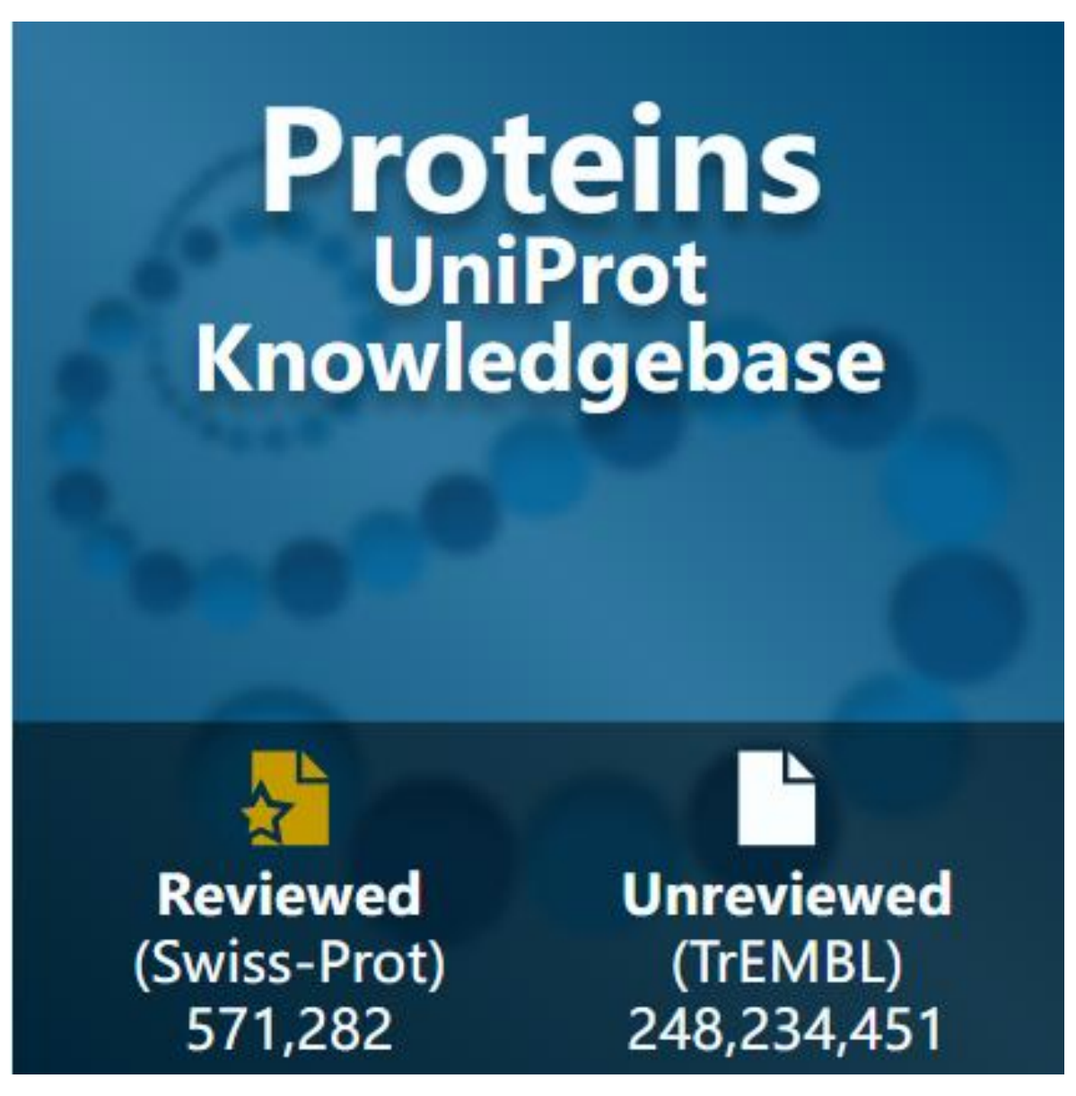
RESEARCH AND APPLICATION OF PROTEIN FUNCTION PREDICTION ALGORITHMS

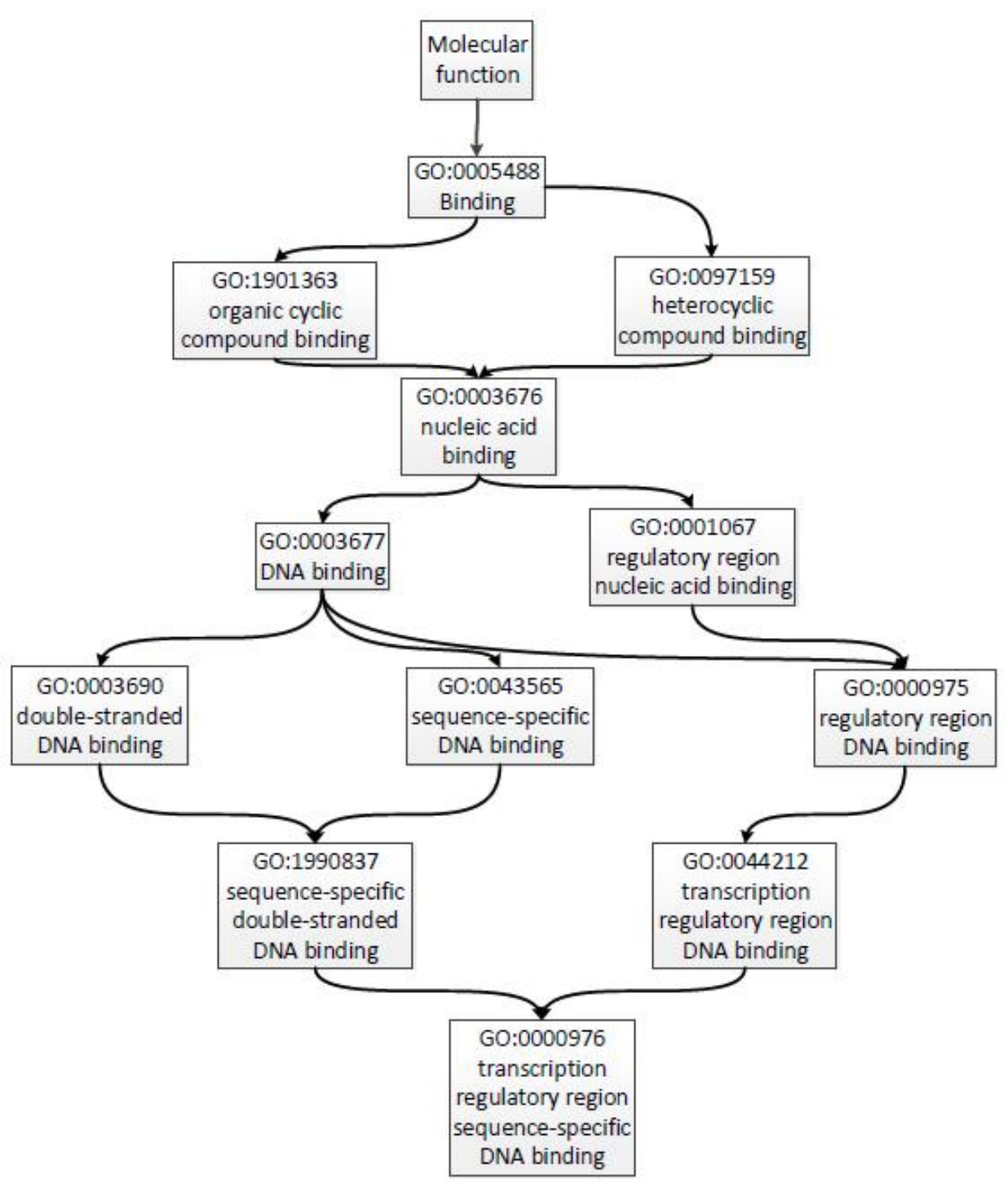
Liu Hancheng[1] & Dong Yibo[1], Haolan Yang[2], Yuzheng Wu[3], Alban Malaj[4], Chechic Lucía Florencia[5]

[1] Institute of Science and Technology for Brain-Inspired Intelligence , Fudan University, Shanghai 200433, China, [2] Nanyang Technological University, Singapore, [3] Fudan University, China, [4] Institute of Psychiatry, Psychology and Neuroscience King's Collage London, [5] Universidad de Buenos Aires, Buenos Aires, Argentina

Introduction: In modern biological and medical research, grasping protein functionality is essential for uncovering disease mechanisms and developing proteins and pharmaceuticals. As illustrated on the right, Gene Ontology (GO) offers a well-organized and extensive framework for categorizing these functions. Nevertheless, as sequence databases expand rapidly, numerous proteins remain without comprehensive experimental annotations. Therefore, it is necessary to develop high-performance function prediction algorithms to bridge the gap between the increasing volume of protein sequences and the limited understanding of their functions.

1 Background

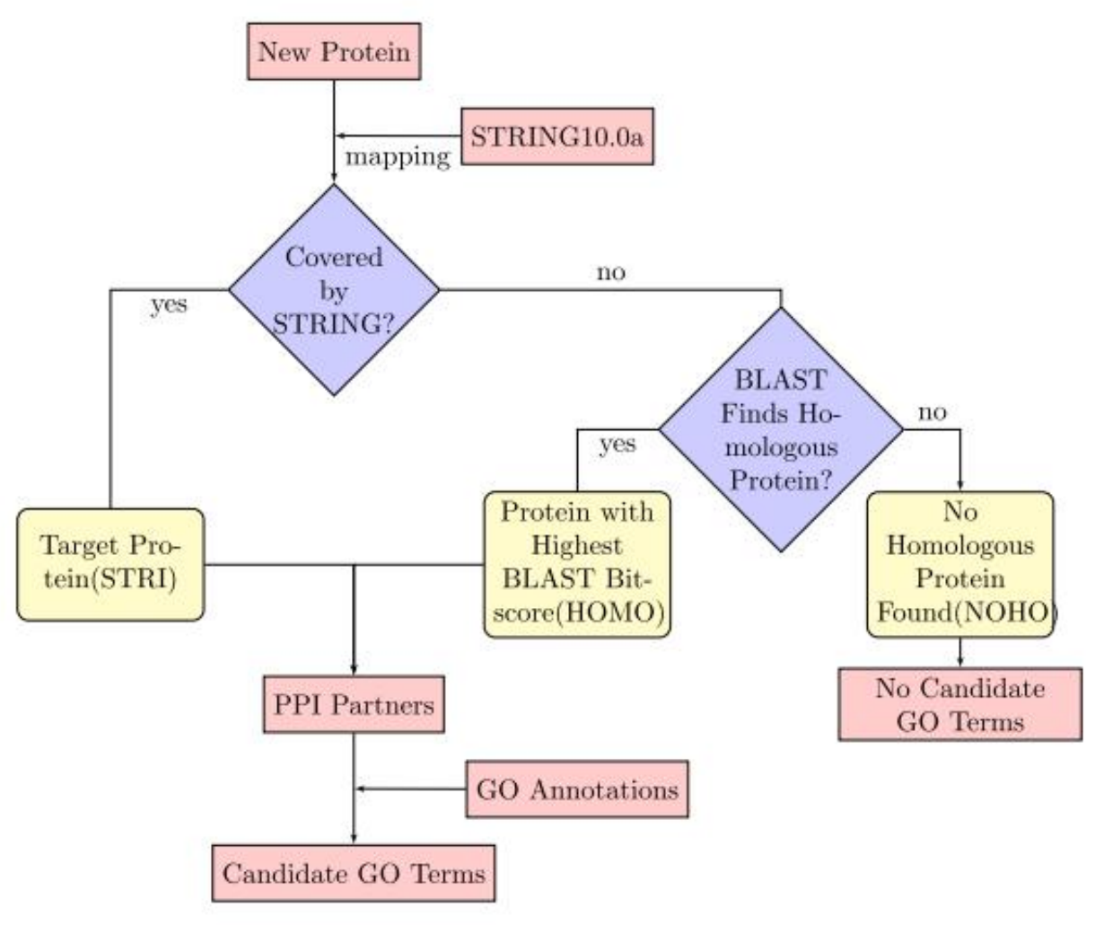



Molecular function
GO:0005488 Binding
GO:1901363 organic cyclic compound binding
GO:0097159 heterocyclic compound binding
GO:0003676 nucleic acid binding
GO:0003677 DNA binding
GO:0001067 regulatory region nucleic acid binding
GO:0003690 double-stranded DNA binding
GO:0043565 sequence-specific DNA binding
GO:0000975 regulatory region DNA binding
GO:1990837 sequence-specific double-stranded DNA binding
GO:0044212 transcription regulatory region DNA binding
GO:0000976 transcription regulatory region sequence-specific DNA binding

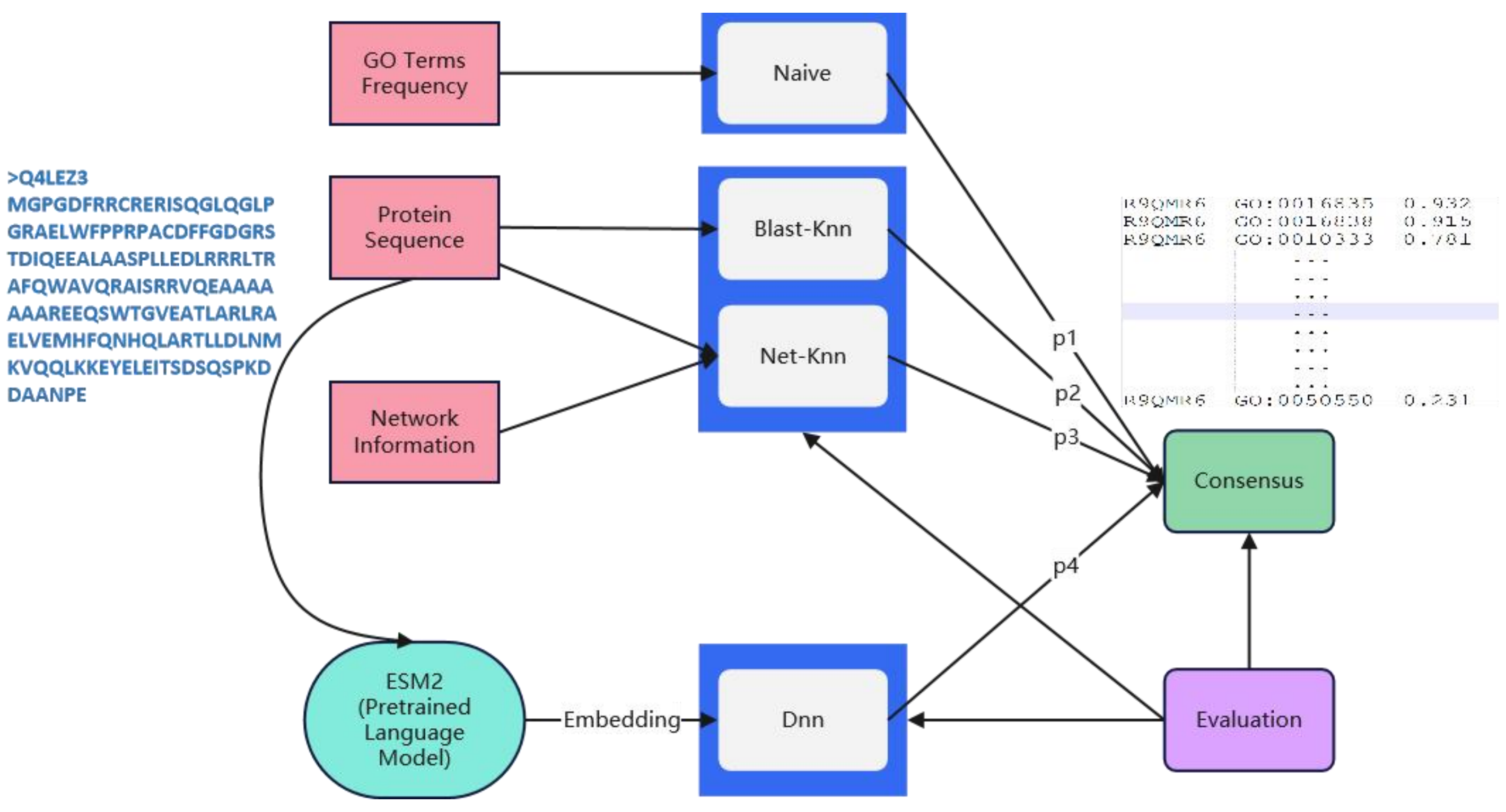
- **Significance**：Elucidating disease mechanisms，Designing proteins and drugs.
- **Challenges**：The gap between the growing number of protein sequences and limited known functional annotations.

2 Methods

- **Blast:** BLAST operates by dividing the query sequence into smaller segments and looking for corresponding sequences in the database. It employs a scoring system to assess similarity, incorporating substitution matrices such as PAM or BLOSUM and imposing gap penalties for insertions or deletions. The outcomes are ranked according to the E-value, which reflects the statistical significance of the matches.
- **Net-KNN**: Identifies GO candidates by using network information. Similar to Blast but replaces sequence similarity with association score (edge weight) in a network.

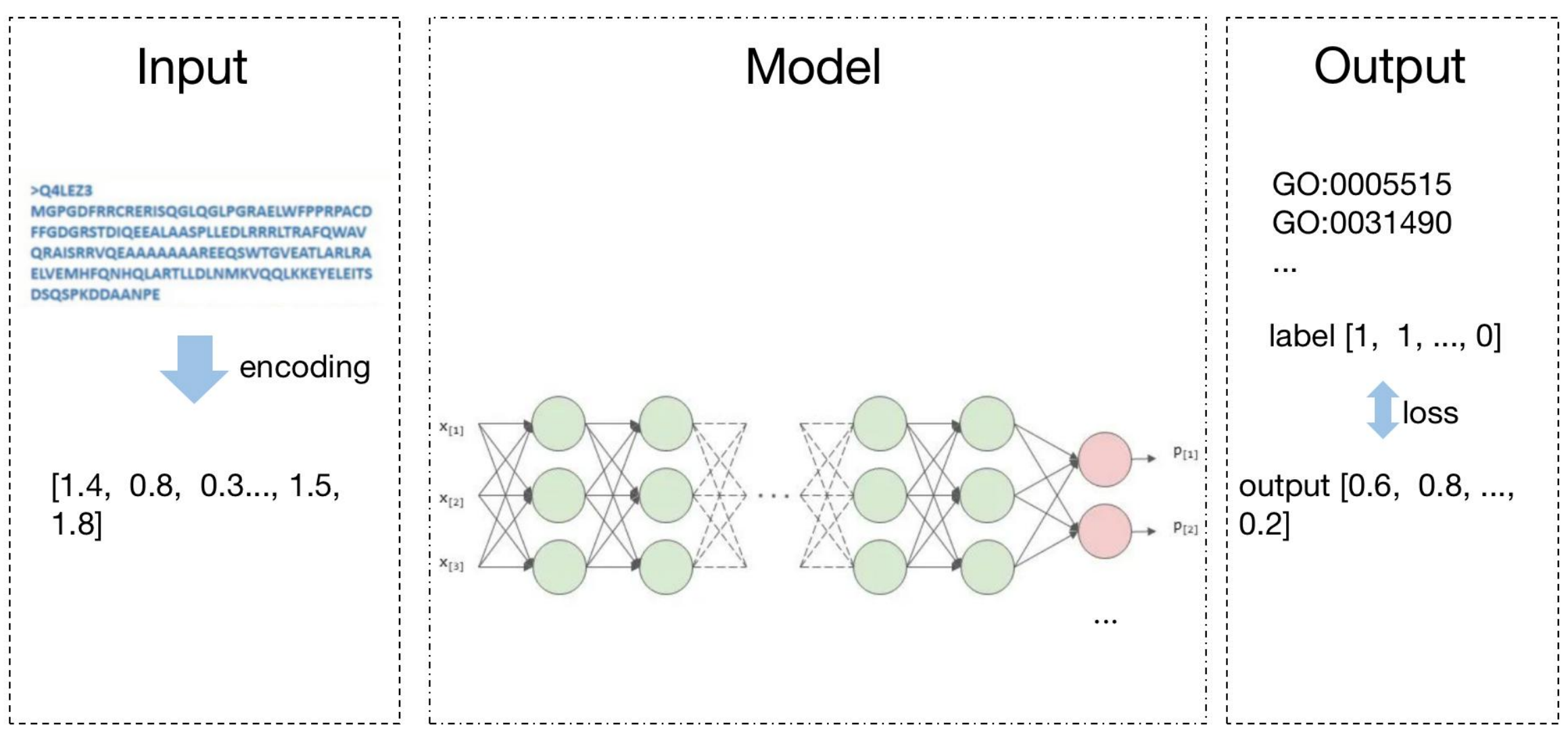


- **Naive:** Predicts PFP only from relative Gene Ontology (GO) frequency in training data. No sophisticated methods or optimization
- **ESM2 (LLM)**



Performance measurements: **FMax** and **AUPR**

- **Deep Machine Learning and Deep Neural Network**


Input
>Q4LEZ3
MGPGDFRRCRERISQGLQGLPGRAELWFPPRPACD
FFGDGRSTDIQEEALAASPLLEDLRRRLTRAFQWAV
QRAISRRVQEAAAAAAAREEQSWTGVEATLARLRA
ELVEMHFQNHQLARTLLDLNMKVQQLKKEYELEITS
DSQSPKDDAANPE
encoding
[1.4, 0.8, 0.3..., 1.5, 1.8]
Model
x[1]
x[2]
x[3]
p[1]
p[2]
...
Output
GO:0005515
GO:0031490
...
label [1, 1, ..., 0]
loss
output [0.6, 0.8, ..., 0.2]

3 Results

	Train Data			Test Data		
	mf	bp	cc	mf	bp	cc
Protein	79866	90014	95988	974	971	960
Protein Function Label	606082	3405279	1145331	9218	35449	10996
Average Label	7.6	37.8	11.9	9.5	36.5	11.5

	Fmax			AUPR		
	mf	bp	cc	mf	bp	cc
naive	0.222	0.222	0.348	0.191	0.122	0.461
blast_knn	0.646	0.478	0.601	0.615	0.340	0.629
net_knn	0.394	0.451	0.583	0.350	0.347	0.695
dnn_esm2	0.636	0.460	0.617	**0.676**	0.382	0.722
consensus	**0.659**	**0.515**	**0.654**	0.669	**0.442**	**0.760**

Conclusions: After analysing all the previously mentioned methods we can said that overall the Naive algorithm had the worse performance, followed by NetKNN. Both BlastKNN and ESM2 had a better performance, being ESM2 the best method for mf, according to AUPR metric.
The best method in all the other cases vas consensus, which uses information of all the previously mentioned techniques

Selected references

You R, Yao S, Xiong Y, et al. NetGO: improving large-scale protein function prediction with massive network information[J]. Nucleic acids research, 2019, 47(W1): W379-W387.

Wenkang Wang, Yunyan Shuai, Qirong Yang, Fuhao Zhang, Min Zeng, Min Li, A comprehensive computational benchmark for evaluating deep learning-based protein function prediction approaches, Briefings in Bioinformatics, Volume 25, Issue 2, March 2024, bbae050, https://doi.org/10.1093/bib/bbae050

Further Information:
lucia.chechic@gmail.com,
alban.malaj0406@gmail.com,
haolan.yang2002@gmail.com