TU Delft

UNIVERSIDAD DE GRANADA

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

Amelia Villegas-Morcillo

ISMB 2020

Automatic function prediction

Amino acid sequence → 3D structure → Protein function

MQIFVKTLTGKTITLE
VEPSDTIENVKAKIQ
DKEGIPPDQQRLIFA
GKQLEDGRTLSDYN
IQKESTLHLV…

Ubiquitin

UniProtKB
SwissProt

PDB
SCOPe / CATH

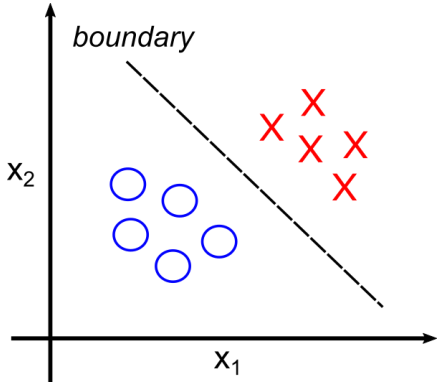
Gene Ontology

- **Molecular Function**
- Biological Process
- Cellular Component

Automatic function prediction

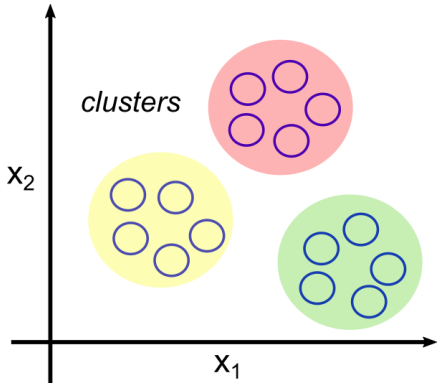
Supervised training

- Value / Class guided
- Dependent on the task
- Limited amount of labeled proteins (thousands)



Unsupervised pre-training

- Self-guided
- Independent of the task
- Millions of (unlabeled) sequences
  - >175M UniProtKB
- Adapt to the task → Fine-tuning / Transfer-learning



Language models on proteins

UNIVERSIDAD DE GRANADA TUDelft

UDSMProt: universal deep sequence models for protein classificat

Nils Strodthoff ✉, Patrick Wag

Bioinformatics, Volume 36, Issu

https://doi.org/10.1093/bioinf

Published: 08 January 2020

Article | Published: 21 October 2019

Unified rational protein engineering with sequence-based deep representation

hammed AlQuraishi & George M.

ticle

Heinzinger *et al. BMC Bioinformatics* (2019) 20:723
https://doi.org/10.1186/s12859-019-3220-8

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Modeling aspects of the language of life through transfer-learning protein sequences

Michael Heinzinger[1,2*†], Ahmed Elnaggar[1,2†], Yu Wang[3], Christian Dallago[1,2], Dmitrii Nechaev[1,2], Florian Matthes[4] and Burkhard Rost[1,5,6,7]

Check for updates

dor ure

and

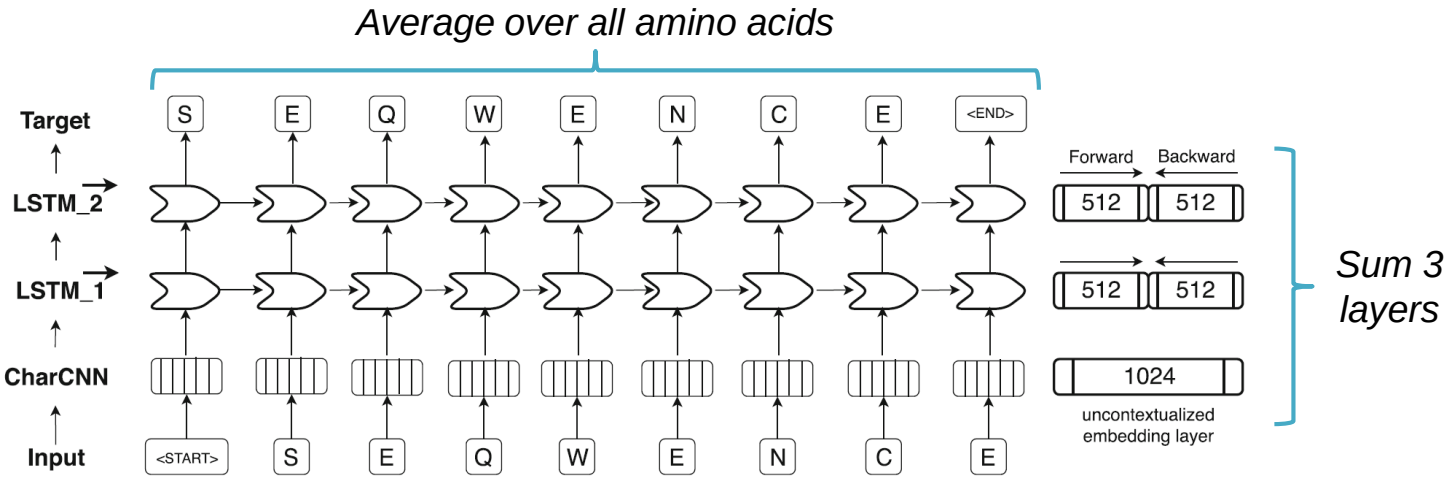
architectures

Damianos P. Melidis, Brandon Malone, Wolfgang Nejdl

doi: https://doi.org/10.1101/2020.03.17.995498

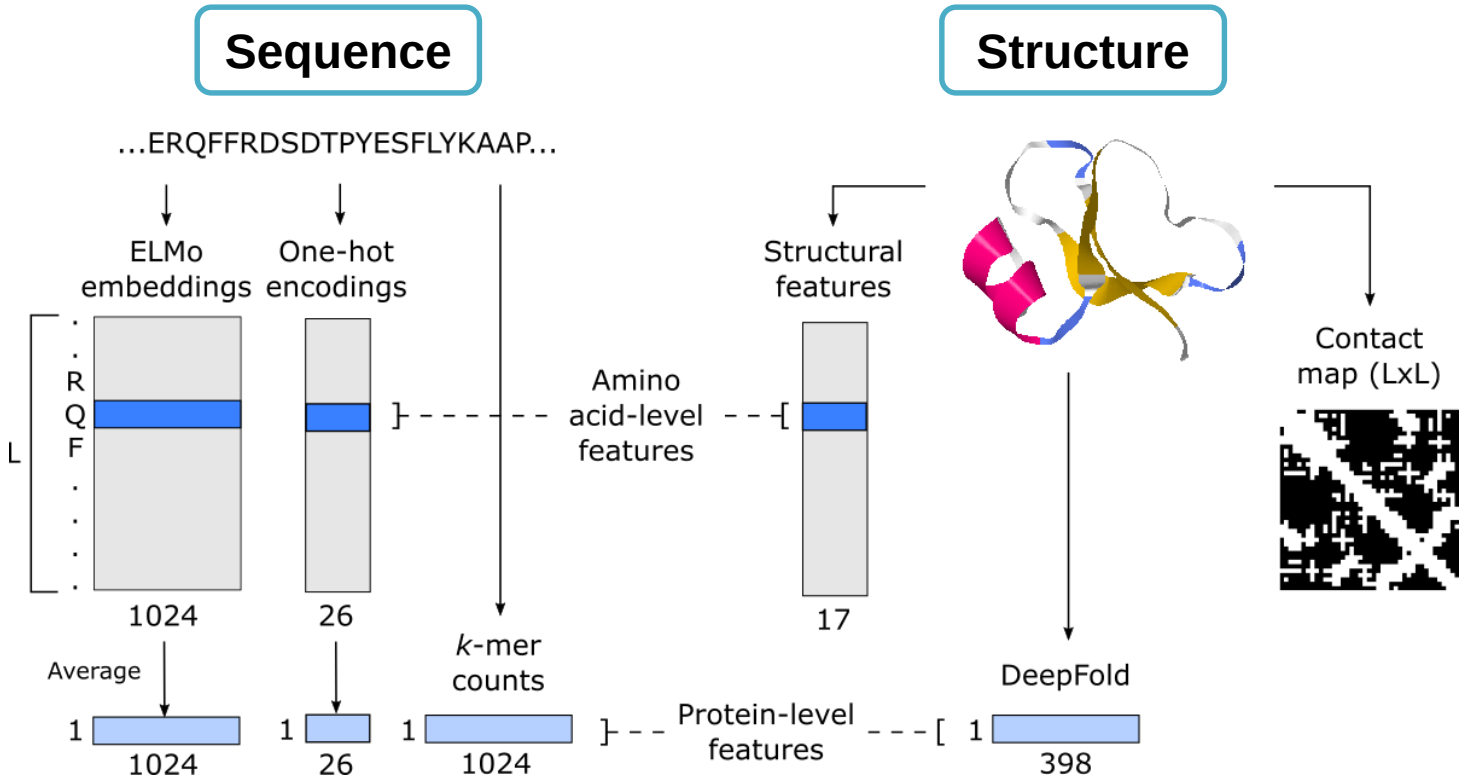
ELMo embeddings

- ELMo model from NLP
  - Predict the next amino acid in a sequence given all previous ones
- SeqVec* model trained on UniRef50 (33M proteins)



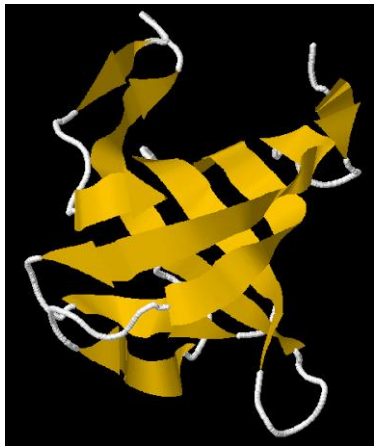
* Heinzinger et al. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*.

How to represent the protein?



Structural representations

3D structure

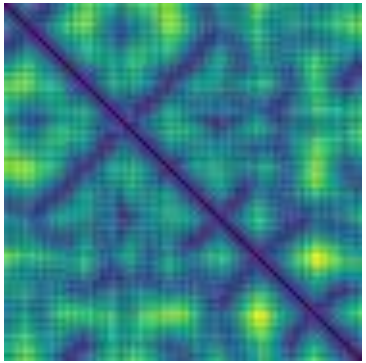


3D atom coordinates

Secondary structure and backbone angles

Euclidean distance Cb-Cb

Distance map

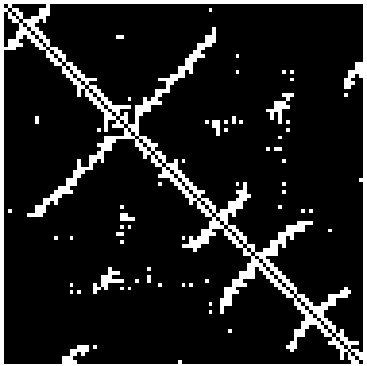


Real-valued matrix (LxL)

DeepFold*

<10Å threshold

Contact map



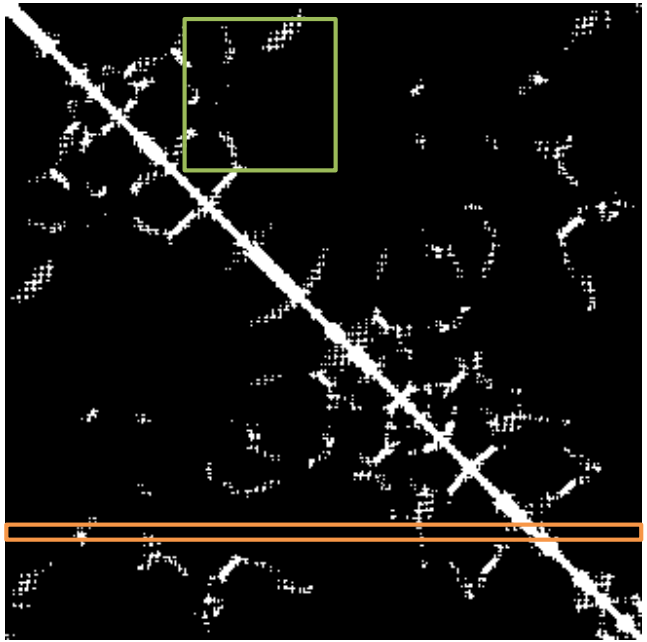
Binary matrix (LxL)

2D-CNN / GraphCN**

* Liu et al. (2018). Learning structural motif representations for efficient protein structure search. *Bioinformatics*.

** Gligorijevic et al. (2019). Structure-Based Function Prediction using Graph Convolutional Networks. *bioRxiv*.

Contact map processing



Image

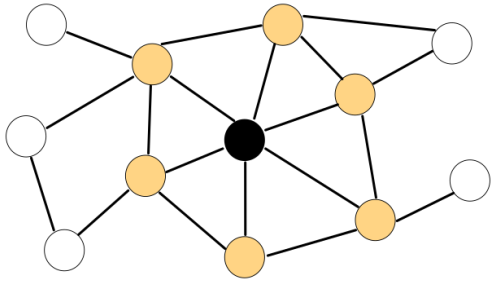
1x1	1x0	1x1	0	0
0x0	1x1	1x0	1	0
0x1	0x0	1x1	1	1
0	0	1	1	0
0	1	1	0	0

2D-CNN

4		

Graph

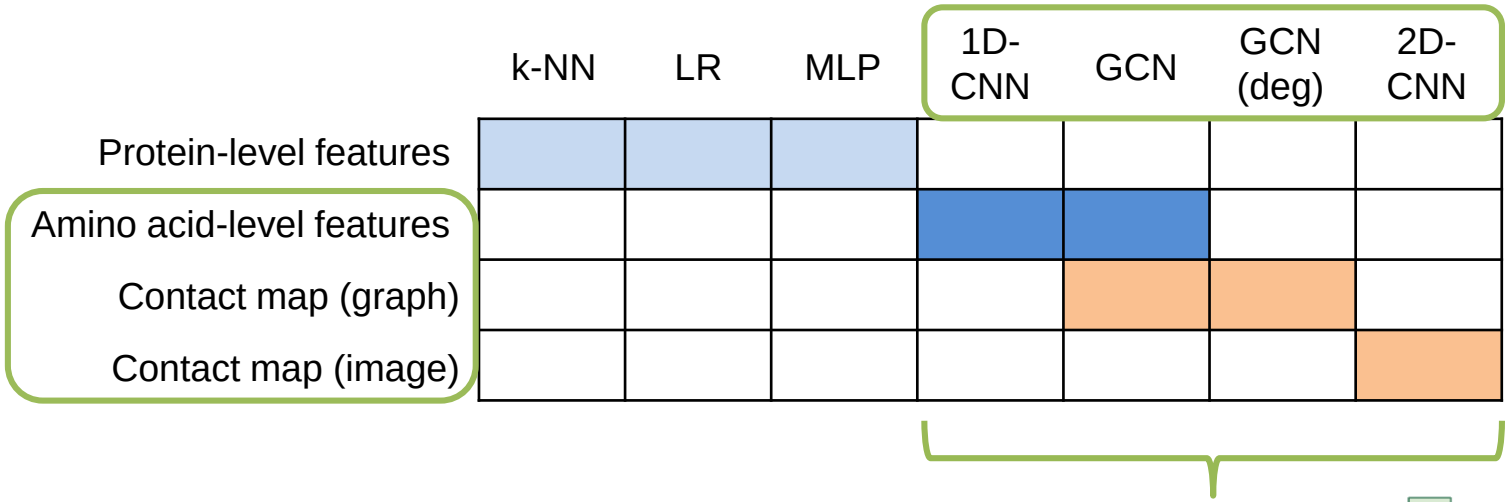
GraphCN*



$$Z = \mathrm{ReLU}((I+A)XW)$$

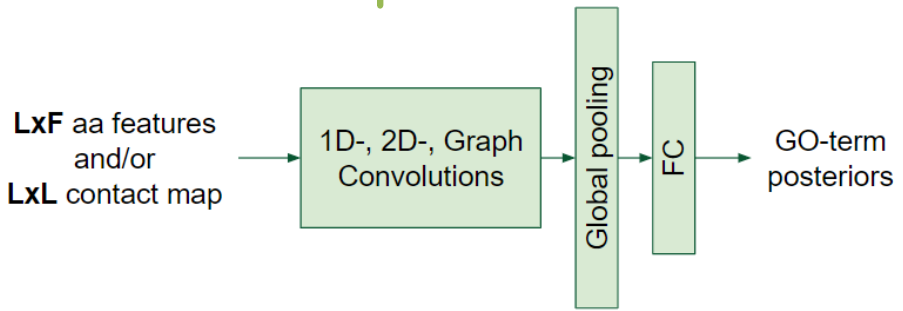
* Kipf and Welling (2019). Semi-supervised classification with graph convolutional networks. *ICLR 2017*.

Function prediction methods



Network architectures:

- MLP → 1 hidden layer
- 1D-CNN / 2D-CNN → 2 conv layers
- GCN → 1 graph conv layer



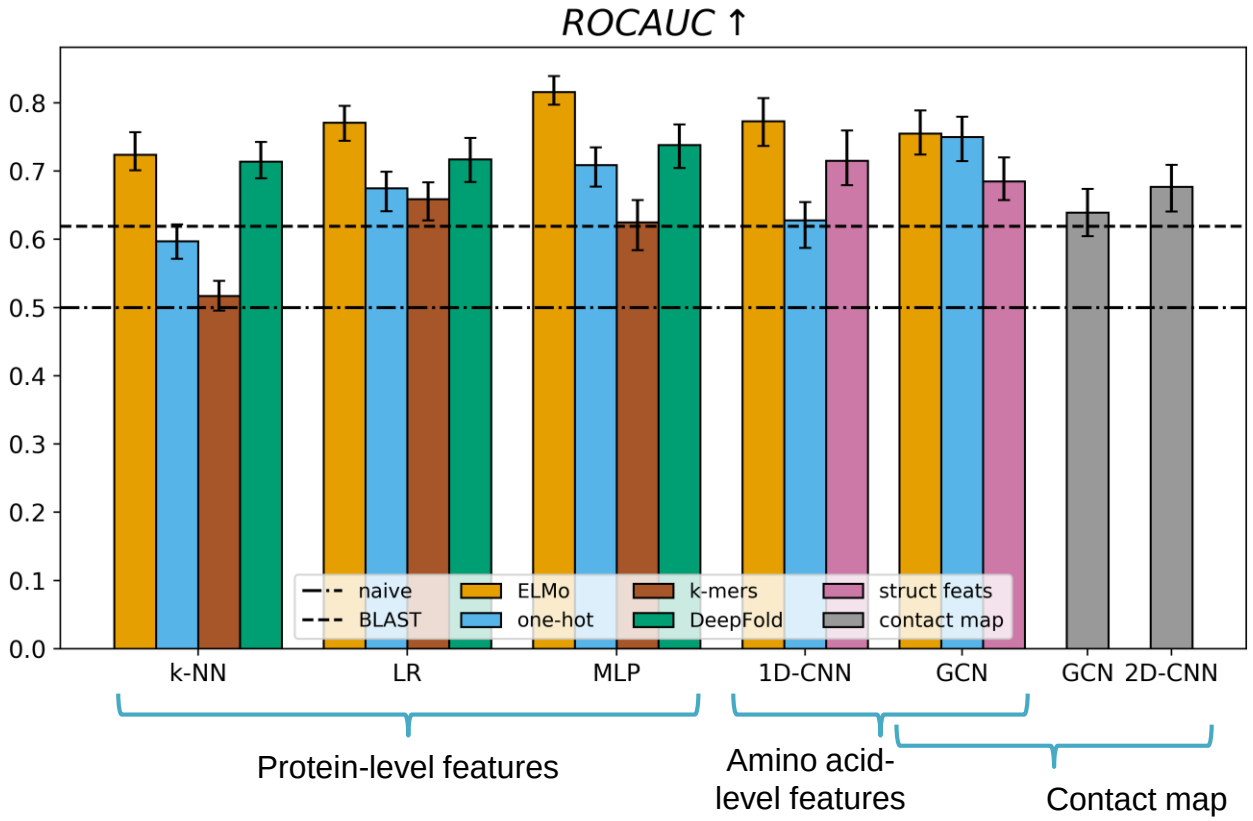
PDB dataset

- Protein **structures** annotated with non-computational codes
  - MFO, BPO, CCO

- Train / validation / test split
  - Training → ~9k, ~8k, ~7k proteins, respectively
  - Validation → ~1k proteins
  - Test (max 30% identity) → ~400 proteins

- GO terms
  - ~250 for MFO and CCO
  - ~1k for BPO

MFO prediction results

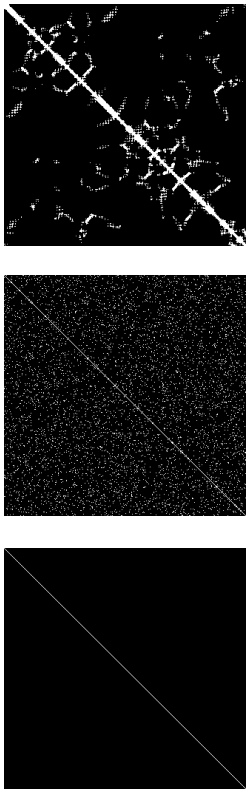


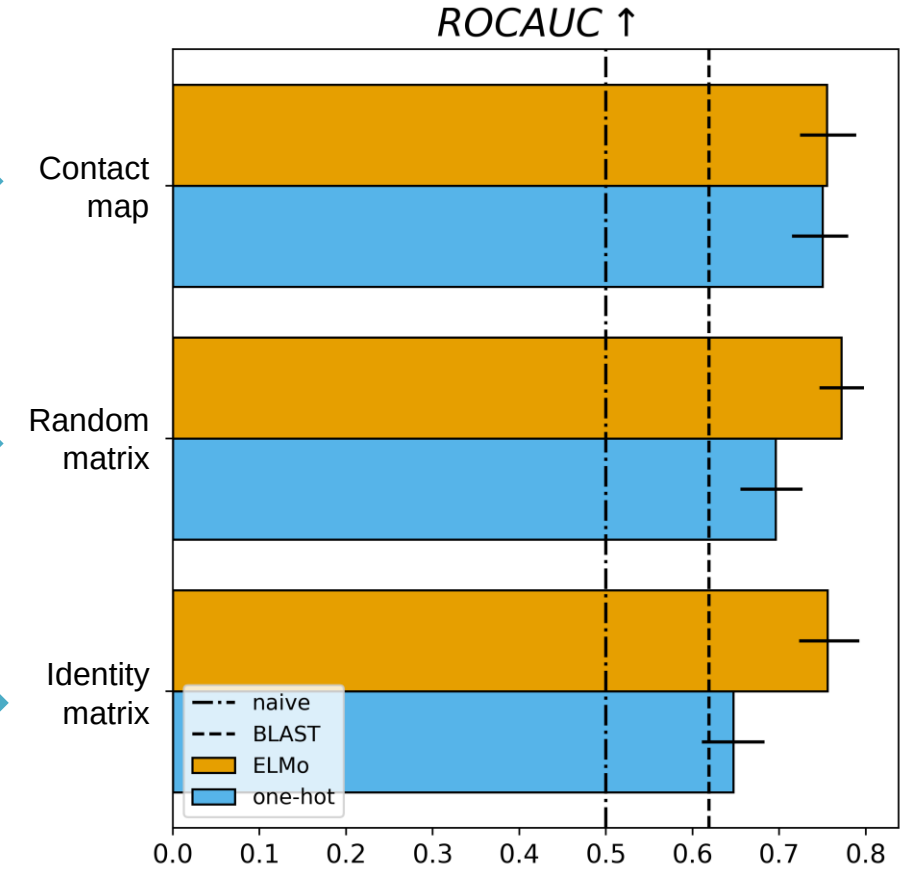
Observations

- Pre-trained features (ELMo and DeepFold) outperform other representations
- Contact map helps when using one-hot features (GCN)

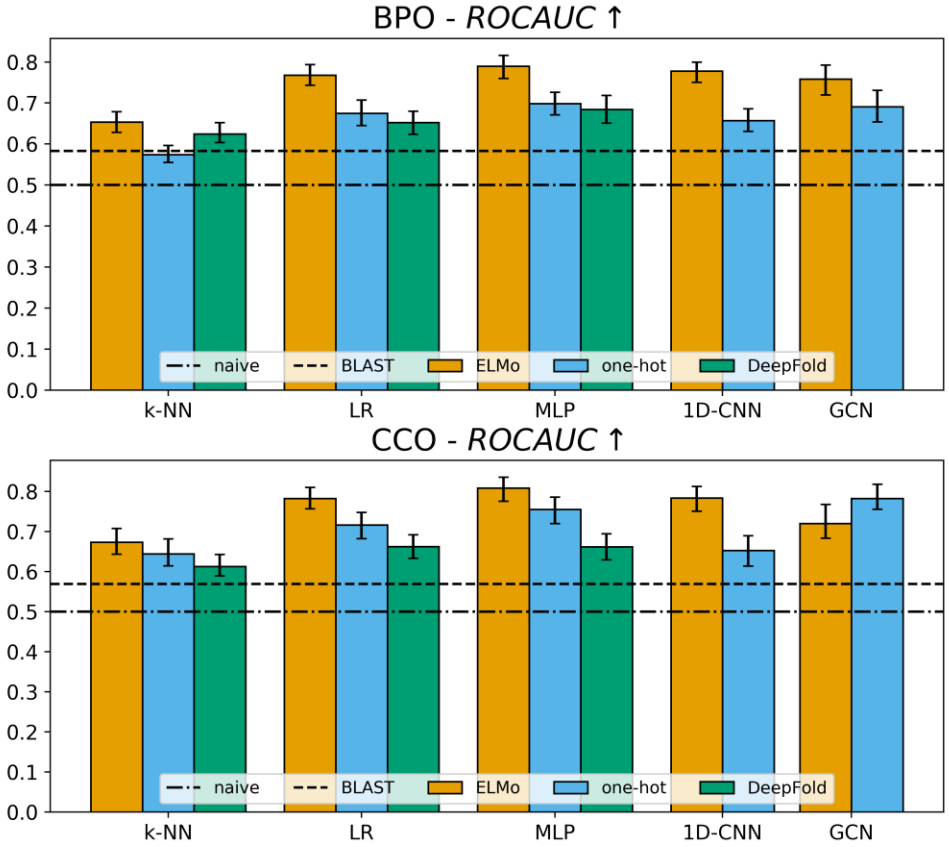
Changing the graph

Adjacency matrices (GCN)





BPO / CCO results



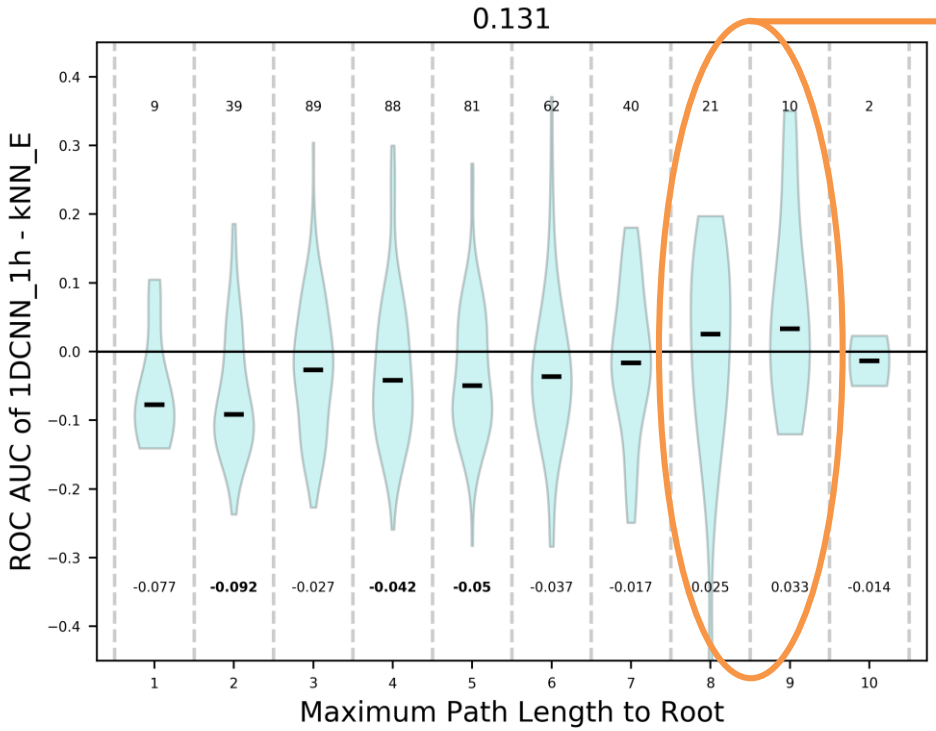
No clear superiority of DeepFold features

MFO terms specificity

- Larger dataset with sequences from SwissProt:
  - ~64k training, ~8k validation and ~3.5k test proteins
  - 441 MFO terms
- ELMo embeddings vs one-hot encodings
- Sequence-based models (no structure):
  - Protein-level → k-NN, LR, MLP
  - Amino acid-level → 1D-CNN
- Top method → MLP with ELMo (ROC AUC=0.87)

MFO terms specificity

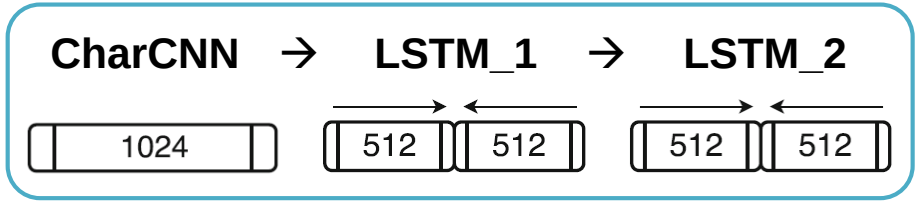
- k-NN with ELMo > 1D-CNN with one-hot (in ROC AUC)



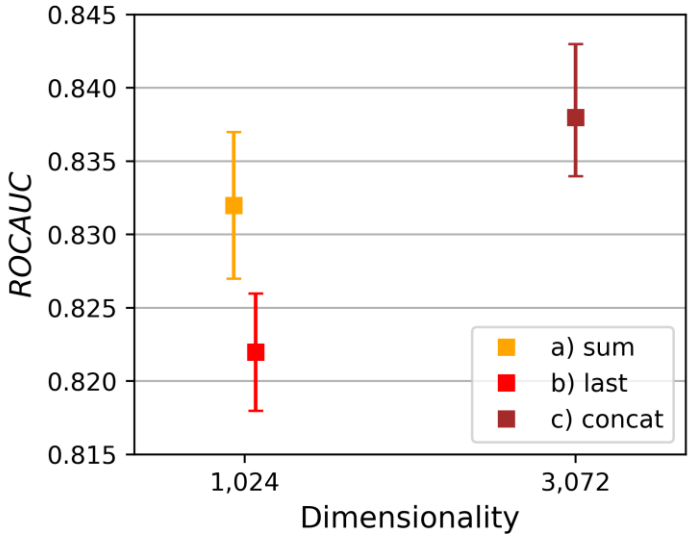
Further supervised training (1D-CNN) helps for more specific terms

ELMo layers performance

- SeqVec model layers
  (averaged over all amino acids, protein-level)



- Logistic Regression results
  a) sum → CharCNN + LSTM_1 + LSTM_2
  b) last layer → LSTM_2
  c) concat → [CharCNN, LSTM_1, LSTM_2]



CAFA3 results

Method	Features	Fmax
Naive*	----	0.33
BLAST*	----	0.42
k-NN	**Protein-level ELMo embeddings**	0.50
LR		0.51
MLP		**0.55**
1D-CNN	Amino acid-level ELMo embeddings	0.53
DeepGOCNN	Amino acid-level one-hot encodings	0.43
CAFA3 rank 1*	----	0.62
CAFA3 rank 2*	----	0.61
CAFA3 rank 3*	----	0.61
CAFA3 rank 4*	----	0.61
CAFA3 rank 5*	----	0.54

- MFO
  - Rank 5th out of 146 methods
- k-NN using ELMo:
  - BPO: Fmax=0.34 (44/146)
  - CCO: Fmax=0.60 (10/146)

* Zhou et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology.*

Summary

- Pre-trained features outperform hand-crafted representations
- ELMo contain functional information
  - Not only for MFO, but also for BPO and CCO
- Adding structure information to ELMo does not improve performance
  - It helps with simpler representations as one-hot encodings
- Supervised training helps for more specific terms
- Systematic comparison with other protein sequence embedders

Acknowledgements

- **Stavros Makrodimitris**
- Marcel Reinders
- Roeland van Ham
- Victoria Sánchez
- Ángel Gómez
- Elvin Isufi
- Chirag Raman
- Irene van den Bent

Keygene N.V., The Netherlands

Spanish MINECO/FEDER Project TEC2016-80141-P and FPI grant BES-2017-079792