\geq 20\%$ PIDE to any protein in any other set. Next, we mapped location classes from DeepLoc to SwissProt, merged duplicates, and removed multi-localized proteins (protein X both in class Y and Z). Finally, we clustered at $\geq 20\%$ PIDE leaving only one representative of each cluster in the new, more challenging test set (dubbed *setHARD*; 490 proteins; *Appendix: Datasets*).

3.2. Models

Input embeddings. As input to the Light Attention (LA) architectures, we extracted *frozen* embeddings from pLMs, i.e. without fine-tuning for location prediction (details below). We compared embeddings from five main and a sixth additional pre-trained pLMs (Table 1): (1) *SeqVec* (Heinzinger et al., 2019) is a bidirectional LSTM based on on ELMo (Peters et al., 2018) that was trained on UniRef50 (Suzek et al., 2015). (2) *ProtBert* (Elnaggar et al., 2021) is an encoder-only model based on BERT (Devlin et al., 2019) that was trained on BFD (Steinegger & Söding, 2018). (3) ProtT5-XL-UniRef50 (Elnaggar et al., 2021) (for simplicity: *ProtT5*) is an encoder-only model based on T5 (Raffel et al., 2020) that was trained on BFD and fine-tuned on Uniref50. (4) *ESM-1b* (Rives et al., 2021) is a transformer model that was trained on UniRef50. (5) *UniRep* (Alley et al., 2019) is a multiplicative LSTM (mLSTM)-based model trained on UniRef50. (6) Bepler&Berger (dubbed *BB*) is a bidirectional LSTM by (Bepler & Berger, 2019), which fused modelling the protein language with learning information about protein structure into a single pLM. Due to different training objectives, this pLM was expected suboptimal for our task. As results confirmed this expectation, we confined these to *Appendix: Additional Results*.

Frozen embeddings were preferred over fine-tuned embeddings as the latter previously did not improve (Elnaggar et al., 2021) and consumed more resources/energy. *ProtT5* was instantiated at half-precision (float16 weights instead of float32) to ensure the encoder could fit on consumer GPUs with limited vRAM. Due to model limitations, for ESM-1b, only proteins with fewer than 1024 residues were used for training and evaluation (*Appendix: Datasets*).

Embeddings for each residue (NLP equivalent: word)

Table 1. Implementation details for SeqVec (Heinzinger et al., 2019), ProtBert (Elnaggar et al., 2021), ProtT5 (Elnaggar et al., 2021), ESM-1b (Rives et al., 2021), UniRep (Alley et al., 2019) and BB (Bepler & Berger, 2019). Estimates marked by *; differences in the number of proteins (*Sequences*) for the same set (*Dataset*) originated from versioning. The embedding time (in seconds) was averaged over 10 000 proteins taken from the PDB (Berman et al., 2000) using the embedding models taken from *bio-embeddings* (Dallago et al., 2021).

	SeqVec	ProtBert	ProtT5	ESM-1b	UniRep	BB
Parameters	93M	420M	3B	650M	18.2M	90M*
Dataset	UniRef50	BFD	BFD	UniRef50	UniRef50	Pfam
Sequences	33M	2.1B	2.1B	27M	27M	21M
Embed time (s)	0.03	0.06	0.1	0.09	2.1	0.1
Attention heads	0	16	32	20	0	0
Bits per Float	32	32	16	32	32	32
Size (GB)	0.35	1.6	3.6	7.3	0.06	0.12

in a protein sequence (NLP equivalent: document) were obtained using the bio-embeddings software (Dallago et al., 2021). For *SeqVec*, the per-residue embeddings were generated by summing the representations of each layer. For all other models, the per-residue embeddings were extracted from the last hidden layer. Finally, the inputs obtained from the pLMs were of size $d_{in} \times L$, where L is the length of the protein sequence, while d_{in} is the size of the embedding.

Implementation details. The LA models were trained using filter size $s = 9$, $d_{out} = 1024$, the Adam (Kingma & Ba, 2015) optimizer (learning rate 5×10^{-5}) with a batch size of 150, and early stopping after no improvement in validation loss for 80 epochs. We selected the hyperparameters via random search (*Appendix: Hyperparameters*). Models were trained either on an Nvidia Quadro RTX 8000 with 48GB vRAM or an Nvidia GeForce GTX 1060 with 6GB vRAM.

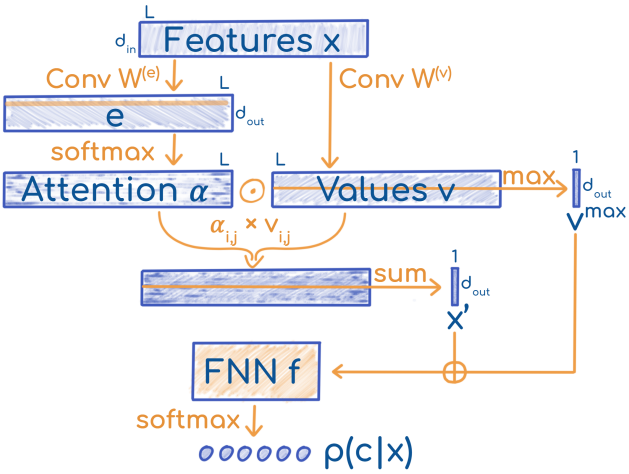

Figure 1. **Sketch of Light Attention (LA).** The LA architecture was parameterized by two weight matrices $W^{(e)}, W^{(v)} \in \mathbb{R}^{s \times d_{in} \times d_{out}}$ and the weights of an FNN $f : \mathbb{R}^{2d_{out}} \mapsto \mathbb{R}^{d_{class}}$.

Light Attention (LA) architecture. The input to the light attention (LA) classifier (Fig. 1) was a protein embedding $x \in \mathbb{R}^{d_{in} \times L}$ where L is the sequence length, while d_{in} is the size of the embedding (which depends on the model, Table 1). The input was transformed by two separate 1D convolutions with filter sizes s and learned weights $W^{(e)}, W^{(v)} \in \mathbb{R}^{s \times d_{in} \times d_{out}}$. The convolutions were applied over the length dimension to produce attention coefficients and values $e, v \in \mathbb{R}^{d_{out} \times L}$

$$e_{i,j} = b_i + \sum_{k=1}^{d_{in}} \sum_{l=-\lfloor \frac{s}{2} \rfloor}^{\lceil \frac{s}{2} \rceil} W^{(e)}_{l,k,i} x_{k,j+l} \tag{1}$$

where $b \in \mathbb{R}^{d_{out}}$ is a learned bias. For $j \notin [0, L)$, the $x_{:,j}$ were zero vectors. To use the coefficients as attention distributions over all j, we *softmax-normalized them over the length dimension*, i.e. the attention weight $\alpha_{i,j} \in \mathbb{R}$ for the j-th residue and the i-th feature dimension was calculated as:

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum_{l=1}^{L} exp(e_{i,l})} \tag{2}$$

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

$$x'_i = \sum_{j=1}^{L} \alpha_{i,j} v_{i,j} \tag{3}$$

$$p(c|x) = softmax_c(f(x' \oplus m)) \tag{4}$$

Methods used for comparison. For comparison, we trained a two-layer feed-forward neural network (FNN) proposed previously (Heinzinger et al., 2019). Instead of per-residue embeddings in $\mathbb{R}^{d_{in} \times L}$, the FNNs used

sequence-embeddings in $\mathbb{R}^{d_{in}}$, which derived from residue-embeddings averaged over the length dimension (i.e. mean pooling). Furthermore, for these representations, we performed embeddings distance-based annotation transfer (dubbed EAT) (Littmann et al., 2021). In this approach, proteins in *setDeepLoc* and *setHARD* were annotated by transferring the location from the nearest neighbor (L1 embedding distance) in the training set.

For ablations on the architecture, we tested LA without the softmax aggregation (LA w/o Softmax) that previously produced x', by replacing it with averaging of the coefficients e. Then, with LA w/o MaxPool, we discarded the max-pooled values v^{max} as input to the FNN instead of concatenating them with x'. With Attention from v, we computed the attention coefficients e via a convolution over the values v instead of over the inputs x. Additionally, we tested using a simple stack of convolutions (kernel-size 3, 9, and 15) followed by adaptive pooling to a length of 5 and an FNN instead of LA (Conv + AdaPool). Similarly, Query-Attention replaces the whole LA architecture with a transformer layer that used a single learned vector as query to summarize the whole sequence. As the last alternative operating on LM representations, we considered the DeepLoc LSTM (Almagro Armenteros et al., 2017) with *ProtT5* embeddings instead of MSAs.

To evaluate how traditional representations stack up against pLM embeddings, we evaluated MSAs (LA(MSA)) and one-hot encodings of amino acids (LA(OneHot)) as inputs to the LA model.

3.3. Evaluation.

Following previous work, we assessed performance through the mean ten-class accuracy (Q10), giving the percentage of correctly predicted proteins in one of ten location classes. As additional measures tested (i.e., F1 score and Matthew correlation coefficient (MCC) (Gorodkin, 2004)) did not provide any novel insights, these were confined to the *Appendix: Additional Results*. Error estimates were calculated over ten random seeds on both test sets. For previous methods (DeepLoc and DeepLoc62 (Almagro Armenteros et al., 2017), LocTree2 (Goldberg et al., 2012), MultiLoc2 (Blum et al., 2009), SherLoc2 (Briesemeister et al., 2009), CELLO (Yu et al., 2006), iLoc-Euk (Chou et al., 2011), YLoc (Briesemeister et al., 2010) and WoLF PSORT (Horton et al., 2007)) published performance values were used (Almagro Armenteros et al., 2017) for *setDeepLoc*. For *setHARD*, the webserver for DeepLoc[1] was used to generate predictions using either profile or BLOSUM inputs, whose results were later evaluated in Q10 and MCC. As a naive baseline, we implemented a method that predicted the same location class for all proteins, namely the one most often observed (in results: Majority). We provided code to reproduce all results[2].

4. Results

Embeddings outperformed MSAs. The simple EAT (embedding-based annotation transfer) already outperformed some advanced methods using MSAs (Fig. 2). The FNNs trained on *ProtT5* (Elnaggar et al., 2021) and ESM-1b (Rives et al., 2021) outperformed the SOTA *DeepLoc* (Almagro Armenteros et al., 2017) (Fig. 2). Methods based on *ProtT5* embeddings consistently reached higher performance values than other embedding-based methods (**ProtT5* vs. rest in Figure 2). Results on Q10 were consistent with those obtained for MCC (*Appendix: Additional Results*).

LA architecture best. The light attention (LA) architecture introduced here consistently outperformed other embedding-based approaches for all pLMs tested (LA* vs. EAT/FNN* in Fig. 2). Using *ProtBert* embeddings, LA outperformed the SOTA (Almagro Armenteros et al., 2017) by 1 and 2 percentage points on *setHARD* and *setDeepLoc* (*LA(ProtBert)* Fig. 2). For both test sets, LA improved the previous best on either set by around eight percentage points with *ProtT5* embeddings.

Standard data set over-estimated performance. The substantial drop in performance measures (by about 22 percentage points) between the standard *setDeepLoc* and the new challenging *setHARD* (Fig. 2: light-gray vs. dark-gray, respectively) suggested substantial over-fitting. Mimicking the class distribution from *setDeepLoc* by sampling with replacement from *setHARD* led to higher values (Q10: *DeepLoc62*=63%; *DeepLoc*=54%; *LA(ProtBert)*=62%; *LA(ProtT5)*=69%). *DeepLoc* performed worse on *setHARD* with than without MSAs (only BLOSUM; Fig. 2: *DeepLoc* vs. *DeepLoc62*). Otherwise, the relative ranking and difference of models largely remained consistent between the two data sets *setDeepLoc* and *setHARD*.

Low performance for minority classes. The confusion matrix of predictions for *setDeepLoc* using LA(*ProtT5*) highlighted how many proteins were incorrectly predicted to be in the second most prevalent class (*cytoplasm*), and that the confusion of the two most common classes mainly occurred between each other (Fig. 3: *nucleus* and *cytoplasm*). As for other methods, including the previous SOTA (Almagro Armenteros et al., 2017), performance was particularly low for the most under-represented three classes (*Golgi apparatus*, *lysosome/vacuole*, and *peroxisome*) that,

[1] http://www.cbs.dtu.dk/services/DeepLoc

[2] https://github.com/HannesStark/protein-localization

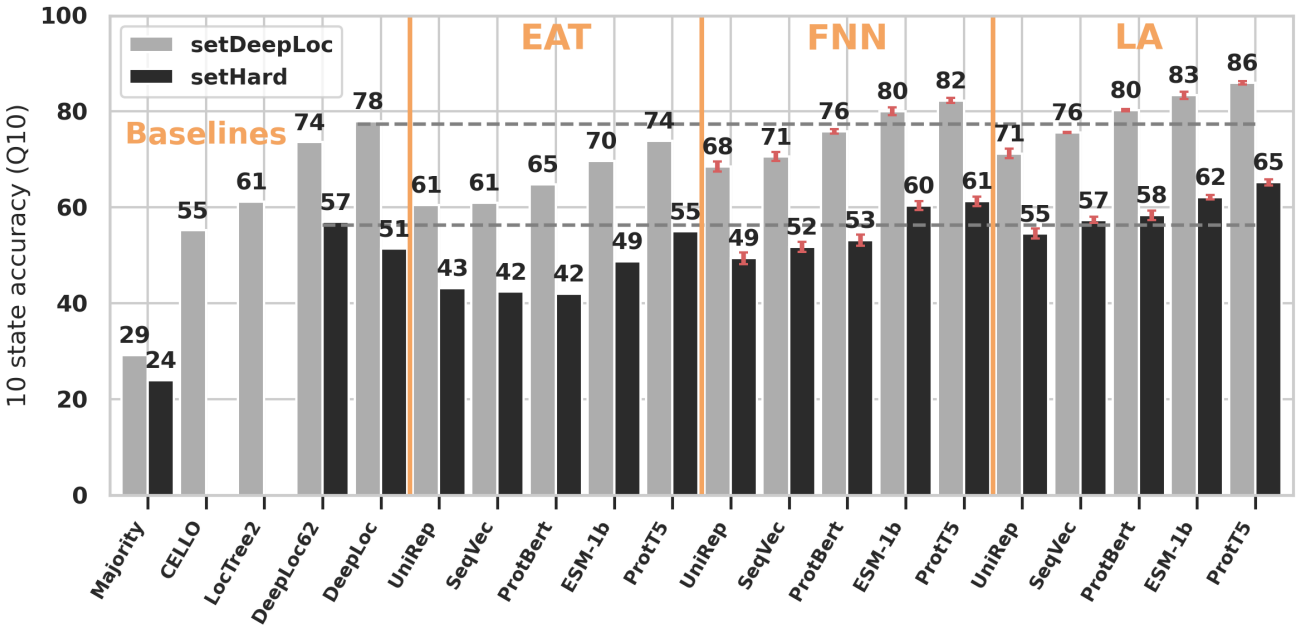

Figure 2. **LA architectures performed best.** Performance: Bars give the ten-class accuracy (Q10) assessed on *setDeepLoc* (light-gray bars) and *setHARD* (dark-gray bars). Methods: *Majority*, *CELLO**, *LocTree2**, *DeepLoc**, *DeepLoc62*; MSA-based methods marked by star. EAT used the mean-pooled pLM embeddings to transfer annotation via distance, while FNN(pLM) used the mean-pooled embeddings as input to a feed-forward neural network. LA(pLM) marked predictions using light attention on top of the pLMs from: UniRep (Alley et al., 2019), SeqVec (Heinzinger et al., 2019), ProtBert (Elnaggar et al., 2021), ESM-1b (Rives et al., 2021), ProtT5 (Elnaggar et al., 2021). Horizontal gray dashed lines mark the previous SOTA (*DeepLoc* and *DeepLoc62)* on either set. Estimates for standard deviations are marked in red for the new methods. Overall, LA significantly outperformed the SOTA without using MSAs, and values differed substantially between the two data sets (light vs. dark gray).

together accounted for 6% of the data. To attempt boosting performance for minority classes, we applied a *balanced loss*, assigning a higher weight to the contributions of underrepresented classes. This approach did not raise accuracy for the minority classes but lowered the overall accuracy, thus it was discarded.

Table 2. Comparison of LA(ProtT5) to different architectures and inputs. Methods described in Section 3.2. Standard deviations are estimated from 10 runs with different weight initializations.

METHOD	SETDEEPLOC	SETHARD
LA(PROTT5)	**86.0**± 0.3	**65.2**± 0.6
LA W/O SOFTMAX	85.3± 0.3	64.7± 0.7
LA W/O MAXPOOL	84.7± 0.2	63.8± 0.7
ATTENTION FROM v	85.4± 0.3	64.7± 0.9
DEEPLOC LSTM	79.4± 0.9	59.3± 0.8
CONV + ADAPOOL	82.0± 0.9	60.7± 2.0
QUERY-ATTENTION	75.3± 0.5	52.41± 0.4
LA(ONEHOT)	43.5± 1.5	32.5± 2.4
LA(MSA)	43.7± 1.3	33.3± 1.8

Light attention (LA) mechanism crucial. To probe the effectiveness of the LA aggregation mechanism on *ProtT5* we considered several alternatives for compiling the attention (*LA w/o Softmax* & *LA w/o MaxPool* & *Attention from v* & *DeepLoc LSTM* & *Conv + AdaPool*), and used the LA mechanism with non-embedding input (*LA(OneHot)* & *LA(MSA)*). Q10 dropped substantially without softmax- or max-aggregation. Furthermore, inputting *traditional* protein representations (*one-hot encoding*, i.e. representing the 20 amino acids by a 20-dimensional vector with 19 zeroes) or MSAs, the LA approach did not reach the heights of using pLM embeddings (Table 2: *LA(OneHot)* & *LA(MSA)*).

Model trainable on consumer hardware. Extracting ProtT5 pLM embeddings for all proteins used for evaluation took 21 minutes on a single Quadro RTX 8000 with 48GB vRAM. Once those input vectors had been generated, the final LA architecture, consisting of 19 million parameters, could 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.

5. Discussion

LA predicting location: beyond accuracy, four observations for machine learning in biology. The LA approach introduced here constituted possibly the largest margin to date of pLM embeddings improving over SOTA methods

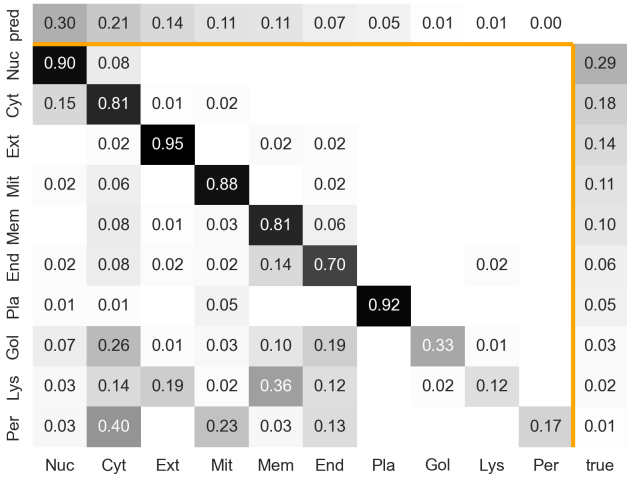

Figure 3. **Mostly capturing majority classes.** Confusion matrix of LA predictions on ProtT5 (Elnaggar et al., 2021) embeddings for *setDeepLoc* (Almagro Armenteros et al., 2017) (see *Appendix: Additional Results* for *setHARD*). Darker color means higher fraction; the diagonal indicates accuracy for the given class; vertical axis: true class; horizontal axis: predicted class. Labels are sorted according to prevalence in ground truth with the most common class first (left or top). Labels: Nuc=**Nuc**leus; Cyt=**Cyt**oplasm; Ext=**Ext**racellular; Mit=**Mit**ochondrion; Mem=cell **Mem**brane; End=**End**oplasmatic Reticulum; Pla=**Pla**stid; Gol=**Gol**gi apparatus; Lys=**Lys**osome/vacuole; Per=**Per**oxisome; **pred**=distribution for predicted (proteins predicted in class X / total number of proteins); **true**=distribution for ground truth (proteins in class X / total number of proteins)

using MSAs. Although this improvement might become crucial to revive location prediction, ultimately this work might become even more important for other lessons learned:

(i) The LA solution improved substantially over all previous approaches to aggregate per-residue embeddings into per-protein embeddings for predictions. Many protein function tasks require per-protein representations, e.g. predictions of Gene Ontology (GO), Enzyme Classifications (E.C.), binary protein-protein interactions (to bind or not), cell-specific and pathway-specific expression levels. Indeed, LA might help in several of these tasks, too.

(ii) Although static, standard data sets (here the *DeepLoc* data) jumpstart advances and help in comparisons, they may become a trap for performance over-estimates through over-fitting. Indeed, the substantial difference in performance between *setDeepLoc* and *setHARD* highlighted this effect dramatically. Most importantly, our results underlined that claims of the type *"method NEW better than SOTA"* should not necessarily constitute wedges for advancing progress. For instance, *NEW* on *setStandard* reaching P(NEW)>P(SOTA) does not at all imply that *NEW* outperformed SOTA. Instead, it might point more to *NEW* over-fitting *setStandard*.

(iii) The new data *setHARD* also pointed to problems with creating too well-curated data sets such as *setDeepLoc*: one aim in selecting a *good* data set is to use only the most reliable experimental results. However, those might be available for only some subset of proteins with particular features (e.g. short, well-folded). Experimental data is already extremely biased for the classes of location annotated (Marot-Lassauzaie et al., 2021). Cleaning up might even increase this bias and thereby limit the validity of prediction methods optimized on those data. Clearly, existing location data differ substantially from entire proteomes (Marot-Lassauzaie et al., 2021).

(iv) *setHARD* also demonstrated that, unlike the protein structure prediction problem (Jumper et al., 2021), the location prediction problem remains unsolved: while Q10 values close to 90% for *setDeepLoc* might have suggested levels close to - or even above - the experimental error, *setHARD* revealed values of Q10 below 70%. In fact, while most proteins apparently mostly locate in one compartment, for others the multiplicity of locations is key to their role. This issue of *travellers vs. dwellers*, implies that Q10 cannot reach 100% as long as we count only one class as correctly predicted for each protein, and if we dropped this constraint, we would open another complication (Marot-Lassauzaie et al., 2021). In short, the new data set clearly generated more realistic performance estimates.

Light attention (LA) beats pooling. The central challenge for the improvement introduced here was to convert the per-residue embeddings (NLP equivalent: word embeddings) from pLMs (*BB* (Bepler & Berger, 2019), *UniRep* (Alley et al., 2019), *SeqVec* (Heinzinger et al., 2019), *ProtBert* (Elnaggar et al., 2021), *ESM-1b* (Rives et al., 2021), and *ProtT5* (Elnaggar et al., 2021)) to meaningful per-protein embeddings (NLP equivalent: document). Qualitatively inspecting the influence of the light attention (LA) mechanism through a UMAP comparison (Fig. 4) highlighted the basis for the success of the LA. The embedding-based annotation transfer (EAT) surpassed some MSA-based methods without any optimization of the underlying pLMs (Fig. 2). In turn, inputting *frozen* pLM embeddings averaged over entire proteins into FNNs surpassed EAT and MSA-based methods (Fig. 2). The simple FNNs even improved over the SOTA, *DeepLoc*, for some pLMs (Fig. 2). However, LA consistently distilled more information from the embeddings. Most likely, the improvement can be attributed to LA coping better with the immense variation of protein length (varying from 30 to over 30 000 residues (Consortium, 2021)) by learning attention distributions over the sequence positions. LA models appeared to have captured relevant long-range dependencies while retaining the ability to focus on specific sequence regions such as beginning and end, which play a

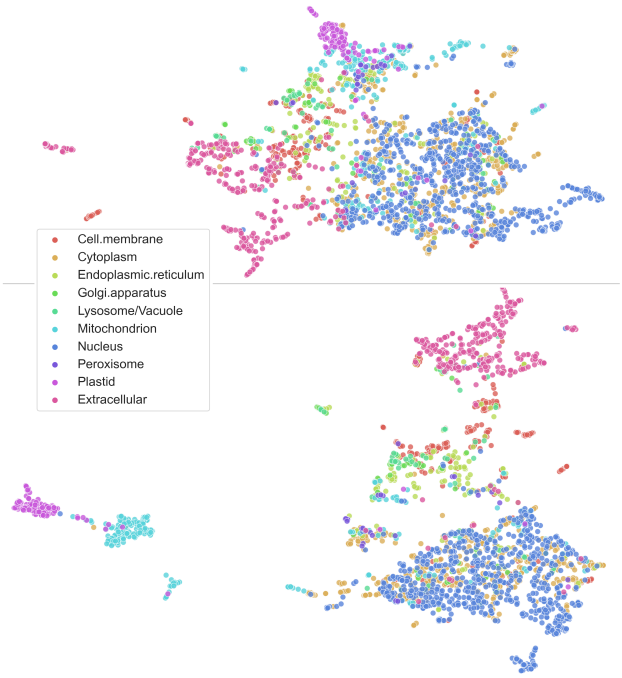

Figure 4. **Qualitative analysis confirmed LA to be effective.** UMAP (McInnes et al., 2018) projections of per-protein embeddings colored according to subcellular location (*setDeepLoc*). Both plots were created with the same default values of the python *umap-learn* library. Top: ProtT5 embeddings (LA input; x) mean-pooled over protein length (as for FNN/EAT 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).

particularly important role in determining protein location for some proteins (Nair & Rost, 2005; Almagro Armenteros et al., 2017).

Embeddings outperformed MSA: first for function. Effectively, LA trained on pLM embeddings from *ProtT5* (Elnaggar et al., 2021) was at the heart of the first method that clearly appeared to outperform the best existing method (*DeepLoc*, (Almagro Armenteros et al., 2017; Heinzinger et al., 2019)) in a statistically significant manner on a new representative data set not used for development (Fig. 2). To the best of our knowledge, it was also the first in outperforming the MSA-based SOTA in the prediction of subcellular location in particular, and of protein function in general. Although embeddings have been extracted from pLMs trained on large databases of unannotated (unlabelled) protein sequences that evolved, the vast majority of data learned originated from much more generic constraints informative of protein structure and function. Clearly, pre-trained pLMs never had the opportunity to learn protein family constraints encoded in MSAs.

Better and faster than MSAs. When applying our solution to predicting location for new proteins (or *at inference*), the embeddings needed as input for the LA models come with three advantages over the historically most informative MSAs that were essential for methods such as *DeepLoc* (Almagro Armenteros et al., 2017) to become top. Most importantly, embeddings can be obtained in far less time than is needed to generate MSAs 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), in our experience, required about 0.3 seconds per protein to generate MSAs for a large set of 10 000 proteins. One of the slowest but most informative pLMs (*ProtT5*) is three times faster, while the third most informative (*ProtBert*) is five times faster (Table 1). Moreover, these MMseqs2 stats derive from runs on a machine with $>$ 300GB of RAM and 2x40cores/80threads CPUs, while generating pLM embeddings required only a moderate machine (8 cores, 16GB RAM) equipped with a modern GPU with >7GB of vRAM. Additionally, the creation of MSAs relied on tools such as MMseqs2 that are sensitive to parameter changes, ultimately an extra complication for users. In contrast, generating embeddings required no parameter choice for users beyond the choice of the pLM (best here *ProtT5*). However, retrieving less specific evolutionary information (e.g. BLOSUM (Henikoff & Henikoff, 1992)) constituted a simple hash-table lookup. Computing such input could be instantaneous, beating even the fastest pLM *SeqVec*. Yet, these generic substitution matrices have rarely ever been competitive in predicting function (Ng & Henikoff, 2003; Bromberg et al., 2008). One downside to using embeddings is the one-off expensive pLM pre-training (Elnaggar et al., 2021; Heinzinger et al., 2019). In fact, this investment pays off if and only if the resulting pLMs are not retrained. If they are used unchanged - as shown here - the advantage of embeddings over MSA is increasing with every single new prediction requested by users (over 3,000/months just for *PredictProtein* (Bernhofer et al., 2021)). In other words, every day, embeddings save more over MSAs.

Overfitting through standard data set? For location prediction, the *DeepLoc* data (Almagro Armenteros et al., 2017) has become a standard. Static standards facilitate method comparisons. To solidify performance estimates, we created a new test set (*setHARD*), which was redundancy-reduced both with respect to itself and all proteins in the *DeepLoc* data (comprised of training plus testing data, the latter dubbed *setDeepLoc*). For *setHARD*, the 10-state accuracy (Q10) dropped, on average, 22 percentage points with respect to the static standard, *setDeepLoc* (Fig. 2). We argue that this large margin may be attributed to some combination of the following coupled effects.

(i) Previous methods may have been substantially overfitted to the static data set, e.g., by misusing the test set to optimize hyperparameters. This could explain the increase in performance on *setHARD* when mimicking the class distributions in the training set and *setDeepLoc*.

(ii) The static standard set allowed for some level of sequence-redundancy (information leakage) at various levels: certainly within the test set, which had not been redundancy reduced to itself (data not shown), maybe also between train and test set. Methods with many free parameters might more easily exploit such residual sequence similarity for prediction because proteins with similar sequences locate in similar compartments. In fact, this may explain the somewhat surprising observation that *DeepLoc* appeared to perform worse on *setHARD* using MSAs than the generic BLOSUM62 (Fig. 2: *DeepLoc62* vs. *DeepLoc*). Residual redundancy is much easier to capture by MSAs than by BLOSUM (Henikoff & Henikoff, 1992) (for computational biologists: the same way in which PSI-BLAST can outperform pairwise BLAST (Altschul et al., 1997)).

(iii) The confusion matrix (Fig. 3) demonstrated how classes with more experimental data tended to be predicted more accurately. As *setDeepLoc* and *setHARD* differed in their class composition, even without overfitting and redundancy, prediction methods would perform differently on the two. In fact, this can be investigated by recomputing the performance on a similar class-distributed superset of *setHARD*, on which performance dropped only by 11, 24, 18, and 17 percentage points for *DeepLoc62*, *DeepLoc*, *LA(ProtBert)*, and *LA(ProtT5)*, respectively.

Possibly, several effects contributed to the performance from standard to new data set. Interestingly, different approaches behaved alike: both for alternative inputs from pLMs (*SeqVec*, *ProtBert*, *ProtT5*) and for alternative methods (EAT, FNN, LA), of which one (EAT) refrained from weight optimization.

What accuracy to expect for the next 10 location predictions? If the top accuracy for one data set was Q10 $\sim 60\%$ and Q10 $\sim 80\%$ for the other, what could users expect for their next ten queries: either six correct or eight, or between six and eight? The answer depends on the query: if those proteins were sequence similar to proteins with known location (case: redundant): the answer would be eight. Conversely, for new proteins (without homologs of known location), six in ten will be correctly predicted, on average. However, this assumes that the ten sampled proteins follow somehow similar class distributions to what has been collected until today. In fact, if we applied *LA(ProtT5)* to a hypothetical new proteome similar to existing ones, we can expect the distribution of proteins in different location classes to be relatively similar (Marot-Lassauzaie et al., 2021). Either way, 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 MSAs.

6. Conclusion

We presented a light attention mechanism (LA) in an architecture operating on embeddings from several pLMs (*BB*, *UniRep*, *SeqVec*, *ProtBert*, *ESM-1b*, and *ProtT5*. LA efficiently aggregated information and coped with arbitrary sequence lengths, thereby mastering the enormous range of proteins spanning from 30-30 000 residues. By implicitly assigning a different importance score for each sequence position (each residue), the method succeeded in predicting protein subcellular location much better than methods based on simple pooling. More importantly, for three pLMs, LA succeeded in outperforming the SOTA without using MSA-based inputs, i.e., the single most important input feature for previous methods. This constituted an important breakthrough: although many methods had come close to the SOTA using embeddings instead of MSAs (Elnaggar et al., 2021), none had ever overtaken as the methods presented here. Our best method, *LA(ProtT5)*, was based on the largest pLM, namely on *ProtT5* (Fig. 2). Many methods were assessed on a standard data set (Almagro Armenteros et al., 2017). Using a new, more challenging data set (*setHARD*), the performance of all methods appeared to drop by around 22 percentage points. While class distributions and data set redundancy (or homology) may explain some of this drop, over-fitting might have contributed more. Overall, the drop underlined that many challenges remain to be addressed by future methods. For the time being, the best method *LA(ProtT5)* is freely available via a webserver (embed.protein.properties) and as part of a high-throughput pipeline (Dallago et al., 2021). Predictions for the human proteome are available via Zenodo https://zenodo.org/record/5047020.

Acknowledgements

Thanks to Tim Karl (TUM) for help with hardware and software; to Inga Weise (TUM) for support with many other aspects of this work. Thanks to the Rostlab for constructive conversations and to the anonymous reviewers for constructive criticism. Thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases. In particular, thanks to the Ioanis Xenarios (SIB, Univ. Lausanne), Matthias Uhlen (Univ. Upssala), and their teams at Swiss-Prot and HPA. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) – project number RO1320/4-1, by the Bundesministerium für Bildung und Forschung (BMBF) – project number 031L0168, and by the BMBF through the program "Software Campus 2.0 (TU München)" – project number 01IS17049.

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, December 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.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25 (17):3389–3402, September 1997. ISSN 0305-1048. doi: 10.1093/nar/25.17.3389. URL https://doi.org/10.1093/nar/25.17.3389.

Bahdanau, D., Cho, K., and Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Bepler, T. and Berger, B. Learning protein sequence embeddings using information from structure. *arXiv:1902.08661 [cs, q-bio, stat]*, October 2019. URL http://arxiv.org/abs/1902.08661. arXiv: 1902.08661.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Bernhofer, M., Dallago, C., Karl, T., Satagopam, V., Heinzinger, M., Littmann, M., Olenyi, T., Qiu, J., Schütze, K., Yachdav, G., Ashkenazy, H., Ben-Tal, N., Bromberg, Y., Goldberg, T., Kajan, L., O'Donoghue, S., Sander, C., Schafferhans, A., Schlessinger, A., Vriend, G., Mirdita, M., Gawron, P., Gu, W., Jarosz, Y., Trefois, C., Steinegger, M., Schneider, R., and Rost, B. PredictProtein - predicting protein structure and function for 29 years. *Nucleic Acids Research*, 49(W1):W535–W540, May 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab354. URL https://doi.org/10.1093/nar/gkab354.

Bhattacharya, N., Thomas, N., Rao, R., Dauparas, J., Koo, P. K., Baker, D., Song, Y. S., and Ovchinnikov, S. Single Layers of Attention Suffice to Predict Protein Contacts. *bioRxiv*, pp. 2020.12.21.423882, December 2020. doi: 10.1101/2020.12.21.423882. URL https://www.biorxiv.org/content/10.1101/2020.12.21.423882v2. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Blum, T., Briesemeister, S., and Kohlbacher, O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC bioinformatics*, 10(1):274, 2009. Publisher: Springer.

Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., and Shatkay, H. SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *Journal of proteome research*, 8(11):5363–5366, 2009. Publisher: ACS Publications.

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. YLoc—an interpretable web server for predicting subcellular localization. *Nucleic acids research*, 38(suppl_2): W497–W502, 2010. Publisher: Oxford University Press.

Bromberg, Y., Yachdav, G., and Rost, B. SNAP predicts effect of mutations on protein function. *Bioinformatics (Oxford, England)*, 24(20):2397–2398, 2008. Publisher: Oxford University Press.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C., Alexandari, A. M., Lu, Z., Harris, D. J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Segler, M. H. S., Boca, S. M., Swamidass, S. J., Huang, A., Gitter, A., and Greene, C. S. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15 (141):20170387, April 2018. doi: 10.1098/rsif.2017.0387. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0387. Publisher: Royal Society.

Chou, K.-C., Wu, Z.-C., and Xiao, X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PloS one*, 6(3):e18258, 2011. Publisher: Public Library of Science.

Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 2021. doi: 10.1093/nar/gkaa1100. URL https://academic.o

up.com/nar/advance-article/doi/10.1093/nar/gkaa1100/6006196.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00994018. URL http://link.springer.com/10.1007/BF00994018.

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., Yang, K. K., Min, S., Yoon, S., Morton, J. T., and Rost, B. Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113, 2021. doi: 10.1002/cpz1.113. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.113.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. 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.

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.

Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367 – 374, 2004. ISSN 1476-9271. doi: https://doi.org/10.1016/j.compbiolchem.2004.09.006. URL http://www.sciencedirect.com/science/article/pii/S1476927104000799.

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.

Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, November 1992. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.89.22.10915. URL http://www.pnas.org/cgi/doi/10.1073/pnas.89.22.10915.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Nakai, K. WoLF PSORT: protein localization predictor. *Nucleic acids research*, 35 (suppl_2):W585–W587, 2007. Publisher: Oxford University Press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., and Rost, B. Embeddings from deep learning transfer GO

annotations beyond homology. *Scientific Reports*, 11(1): 1160, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-80786-0. URL https://www.nature.com/articles/s41598-020-80786-0. Number: 1 Publisher: Nature Publishing Group.

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.

Marot-Lassauzaie, V., Goldberg, T., Armenteros, J. J. A., Nielsen, H., and Rost, B. Spectrum of protein location in proteomes captures evolutionary relationship between species. *Journal of molecular evolution*, pp. 1–10, 2021. Publisher: Springer.

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.

Nair, R. and Rost, B. Sequence conserved for subcellular localization. *Protein Science*, 11(12):2836–2847, 2002. ISSN 1469-896X. doi: https://doi.org/10.1110/ps.0207402. URL https://onlinelibrary.wiley.com/doi/abs/10.1110/ps.0207402.

Nair, R. and Rost, B. Mimicking cellular sorting improves prediction of subcellular localization. *Journal of Molecular Biology*, 348(1):85–100, 2005. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2005.02.025. URL https://www.sciencedirect.com/science/article/pii/S0022283605001774.

Ng, P. C. and Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, July 2003. ISSN 0305-1048. doi: 10.1093/nar/gkg509. URL https://doi.org/10.1093/nar/gkg509.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, 22(14):e408–416, July 2006.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

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.

Rao, R., Ovchinnikov, S., Meier, J., Rives, A., and Sercu, T. Transformer protein language models are unsupervised structure learners. *bioRxiv*, pp. 2020.12.15.422761, December 2020. doi: 10.1101/2020.12.15.422761. URL https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, 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. *Proceedings of the National Academy of Sciences*, 118 (15), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/content/118/15/e2016239118.

Rost, B. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94, February 1999. ISSN 1741-0126. doi: 10.1093/protein/12.2.85. URL https://doi.org/10.1093/protein/12.2.85.

Rost, B. Enzyme Function Less Conserved than Anticipated. *Journal of Molecular Biology*, 318(2):595–608, April 2002. ISSN 0022-2836. doi: 10.1016/S0022-2836(02)00016-5. URL http://www.sciencedirect.com/science/article/pii/S0022283602000165.

Rost, B. and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*, 232(2):584–599, 1993. doi: 10.1006/jmbi.1993.1413.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, 60(12):2637–2650, 2003. doi: 10.1007/s00018-003-3114-8. URL http://www.rostlab.org/papers/2003_rev_func/. Type: Journal article.

Sander, C. and Schneider, R. Database of homology-derived protein structures and the structural meaning

of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9(1):56–68, 1991. ISSN 1097-0134. doi: https://doi.org/10.1002/prot.340090107. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340090107.

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments. *Bioinformatics*, 33(3):347–353, 2017.

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Research*, 46(W1):W459–W466, 2018. ISSN 0305-1048. doi: 10.1093/nar/gky320. URL https://doi.org/10.1093/nar/gky320.

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.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, March 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Urban, G., Torrisi, M., Magnan, C. N., Pollastri, G., and Baldi, P. Protein profiles: Biases and protocols. *Computational and Structural Biotechnology Journal*, 18:2281 – 2289, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2020.08.015. URL http://www.sciencedirect.com/science/article/pii/S2001037020303688.

Weißenow, K., Heinzinger, M., and Rost, B. Protein language model embeddings for fast, accurate, alignment-free protein structure prediction. *bioRxiv : the preprint server for biology*, 2021. doi: 10.1101/2021.07.31.454572.

Yu, C.-S., Chen, Y.-C., Lu, C.-H., and Hwang, J.-K. Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, 64(3):643–651, 2006.