Pretrained embeddings in protein function prediction

Tomasz Cheda
Uniwersytet Warszawski
tomasz.cheda@students.mimuw.edu.pl

Jakub Stępka
Uniwersytet Warszawski
j.stepka@uw.student.uw.edu.pl

Jan Dziuba
Uniwersytet Warszawski
jd406140@students.mimuw.edu.pl

Supervisor
Grzegorz Preibisch
Uniwesytet Warszawski, Deepflare

Abstract

This paper targets the protein function prediction problem as formulated in the CAFA 3 challenge (Zhou et al., 2019). We reproduce and extend the results achieved by goPredSim (Littmann et al., 2021). We compare the results achieved with embeddings generated from ProteinBert (Brandes et al., 2022), ESM2 (Lin et al., 2022), SeqVec (Heinzinger et al., 2019), ProtBert and ProtT5-XL (Elnaggar et al., 2022). Our goal is to investigate the dependence of annotation transfer performance on embedding models and distance functions.

1 Introduction

Proteins are the molecular machinery powering all organisms. Determining how they work is crucial to understanding life at the molecular level. Their chemical structure seems simple - they are built from a set of just 20 amino acids linked in linear chains, generally 50 to 2000 units long. However, elements of these chains interact with each other (and their environment) in non-obvious ways to form a variety of 3D structures. They are also not rigid and may transition between different shapes as part of their function. Different proteins, or copies of the same protein bind together to perform tasks. Changing just one of the amino acids may disturb these processes, a well-known example being a mutation in human hemoglobin causing sickle cell anemia. Proteins may have multiple roles and similar proteins may play disparate roles in different organisms.

These complexities make determining the task a given protein performs, referred to as functional annotation, crucial, but extremely challenging. While recent advances in sequencing technology made determining the amino acid sequence cheap, with the UniProtKB database containing more than 245 million sequences, there have not yet been corresponding improvements in the determination of protein function. Transferability of in-vitro results is limited, there are no high-throughput assays for experimental function determination, and scientists still rely mainly on manual annotation. This is reflected by the size of the annotated portion of UniProtKB - roughly 570000 sequences. This gap is growing exponentially.

Aiming to address this problem, automatic annotation, or protein function prediction, is an active area of research. The problem is usually formulated as predicting the consistent subgraph of ontologized annotation terms applicable to a new protein given its sequence.

Evaluation of protein function prediction is notoriously difficult. As the existing, ground-truth annotations are incomplete, the absence of evidence for a given function may not be taken as evidence for the absence of that function. Not all species are sequenced and annotated uniformly.

Aiming to address this, the Function Special Interest Group periodically conducts the open Critical Assessment of protein Function Annotation, or CAFA, challenge. During the prediction phase, teams submit their predictions of function annotations for a set of proteins. The prediction phase ends, and in the following months newly experimentally determined annotations are collected. At the end of the collection period, earlier submissions are evaluated, providing a realistic assessment scenario. The third edition report (Zhou et al., 2019) was published in 2019 and CAFA5 is ongoing on Kaggle.

The recent NLP revolution, with better models for processing sequences (Vaswani et al., 2017) and the successes of pretraining in tasks with limited labeled data (Brown et al., 2020), inspired the application of similar techniques in bioinformatics. Representable as strings of amino-acid letters, proteins are a natural fit for methods derived from NLP. Although the analogies only go so far, many algorithms have been successfully transferred to this domain. (Ofer et al., 2021)

Many of the current methods for protein function annotation already use embeddings from pretrained language models, either to enrich the data (Gligorijević et al., 2021), to transfer annotations from nearest neighbors in the embedding space (Littmann et al., 2021)), or even to directly predict function(Dohan et al., 2021).

This paper expands on goPredSim, which transfers annotations from the most similar protein, as measured by Euclidean or cosine distance in the embedding space. We transfer from experimentally determined annotations available at the commencement of the CAFA3 challenge and evaluate on proteins which received new annotations in the challenge review period. In addition to SeqVec (Heinzinger et al., 2019), ProtBert-BFD and ProtT5-XL (Elnaggar et al., 2022) previously used by the authors, we compute embeddings from ESM2 (Lin et al., 2022), ProteinBert (Brandes et al., 2022), and ProtBert-UR100 (Elnaggar et al., 2022). We compare the maximum F-measure in no-knowledge and limited-knowledge settings on the BPO, MFO, and CCO ontologies across embedding models and distance functions and discuss our findings.

2 Related Work

Over the past decade, extensive research has been dedicated to achieving remarkable performance in protein function prediction. These methods often integrate multiple data sources, such as protein-protein interaction databases(Xu and Wang, 2022), 3D structures resolved in-silico or experimentally (Gligorijević et al., 2021), or similarity to proteins with previously determined annotations(Littmann et al., 2021)(Törönen and Holm, 2021).

Methods for protein function prediction that utilize annotation transfer from other proteins employ diverse strategies for identifying similar protein sequences. The homology-based transfer involves using alignment algorithms, including popular methods like BLAST, PSI-BLAST (You et al., 2019), and tools like SANSparallel (Törönen and Holm, 2021), to identify similarities directly between aminoacid sequences. Some methods use sequence embedding to determine similarity. In particular, goPredSim (Littmann et al., 2021) uses pretrained language models, either bi-directional LSTM (Heinzinger et al., 2019) or transformer (ProtBert), to generate embeddings, and then the kNN algorithm to determine the closest proteins in embedding space and transfer its embeddings.

Among models which do not directly transfer annotations, we can distinguish between two groups: one consists of approaches that utilize engineered features, such as amino acid frequencies, motif and domain occurrences, and biophysical properties of proteins (You et al., 2019). The other subgroup comprises methods that leverage deep learning to extract features automatically, which the model subsequently employs for label assignment. These approaches have proliferated in recent years, employing various model architectures including CNNs (Kulmanov and Hoehndorf, 2019) (Bileschi et al., 2022), RNNs (Xia et al., 2022), and transformers (Brandes et al., 2022) (Dohan et al., 2021). An interesting approach was demonstrated in ProLanGO (Cao et al., 2017), which approaches the task as a translation problem. A very recent and promising addition is ProteInfer (Sanderson et al., 2023)

It is also worth noting that approaches based on the sequence are often combined with other protein data - protein structures or networks of interaction (Xu and Wang, 2022). An example of such an approach can be DeepFRI which combines protein structure with sequence embedding in a graph convolutional network to predict protein function (Gligorijević et al., 2021). Another approach for predicting protein function is the prediction of protein names, specifically used in UniProt's Automatic Annotation pipeline for unnamed proteins (Andreea Gane).

3 Method

3.1 goPredSim

goPredSim is designed to generate GO predictions for a set of target proteins using a lookup dataset of proteins that already have known GO annotations. Both the lookup dataset and the target proteins are represented as embeddings. The model calculates pairwise distances between proteins in the lookup set and the target set. By identifying similar proteins in the lookup set, the GO annotations associated with those proteins are transferred to the target proteins. The reliability of predictions is quantified with a 0-1 similarity score calculated from the distance.

3.2 goPredSim + ESM-2 (added)

The ESM-2 (Lin et al., 2022) is a general-purpose protein language model introduced in August 2022

by Meta for predicting structure, function, and others.

It is a transformer trained with a masked language modeling objective on approximately 65 million sequences from the UniRef (32) protein sequence database.

We used the `esm2_t33_650M_UR50D` version of ESM-2. It has 650M parameters, 33 layers, and an embedding dimension of 1280. It was trained on the 2021_04 version of UR50D.

Our max sequence length is 1022, cropping longer sequences.

The per-protein embedding we use is extracted from the 33rd layer of the model and mean-pool over positions.

3.3 goPredSim + ProteinBert (added)

ProteinBert (Brandes et al., 2022) is a Bert derivative with several innovations, pretrained on two tasks: language modeling and GO annotation prediction. It was introduced in January 2022.

The dataset it uses consists of approximately 106 million proteins sourced from UniRef90, with corresponding UniProtKB GO annotations on selected terms.

The model is trained for about 6 epochs. Both objectives are optimized jointly.

Our max sequence length is 1024, cropping longer sequences. This could be changed in future iterations, as ProteinBert incorporates mechanisms for the effective processing of large sequences.

For embedding we use the global representation of the model.

3.4 goPredSim + ProtBert-BFD (pre-existing)

ProtBert (Elnaggar et al., 2022) is one of the protein language models introduced in the ProtTrans in July 2020, utilizing a masked language modeling objective similar to Bert. It increases the number of layers, compared to the original Bert.

ProtBert-BFD was trained for 800k steps on proteins with a maximum length of 512, then for another 200k steps on sequences with a maximum length of 2000 to maximize the sample efficiency. It used the BFD-100(Steinegger and Söding, 2018) dataset.

We take embeddings from the last hidden layer state and mean-pool over sequence length. There is no set maximum sequence length. The embedding dimension is 1024.

3.5 goPredSim + ProtBert-UR100 (added)

This is a different version of ProtBert, also introduced in the July 2020 ProtTrans paper.

It shares the architecture of ProtBert-BFD but is instead trained on the UniRef100 dataset, first for 300,000 steps on protein sequences with a maximum length of 512, then 100,000 steps more on proteins with a maximum length of 2000.

We similarly take the last hidden layer state and mean-pool over length to get an embedding of dimension 1024. However, our maximum sequence length is 1200, and longer sequences are cropped.

3.6 goPredSim + ProtT5 (pre-existing)

ProtT5-XL-U50 (Elnaggar et al., 2022) is another model introduced in ProtTrans in July 2020, this time derived from T5.

It's trained with a denoising objective, first for 1.2 million steps on BFD and next for 991000 on UniRef50.

We take embeddings from the last hidden state layer and mean-pool over sequence length. There is no set maximum sequence length. The embedding dimension is 1024.

3.7 goPredSim + SeqVec (pre-existing)

SeqVec (Heinzinger et al., 2019) is a modification of the ELMo architecture adapted to protein sequences and introduced in April 2019. It has 1 CharCNN and 2 LSTM layers.

The model was trained on UniRef50.

We take a 1024-dimensional output of each of the three layers and sum position-wise.

There is no set maximum sequence length.

4 Experimental Set-up

4.1 Code availability

The code used to conduct these experiments is available at `https://github.com/chedatomasz/annotation_transfer`.

4.2 Dataset preparation

We determined the set of protein ids for training (computing embeddings and downloading annotations for transfer) by taking the intersection of the ids in the `uniprot_sprot_exp.fasta` file provided by CAFA3 organizers and the `goa_annotations_exp_2017.txt` file provided by goPredSim developers. This represents the set of proteins, for which experimentally determined annotations were present in the GOA database at the end

of the submission period of CAFA3. We derived the annotations from the goa_annotations_exp_2017.txt file. This similarly represents the state of knowledge available to CAFA3 participants.

The set of protein ids for testing is derived from the cafa3_targets.fasta file from goPredSim or equivalently the ground truth lists from the CAFA assessment tool. This represents the set of proteins, for which new annotations were experimentally determined in the evaluation period of CAFA3.

This lets us achieve a temporal split between the training and testing set, providing a realistic assessment scenario.

4.3 Embedding preparation

The embeddings for preexisting models were derived from the files made available by goPredSim developers at https://rostlab.org/public/goPredSim/ and in the data/target_embeddings folder.

The embeddings for ESM2 and ProtBert-UR100 were computed using the Entropy cluster at the MIM faculty of the University of Warsaw. The embeddings for ProteinBert were computed using a personal RTX3070 GPU with 8GB of VRAM, inferencing in-framework with Tensorflow mixed precision.

4.4 Prediction and evaluation

The predictions were computed using a personal system with an RTX3070 GPU with 8GB of VRAM, 32GB system RAM, and an intel i7 4770K CPU. The goPredSim script, available at https://github.com/Rostlab/goPredSim, was modified to be compatible with the CAFA assessment tool (Ashleyzhou972 et al., 2019), fixing minor bugs. The CAFA assessment tool, available at https://github.com/ashleyzhou972/CAFA_assessment_tool, was used to evaluate the predictions.

4.5 Evaluation Metrics

For evaluation, we report the maximum F-measure as the performance metric. This measure is based on weighted precision and recall calculations performed on each of the three test sets. Detailed formulas for computing the weighted F-measures can be found in the referenced paper.(et al., 2016)

4.6 No Knowledge vs Limited Knowledge

In keeping with CAFA3, we report results for No Knowledge and Limited Knowledge evaluation. A protein is included in the No Knowledge evaluation only if it previously had no experimentally determined annotation for a given ontology.

5 Results and Discussion

ontology model	BPO	CCO	MFO	mean
cosine/prott5	0.3219	0.5862	0.5412	0.4831
euclidean/prott5	0.3243	0.5886	0.5312	0.4814
cosine/esm2	0.3192	0.5839	0.5108	0.4713
cosine/proteinbert	0.3090	0.5730	0.5225	0.4682
euclidean/esm2	0.3209	0.5752	0.5044	0.4669
euclidean/proteinbert	0.3110	0.5714	0.5117	0.4647
cosine/seqvec	0.3119	0.5581	0.5212	0.4638
euclidean/seqvec	0.3056	0.5611	0.4963	0.4543
euclidean/protbert-bfd	0.2979	0.5670	0.4728	0.4459
cosine/protbert-bfd	0.2976	0.5619	0.4706	0.4434
cosine/protbert-ur100	0.2641	0.5372	0.3980	0.3998
euclidean/protbert-ur100	0.2601	0.5382	0.3987	0.3990

Table 1: Comparison of embedding models and distance functions using the F measure in the 'No knowledge' evaluation for BPO, CCO, and MFO ontologies, sorted by mean of these scores

ontology model	BPO	CCO	MFO	mean
cosine/esm2	0.3432	0.5633	0.4546	0.4537
euclidean/prott5	0.3320	0.5661	0.4624	0.4535
euclidean/proteinbert	0.3623	0.5434	0.4520	0.4526
cosine/prott5	0.3328	0.5630	0.4570	0.4509
euclidean/esm2	0.3402	0.5701	0.4416	0.4507
cosine/proteinbert	0.3673	0.5363	0.4449	0.4495
euclidean/protbert-bfd	0.3309	0.5386	0.4317	0.4338
cosine/protbert-bfd	0.3292	0.5452	0.4183	0.4309
cosine/seqvec	0.2983	0.5326	0.4286	0.4198
euclidean/seqvec	0.2920	0.5403	0.4219	0.4181
euclidean/protbert-ur100	0.3154	0.5371	0.3481	0.4002
cosine/protbert-ur100	0.3146	0.5344	0.3507	0.3999

Table 2: Comparison of embedding models and distance functions using the F measure in the 'Limited knowledge' evaluation for BPO, CCO, and MFO ontologies, sorted by mean of these scores

5.1 Limited transfer pool

The chosen set of training ids, being limited to purely experimentally determined annotations, is fairly limited. Extending this set may shift the balance between embedding models and is an interesting direction for further research.

5.2 Temporal vs clustered split

As we follow the temporal train/test split resulting from CAFA's blind assessment, we are limited in what other models we could compare our results

to. Many models report results for a random split (corrected by clustering sequences with very high homology).

5.3 Protein cropping

ProtBert-UR100 consistently underperforms compared to ProtBert-BFD. While this may also be a case of innate differences between the model, it is also likely to be the result of cropping the proteins for processing by the UR100 version. Some protein functions are closely connected to a specific protein domain, which may lay entirely outside the processed segment. A focused analysis, comparing embeddings for a set model in a crop/no crop scenario, could quantify this effect

5.4 Choice of distance metrics and similarity

It is not immediately obvious that Euclidean similarity between high-dimensional embeddings will work sufficiently well for identifying neighbors, especially with no special properties induced in the space.

However, our findings seem to support the conclusion of goPredSim authors that there is no discernible difference between cosine and Euclidean distance, at least for the 1-nearest-neighbor case. Extending the model with a trainable distance metric would be an interesting avenue to explore.

5.5 Potential temporal split violation

Not all embedding models have a known date of data capture (UniRef is a rolling dataset family refreshed every few weeks). This may cause the inclusion in the pretraining process of data which could be known at the training set cutoff. ProteinBert explicitly includes GO annotations in its pretraining. Although no cutoff is known, given the recent publishing date it is likely that some of the CAFA3 targets were included in that pretraining process, thus constituting a data leak. A more fair assessment would require determining the knowledge cutoff date of ProteinBert and re-evaluating all methods on an appropriately constructed temporal split dataset.

5.6 Lack of confidence intervals

The differences between the F-max of different configurations are small. To conclude whether the differences are statistically significant, we would need to perform a bootstrapping analysis.

5.7 Methods incorporating other data

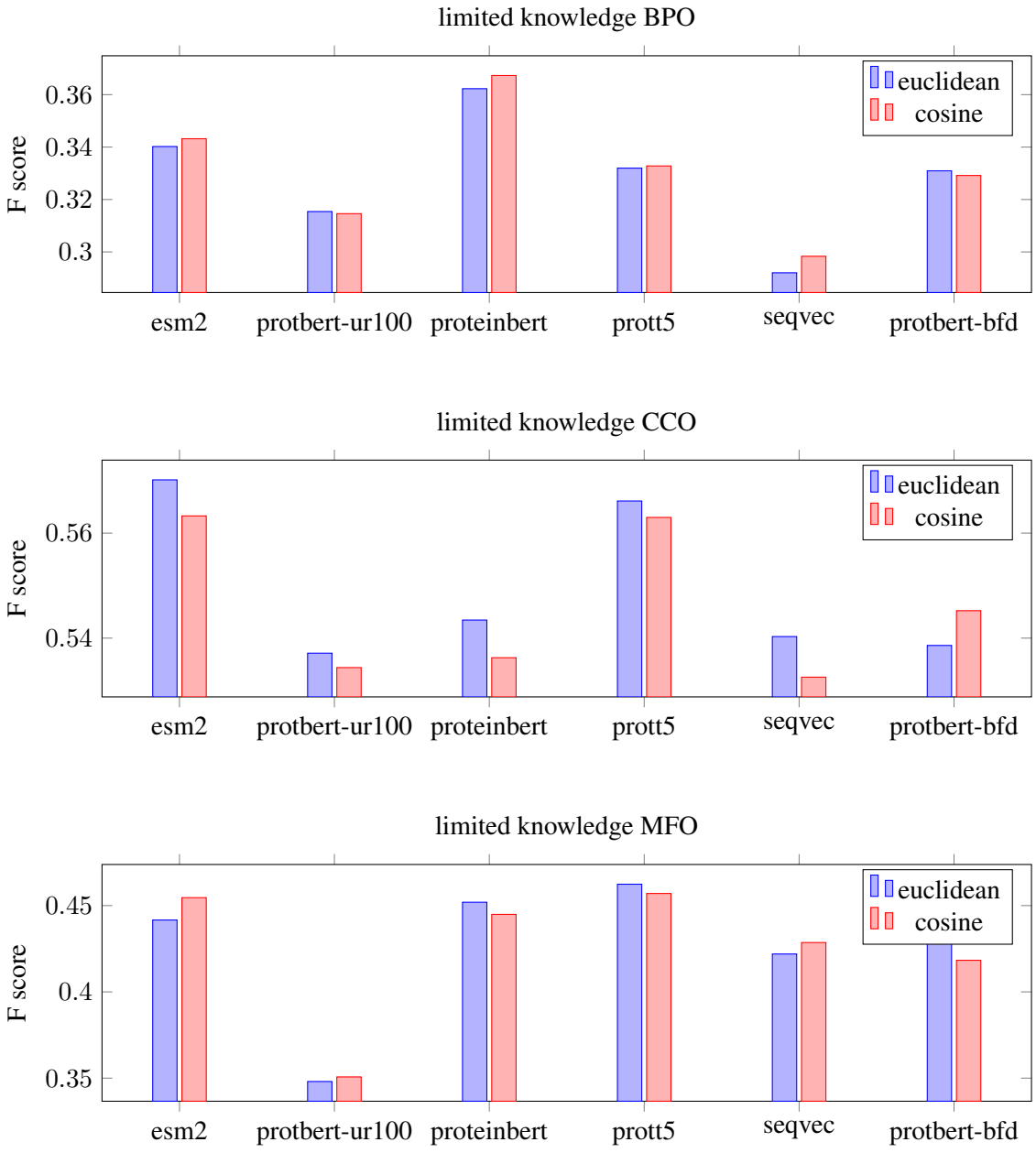
goPredSim is a relatively simple model, in that it relies solely on embedding similarity to perform annotation transfer. Models incorporating other data sources, such as DeepFRI which we originally also intended to assess, may show less improvement from more powerful embeddings. DeepFRI and some other models also require per-residue instead of per-protein embeddings. Most embedding models surveyed here are capable of producing these, but a re-computation of embeddings provided by goPredSim would be required.

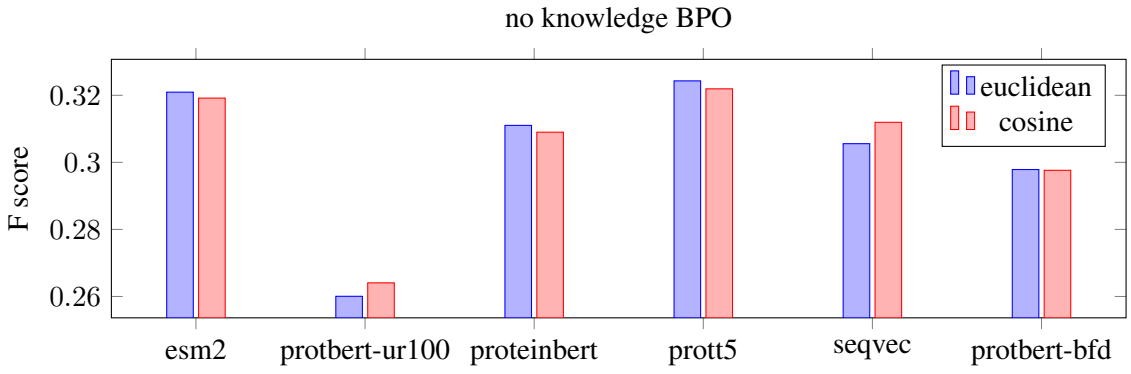
5.8 Comparison to goPredSim

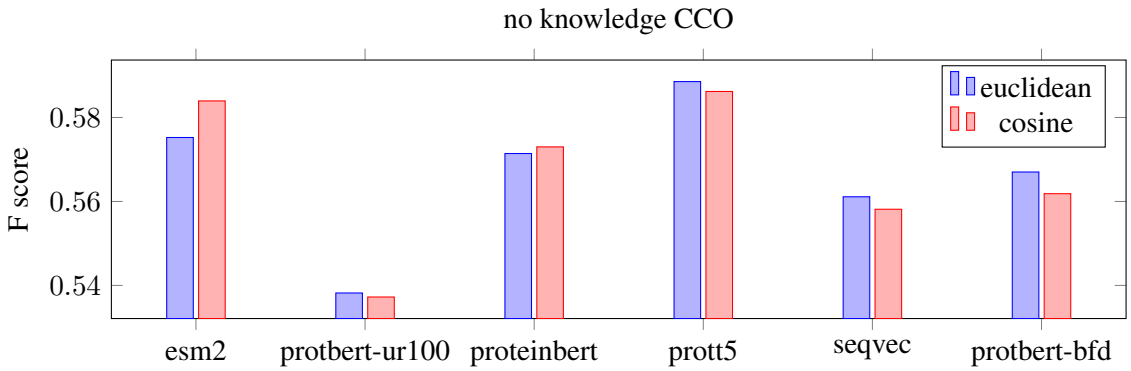
Compared to the results reported by goPredSim on the GOA2017 set, we observe the SeqVec embeddings underperforming relative to ProtBert-BFD and ProtT5. We conjecture that with a more limited lookup set, more distant neighbors need to be identified, and more powerful embeddings handle these situations better, similar to how embeddings outperform homology-based methods.

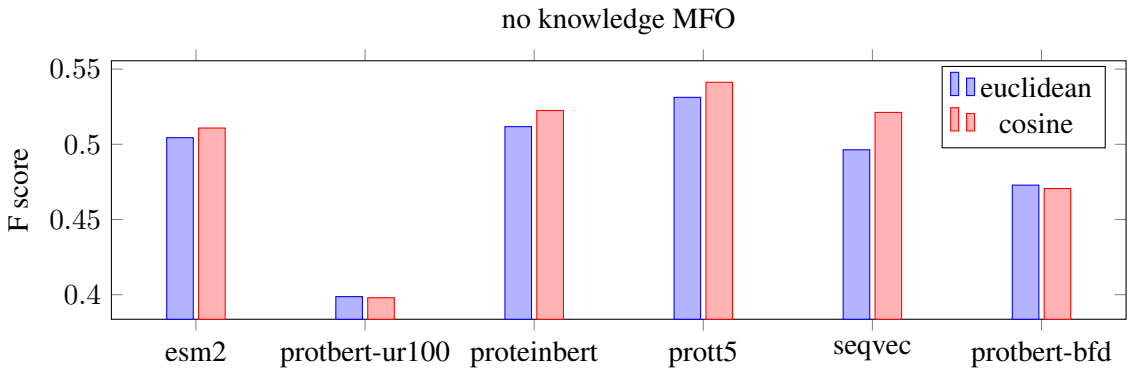
We share the finding that ProtT5 exhibits good performance across the board.

5.9 Plots


limited knowledge BPO
F score
0.36
0.34
0.32
0.3
euclidean
cosine
esm2
protbert-ur100
proteinbert
prott5
seqvec
protbert-bfd
limited knowledge CCO
F score
0.56
0.54
euclidean
cosine
esm2
protbert-ur100
proteinbert
prott5
seqvec
protbert-bfd
limited knowledge MFO
F score
0.45
0.4
0.35
euclidean
cosine
esm2
protbert-ur100
proteinbert
prott5
seqvec
protbert-bfd


no knowledge BPO
F score
0.32
0.3
0.28
0.26
euclidean
cosine
esm2
protbert-ur100
proteinbert
prott5
seqvec
protbert-bfd


no knowledge CCO
F score
0.58
0.56
0.54
euclidean
cosine
esm2
protbert-ur100
proteinbert
prott5
seqvec
protbert-bfd


no knowledge MFO
F score
0.55
0.5
0.45
0.4
euclidean
cosine
esm2
protbert-ur100
proteinbert
prott5
seqvec
protbert-bfd

6 Conclusions

ProteinBert and ESM2 embeddings are promising candidates for annotation transfer methods. While the effect size is uncertain, more powerful embedding models are a promising direction for embedding-similarity-based annotation transfer.

References

David Dohan Elena Speretta Amélie Héliou Laetitia Meng-Papaxanthos Hermann Zellner Eugene Brevdo Ankur Parikh Maria J. Martin Sandra Orchard UniProt Collaborators Lucy J. Colwell Andreea Gane, Maxwell L. Bileschi. Protnlm: Model-based natural language protein annotation.

Ashleyzhou972, Michael Gerten, and Iddo Friedberg. 2019. ashleyzhou972/cafa$_a ssessment_t ool$: $Cafa3$.

Maxwell L. Bileschi, David Belanger, Drew H. Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Alex Bateman, Mark A. DePristo, and Lucy J. Colwell. 2022. Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40(6):932–937.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Renzhi Cao, Colton Freitas, Leong Chan, Miao Sun, Haiqing Jiang, and Zhangxin Chen. 2017. ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network. *Molecules*, 22(10):1732.

David Dohan, Andreea Gane, Maxwell L. Bileschi, David Belanger, and Lucy Colwell. 2021. Improving protein function annotation via unsupervised pretraining: Robustness, efficiency, and insights. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp Data Mining*. ACM.

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. 2022. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.

Jiang Y et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(184).

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1).

Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1).

Maxat Kulmanov and Robert Hoehndorf. 2019. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2022. Evolutionary-scale prediction of atomic level protein structure with a language model.

Maria Littmann, Michael Heinzinger, Christian Dallago, Tobias Olenyi, and Burkhard Rost. 2021. Embeddings from deep learning transfer GO annotations beyond homology. *Scientific Reports*, 11(1).

Dan Ofer, Nadav Brandes, and Michal Linial. 2021. The language of proteins: Nlp, machine learning protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. 2023. ProteInfer, deep neural networks for protein functional inference. *eLife*, 12.

Martin Steinegger and Johannes Söding. 2018. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1).

Petri Törönen and Liisa Holm. 2021. scpPANNZER/scp—a practical tool for protein function prediction. *Protein Science*, 31(1):118–128.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Weiqi Xia, Lingyan Zheng, Jiebin Fang, Fengcheng Li, Ying Zhou, Zhenyu Zeng, Bing Zhang, Zhaorong Li, Honglin Li, and Feng Zhu. 2022. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Computers in Biology and Medicine*, 145:105465.

Hanwen Xu and Sheng Wang. 2022. Protranslator: zero-shot protein function prediction using textual description.

Ronghui You, Shuwei Yao, Yi Xiong, Xiaodi Huang, Fengzhu Sun, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Research*, 47(W1):W379–W387.

Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S Rifaioglu, Alperen Dalkiran, Rengul Cetin-Atalay, Chengxin Zhang, Rebecca L Hurto, Peter L Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M Fernández, Branislava Gemovic, Vladimir R Perovic, Radoslav S Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad RK Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Indika Kahanda, Natalie Thurlby, Alice C McHardy, Alexandre Renaux, Rabie Saidi, Julian Gough, Alex A Freitas, Magdalena Antczak, Fabio Fabris, Mark N Wass, Jie Hou, Jianlin Cheng, Jie Hou, Zheng Wang, Alfonso E Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po-Han Chi, Wei-Cheng Tseng, Michal Linial, Peter W Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Wiarwick Vesztrocy, Jose Manuel Rodriguez, Michael L Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B Roche, Jonas Reeb, David W Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang, Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio CE Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S Black, Dane Jo, Dallas J Larsen, Ashton R Omdahl, Luke W Sagers, Erica Suh, Jonathan B Dayton, Liam J McGuffin, Danielle A Brackenridge, Patricia C Babbitt, Jeffrey M Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia-Ming Chang, Wen-Hung Liao, Yi-Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E E Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E Brenner, Christine A Orengo, Constance J Jeffery, Giovanni Bosco, Deborah A Hogan, Maria J Martin, Claire O'Donovan, Sean D Mooney, Casey S Greene, Predrag Radivojac, and Iddo Friedberg. 2019. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens.