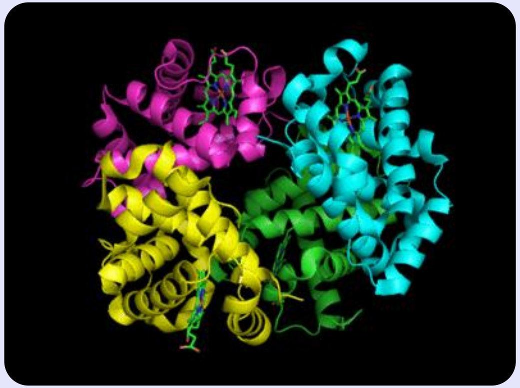
BMEG 3102 Bioinformatics

Protein Function Prediction

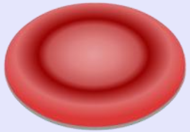
Cheung Ho Lun 1155174348
Chan Cheuk Ka 1155174356

Protein Function Prediction

Background



Haemoglobin



- Red blood cell
- Transport of oxygen

Structural features of protein

- Understanding life process
- Drug development
- Personalised medicine

https://microbenotes.com/hemoglobin/

Problem

Incomplete annotation (< 1%)

- Sequence similarity
- Subtle differences
- Lack of features
- Multi-functional proteins

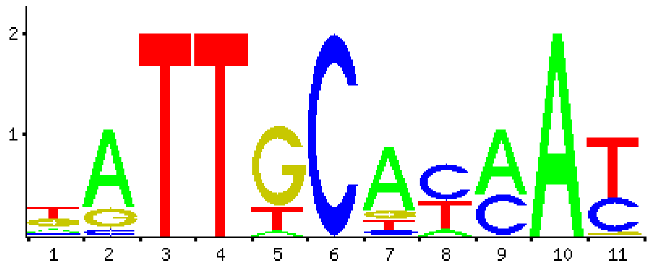
- Accuracy limitations
- Difficult to predict rare functions

Fujita, S., & Terada, T, *Computational and Structural Biotechnology Journal*, 2024

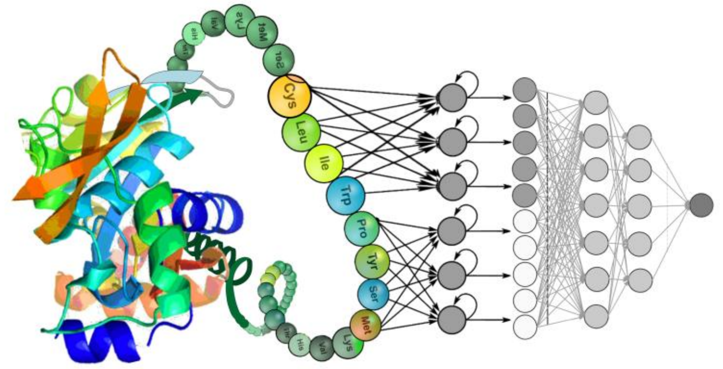
Jeffery, C. J. , *Frontiers in Bioinformatics*, 2023

Bioinformatics

- Motif-based methods
- Deep learning frameworks
- Protein language models
- Gene Ontology (GO)



1 Transcription factor motifs. *Nature,* 2019



Ingrid Fadelli , Phys.org, 2022

DEEPred [a]

Protein sequence

Amino acid sequence

In → ? → ? → Out

GO terms

with confidence values

[a] A. Sureyya Rifaioglu et al., Scientific Reports, 2019

DEEPred [a]

Protein sequence

Amino acid sequence

In → F → ? → Out

Feature extraction

Preprocess feature vectors

GO terms

with confidence values

DEEPred [a]

[Suppl. Info]

F

Feature extraction

Preprocess feature vectors

DEEPred [a]

[Suppl. Info]

F

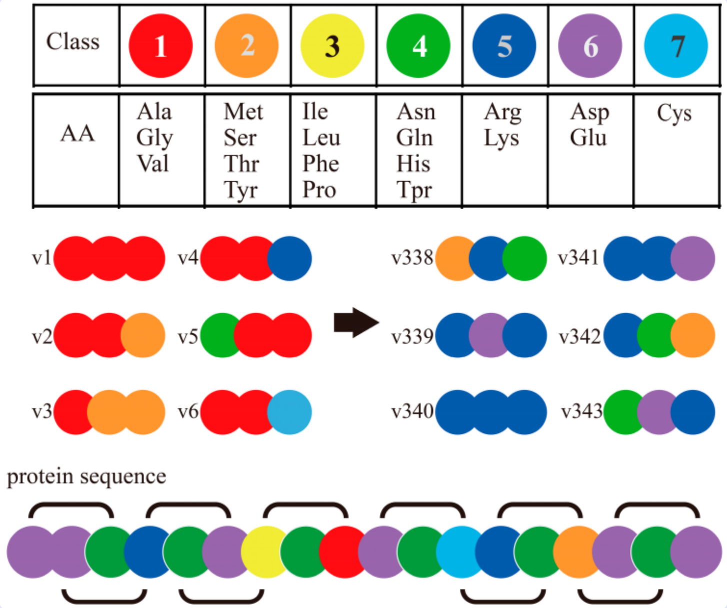
Feature extraction

Preprocess feature vectors

1 **Assign class**

2 **Record triplet frequency**

Conjoint Triad



J.-W. Chang et al., *International Journal of Molecular Sciences*, 2016

DEEPred[a]

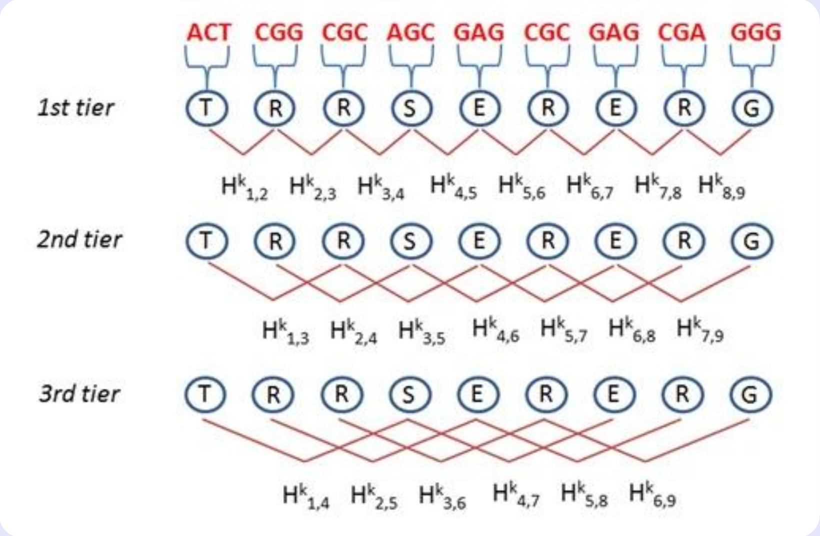
[Suppl. Info]

F

Feature extraction

Preprocess feature vectors

Pseudo-Amino Acid Composition (PACC)



I. Limongelli, S. Marini et al., BMC Bioinformatics, 2015

DEEPred [a]

[Suppl. Info]

F

Feature extraction

Preprocess feature vectors

Subsequence profile map (SPMap)

… M K L R F T A I S H G W Q N E V P T Y A L …

↓ **Subsequences**

M K L R F T
F T A I S H
Q N E V P
…

↓

Clustering information

DEEPred [a]

F

Feature extraction

Preprocess feature vectors

Model & GO level	GO term id	GO description	# of annotated proteins	Predictive performance (F1-score)		
				SPMap	Pseudo-amino acid composition	Conjoint triad
Model 1 (GO level: 2)	GO:0036094	small molecule binding	1 847	0.49	0.29	0.23
	GO:0003700	DNA binding transcription factor activity	1 652			
	GO:0004872	receptor activity	1 332			
	GO:0044877	protein-containing complex binding	1 296			
	GO:0097367	carbohydrate derivative binding	1 252			
Model 2 (GO level: 4)	GO:0004529	exodeoxyribonuclease activity	50	0.68	0.53	0.38
	GO:0045309	protein phosphorylated amino acid binding	50			
	GO:0008395	steroid hydroxylase activity	49			
	GO:0008649	rRNA methyltransferase activity	49			
	GO:0015645	fatty acid ligase activity	49			
Model 3 (GO level: 7)	GO:0001012	RNA polymerase II regulatory region DNA binding	818	0.74	0.53	0.47
	GO:0016887	ATPase activity	764			
	GO:0046873	metal ion transmembrane transporter activity	685			
	GO:0001159	core promoter proximal region DNA binding	504			
	GO:0015077	monovalent inorganic cation transmembrane transporter activity	480			

DEEPred [a]

Protein sequence

Amino acid sequence

In → F → ? → Out

Feature extraction

Preprocess feature vectors

GO terms

with confidence values

DEEPred[a]

Protein sequence

Amino acid sequence

Deep Neural Network

Multi-task feed-forward DNN stack

In → F → DNN → Out

Feature extraction

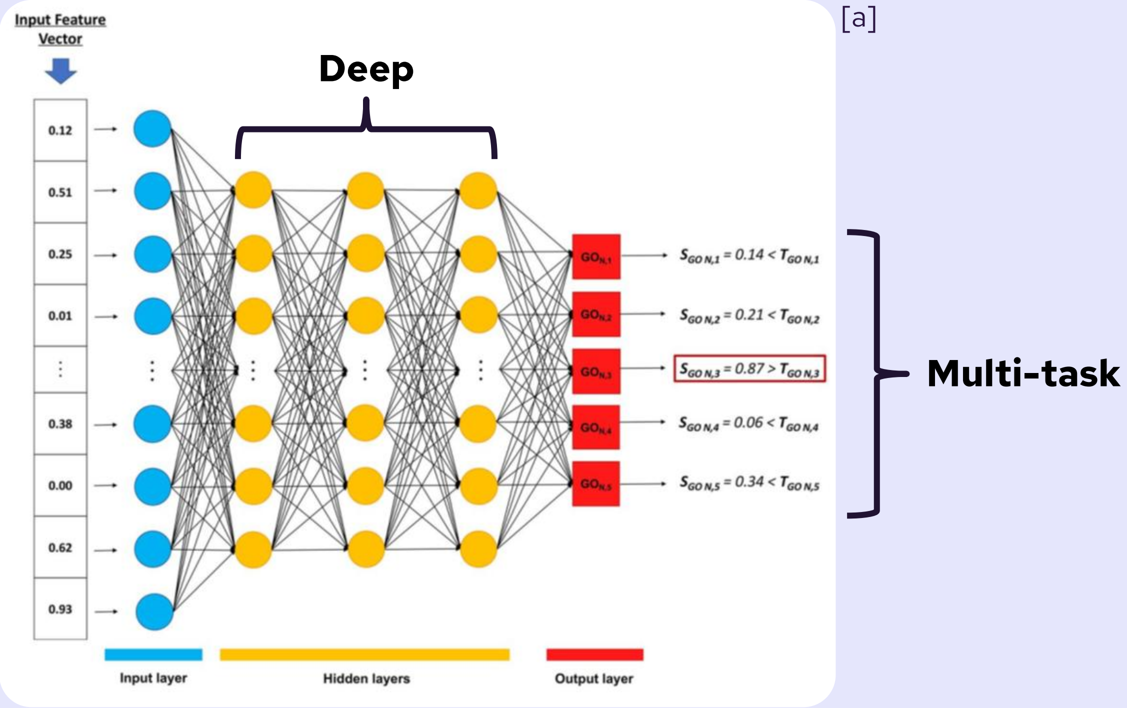
Preprocess feature vectors

GO terms

with confidence values

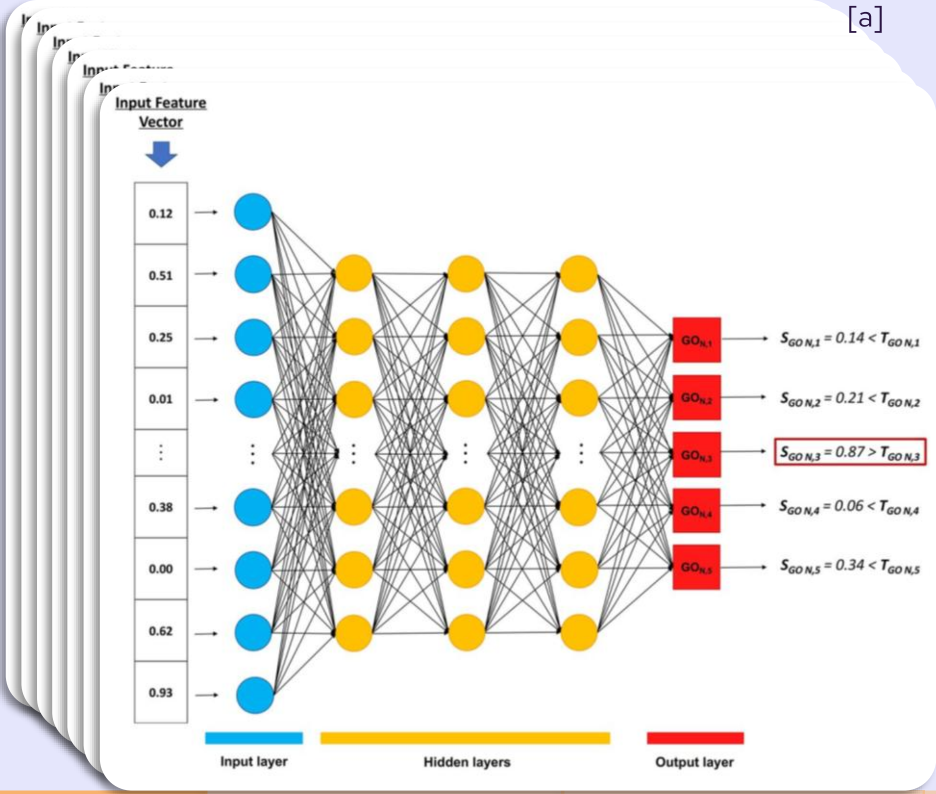
DEEPred architecture

[a]



Feed-forward

DEEPred architecture



[a]

×1101

DEEPred architecture

Different broadness

GO term 1: **Broad (40%)**
GO term 2: **Common (10%)**
GO term 3: **Narrow (5%)**
GO term 4: **Narrow (2%)**
GO term 5: **Very Narrow (1%)**

DEEPred architecture

Different broadness

GO term 1: **Broad (40%)** ← *Always choose this*
GO term 2: **Common (10%)**
GO term 3: **Narrow (5%)**
GO term 4: **Narrow (2%)**
GO term 5: **Very Narrow (1%)**

High accuracy
without learning

DEEPred architecture

Different broadness

GO term 1: **Broad (40%)** ← *Always choose this*
GO term 2: **Common (10%)**
GO term 3: **Narrow (5%)**
GO term 4: **Narrow (2%)**
GO term 5: **Very Narrow (1%)**

→ **High accuracy** without learning

Same broadness

GO term 1: **Common (8%)**
GO term 2: **Common (10%)**
GO term 3: **Common (9%)**
GO term 4: **Common (11%)**
GO term 5: **Common (7%)**

DEEPred architecture

Different broadness

GO term 1: **Broad (40%)** ← *Always choose this*
GO term 2: **Common (10%)**
GO term 3: **Narrow (5%)**
GO term 4: **Narrow (2%)**
GO term 5: **Very Narrow (1%)**

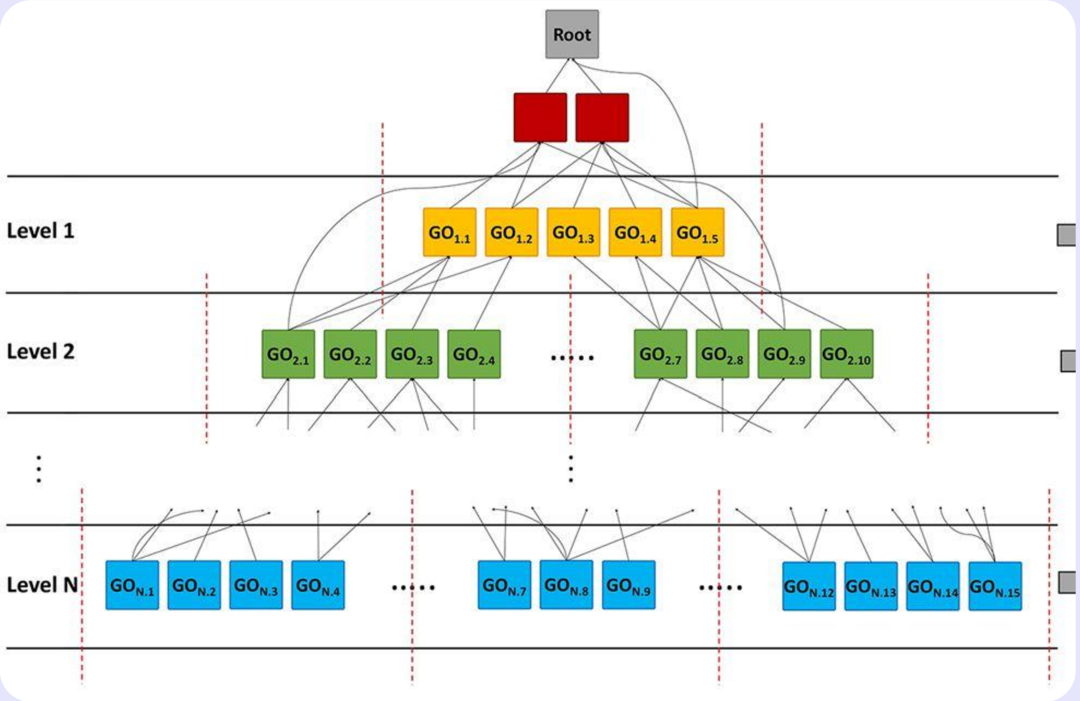
→ **High accuracy** without learning

Same broadness

GO term 1: **Common (8%)** ← *Always choose this*
GO term 2: **Common (10%)**
GO term 3: **Common (9%)**
GO term 4: **Common (11%)**
GO term 5: **Common (7%)**

→ **LOW accuracy** without learning

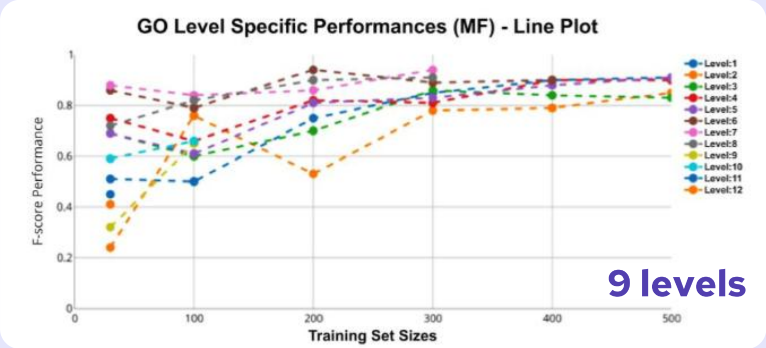
DEEPred architecture

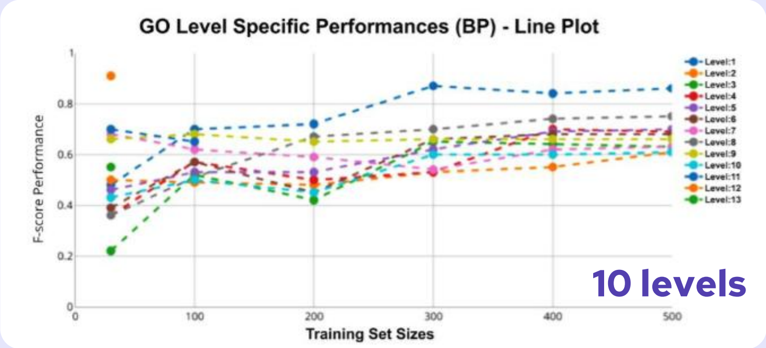


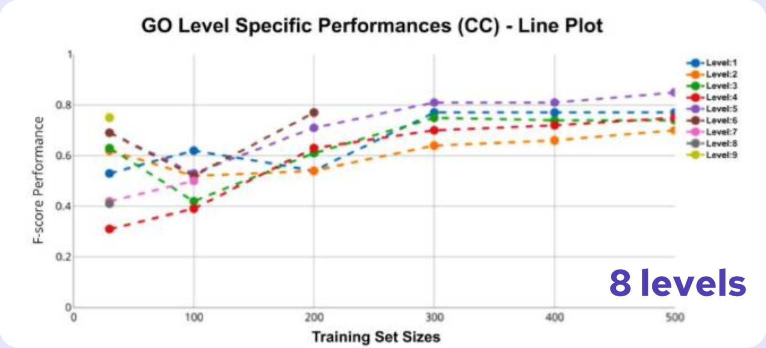
[a]

DEEPred architecture

[a] [Suppl. Info]

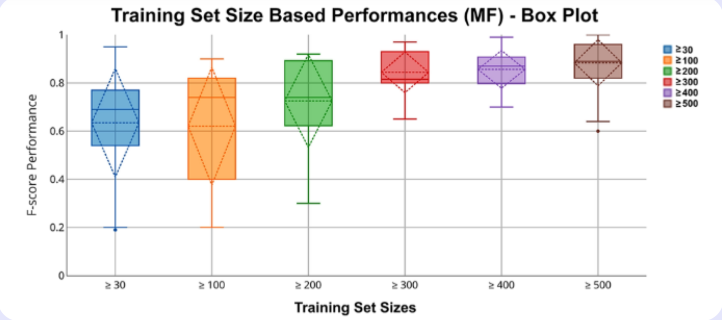




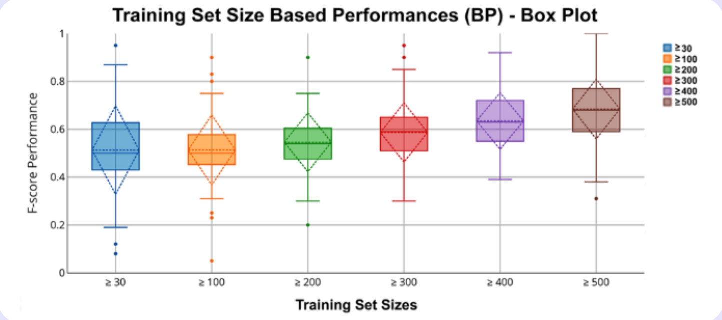


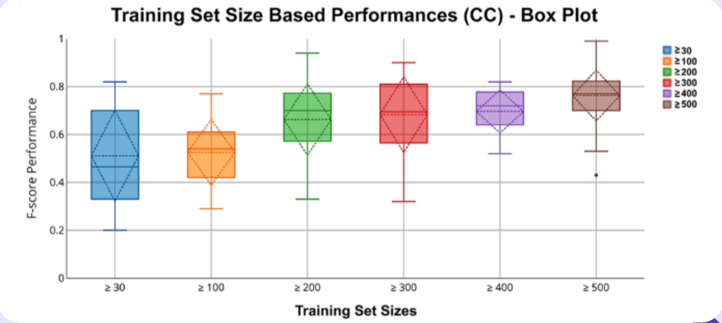
DEEPred results

GO categories	Performance measures (F1-score) for different training dataset sizes					
	≥ 30	≥ 100	≥ 200	≥ 300	≥ 400	≥ 500
Molecular Function	0.66	0.68	0.77	0.82	0.82	0.83
Biological Process	0.42	0.50	0.52	0.52	0.56	0.55
Cellular Component	0.50	0.59	0.64	0.63	0.64	0.65

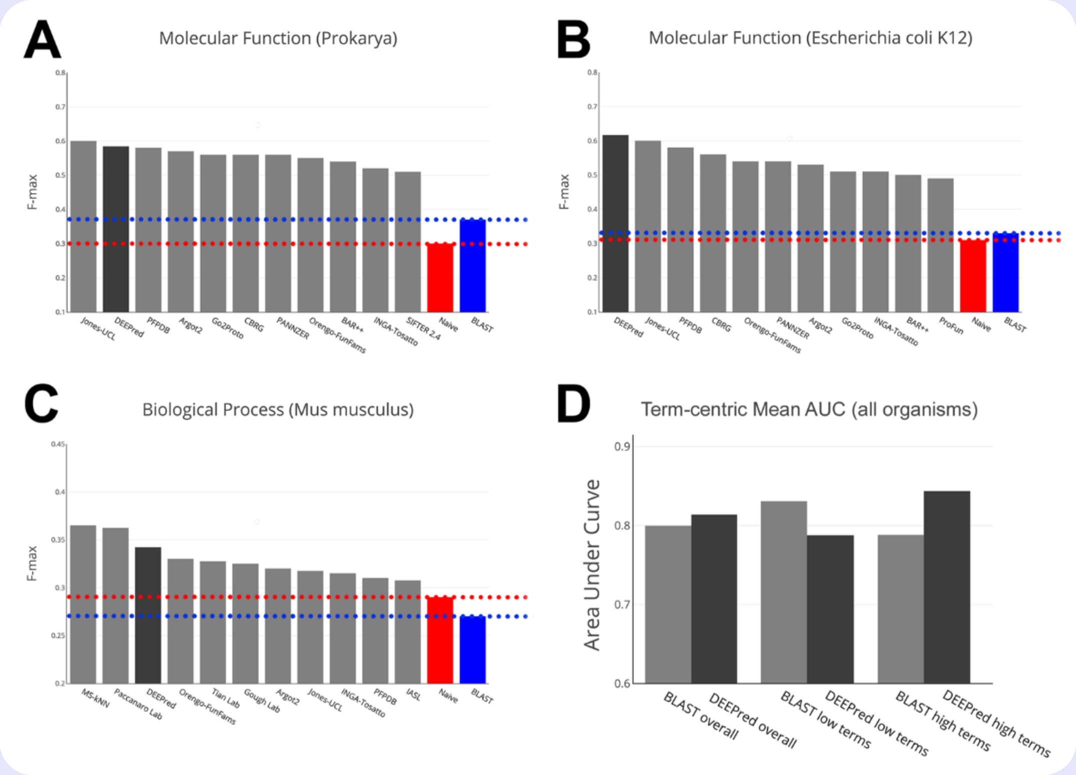


[a]





DEEPred results



[a]

why DEEPred?

Hyper-optimised [a] [Suppl. Info]

Tested with 100,000
different hyper-parameters

Noise-tolerant

Trained with noisy data
(Experimental & Electronic)

Scalable

Fast to train
(Parallelisable)

DeepGraphGO[b]

Protein sequence

Amino acid sequence

In → ? → ? → Out

GO terms

with confidence values

[b] R. You et al., Bioinformatics, 2021

DeepGraphGO[b]

Protein sequence

Amino acid sequence

In → F → ? → Out

Feature extraction

InterProScan feature vectors

GO terms

with confidence values

DeepGraphGO[b]

Protein sequence

Amino acid sequence

Graph Neural Network

with Graph Convolutional Layers

In → F → GNN → Out

Feature extraction

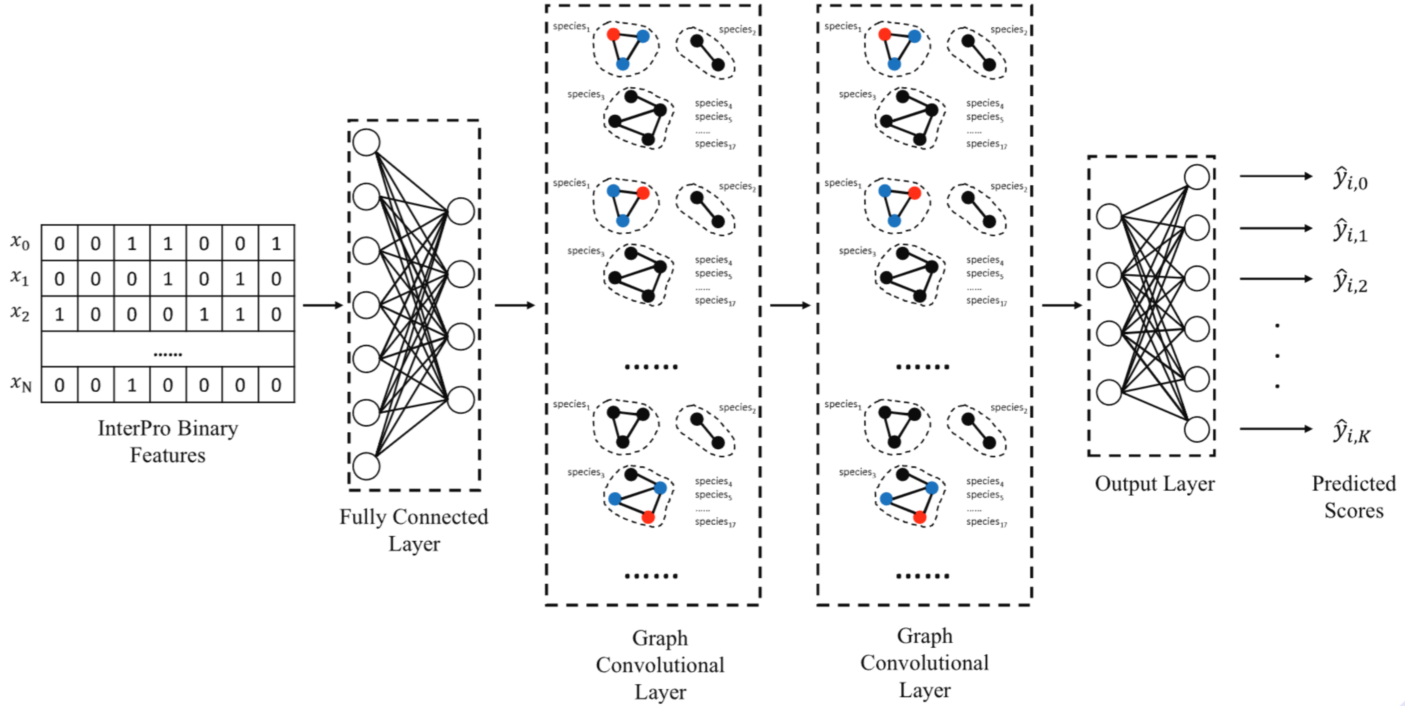
InterProScan feature vectors

GO terms

with confidence values

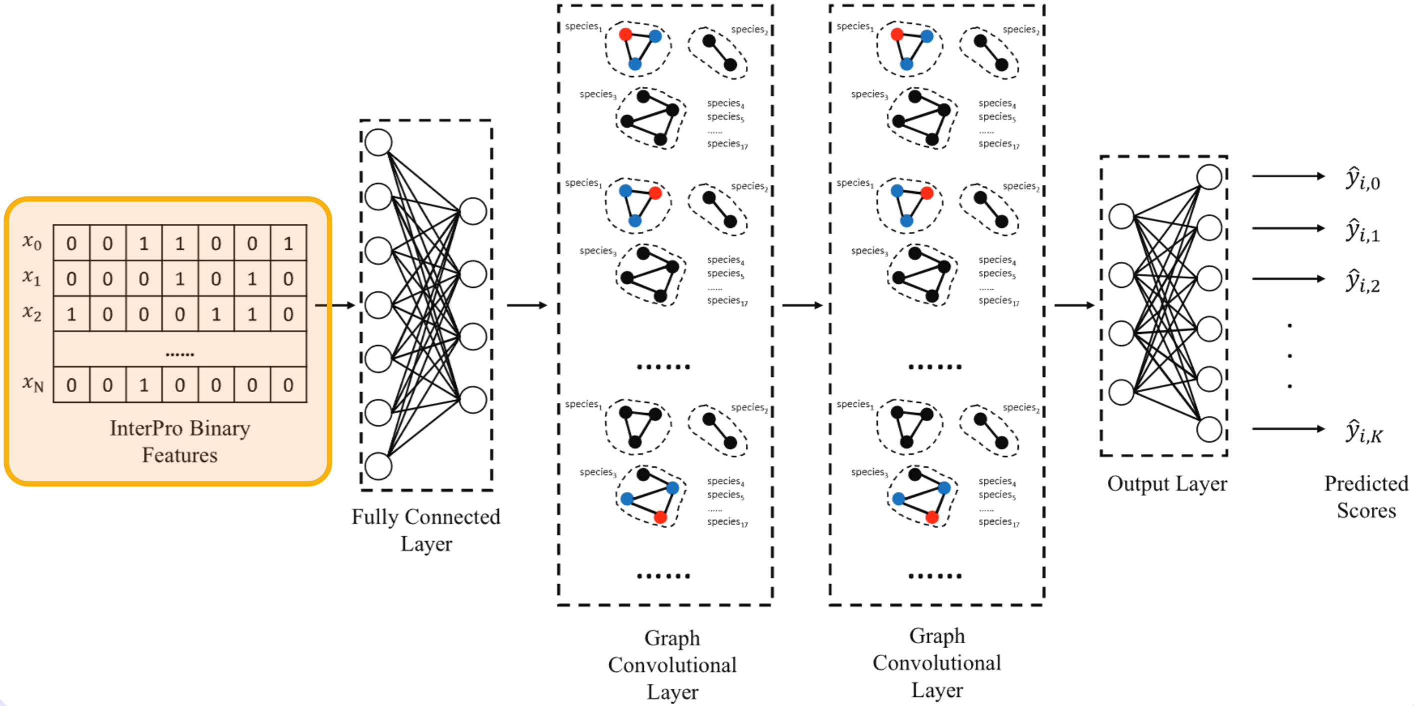
DeepGraphGO architecture

[b]

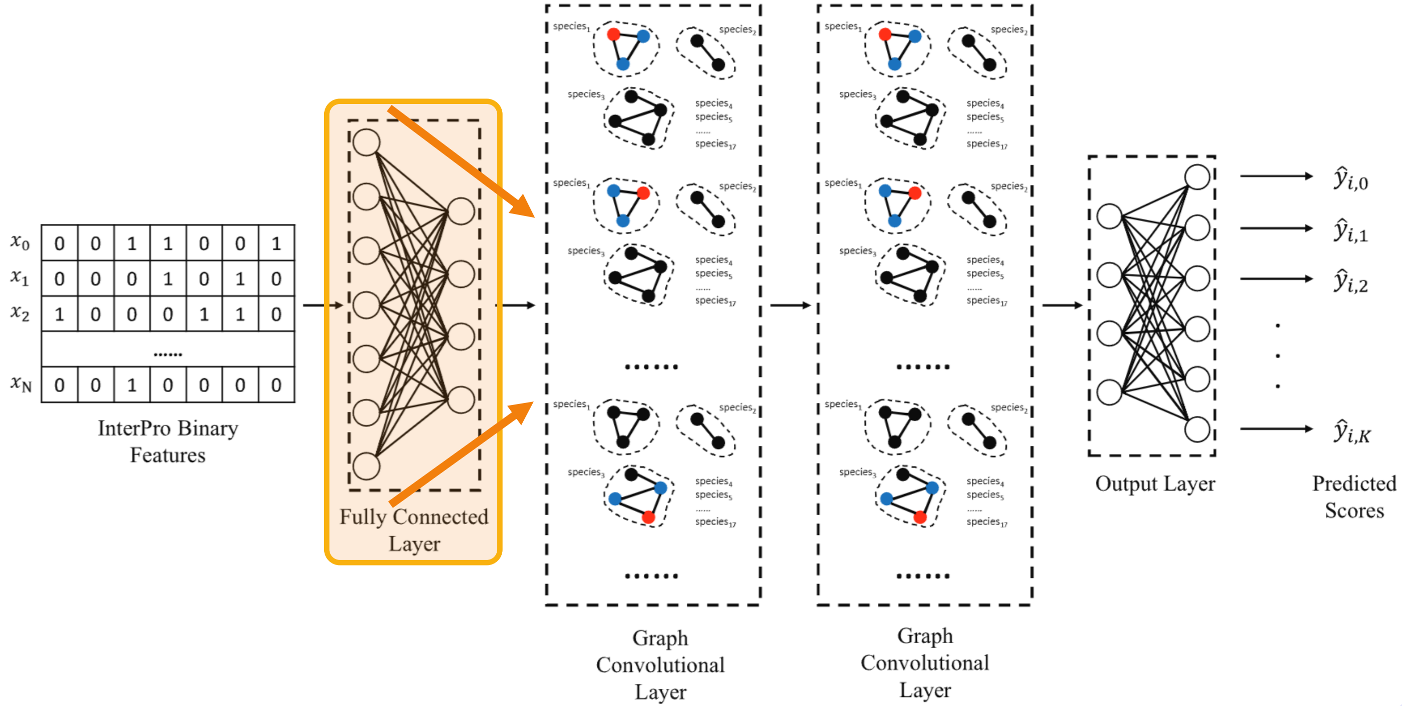


DeepGraphGO architecture

[b]



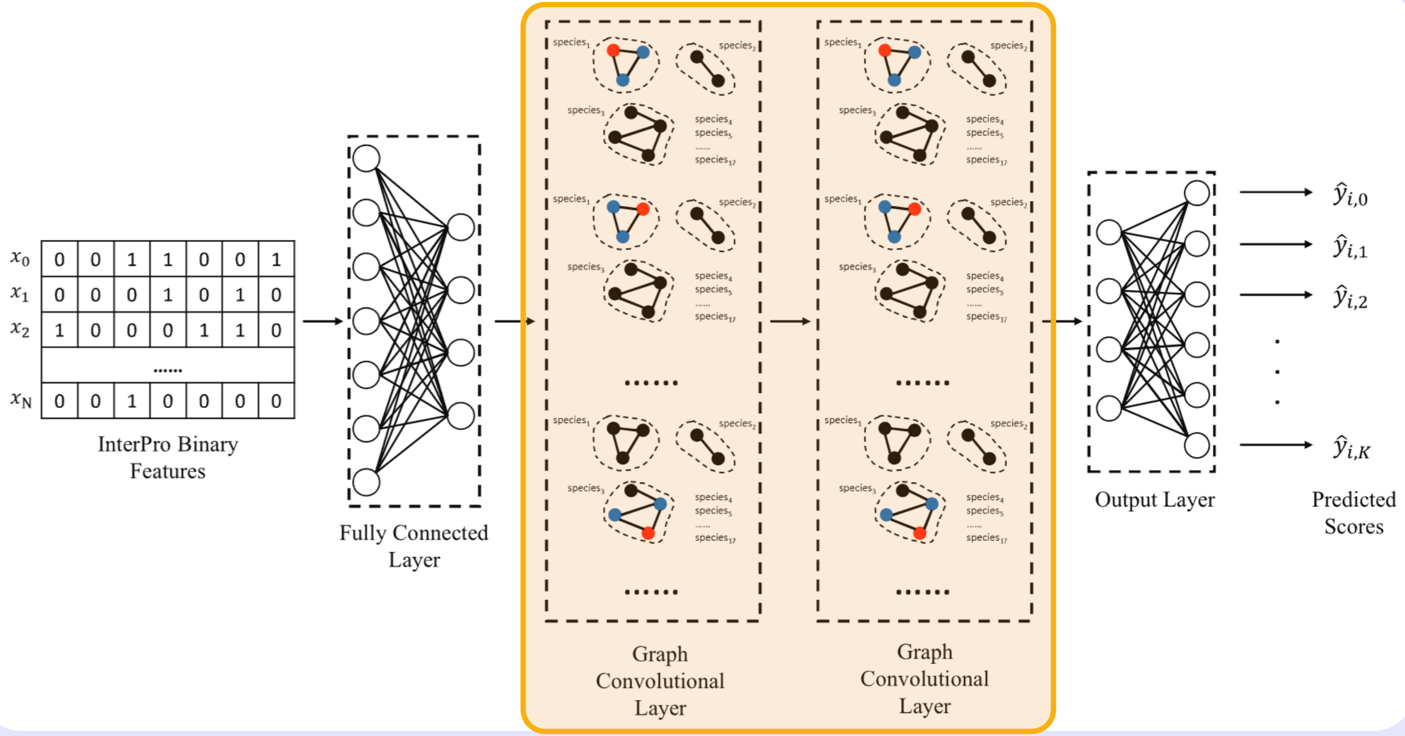
DeepGraphGO architecture



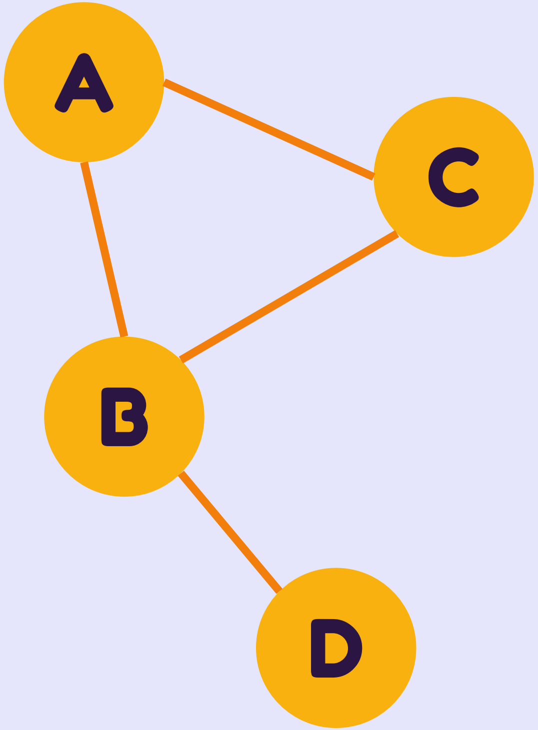
[b]

DeepGraphGO architecture

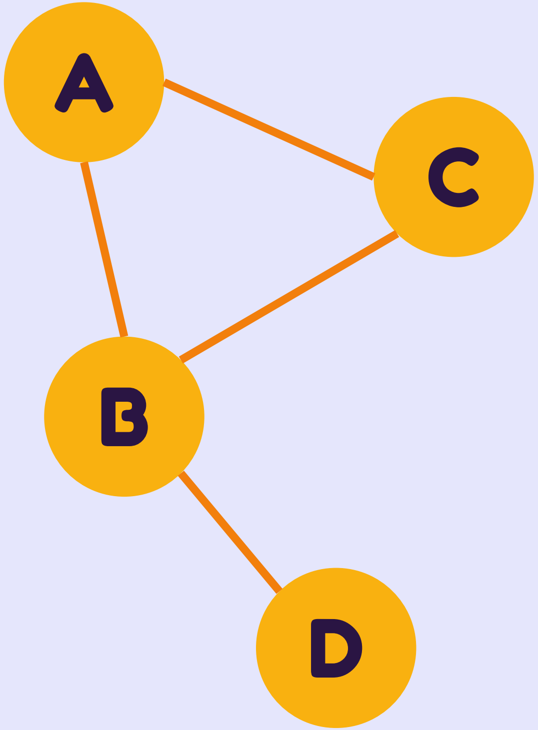
[b]



deepGRAPHgo [b]

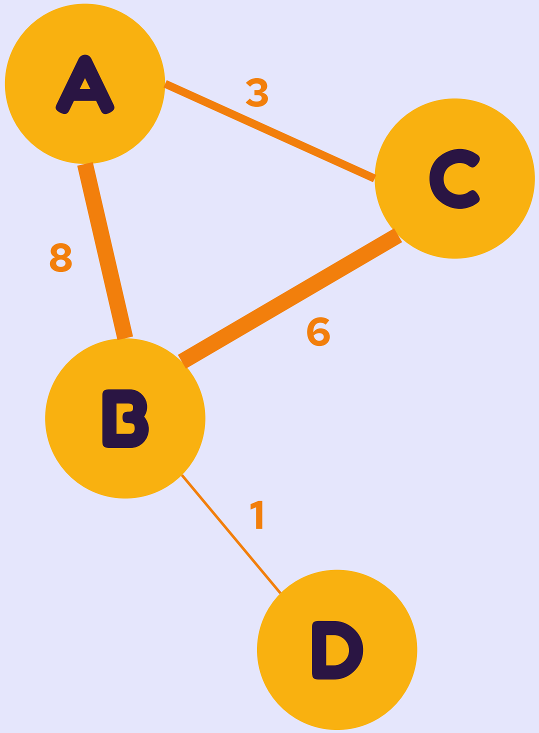


deepGRAPHgo[b]



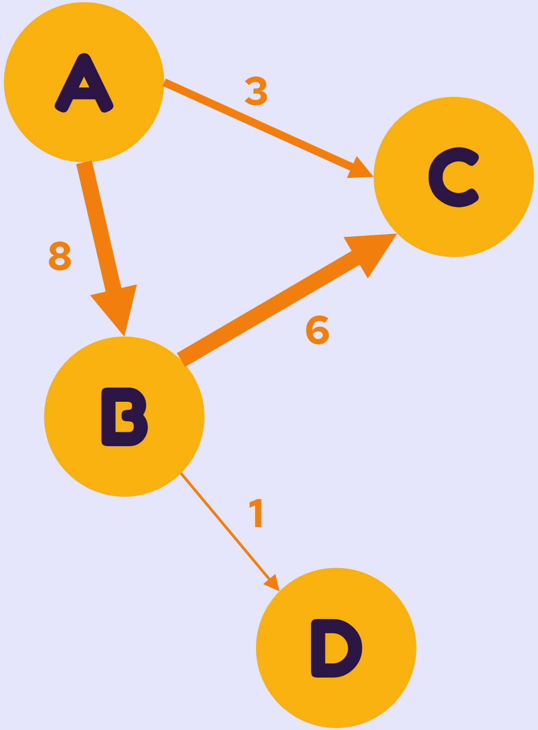
	A	B	C	D
A	–	1	1	0
B	1	–	1	1
C	1	1	–	0
D	0	1	0	–

deepGRAPHgo[b]



	A	B	C	D
A	-	8	3	0
B	8	-	6	1
C	3	6	-	0
D	0	1	0	-

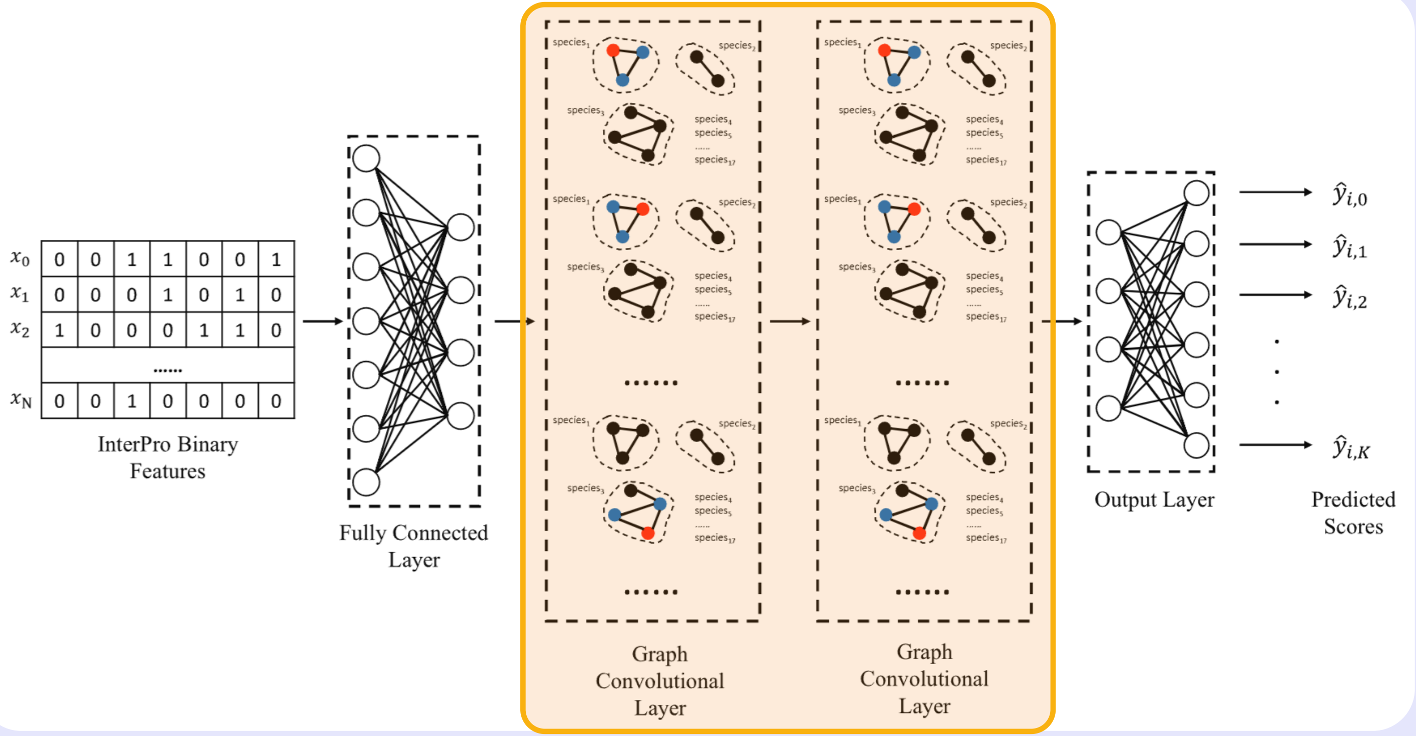
deepGRAPHgo[b]



	A	B	C	D
A	-	8	3	0
B	-8	-	6	1
C	-3	-6	-	0
D	0	-1	0	-

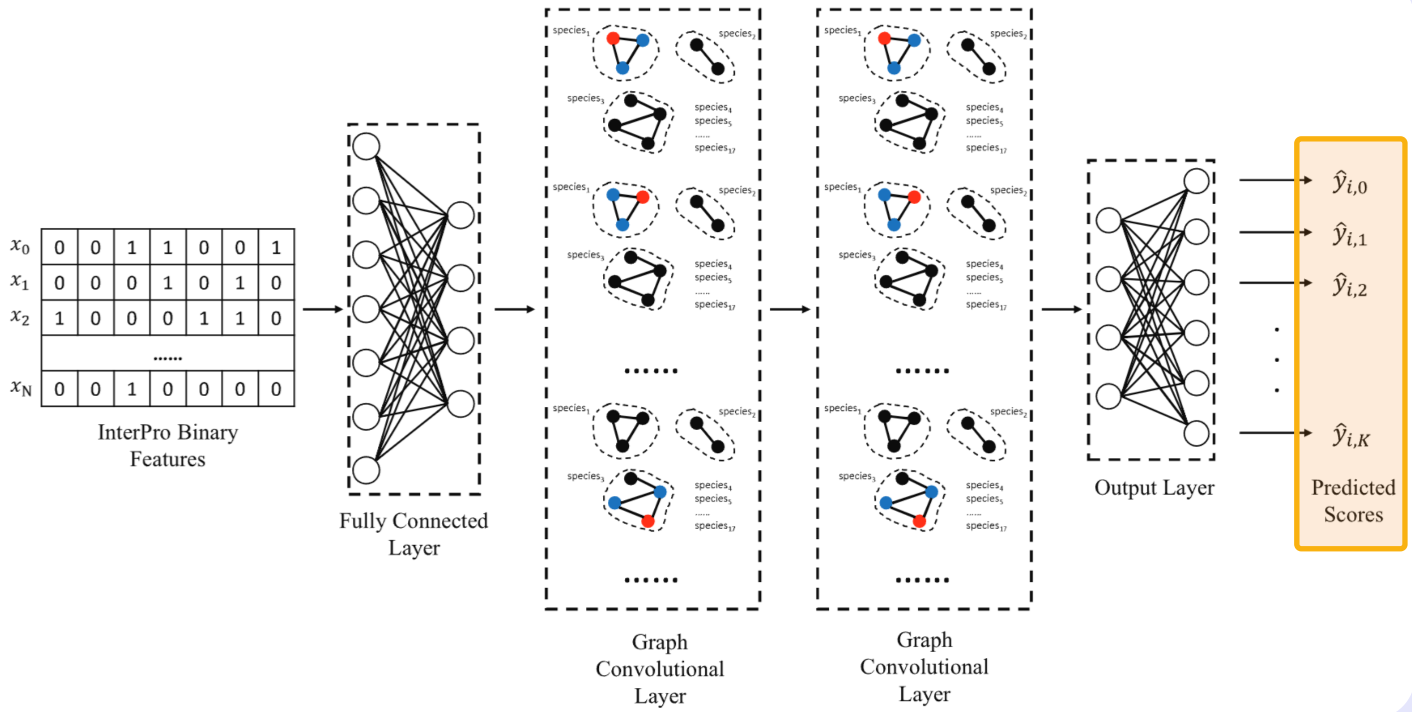
DeepGraphGO architecture

[b]

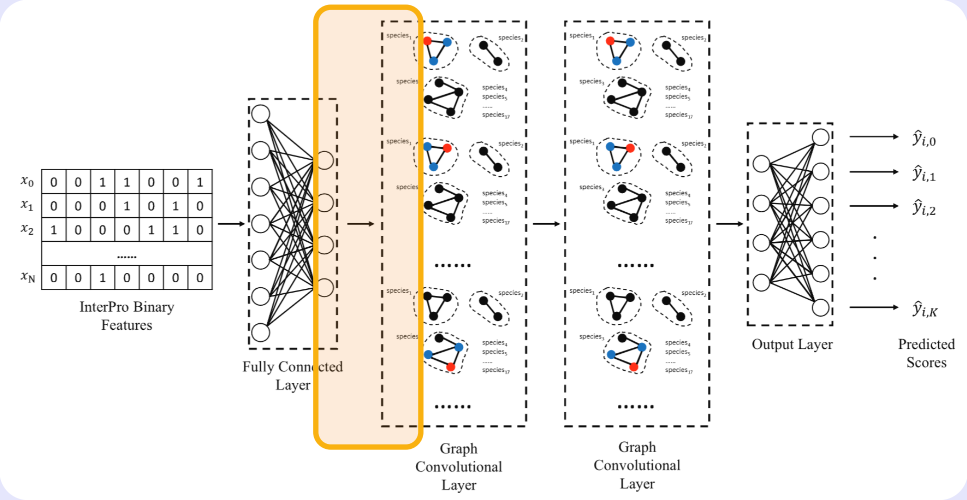


DeepGraphGO architecture

[b]

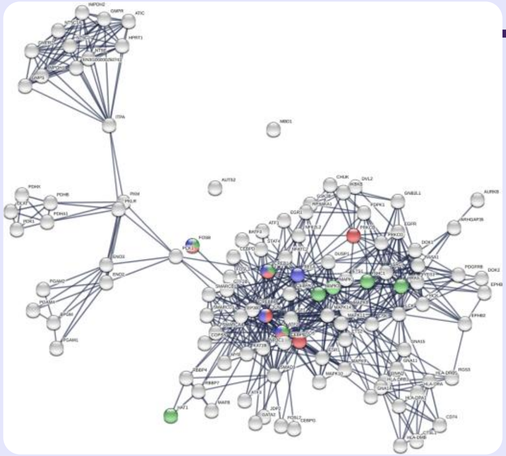


DeepGraphGO architecture



[b]

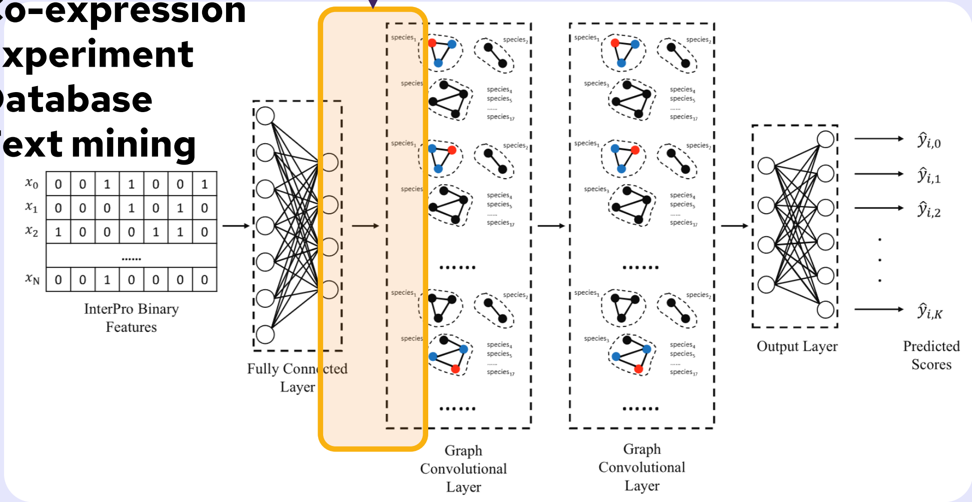
DeepGraphGO architecture



S.-J. Chen et al., Scientific Reports, 2019

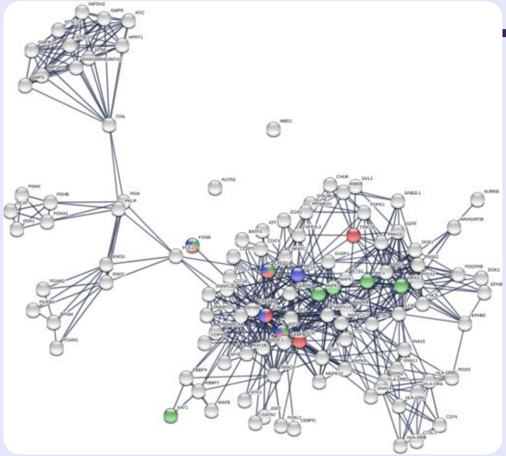
STRING database

1. **Neighbourhood**
2. **Fusion**
3. **Co-occurrence**
4. **Co-expression**
5. **Experiment**
6. **Database**
7. **Text mining**



[b]

DeepGraphGO architecture

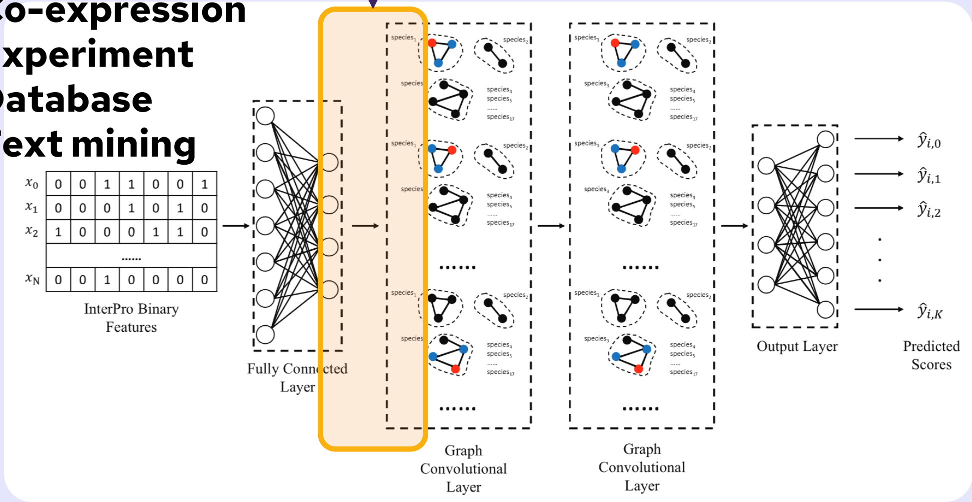


S.-J. Chen et al., Scientific Reports, 2019

STRING database

×17 species (human, mouse, rice, yeast, dog

1. **Neighbourhood**
2. **Fusion**
3. **Co-occurrence**
4. **Co-expression**
5. **Experiment**
6. **Database**
7. **Text mining**



[b]

why DeepGraphGO?

Multi-species

One model fits all

Transfer learning

Easy to expand the PPI network

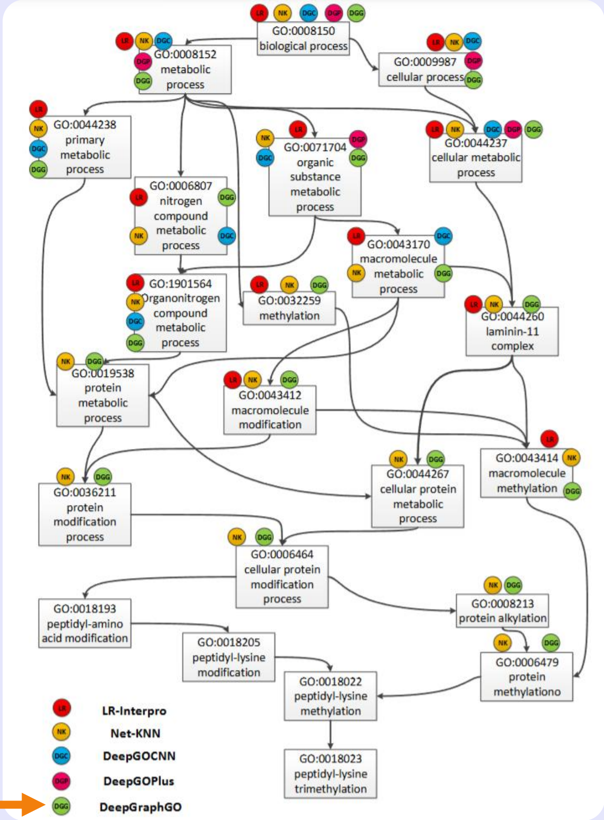
More context

PPI Network information >> Sequence information

DeepGraphGO results

[b] [Suppl. Info]

Method	F_{max}			AUPR		
	MFO	BPO	CCO	MFO	BPO	CCO
BLAST-KNN	0.592	0.274	0.652	0.458	0.114	0.572
	5.22e-52	1.49e-92	9.14e-87	8.68e-76	6.36e-100	3.98e-112
LR-InterPro	0.617	0.280	0.661	0.532	0.145	0.671
	3.04e-14	1.91e-96	6.53e-85	8.11e-20	1.80e-87	5.71e-49
Net-KNN	0.425	0.306	0.667	0.274	0.157	0.642
	7.94e-116	1.57e-59	2.05e-75	2.93e-111	1.02e-66	2.47e-80
DeepGOCNN	0.436	0.248	0.633	0.309	0.102	0.573
	2.30e-111	1.02e-106	1.24e-103	2.46e-108	2.56e-99	1.01e-113
DeepGOPlus	0.597	0.291	0.674	0.402	0.110	0.596
	5.15e-49	1.40e-77	2.14e-57	1.55e-97	4.63e-104	3.48e-108
DeepGraphGO	**0.624**	**0.327**	**0.692**	**0.545**	**0.195**	**0.695**



DeepGraphGO results

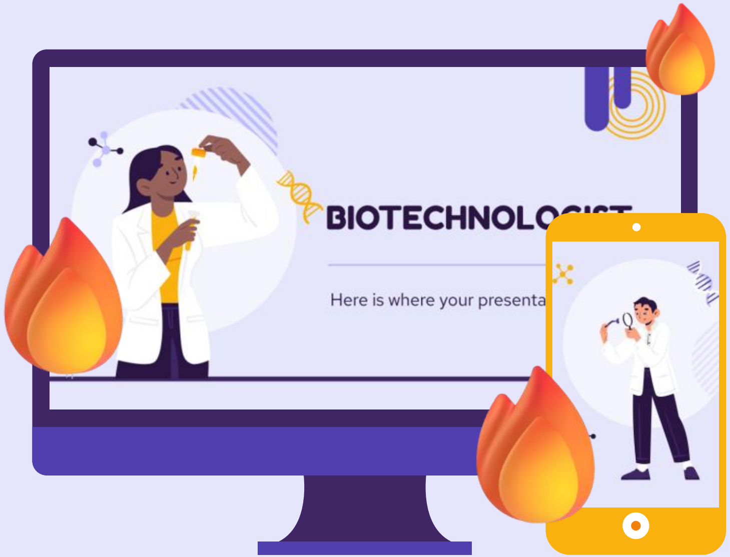
Table 7. Performance comparison on difficult proteins

Method	F_{max}		
	MFO	BPO	CCO
BLAST-KNN	0.534	0.274	0.521
LR-InterPro	0.589	0.275	0.613
Net-KNN	0.404	0.292	0.595
DeepGOCNN	0.406	0.243	0.578
DeepGOPlus	0.564	0.292	0.602
DeepGraphGO	**0.598**	**0.322**	**0.625**

Table 5. Performance comparison on proteins in HUMAN and MOUSE

[b]

Method	F_{max}			AUPR		
	MFO	BPO	CCO	MFO	BPO	CCO
HUMAN (9606)						
BLAST-KNN	0.471	0.241	0.555	0.296	0.074	0.384
LR-InterPro	0.593	0.282	0.650	0.496	0.138	0.603
Net-KNN	0.485	0.261	0.615	0.358	0.143	0.620
DeepGOCNN	0.468	0.263	0.594	0.327	0.114	0.552
DeepGOPlus	0.501	0.277	0.625	0.246	0.088	0.479
DeepGraphGO	**0.633**	**0.320**	**0.655**	**0.520**	**0.178**	**0.642**
MOUSE (10090)						
BLAST-KNN	**0.681**	0.289	0.593	0.593	0.105	0.441
LR-InterPro	0.628	0.312	0.592	0.625	0.175	0.569
Net-KNN	0.420	0.302	0.588	0.319	0.167	0.569
DeepGOCNN	0.475	0.258	0.574	0.405	0.129	0.495
DeepGOPlus	0.634	0.306	0.598	0.550	0.132	0.488
DeepGraphGO	0.650	**0.329**	**0.638**	**0.651**	**0.201**	**0.634**

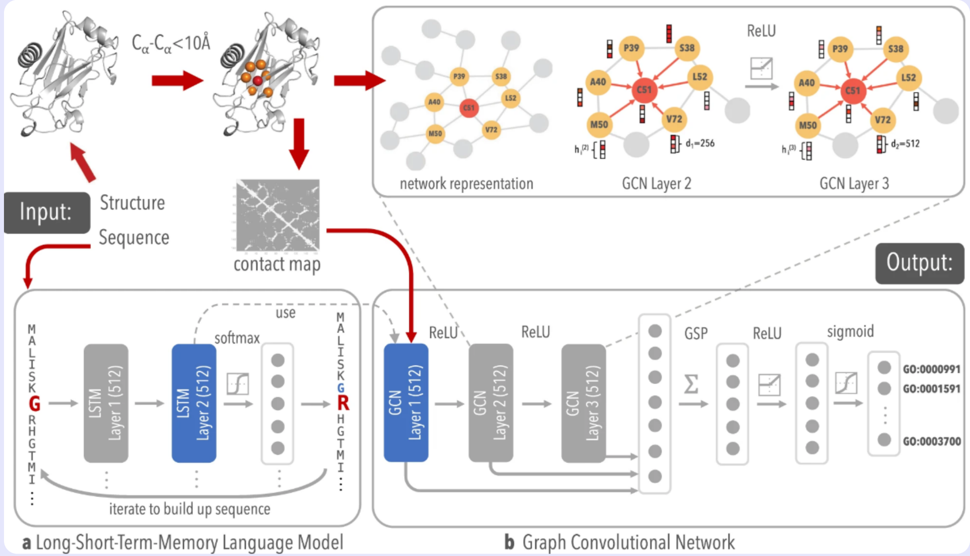


DeepGraphGO limitation

GNNs are very slow to train

DeepFRI – Graph Convolution Network

- Predict protein function by extracting features from sequences and protein structure



Schematic of DeepFRI

Gligorijević, *Nature Communications, 2020*

1. LSTM-LM is pre-trained from protein database
2. Extract residue-level features
3. The extracted features with contact maps are the inputs for second stage
4. Construct protein-level features

DeepFRI performance

Compared to other methods:

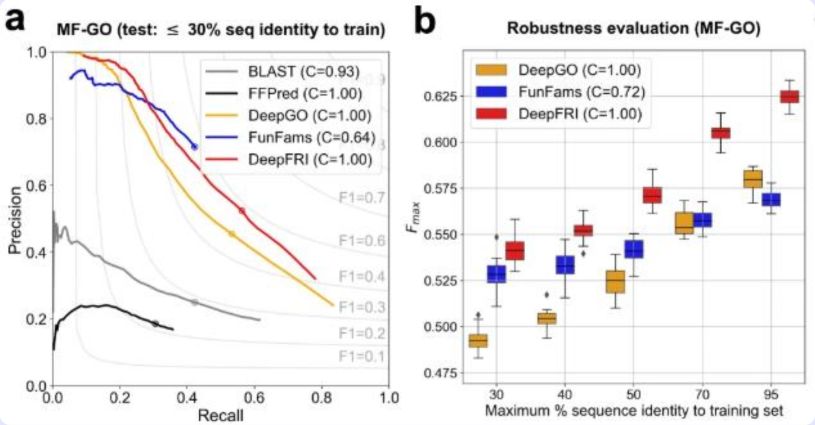
1. 2 sequence-based annotation transfer method (BLAST, FunFams)

2. Deep learning method (DeepGO)

3. Feature engineering-based machine learning method (FFPred)

Gligorijević, *Nature Communications, 2020*

DeepFRI performance



Precision-recall curves showing the performance of different methods

Gligorijević, *Nature Communications, 2020*

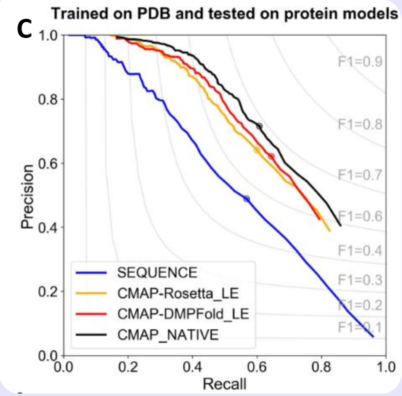
From figure a,

- Better protein-centric F_{max}
- Better performance in Molecular Function (MF) and Biological Process (BP)

From figure b,

- Predict MF-GO proteins with < 30% sequence identity to the training set
- DeepFRI has highest F_{max} (0.545)
- Outperforms FunFams and DeepGO

DeepFRI performance



Precision-recall curves showing the performance of DeepFRI on 700 protein contact maps

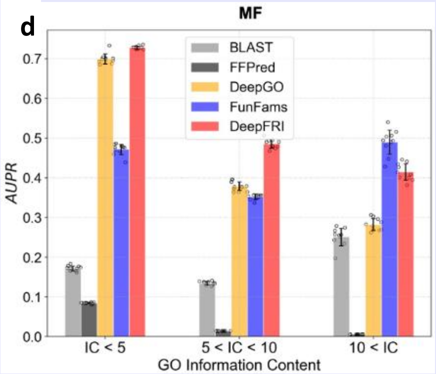
Figure c shows the result of training DeepFRI from Protein Data Bank

- DeepFRI has higher performance for native structures, DMPFold models and Rosetta models
- Significant denoising capability of DeepFRI

Gligorijević, *Nature Communications, 2020*

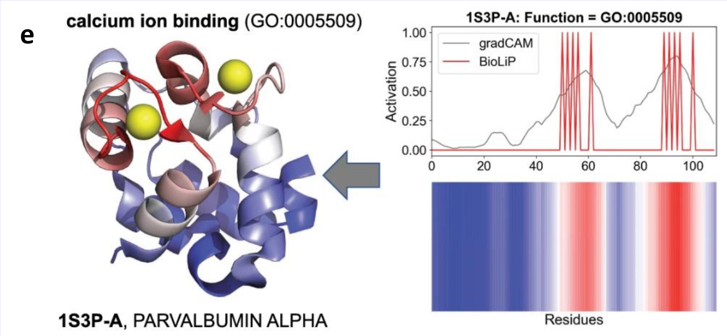
From figure d,

- DeepFRI predicts more specific MF-GO terms with fewer examples
- For proteins well represented in training set, DeepFRI has a comparable performance to FunFams



Distribution of AUPR score on MF-GO terms of different levels of specificities

DeepFRI highlights



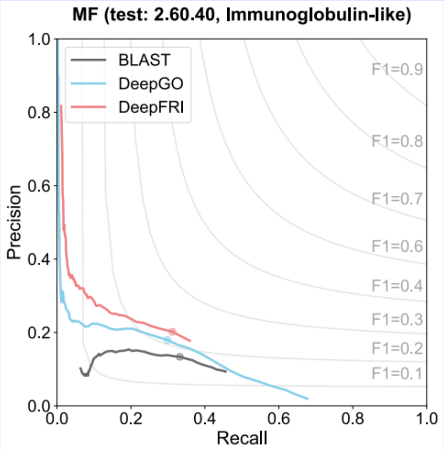
(Right) Gradient-weighted class activation map for calcium ion binding
(Left) 3D structure of a rat protein

From figure e,

- DeeprFRI correctly identify functional sites for calcium ions binding of protein
- The two highest peaks are the calcium-binding residues in the structure of the protein

Gligorijević, *Nature Communications, 2020*

DeepFRI limitation



From *supplementary information,*

Precision-Recall curves showing the performance of DeepFRI compares to DeepGO and BLAST of PDB chains from the top 4 largest CATH folds

1. DeepFRI has lower performance for unseen protein models

2. Limited capture of long-distance structural correlations

Gligorijević, *Nature Communications, 2020*

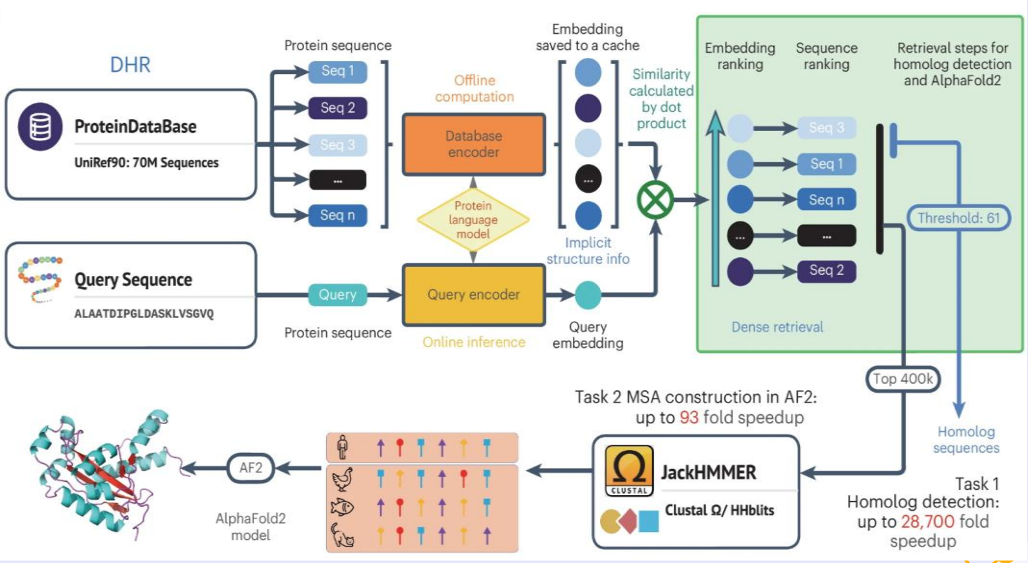
Implications: Role of deep learning

- *"Fast, sensitive detection of protein homologs using deep dense retrieval"*

→ Published in *Nature biotechnology* in 2024, by Prof. Yu Li

In simple words,

1. Convert protein sequences into a special "vector" using a protein language model
2. Compare vectors
3. Skip alignment and just compare the vector representation
4. Contrastive learning to increase accuracy



Summary

- Protein function prediction - Hot research topic
- Deep learning methods >>> Sequence-based methods
- Some limitations are still unsolved

THANK YOU