Deep learning for protein function annotation

PANDA2: protein function prediction using graph neural networks

TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding

Why protein function annotation is important?

- The number protein sequenced each year is too low compared to the annotated protein sequence.
- Protein annotation requires manual and expensive biological experiment.
- The benefit of protein sequence largely depends on knowing its function.

What is protein

- Protein is formed using 20 amino acid
- In sequence representation they are expressed as a sequence of character (ACDEF)
- Proteins also has 3d representation (out of scope)

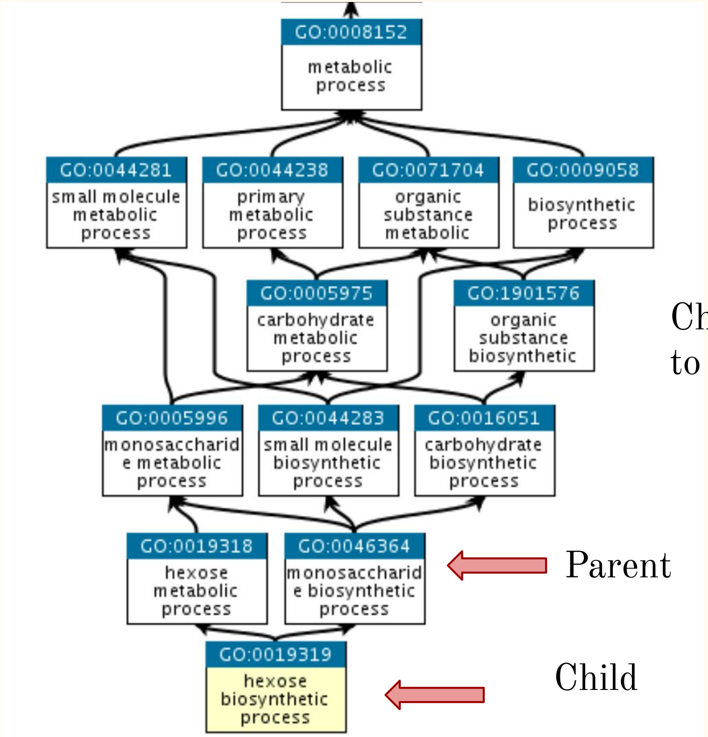
Protein annotation

Gene Ontology consortium defines protein function using three different perspective (ontology)

- Molecular Function Ontology (MFO)
- Biological Process Ontology (BFO)
- Cellular Component Ontology (CCO)

Each protein can have zero or multiple term from each of these ontology

Example of GO term

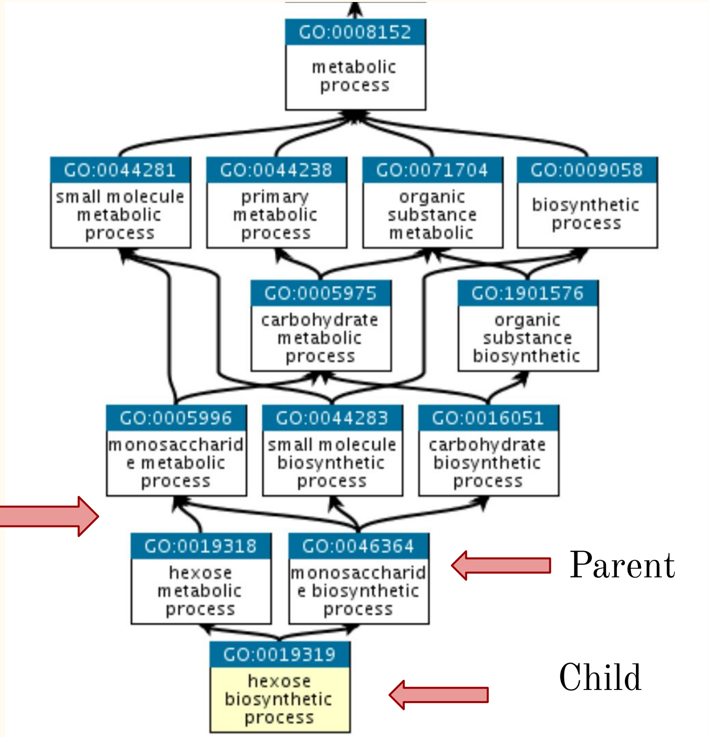


Parent

Child

Children are more specific compared to their parents

Example of GO term



More than one parent

Parent

Child

Terms will form a DAG to root

Protein annotation is more than multi-label classification

A protein is annotated with one GO term, then can also be annotated

by all corresponding ancestral GO terms.

- Sequential property of protein
- Hierarchical property of GO terms

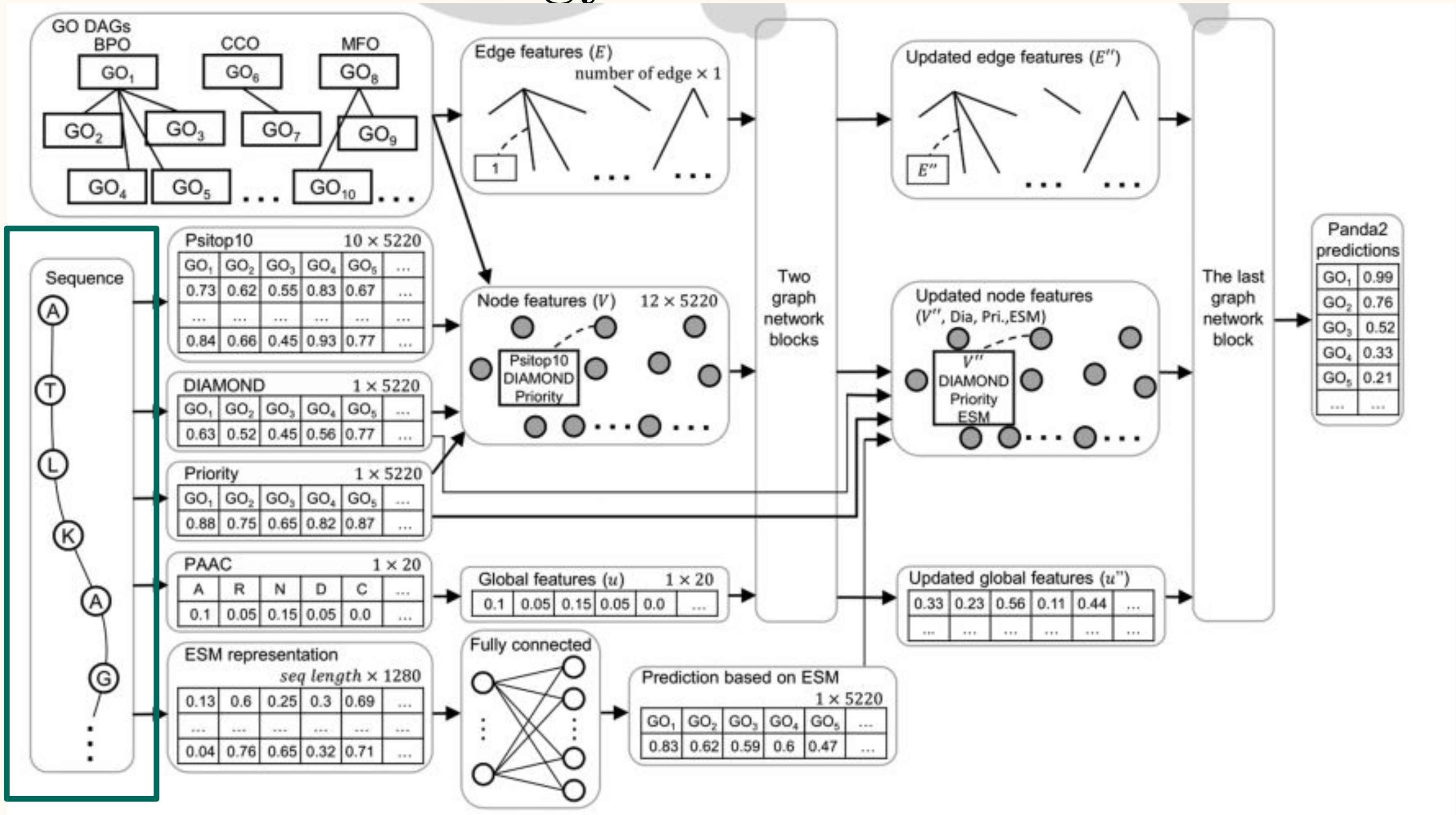
CAFA3 dataset contains 69k annotated protein

Critical Assessment of Functional Annotation (CAFA) largest community driven protein function annotation dataset

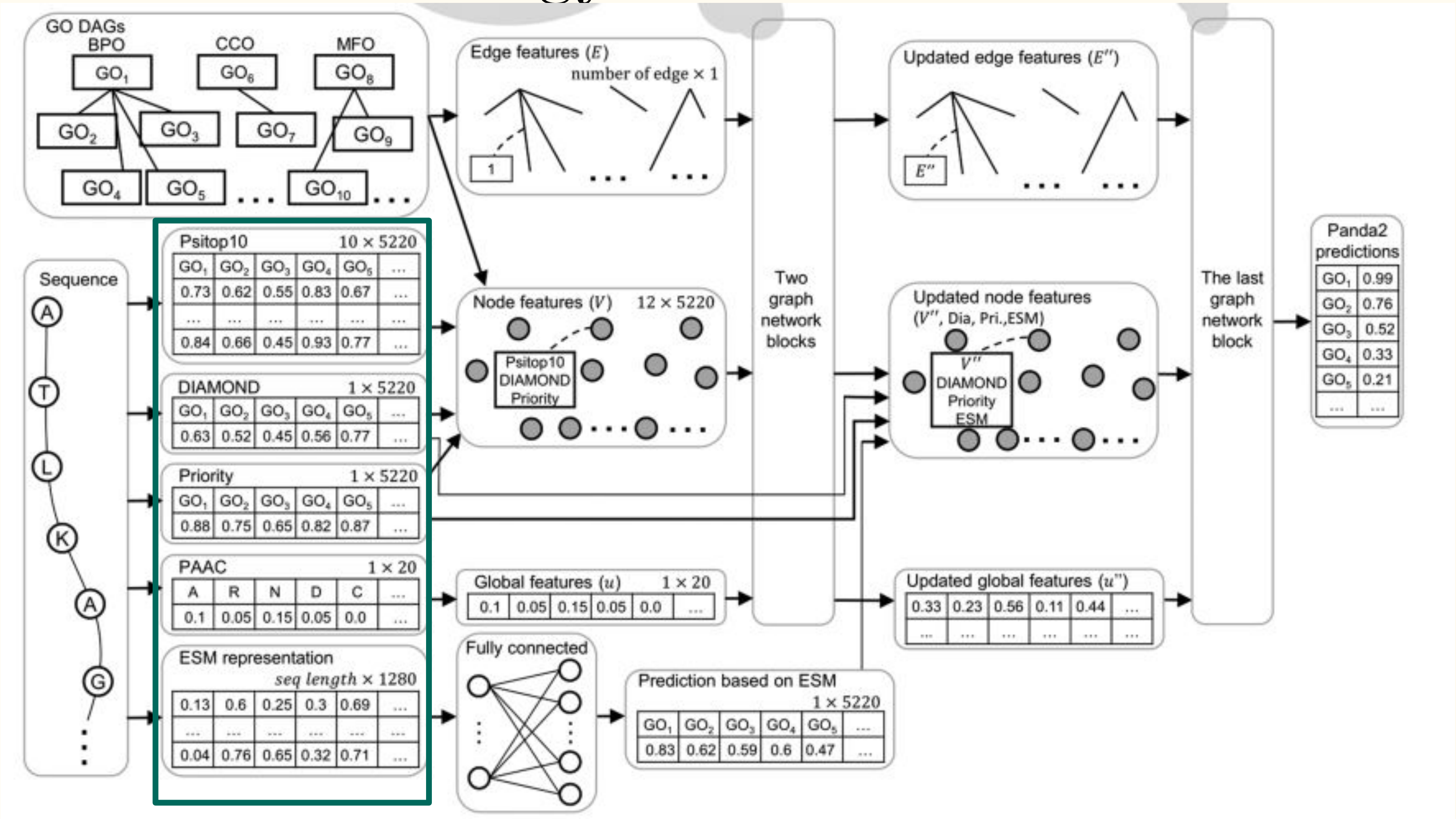
Both the studies used CAFA3 as testing dataset

Dataset	Statistics	MFO	BPO	CPO
CAFA3	Sequence in test set	1137	2392	1265
TALE	Sequence in test set	1916	2836	2084
PANDA2	Sequence in test set	652	3904	545

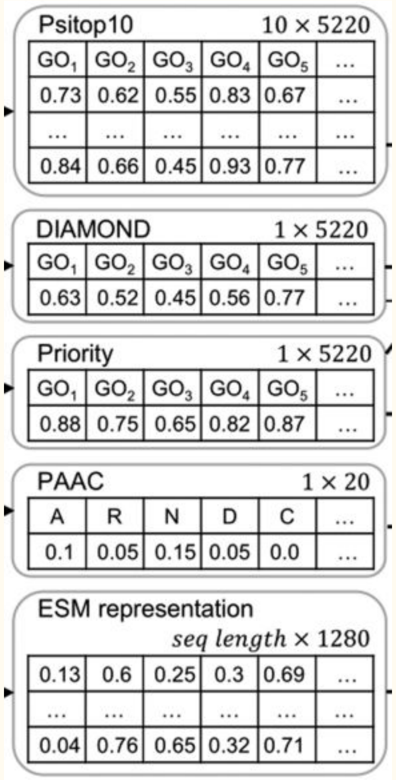
PANDA2 Methodology



PANDA2 Methodology



Features of GO trem

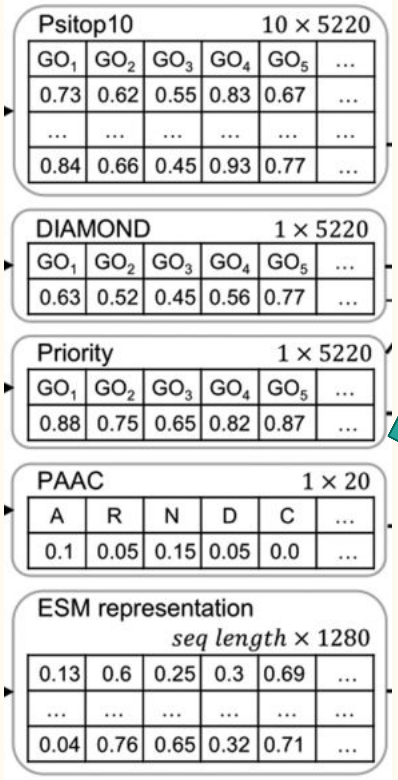


- Identifies the top 10 protein similar (local) to the given protein
- Count of GO term occurrence in top 10 similar protein

- Similar to Psi-top10
- Proteins are identified by BLAST similarity
- Term occurrences are normalized

$$Priority_{GO} = MaxSeqIden(GO) \times \left(\frac{Occurs(GO)}{2 \times Occurs(GO) + 1} + \frac{1}{2} \right)$$

Features of protein sequence

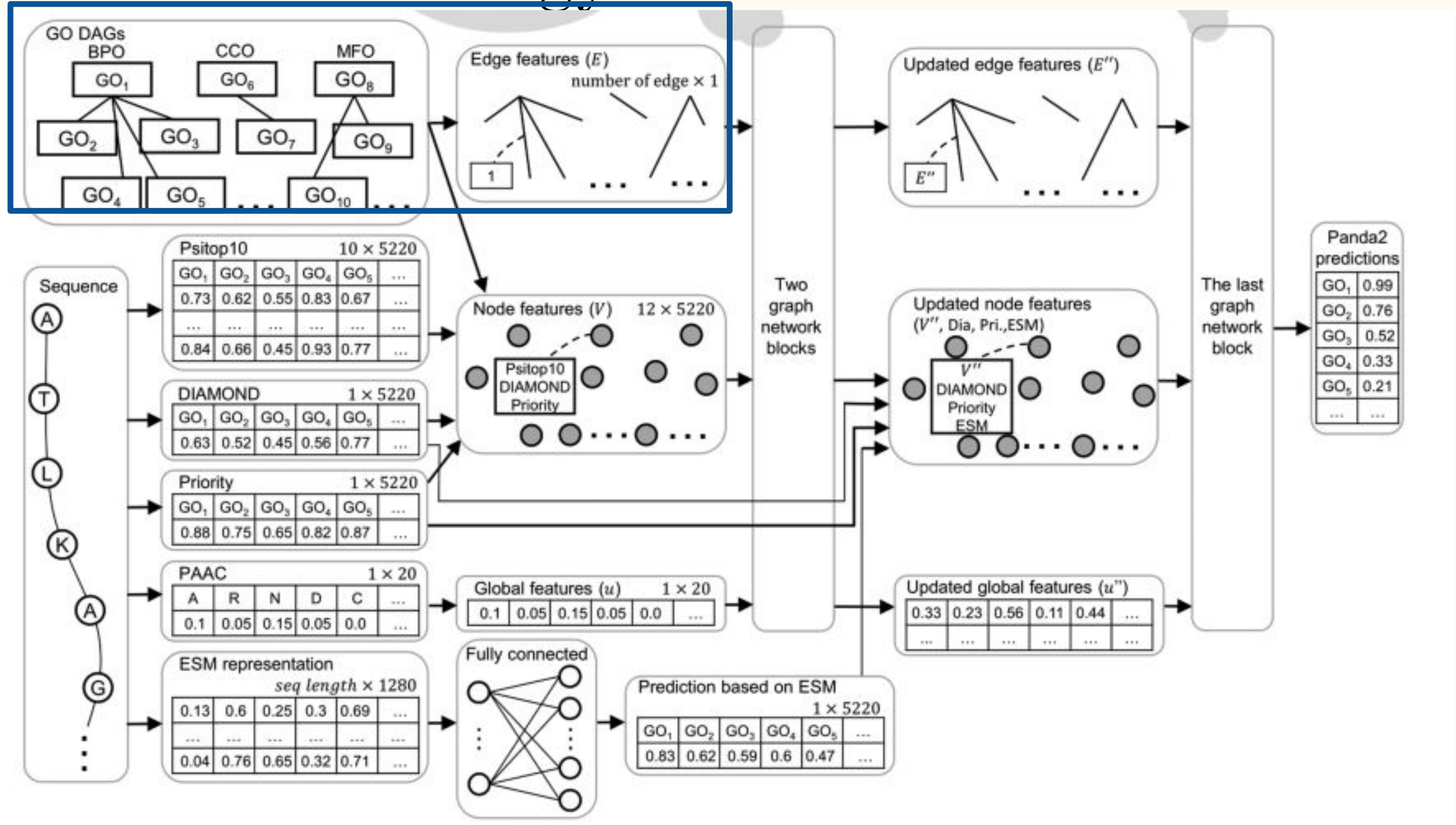


Pseudo amino acid composition (PAAC): Normalized occurrence of of 20 amino acids in the global pool of proteins

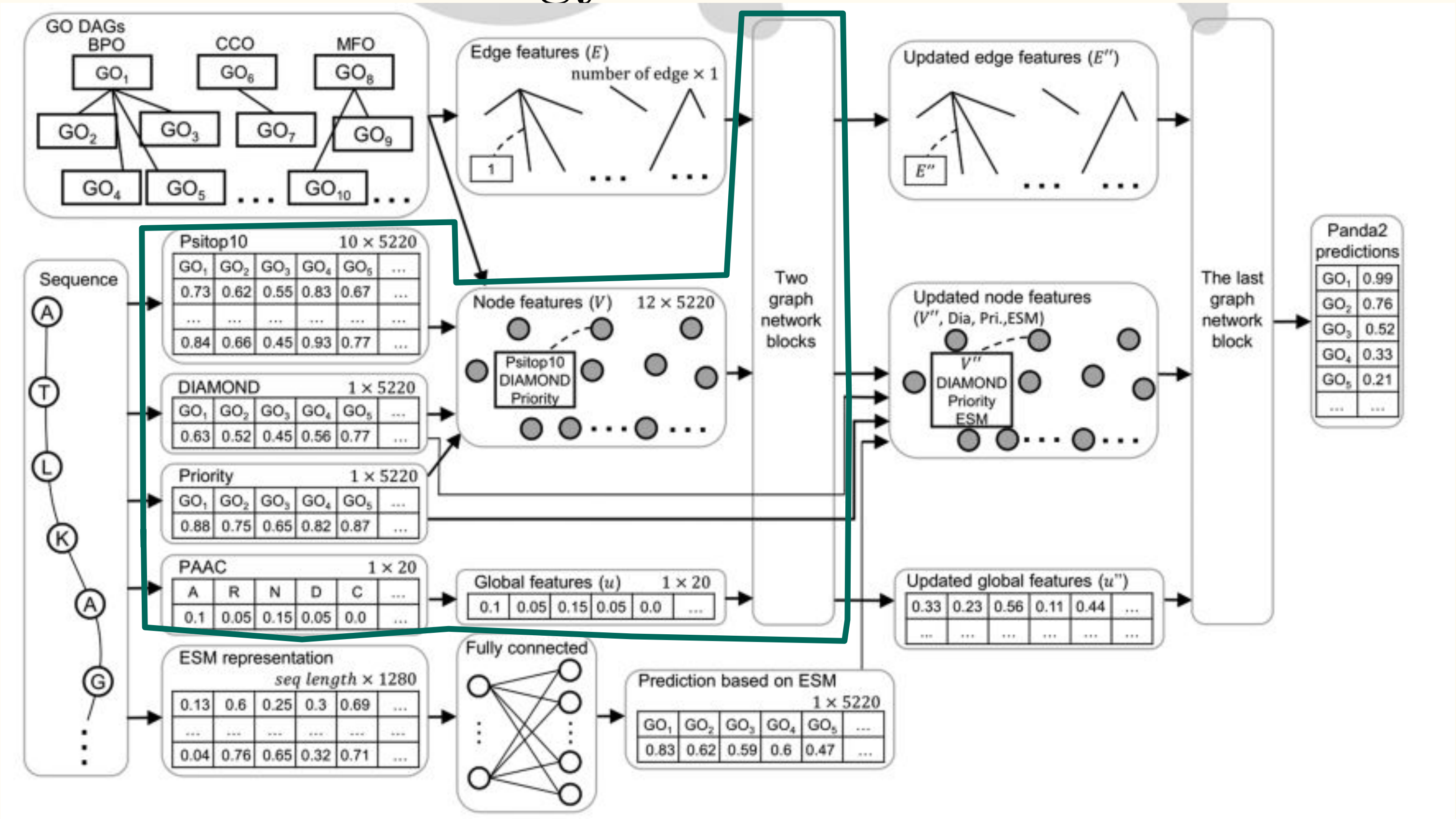
Representation of a protein learned from 250 million of proteins using transformer

SOTA protein representation till 2020 (claimed by **Facebook**)

PANDA2 Methodology



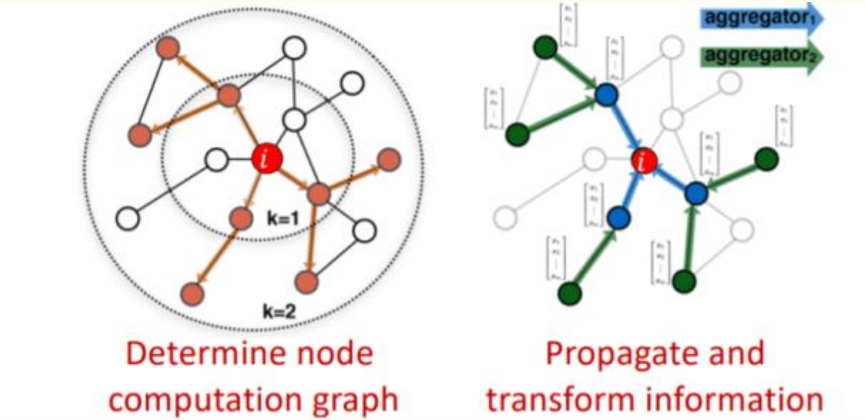
PANDA2 Methodology



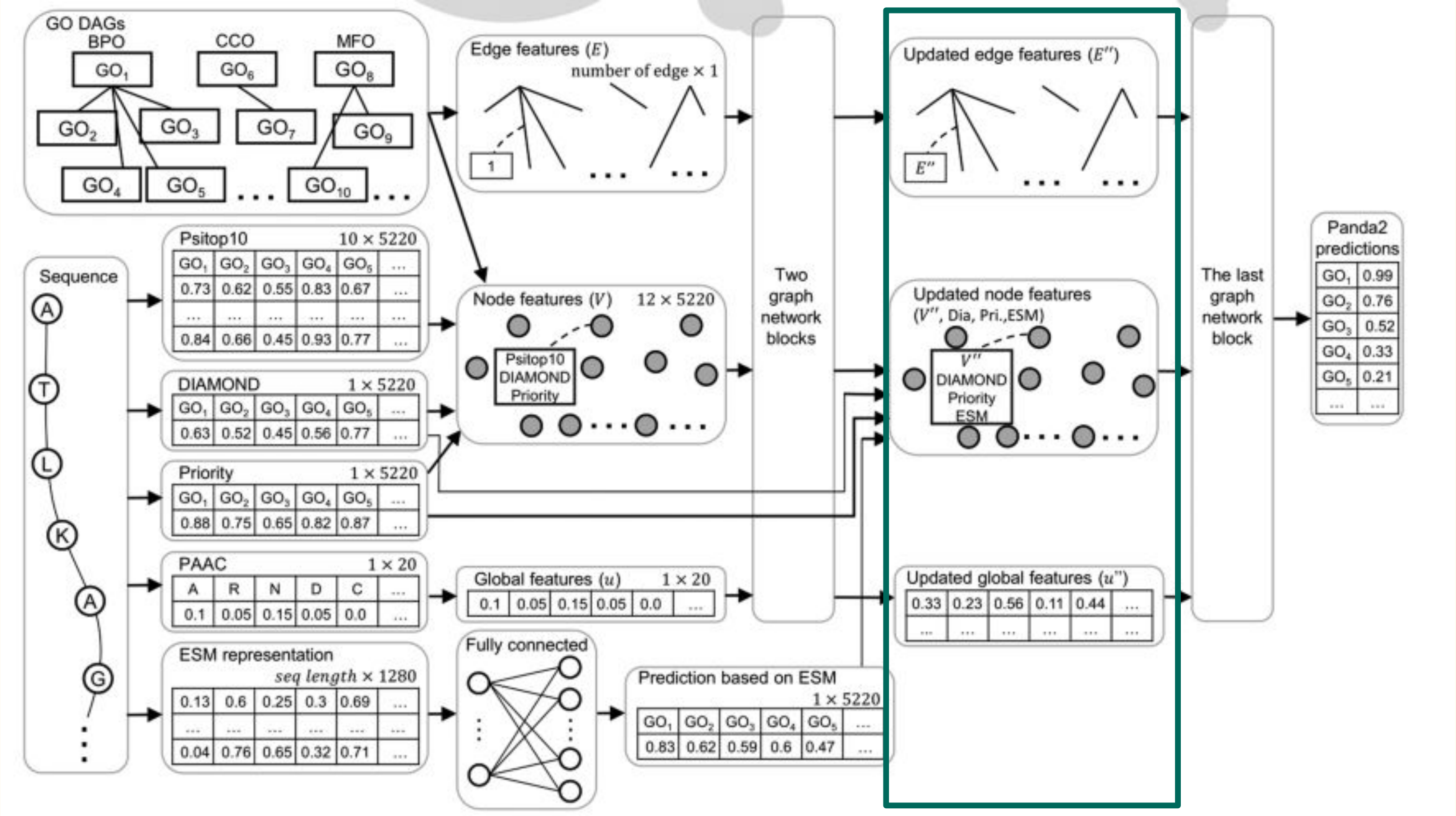
Graph Neural Network

Graph neural network learns the representation of the full graph by

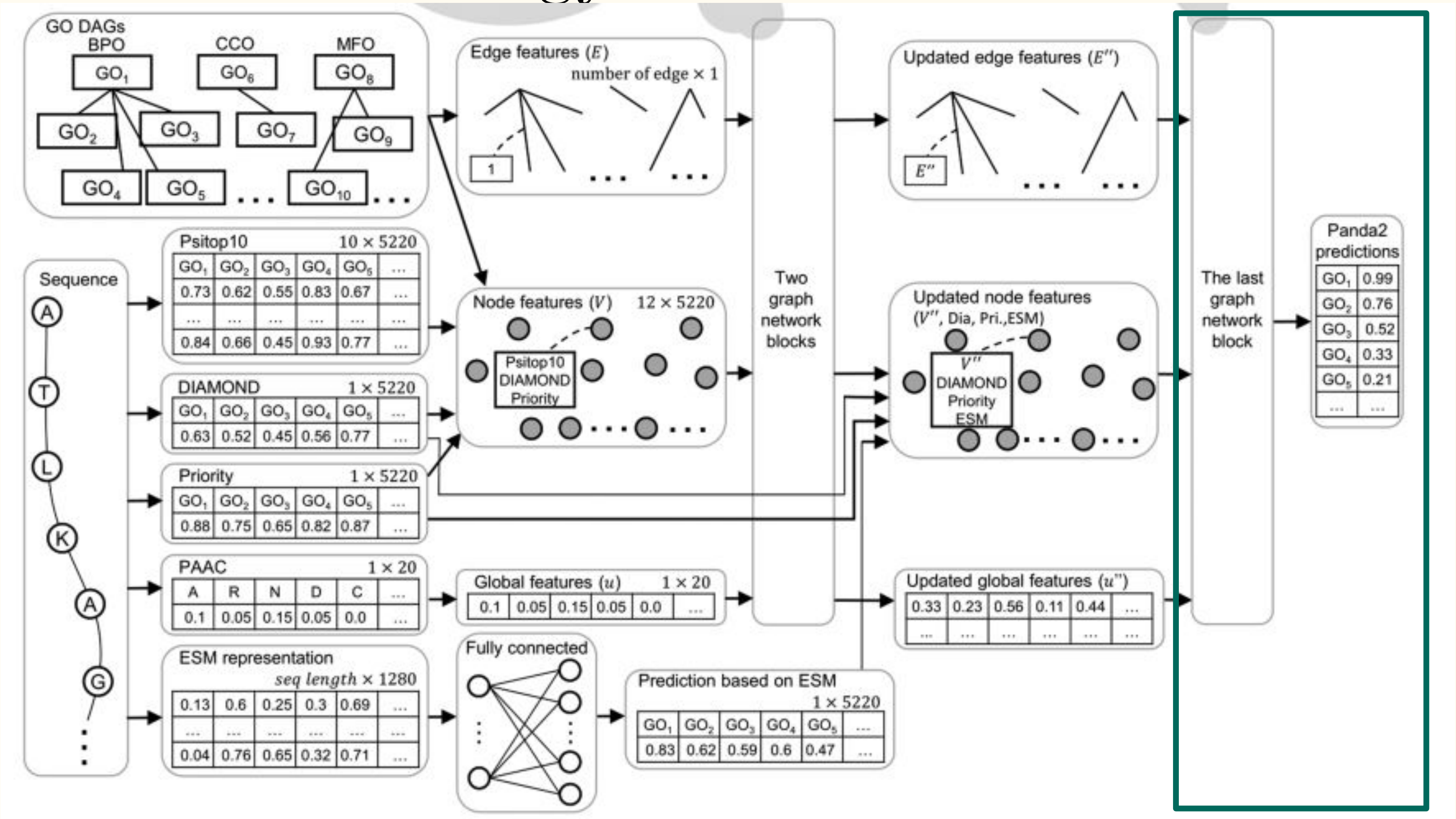
- Learning features Nodes
- Learning features of Edge



PANDA2 Methodology



PANDA2 Methodology

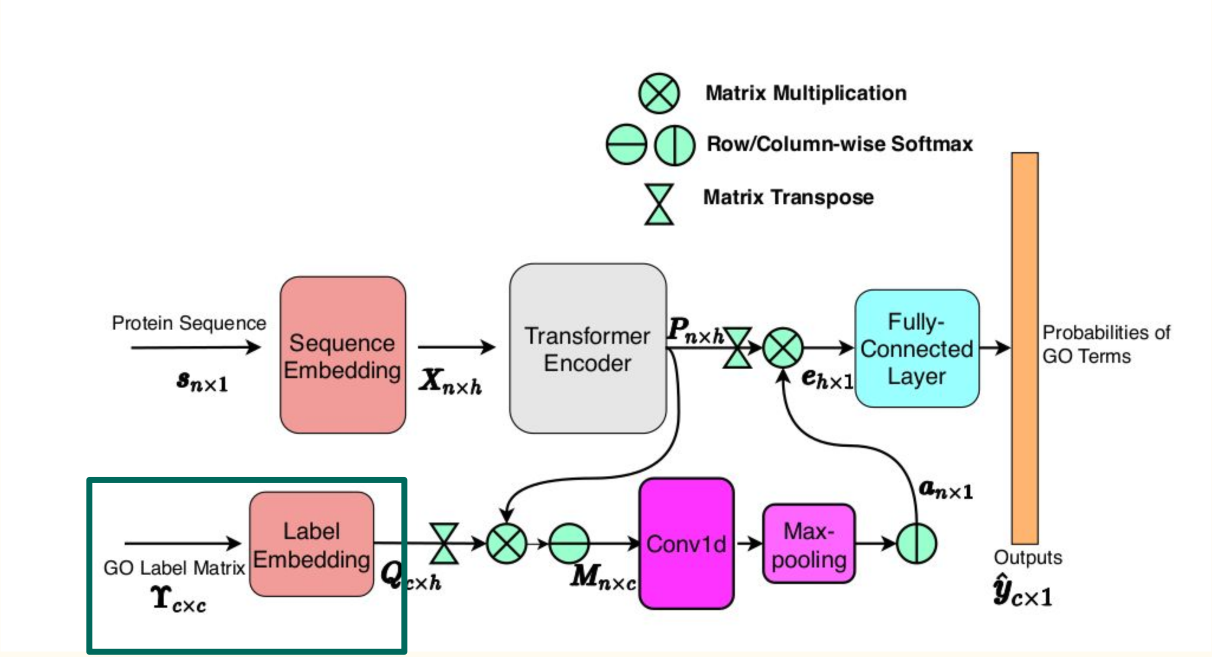


Evaluation

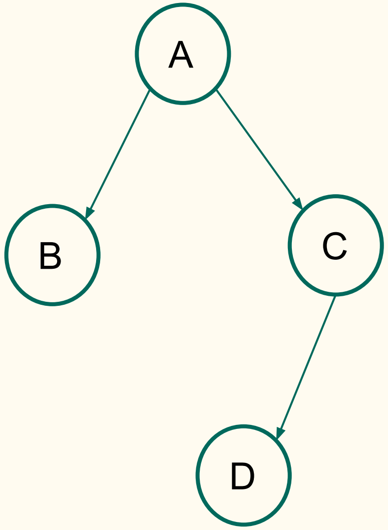
$$S_{min} = \min_{t} \left\{ \sqrt{ru(t)^2 + mi(t)^2} \right\}$$

Method	F_{max} MFO	F_{max} BPO	F_{max} CCO	S_{min} MFO	S_{min} BPO	S_{min} CCO	AUPR MFO	AUPR BPO	AUPR CCO
Naive	0.306	0.318	0.605	12.105	38.890	9.646	0.150	0.219	0.512
DIAMONDBlast	0.525	0.436	0.591	9.291	39.544	8.721	0.101	0.070	0.089
UDSMProt	0.582	0.475	0.697	8.787	33.615	7.618	0.548	0.422	0.728
DeepText2GO	***0.627***	0.441	0.694	5.240	17.713	***4.531***	***0.605***	0.336	0.729
GOLabeler	0.586	0.372	0.691	***5.032***	***15.050***	5.479	0.549	0.236	0.697
DeepGOPlus	0.585	0.474	0.699	8.824	33.576	7.693	0.536	0.407	0.726
PANDA	0.486	0.367	0.520	11.751	45.096	12.723	0.396	0.289	0.394
PANDA2	0.598	***0.478***	***0.709***	9.670	40.229	9.558	0.564	***0.436***	***0.744***

TALE Methodology



Labels are embedded using Graph hierarchy

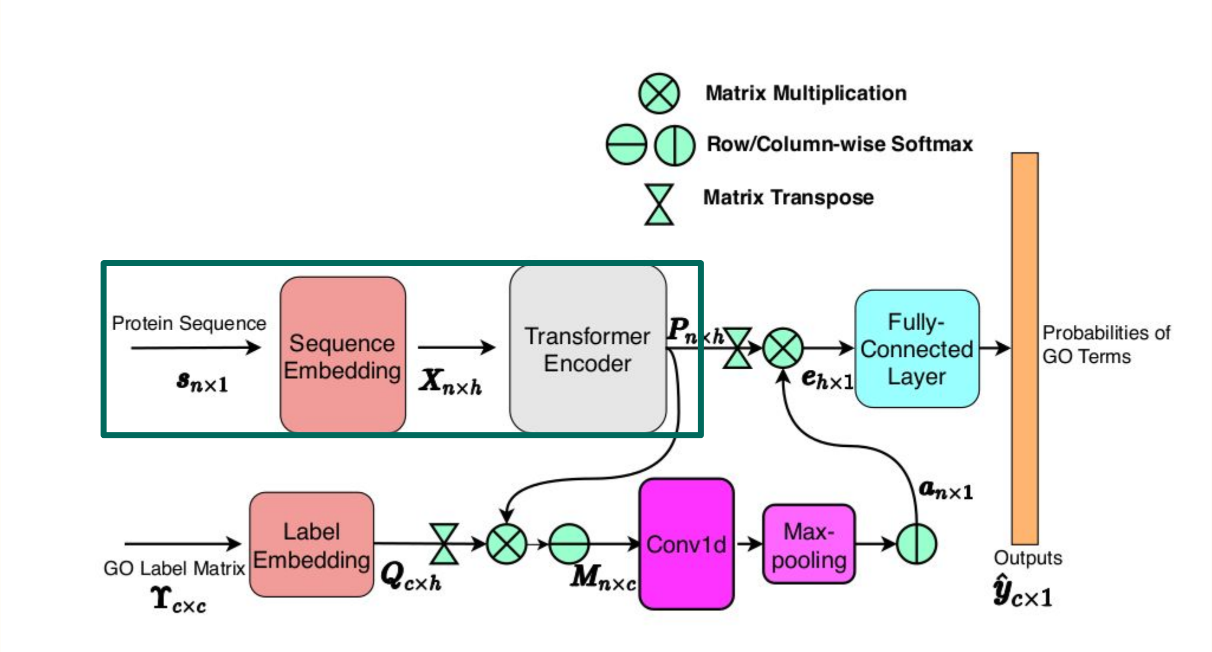


	A	B	C	D
A	**1**	**0**	**0**	**0**
B	**1**	**1**	**0**	**0**
C	**1**	**0**	**1**	**0**
D	**1**	**0**	**1**	**1**

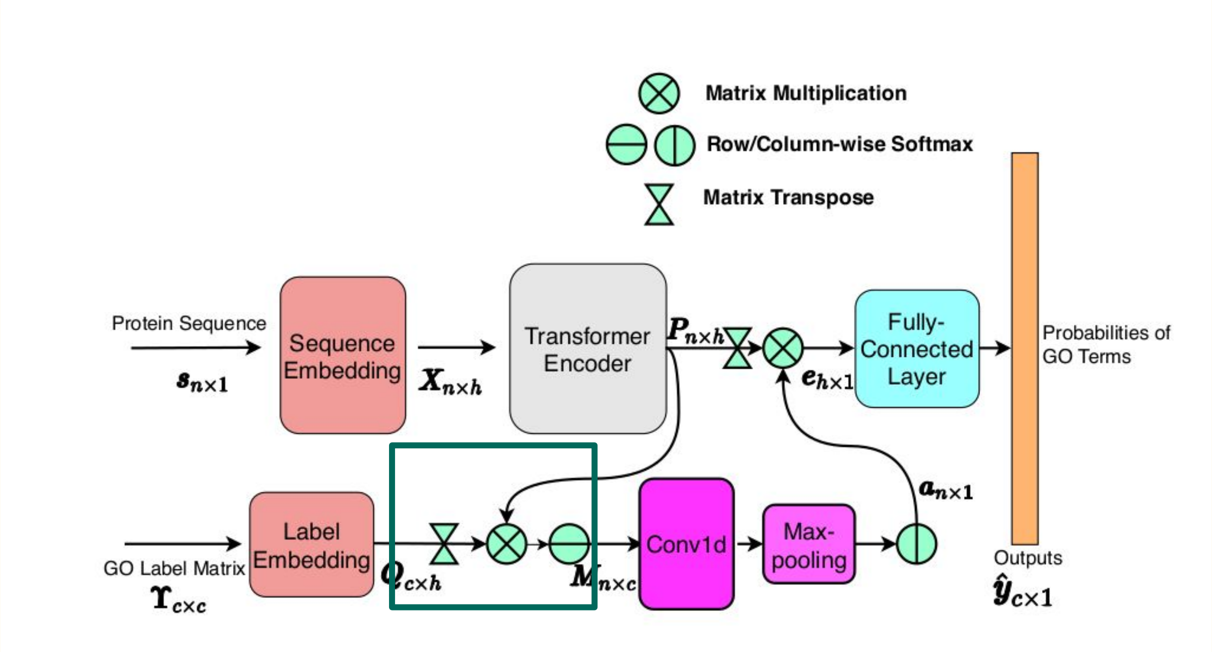
Ancestor

Ancestor

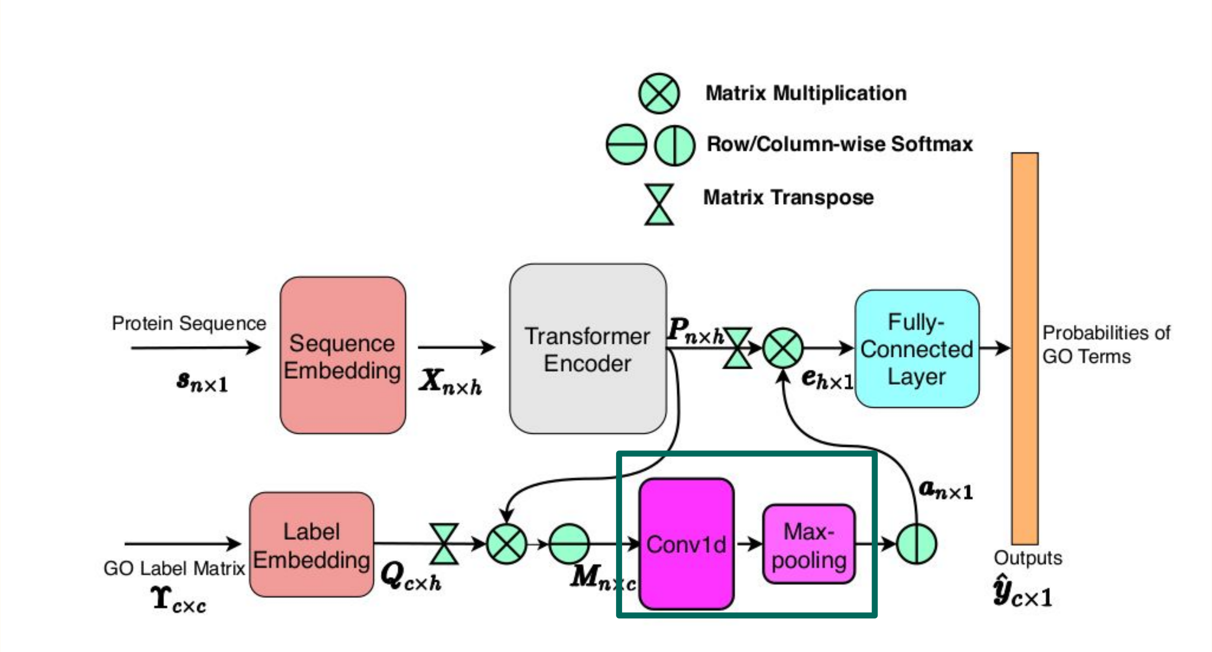
Sequence embedding using Transformer



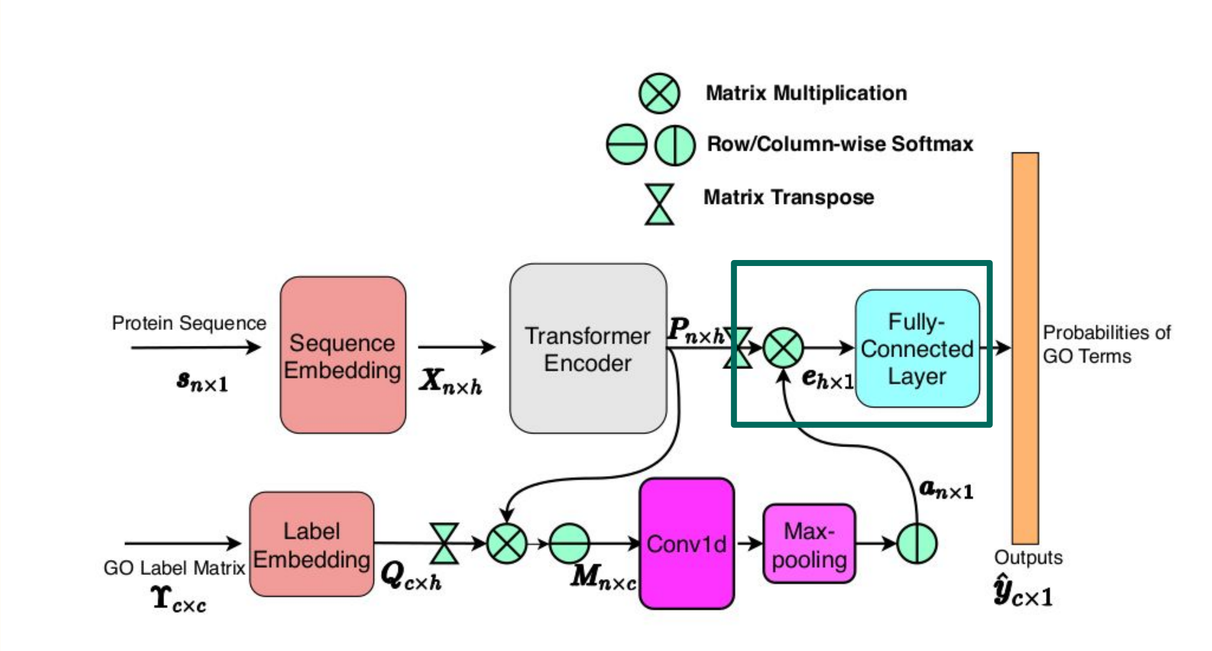
Contribution of each amino acid in each GO label



Influence of other amino acids in GO label



Weighted sequence encoding matrix



Maintaining label hierarchy

A child labels probability can not be high than it's ancestor

$$L' = -\frac{1}{c}\sum_{i=1}^{c} y_i \times \hat{y}_i + (1 - y_i) \times (1 - \hat{y}_i)$$

Binary Cross entropy

$$R = \frac{1}{|E|}\sum_{(i,j)\in E} \max(0, \hat{y}_j - \hat{y}_i) = \frac{1}{|E|}\sum_{(i,j)\in E} \mathrm{ReLU}(\hat{y}_j - \hat{y}_i),$$

$$L = L' + \lambda R,$$

Results

Method	Fmax			AUC		
	MFO	BPO	CCO	MFO	BPO	CCO
Naive	0.306	0.318	0.605	0.15	0.219	0.512
DIAMONDBlast	0.525	0.436	0.591	0.101	0.07	0.089
UDSMProt	0.582	0.475	0.697	0.548	0.422	0.728
DeepText2GO	0.627	0.441	0.694	0.605	0.336	0.729
GOLabeler	0.586	0.372	0.691	0.549	0.236	0.697
DeepGOPlus	0.585	0.474	0.699	0.536	0.407	0.726
PANDA	0.486	0.367	0.52	0.396	0.289	0.394
PANDA2	0.318	**0.478**	**0.709**	**0.564**	**0.436**	**0.744**
TALE	**0.615**	0.431	0.669	0.548	0.37	0.652