XIII Argentine Congress of Bioinformatics and Computational Biology
XIII International Conference of the Iberoamerican Society of Bioinformatics
III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

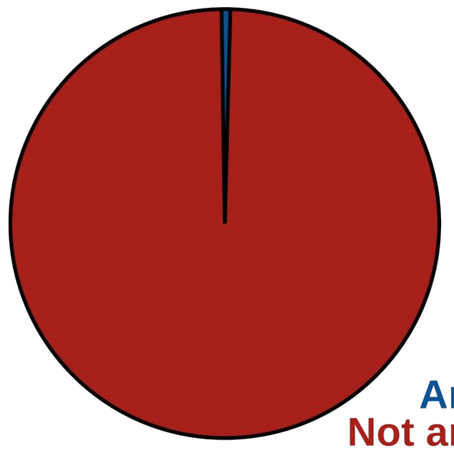
Assessment of Large Language Models for protein domain annotation

Rosario Vitale, Leandro Bugnon, Emilio Fenoy, Diego H. Milone and Georgina Stegmayer
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Santa Fe, Argentina

INTRODUCTION

There are > 248 million protein entries in UniProt. However, < 1% of them have been annotated with one of the more than 17,000 Pfam domains.

Pfam annotations ➡ sequence similarity (BLAST) + manually curated seed alignments of homologous protein regions to get profiles with hidden Markov models (HMMs).

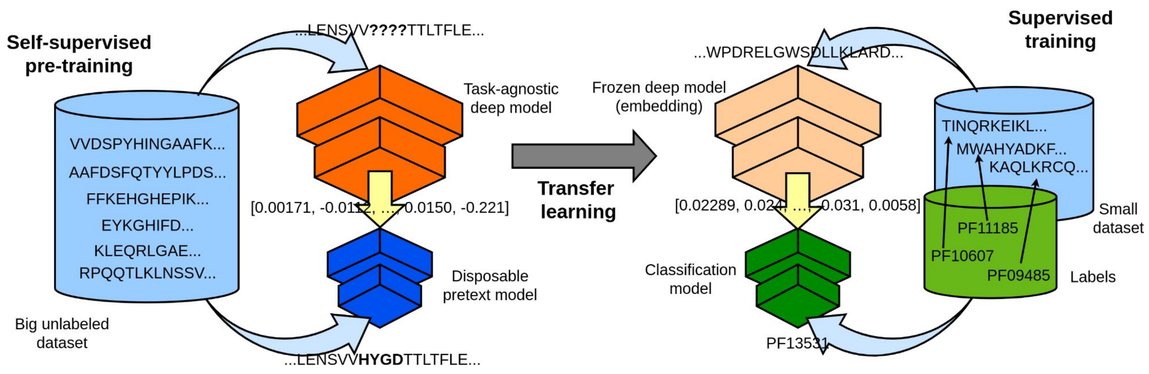


NEW ➡ deep learning (DL) models can learn patterns or hidden conformation rules shared across families. But… need large scale data.

SOLUTION ➡ transfer learning (TL) + large language models (LLMs)

PROPOSAL: TL + LLMs

We propose to use Transfer Learning (TL) from Large Language Model (LLMs).



The TL approach involves self-supervised learning on large and unlabeled protein datasets to generate a numerical embedding for each sequence (left). This representation learned by a LLM (orange) can then be used (transferred) with supervised learning on a small labeled dataset for a specific classification task, such as protein domain classification (right).

MATERIALS AND METHODS

We have compared several ML models (KNN, MLP, CNN and ensembles) using 5 different LLMs:

- ESM models developed by Facebook Research [2] These models use a BERT Transformer that processes sequences of amino acids as input. We have used ESM-1b, ESM-1v and ESM2.
- ProtTrans models developed by Google [3]. These models were trained using several Transformers, including BERT and T5. We have used ProtTransBert and ProtTransT5-XL.

We used Pfam v32 database (https://www.ebi.ac.uk/interpro/entry/pfam/) for classifying proteins into 17,929 families.

Sequences were split by clustering them based on sequence similarity. This provides a hard benchmark task for annotation of protein sequences with remote homology, that is, sequences in the test set have very low (<25%) similarity to the ones in the training set.
The dataset has 1,339,083 training sequences and 21,293 testing sequences.

RESULTS

Table 1 shows the results for the methods without TL. The rows reproduce the results reported in [1] for this same dataset and train/test partition: HMM obtained from the raw sequence, and 2 DL models that use simple one-hot encoding input. Best model in bold.

No TL	Error rate	Errors
HMM	18.10%	3,844
ProtCNN	27.60%	5,882
ProtENN	**12.20%**	**2,590**

Table 1: Results without TL

Tables 2 and 3 show the best results of this study, using TL + LLMs for annotating Pfam v32 with ESM2 and ProtTrans T5-XL LLMs.

ESM 2	Error rate	Errors
KNN	15.55%	3,311
MLP	30.88%	6,576
MLP-Ensemble	18.10%	3,854
CNN	17.30%	3,684
CNN-Ensemble	**7.66%**	**1,631**

Table 2: Results with TL+ESM2

ProtTrans T5-XL	Error rate	Errors
KNN	8.63%	1,838
MLP	26.32%	5,604
MLP-Ensemble	15.08%	3,211
CNN	16.67%	3,549
CNN-Ensemble	**7.23%**	**1,540**

Table 3: Results with TL+ProtTrans T5-XL

Using a CNN ensemble + data embedded with ProtT5-XL achieved the best performance.

If we compare the best model without using TL (ProtENN) and the best model using TL (CNN-Ensemble + ProtTrans T5-XL), the last one is the one with the best results.

These results validate our proposal.

CONCLUSIONS

In this work we have tested cutting-edge transfer learning techniques together with deep learning to improve the actual prediction of protein domain annotations.
The results obtained indicate that this approach has effective predictive advantages over existing methods and it should become part of future Pfam annotation tools.

REFERENCES

[1]Bileschi, M. L., et al. (2022). Using deep learning to annotate the protein universe. Nature Biotechnology, 40(6), 932-937.
[2]Rives, A., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.
[3]Elnaggar, A., et al. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127.
[4]Vitale, R., Stegmayer, G. (2023). Evaluating transfer learning for classification of proteins in bioinformatics. Argentinian Symposium on Artificial Intelligence (ASAI), 1(1), 1-10