Unsupervised protein embeddings outperform handcrafted sequence and structure features at predicting molecular function

Amelia Villegas-Morcillo, Stavros Makrodimitris, Roeland C.H.J. van Ham, Angel M. Gomez, Victoria Sanchez, and Marcel J.T. Reinders

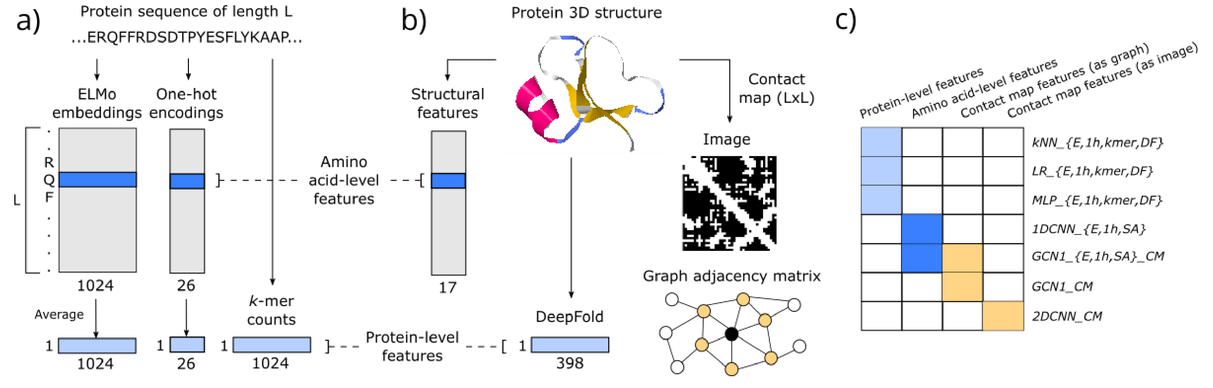
UNIVERSIDAD DE GRANADA

TU Delft

1. INTRODUCTION

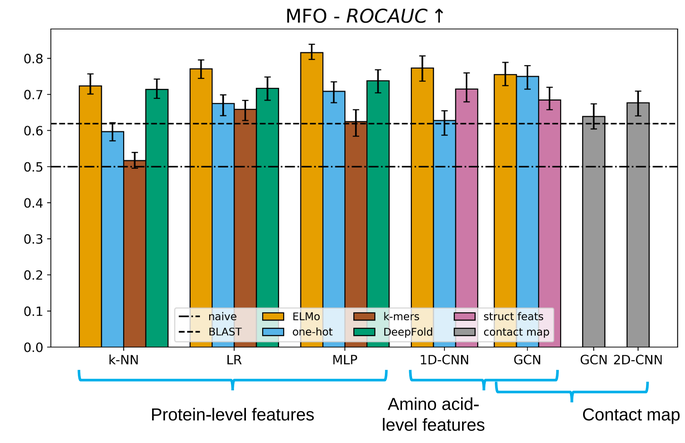
Protein function prediction algorithms are essential, especially for non-model species, as experimental function discovery is simply too costly and time-consuming. Recent advances in machine learning have shifted the focus from homology search and hand-crafting of sequence-based features to automatically learning a useful sequence representation through deep neural networks. However, these deep models require a large amount of labelled training examples, which are not available for the task of function prediction. On the other hand, many protein sequences of unknown function are available. These can be fed into an unsupervised model that learns a general protein representation, which can then be applied to other protein-related tasks, either directly or after further supervised training. Such an unsupervised model (ELMo) was recently published [1], having competitive performance in various protein classification tasks.

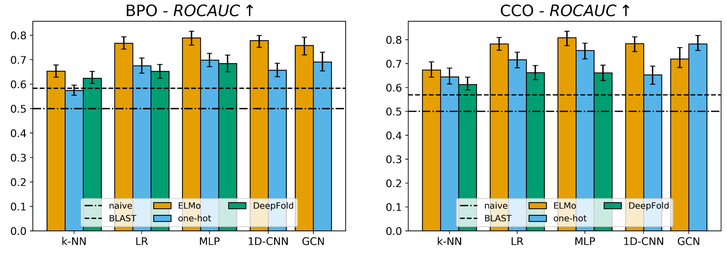
2. COMBINE SEQUENCE AND STRUCTURAL DATA



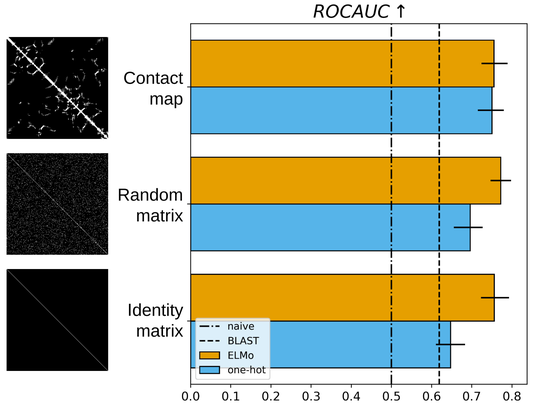
3. PDB RESULTS

- **PDB dataset (structures)**
- ~8k training, ~1k validation, ~400 test (max 30% sequence identity)
- ~250 MFO and CCO, ~1k BPO terms



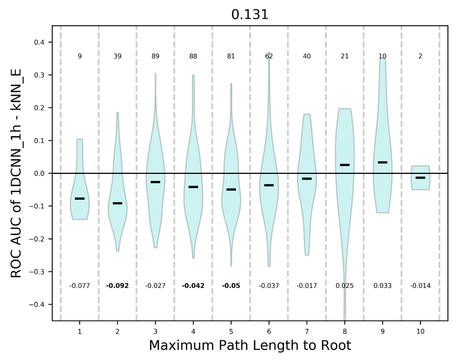


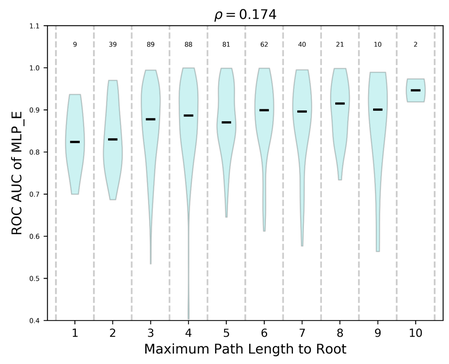
- Changing the GCN adjacency matrix



4. MFO TERMS SPECIFICITY

- **SwissProt dataset (sequences)**
- ~64k training, ~8k validation and ~3.5k test (max 30% sequence identity) / 441 MFO terms





5. CAFA3 RESULTS

Ontology	Method	Fmax (rank)
MFO	k-NN + ELMo	0.50
	LR + ELMo	0.51
	MLP + ELMo	**0.55 (5/146)**
	1D-CNN + ELMo	0.53
	DeepGOCNN + one-hot	0.43
BPO	k-NN + ELMo	0.34 (44/146)
CCO	k-NN + ELMo	0.60 (10/146)

6. CONCLUSIONS

- Pre-trained features outperform hand-crafted representations
- ELMo contain functional information → MFO, BPO, CCO
- Adding structure information to ELMo does not improve performance → It helps with simpler representations as one-hot encodings
- Supervised training is needed for more specific terms

References

[1] Heinzinger (2019). Modeling aspects of the language of life through transfer-learning protein sequences
[2] Liu (2018). Learning structural motif representations for efficient protein structure search
[3] Gligorijevic (2019). Structure-Based Function Prediction using Graph Convolutional Networks

Funding

- Keygene N.V., The Netherlands
- Spanish MINECO/FEDER Project TEC2016-80141-P
- FPI grant BES-2017-079792

Contact: ameliavm@ugr.es

Links: https://github.com/stamakro/GCN-for-Structure-and-Function
https://www.biorxiv.org/content/10.1101/2020.04.07.028373v1