AI驱动的生物分子功能注释

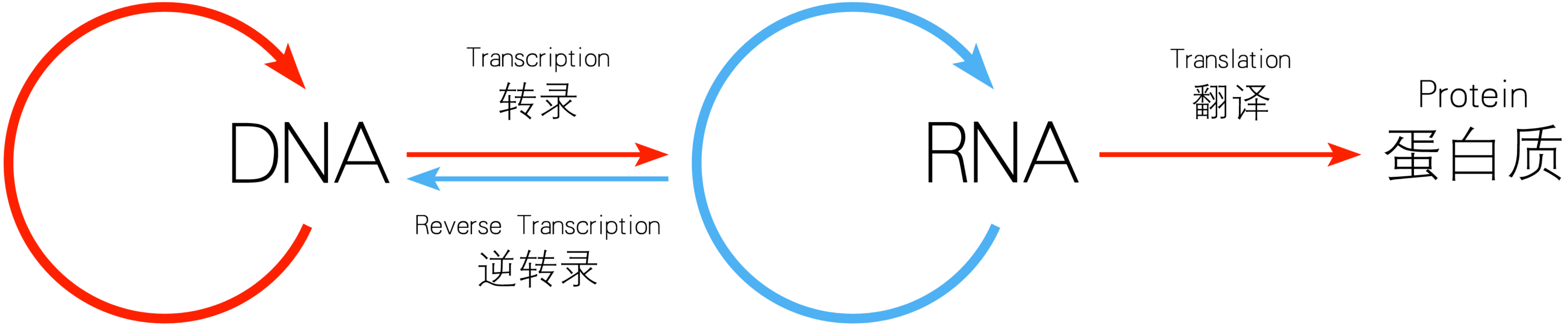
南京农业大学 智慧农业学院（人工智能学院）

汇报人：朱一亨

2025年11月14日

研究背景

- 生物分子（如蛋白质、RNA和DNA）的功能注释是理解生命系统的关键。
- 传统实验注释方法（如酶活性测定、突变实验）耗时、昂贵且无法覆盖所有分子。
- 大规模多组学数据（基因组、转录组、蛋白质组）为AI驱动的生物分子功能预测模型提供了数据基础。



中心法则

研究内容

01 蛋白质功能预测

02 基因功能预测

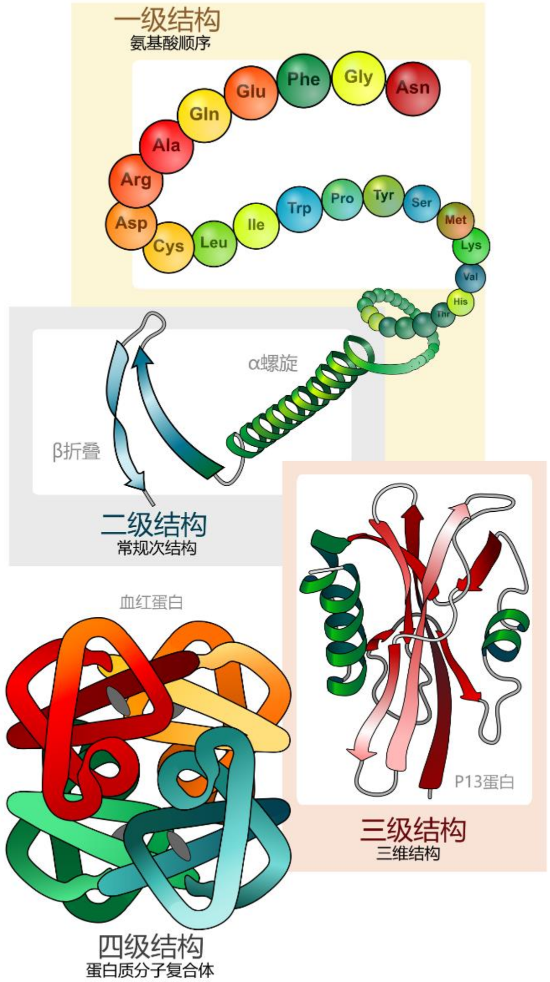
03 蛋白质-配体相互作用预测

04 蛋白质结晶倾向性预测

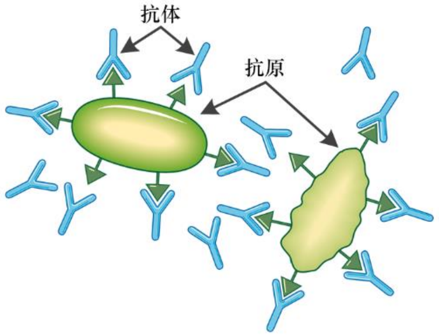
01 Part one

蛋白质功能预测

01 蛋白质的生物功能

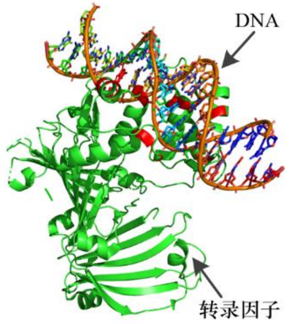


(a) 催化反应



(b) 免疫保护

(c) 运输载体



(d) 基因调控

- ➢ 识别和分析蛋白质的功能有助于解释各种生命活动现象，并阐明相关疾病的发病机理，进而指导相应的药物设计，以期推动智能医疗的发展。

- ➢ 蛋白质功能注释是后基因时代的首要任务之一。

02 蛋白质的功能注释方法

- ➢ 基因本体论 (Gene Ontology, GO)

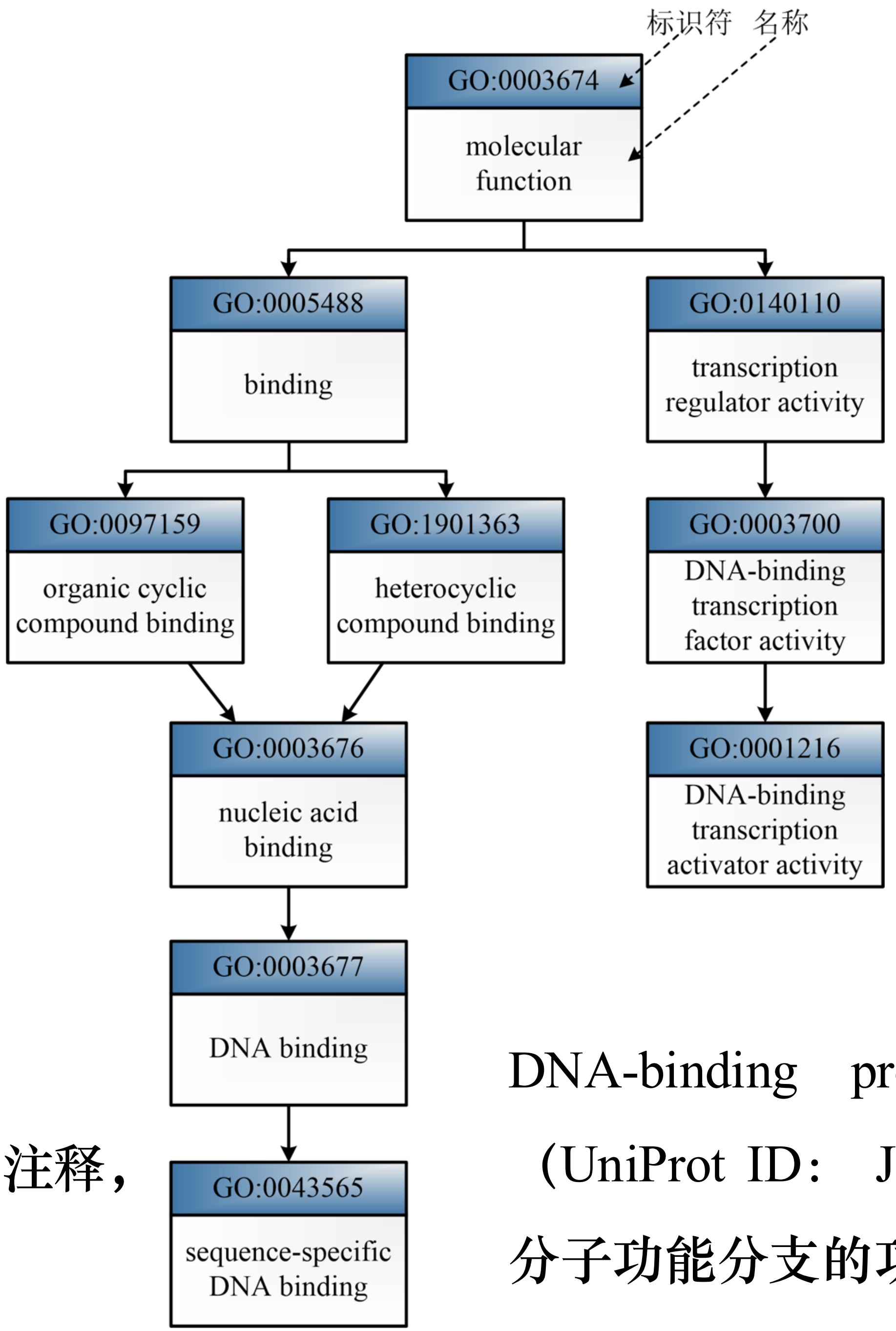
  分子功能 (Molecular Function, MF)

  生物过程 (Biological Process, BP)

  细胞组件 (Cellular Component, CC)

- ➢ 蛋白质功能注释目标

  用GO术语对蛋白质在三个分支下分别进行功能注释，形成三张有向无环图。

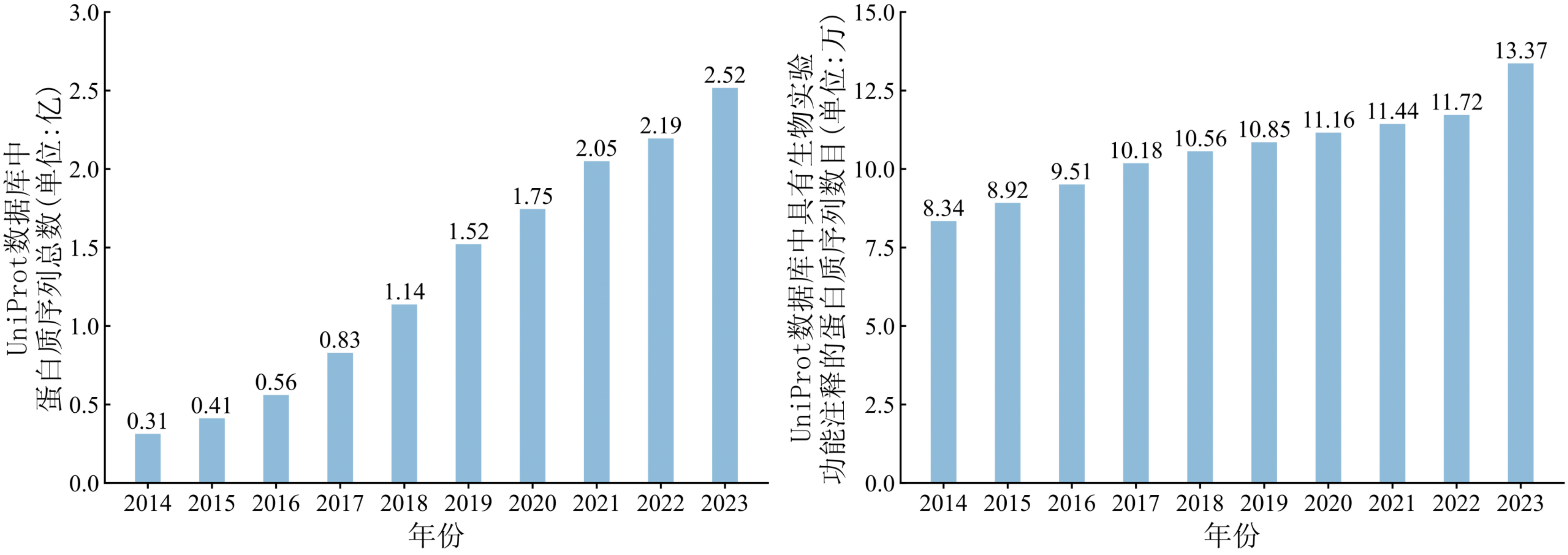


标识符
名称
GO:0003674
molecular function
GO:0005488
binding
GO:0140110
transcription regulator activity
GO:0097159
organic cyclic compound binding
GO:1901363
heterocyclic compound binding
GO:0003700
DNA-binding transcription factor activity
GO:0003676
nucleic acid binding
GO:0001216
DNA-binding transcription activator activity
GO:0003677
DNA binding
GO:0043565
sequence-specific DNA binding

DNA-binding protein LIV4 （UniProt ID： J9VW97）在分子功能分支的功能注释图

03 传统生物实验方法注释蛋白质功能进展缓慢

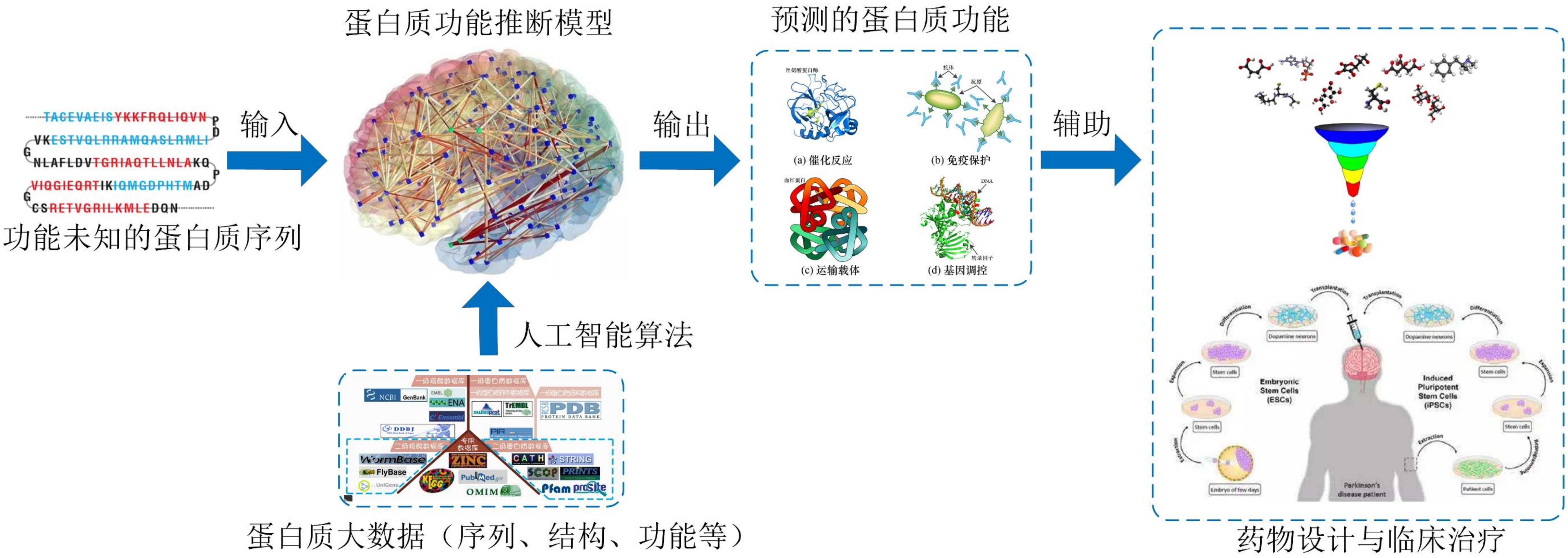
➢ 蛋白质功能注释最可靠的途径是生物实验，但它存在周期长、成本高等缺陷。



UniProt数据库中序列总数和具有生物实验的功能注释的序列数目在近10年的增长趋势图

04 AI算法预测蛋白质功能

- ➢ 研发高效的AI算法来预测蛋白质功能已迫在眉睫。



蛋白质功能推断模型

预测的蛋白质功能

输入

输出

辅助

功能未知的蛋白质序列

人工智能算法

蛋白质大数据（序列、结构、功能等）

药物设计与临床治疗

AI算法预测蛋白质功能框架图

05 蛋白质功能预测研究进展

[1] Yi-Heng Zhu, Chengxin Zhang, Rucheng Diao, Xiaogen Zhou, Peter Freddolino*, Dong-Jun Yu*, Yang Zhang*. MetaGOPlus: *Improving Gene Ontology Prediction of Proteins Using Deep Residual Network with Hierarchical Classification*. **The 28th Conference on Intelligent Systems for Molecular Biology,** 2020.

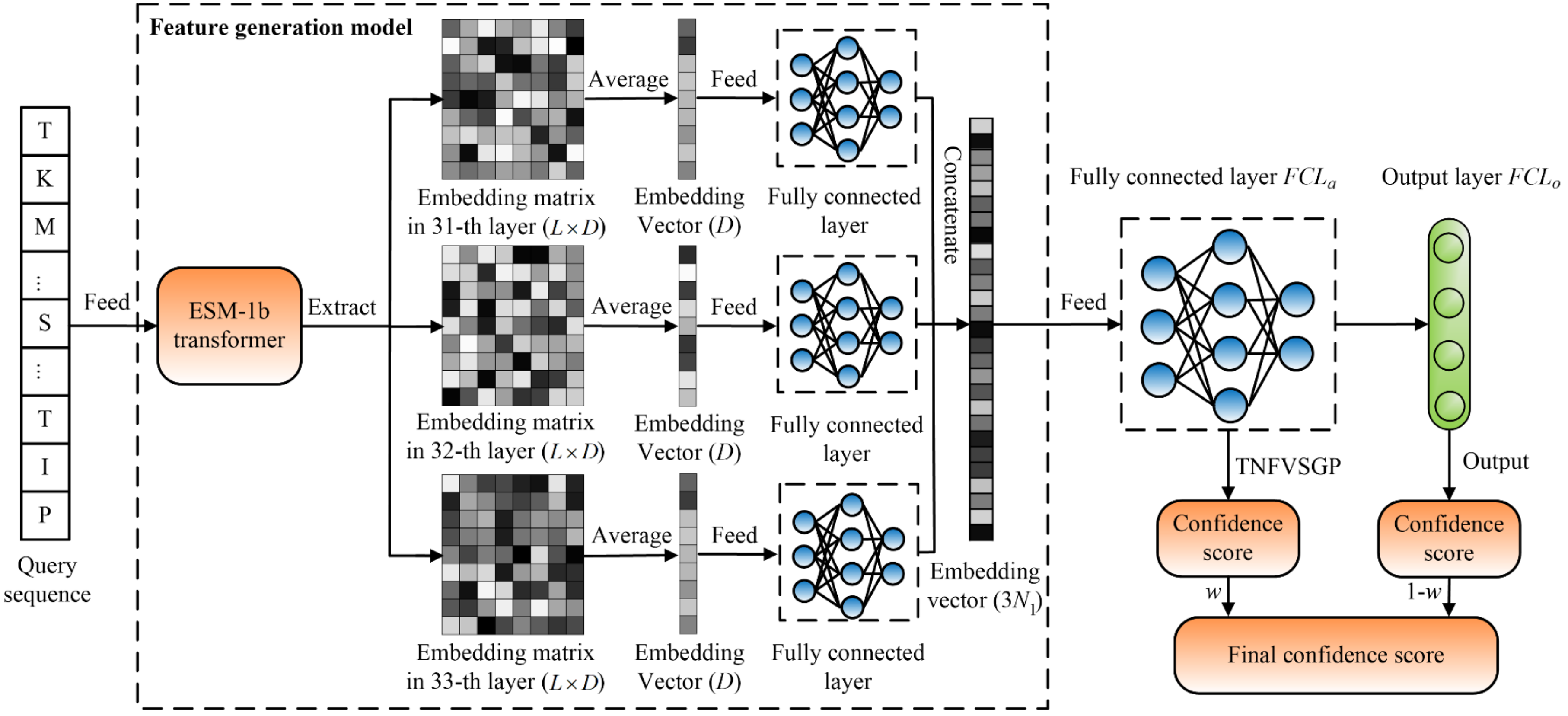
[2] Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu*, Yang Zhang*. *Integrating Unsupervised Language Model with Triplet Neural Networks for Protein Gene Ontology Prediction*. **PLOS Computational Biology**, 2022. (*)

[3] Yi-Heng Zhu, Shuxin Zhu, Xuan Yu, He Yan, Yan Liu, Xiaojun Xie, Dong-Jun Yu*, Rui Ye*. *MKFGO: Integrating Multi-Source Knowledge Fusion with Pre-Trained Language Model for High-Accuracy Protein Function Prediction*. **Briefings in Bioinformatics**, 2025. (*)

[4] Yi-Heng Zhu, Zi Liu, Yu Ding, Zhiwei Ji*, Dong-Jun Yu*. *Machine Learning for Protein Function Prediction*. In: Kurgan, L., Kihara, D. (eds) Protein Function Prediction. **Methods in Molecular Biology**, vol 2947. Humana, New York, NY, 2025.

06 基于无监督语言模型与三元组神经网络的蛋白质功能预测方法ATGO

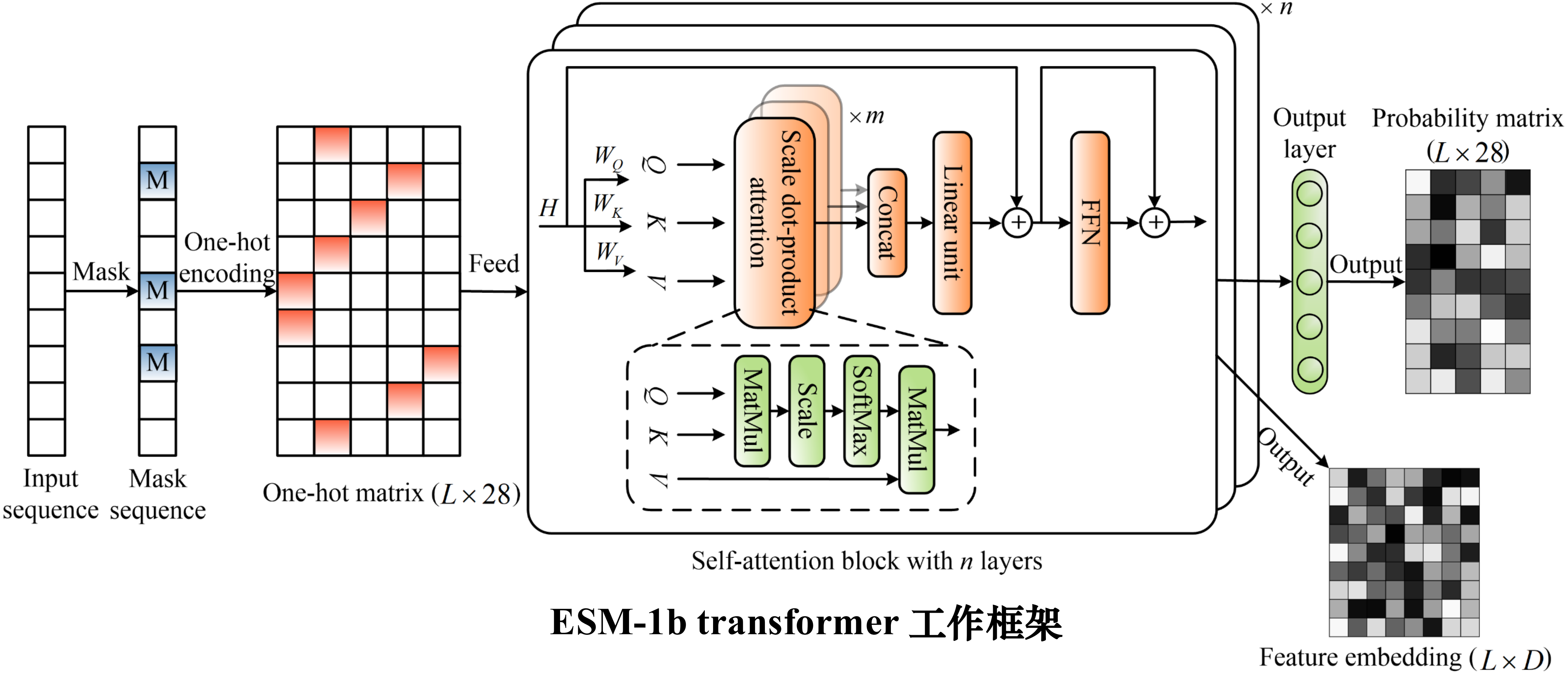
- 主要贡献：首次将计算机视觉领域的无监督语言模型迁移到蛋白质功能预测领域。



ATGO 工作框架

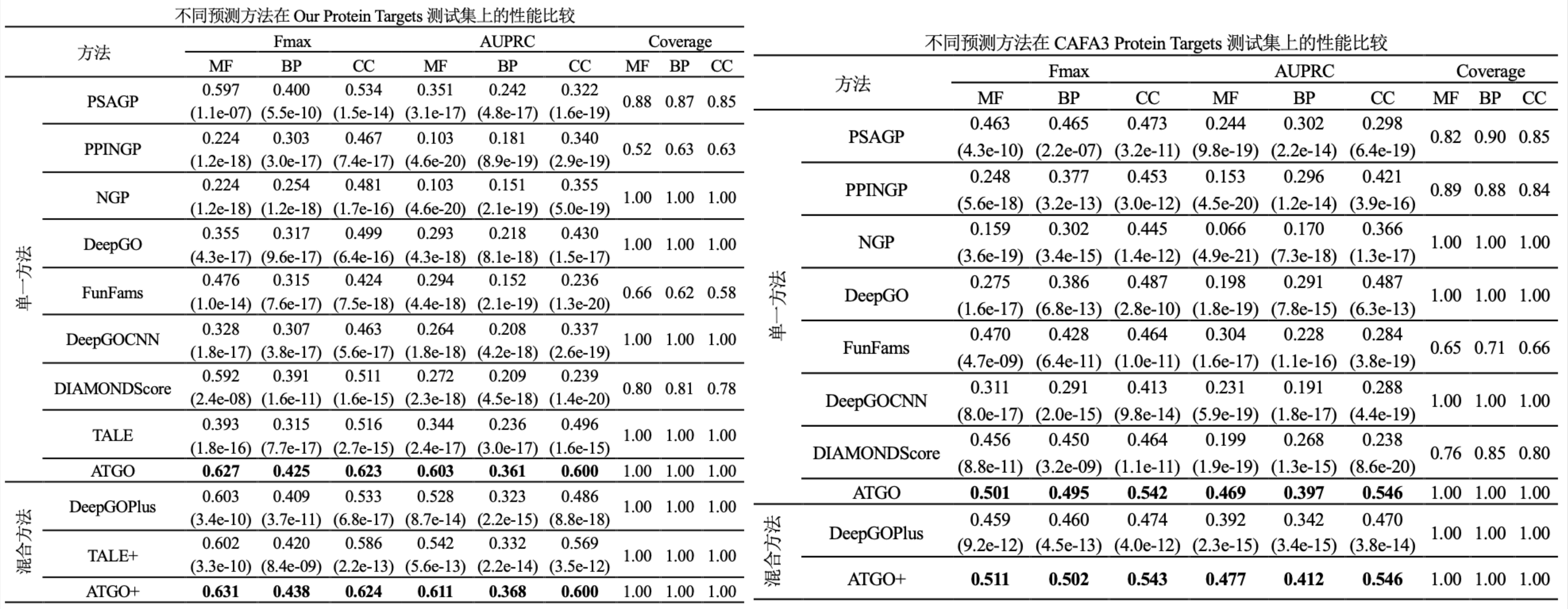
Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu, Yang Zhang. Integrating Unsupervised Language Model with Triplet Neural Networks for Protein Gene Ontology Prediction. **PLOS Computational Biology**. 2022.

06 基于无监督语言模型与三元组神经网络的蛋白质功能预测方法ATGO



ESM-1b transformer 工作框架

06 基于无监督语言模型与三元组神经网络的蛋白质功能预测方法ATGO



不同预测方法在 Our Protein Targets 测试集上的性能比较

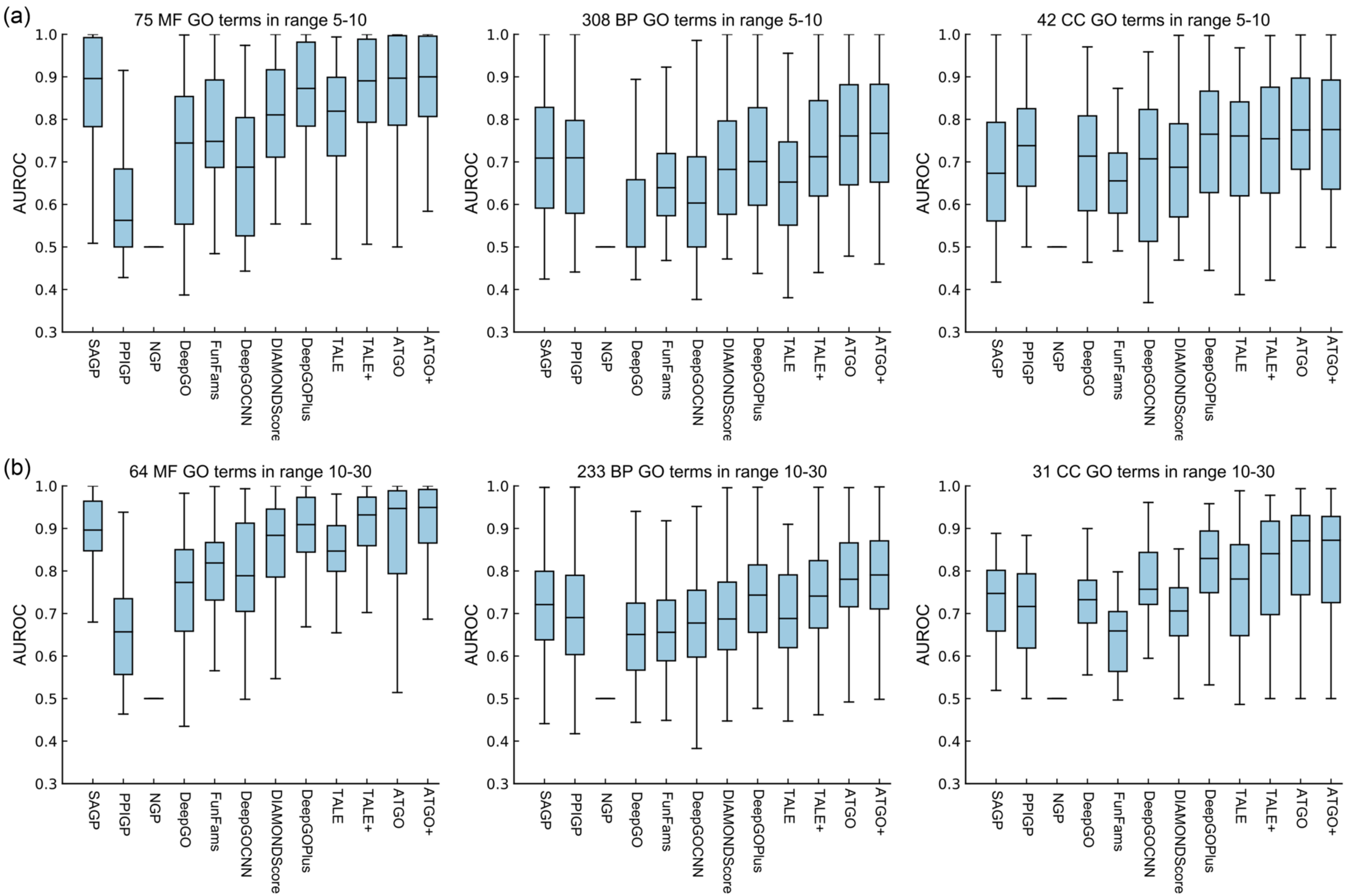
	方法	Fmax			AUPRC			Coverage		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
单一方法	PSAGP	0.597 (1.1e-07)	0.400 (5.5e-10)	0.534 (1.5e-14)	0.351 (3.1e-17)	0.242 (4.8e-17)	0.322 (1.6e-19)	0.88	0.87	0.85
	PPINGP	0.224 (1.2e-18)	0.303 (3.0e-17)	0.467 (7.4e-17)	0.103 (4.6e-20)	0.181 (8.9e-19)	0.340 (2.9e-19)	0.52	0.63	0.63
	NGP	0.224 (1.2e-18)	0.254 (1.2e-18)	0.481 (1.7e-16)	0.103 (4.6e-20)	0.151 (2.1e-19)	0.355 (5.0e-19)	1.00	1.00	1.00
	DeepGO	0.355 (4.3e-17)	0.317 (9.6e-17)	0.499 (6.4e-16)	0.293 (4.3e-18)	0.218 (8.1e-18)	0.430 (1.5e-17)	1.00	1.00	1.00
	FunFams	0.476 (1.0e-14)	0.315 (7.6e-17)	0.424 (7.5e-18)	0.294 (4.4e-18)	0.152 (2.1e-19)	0.236 (1.3e-20)	0.66	0.62	0.58
	DeepGOCNN	0.328 (1.8e-17)	0.307 (3.8e-17)	0.463 (5.6e-17)	0.264 (1.8e-18)	0.208 (4.2e-18)	0.337 (2.6e-19)	1.00	1.00	1.00
	DIAMONDScore	0.592 (2.4e-08)	0.391 (1.6e-11)	0.511 (1.6e-15)	0.272 (2.3e-18)	0.209 (4.5e-18)	0.239 (1.4e-20)	0.80	0.81	0.78
	TALE	0.393 (1.8e-16)	0.315 (7.7e-17)	0.516 (2.7e-15)	0.344 (2.4e-17)	0.236 (3.0e-17)	0.496 (1.6e-15)	1.00	1.00	1.00
	ATGO	**0.627**	**0.425**	**0.623**	**0.603**	**0.361**	**0.600**	1.00	1.00	1.00
混合方法	DeepGOPlus	0.603 (3.4e-10)	0.409 (3.7e-11)	0.533 (6.8e-17)	0.528 (8.7e-14)	0.323 (2.2e-15)	0.486 (8.8e-18)	1.00	1.00	1.00
	TALE+	0.602 (3.3e-10)	0.420 (8.4e-09)	0.586 (2.2e-13)	0.542 (5.6e-13)	0.332 (2.2e-14)	0.569 (3.5e-12)	1.00	1.00	1.00
	ATGO+	**0.631**	**0.438**	**0.624**	**0.611**	**0.368**	**0.600**	1.00	1.00	1.00

不同预测方法在 CAFA3 Protein Targets 测试集上的性能比较

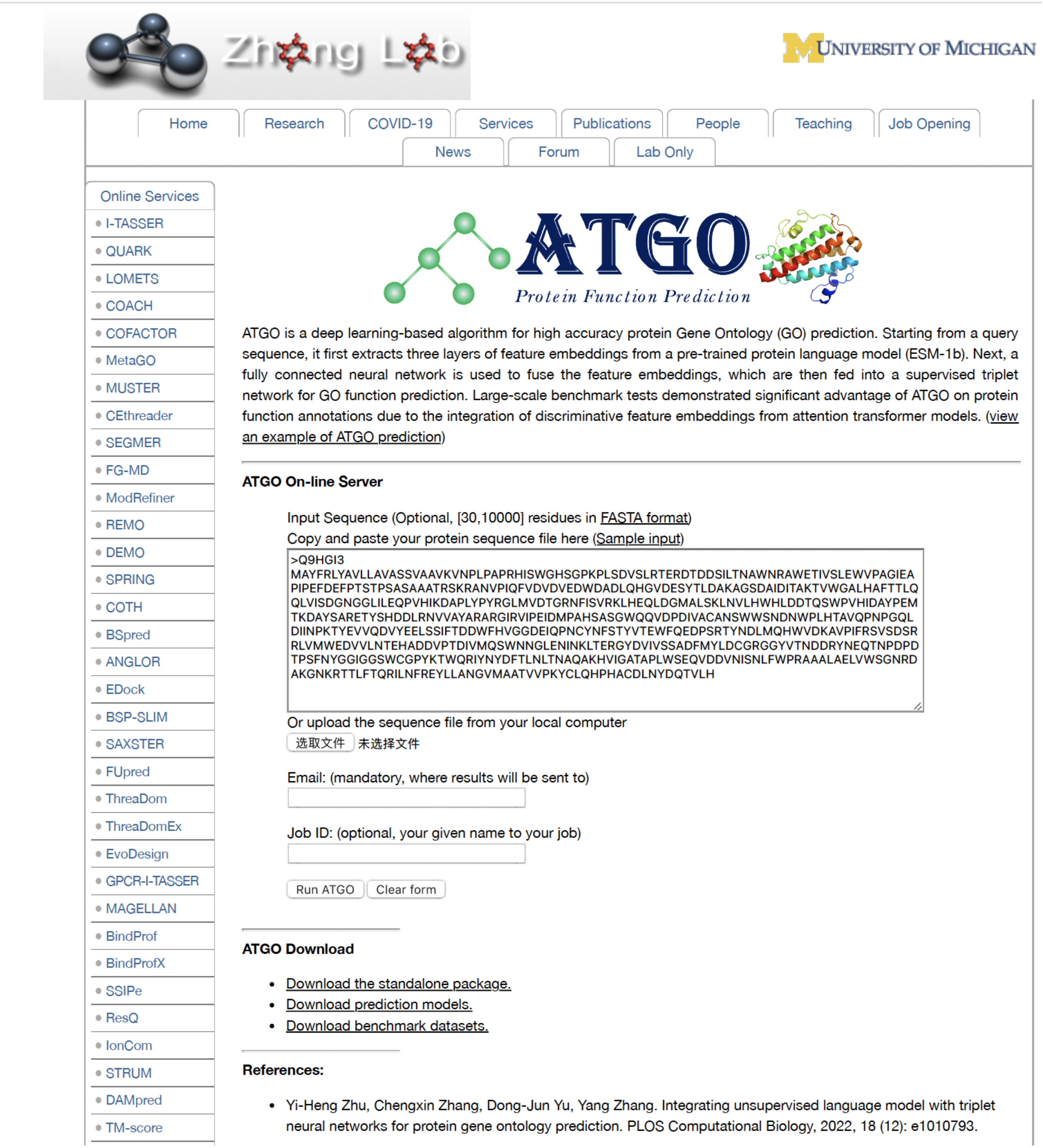
	方法	Fmax			AUPRC			Coverage		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
单一方法	PSAGP	0.463 (4.3e-10)	0.465 (2.2e-07)	0.473 (3.2e-11)	0.244 (9.8e-19)	0.302 (2.2e-14)	0.298 (6.4e-19)	0.82	0.90	0.85
	PPINGP	0.248 (5.6e-18)	0.377 (3.2e-13)	0.453 (3.0e-12)	0.153 (4.5e-20)	0.296 (1.2e-14)	0.421 (3.9e-16)	0.89	0.88	0.84
	NGP	0.159 (3.6e-19)	0.302 (3.4e-15)	0.445 (1.4e-12)	0.066 (4.9e-21)	0.170 (7.3e-18)	0.366 (1.3e-17)	1.00	1.00	1.00
	DeepGO	0.275 (1.6e-17)	0.386 (6.8e-13)	0.487 (2.8e-10)	0.198 (1.8e-19)	0.291 (7.8e-15)	0.487 (6.3e-13)	1.00	1.00	1.00
	FunFams	0.470 (4.7e-09)	0.428 (6.4e-11)	0.464 (1.0e-11)	0.304 (1.6e-17)	0.228 (1.1e-16)	0.284 (3.8e-19)	0.65	0.71	0.66
	DeepGOCNN	0.311 (8.0e-17)	0.291 (2.0e-15)	0.413 (9.8e-14)	0.231 (5.9e-19)	0.191 (1.8e-17)	0.288 (4.4e-19)	1.00	1.00	1.00
	DIAMONDScore	0.456 (8.8e-11)	0.450 (3.2e-09)	0.464 (1.1e-11)	0.199 (1.9e-19)	0.268 (1.3e-15)	0.238 (8.6e-20)	0.76	0.85	0.80
	ATGO	**0.501**	**0.495**	**0.542**	**0.469**	**0.397**	**0.546**	1.00	1.00	1.00
混合方法	DeepGOPlus	0.459 (9.2e-12)	0.460 (4.5e-13)	0.474 (4.0e-12)	0.392 (2.3e-15)	0.342 (3.4e-15)	0.470 (3.8e-14)	1.00	1.00	1.00
	ATGO+	**0.511**	**0.502**	**0.543**	**0.477**	**0.412**	**0.546**	1.00	1.00	1.00

ATGO和SOTA方法在不同数据集上的性能比较

06 基于无监督语言模型与三元组神经网络的蛋白质功能预测方法ATGO



ATGO与SOTA方法在稀有GO术语上的预测性能比较



ATGO is a deep learning-based algorithm for high accuracy protein Gene Ontology (GO) prediction. Starting from a query sequence, it first extracts three layers of feature embeddings from a pre-trained protein language model (ESM-1b). Next, a fully connected neural network is used to fuse the feature embeddings, which are then fed into a supervised triplet network for GO function prediction. Large-scale benchmark tests demonstrated significant advantage of ATGO on protein function annotations due to the integration of discriminative feature embeddings from attention transformer models. (view an example of ATGO prediction)

ATGO On-line Server

Input Sequence (Optional, [30,10000] residues in FASTA format)

Copy and paste your protein sequence file here (Sample input)

```
>Q9HGI3
MAYFRLYAVLLAVASSVAAVKVNPLPAPRHISWGHSGPKPLSDVSLRTERDTDDSILTNAWNRAWETIVSLEWVPAGIEA
PIPEFDEFPTSTPSASAAATRSKRANVPIQFVDVDVEDWDADLQHGVDESYTLDAKAGSDAIDITAKTVWGALHAFTTLQ
QLVISDGNGGLILEQPVHIKDAPLYPYRGLMVDTGRNFISVRKLHEQLDGMALSKLNVLHWHLDDTQSWPVHIDAYPEM
TKDAYSARETYSHDDLRNVVAYARARGIRVIPEIDMPAHSASGWQQVDPDIVACANSWWSNDNWPLHTAVQPNPGQL
DIINPKTYEVVQDVYEELSSIFTDDWFHVGGDEIQPNCYNFSTYVTEWFQEDPSRTYNDLMQHWVDKAVPIFRSVSDSR
RLVMWEDVVLNTEHADDVPTDIVMQSWNNGLENINKLTERGYDVIVSSADFMYLDCGRGGYVTNDDRYNEQTNPDPD
TPSFNYGGIGGSWCGPYKTWQRIYNYDFTLNLTNAQAKHVIGATAPLWSEQVDDVNISNLFWPRAAALAELVWSGNRD
AKGNKRTTLFTQRILNFREYLLANGVMAATVVPKYCLQHPHACDLNYDQTVLH
```

Or upload the sequence file from your local computer

选取文件 未选择文件

Email: (mandatory, where results will be sent to)

Job ID: (optional, your given name to your job)

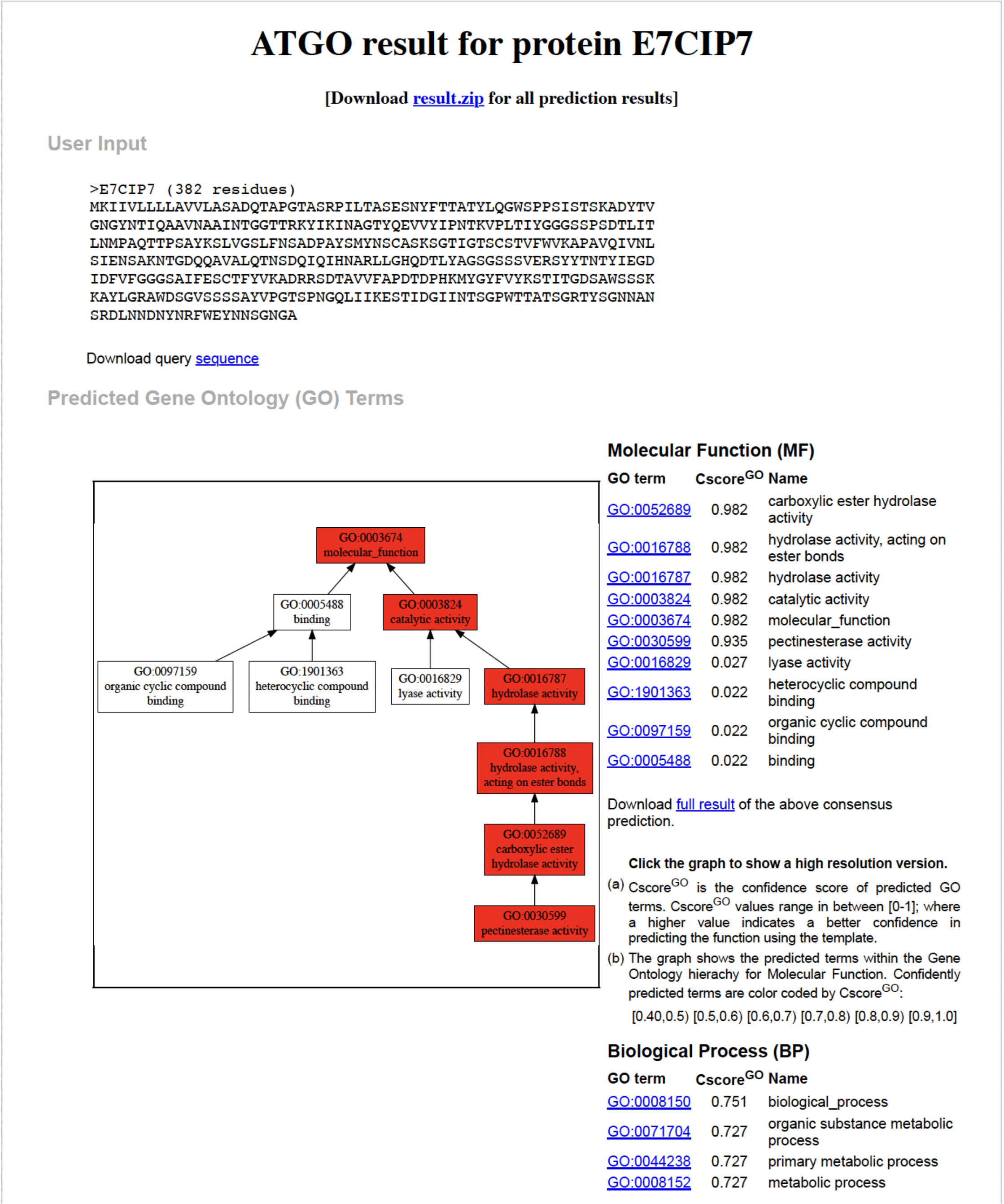
Run ATGO Clear form

ATGO Download

- Download the standalone package.
- Download prediction models.
- Download benchmark datasets.

References:

- Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu, Yang Zhang. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. PLOS Computational Biology, 2022, 18 (12): e1010793.



ATGO result for protein E7CIP7

[Download result.zip for all prediction results]

User Input

```
>E7CIP7 (382 residues)
MKIIVLLLLAVVLASADQTAPGTASRPILTASESNYFTTATYLQGWSPPSISTSKADYTV
GNGYNTIQAAVNAAINTGGTTRKYIKINAGTYQEVVYIPNTKVPLTIYGGGSSPSDTLIT
LNMPAQTTPSAYKSLVGSLFNSADPAYSMYNSCASKSGTIGTSCSTVFWVKAPAVQIVNL
SIENSAKNTGDQQAVALQTNSDQIQIHNARLLGHQDTLYAGSGSSSVERSYYTNTYIEGD
IDFVFGGGSAIFESCTFYVKADRRSDTAVVFAPDTDPHKMYGYFVYKSTITGDSAWSSSK
KAYLGRAWDSGVSSSSAYVPGTSPNGQLIIKESTIDGIINTSGPWTTATSGRTYSGNNAN
SRDLNNDNYNRFWEYNNSGNGA
```

Download query sequence

Predicted Gene Ontology (GO) Terms

Molecular Function (MF)

GO term	Cscore^{GO}	Name
GO:0052689	0.982	carboxylic ester hydrolase activity
GO:0016788	0.982	hydrolase activity, acting on ester bonds
GO:0016787	0.982	hydrolase activity
GO:0003824	0.982	catalytic activity
GO:0003674	0.982	molecular_function
GO:0030599	0.935	pectinesterase activity
GO:0016829	0.027	lyase activity
GO:1901363	0.022	heterocyclic compound binding
GO:0097159	0.022	organic cyclic compound binding
GO:0005488	0.022	binding

Download full result of the above consensus prediction.

Click the graph to show a high resolution version.

(a) Cscore^{GO} is the confidence score of predicted GO terms. Cscore^{GO} values range in between [0-1]; where a higher value indicates a better confidence in predicting the function using the template.

(b) The graph shows the predicted terms within the Gene Ontology hierachy for Molecular Function. Confidently predicted terms are color coded by Cscore^{GO}:

[0.40,0.5) [0.5,0.6) [0.6,0.7) [0.7,0.8) [0.8,0.9) [0.9,1.0]

Biological Process (BP)

GO term	Cscore^{GO}	Name
GO:0008150	0.751	biological_process
GO:0071704	0.727	organic substance metabolic process
GO:0044238	0.727	primary metabolic process
GO:0008152	0.727	metabolic process

在线预测服务 (https://aideepmed.com/ATGO/)

07 基于多源知识融合与无监督语言模型的蛋白质功能预测方法MKFGO

- 多源数据：

（1）蛋白质序列

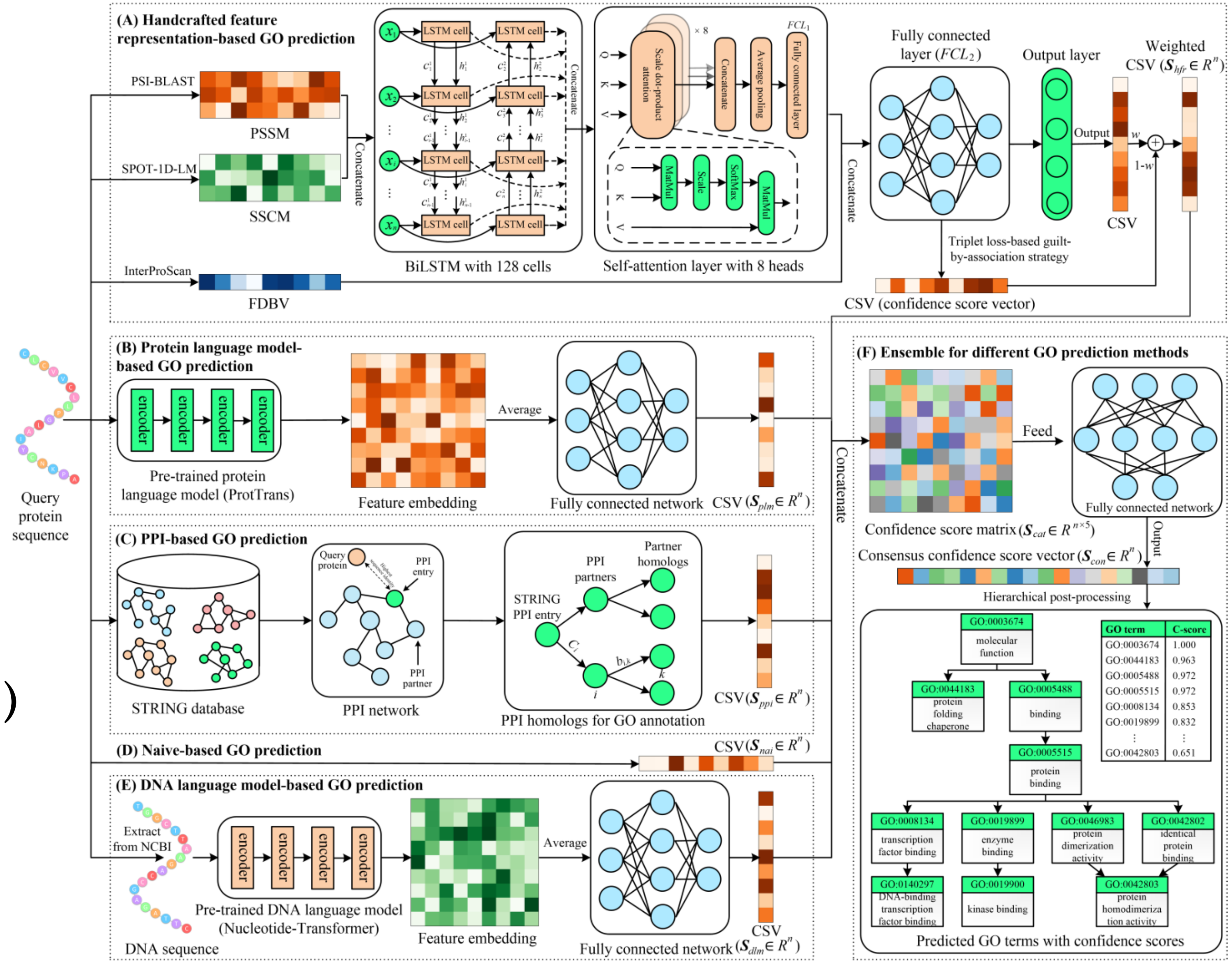
（2）蛋白质相互作用网络

（3）基因序列

- 大语言模型：

（1）蛋白质大语言模型（ProtTrans）

（2）DNA大模型（Nucleotide-Transformer）



(A) Handcrafted feature representation-based GO prediction
PSI-BLAST
PSSM
SPOT-1D-LM
SSCM
InterProScan
FDBV
Concatenate
LSTM cell
BiLSTM with 128 cells
Scale dot-product attention
× 8
Concatenate
Average pooling
Fully connected layer
FCL_1
Q
K
V
MatMul
Scale
SoftMax
MatMul
Self-attention layer with 8 heads
Concatenate
Fully connected layer (FCL_2)
Output layer
Output
CSV
w
1-w
Weighted CSV ($S_{hfr} \in R^n$)
Triplet loss-based guilt-by-association strategy
CSV (confidence score vector)
Query protein sequence
(B) Protein language model-based GO prediction
encoder
Pre-trained protein language model (ProtTrans)
Feature embedding
Average
Fully connected network
CSV ($S_{plm} \in R^n$)
(C) PPI-based GO prediction
STRING database
Query protein
PPI entry
PPI partner
PPI network
STRING PPI entry
PPI partners
Partner homologs
PPI homologs for GO annotation
CSV ($S_{ppi} \in R^n$)
(D) Naive-based GO prediction
CSV ($S_{nai} \in R^n$)
(E) DNA language model-based GO prediction
Extract from NCBI
DNA sequence
encoder
Pre-trained DNA language model (Nucleotide-Transformer)
Feature embedding
Average
Fully connected network
CSV ($S_{dlm} \in R^n$)
Concatenate
(F) Ensemble for different GO prediction methods
Feed
Fully connected network
Confidence score matrix ($S_{cat} \in R^{n \times 5}$)
Output
Consensus confidence score vector ($S_{con} \in R^n$)
Hierarchical post-processing
GO:0003674 molecular function
GO:0044183 protein folding chaperone
GO:0005488 binding
GO:0005515 protein binding
GO:0008134 transcription factor binding
GO:0019899 enzyme binding
GO:0046983 protein dimerization activity
GO:0042802 identical protein binding
GO:0140297 DNA-binding transcription factor binding
GO:0019900 kinase binding
GO:0042803 protein homodimerization activity
GO term | C-score
GO:0003674 | 1.000
GO:0044183 | 0.963
GO:0005488 | 0.972
GO:0005515 | 0.972
GO:0008134 | 0.853
GO:0019899 | 0.832
⋮ | ⋮
GO:0042803 | 0.651
Predicted GO terms with confidence scores

MKFGO工作流程图

Yi-Heng Zhu, Shuxin Zhu, Xuan Yu, He Yan, Yan Liu, Xiaojun Xie, Dong-Jun Yu*, Rui Ye*. *MKFGO: Integrating Multi-Source Knowledge Fusion with Pre-Trained Language Model for High-Accuracy Protein Function Prediction*. **Briefings in Bioinformatics**, 2025.

07 基于多源知识融合与无监督语言模型的蛋白质功能预测方法MKFGO

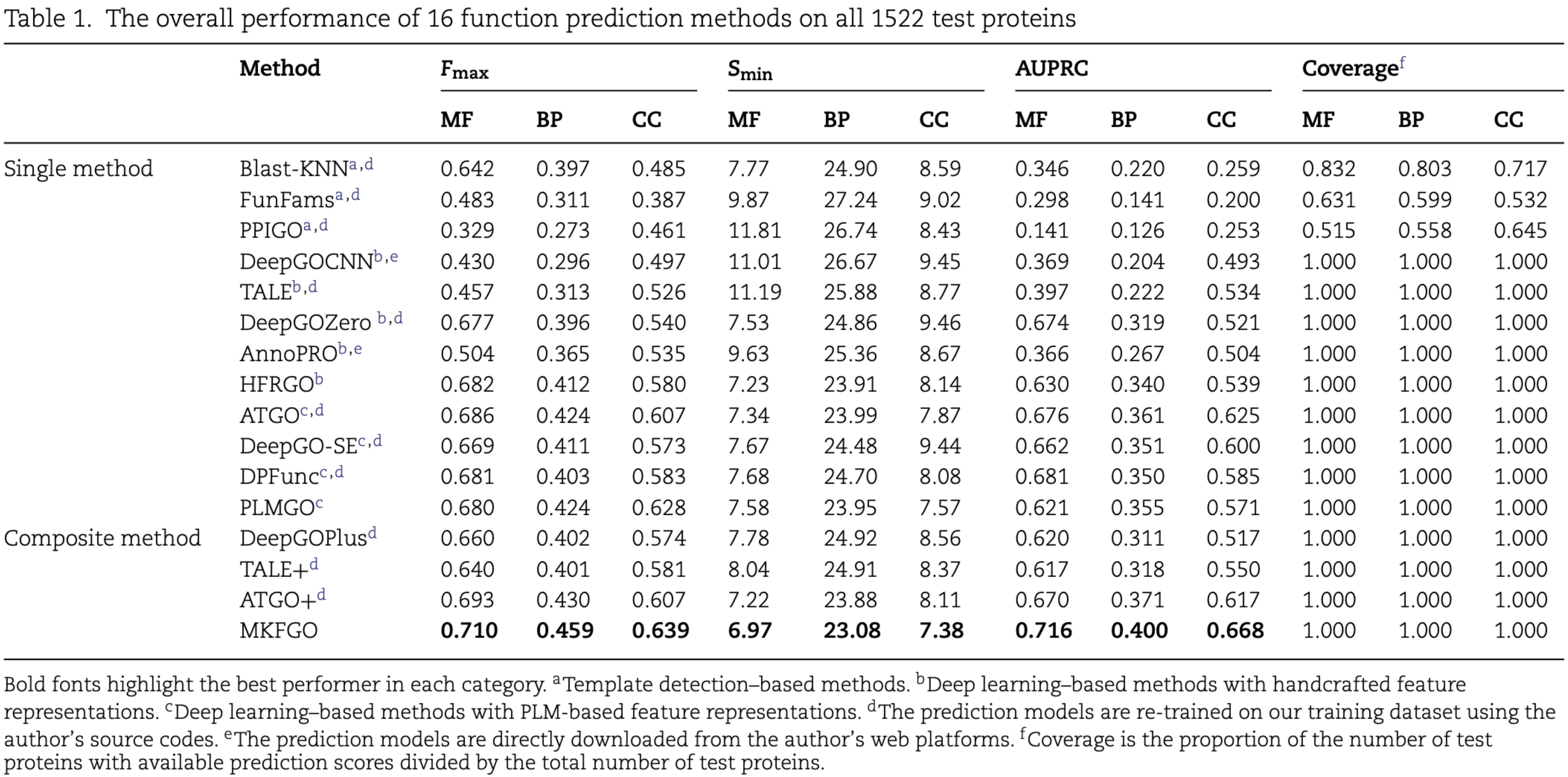


Table 1. The overall performance of 16 function prediction methods on all 1522 test proteins

	Method	F_{max}			S_{min}			AUPRC			Coverage[f]		
		MF	BP	CC	MF	BP	CC	MF	BP	CC	MF	BP	CC
Single method	Blast-KNN[a,d]	0.642	0.397	0.485	7.77	24.90	8.59	0.346	0.220	0.259	0.832	0.803	0.717
	FunFams[a,d]	0.483	0.311	0.387	9.87	27.24	9.02	0.298	0.141	0.200	0.631	0.599	0.532
	PPIGO[a,d]	0.329	0.273	0.461	11.81	26.74	8.43	0.141	0.126	0.253	0.515	0.558	0.645
	DeepGOCNN[b,e]	0.430	0.296	0.497	11.01	26.67	9.45	0.369	0.204	0.493	1.000	1.000	1.000
	TALE[b,d]	0.457	0.313	0.526	11.19	25.88	8.77	0.397	0.222	0.534	1.000	1.000	1.000
	DeepGOZero [b,d]	0.677	0.396	0.540	7.53	24.86	9.46	0.674	0.319	0.521	1.000	1.000	1.000
	AnnoPRO[b,e]	0.504	0.365	0.535	9.63	25.36	8.67	0.366	0.267	0.504	1.000	1.000	1.000
	HFRGO[b]	0.682	0.412	0.580	7.23	23.91	8.14	0.630	0.340	0.539	1.000	1.000	1.000
	ATGO[c,d]	0.686	0.424	0.607	7.34	23.99	7.87	0.676	0.361	0.625	1.000	1.000	1.000
	DeepGO-SE[c,d]	0.669	0.411	0.573	7.67	24.48	9.44	0.662	0.351	0.600	1.000	1.000	1.000
	DPFunc[c,d]	0.681	0.403	0.583	7.68	24.70	8.08	0.681	0.350	0.585	1.000	1.000	1.000
	PLMGO[c]	0.680	0.424	0.628	7.58	23.95	7.57	0.621	0.355	0.571	1.000	1.000	1.000
Composite method	DeepGOPlus[d]	0.660	0.402	0.574	7.78	24.92	8.56	0.620	0.311	0.517	1.000	1.000	1.000
	TALE+[d]	0.640	0.401	0.581	8.04	24.91	8.37	0.617	0.318	0.550	1.000	1.000	1.000
	ATGO+[d]	0.693	0.430	0.607	7.22	23.88	8.11	0.670	0.371	0.617	1.000	1.000	1.000
	MKFGO	**0.710**	**0.459**	**0.639**	**6.97**	**23.08**	**7.38**	**0.716**	**0.400**	**0.668**	1.000	1.000	1.000

Bold fonts highlight the best performer in each category. [a]Template detection–based methods. [b]Deep learning–based methods with handcrafted feature representations. [c]Deep learning–based methods with PLM-based feature representations. [d]The prediction models are re-trained on our training dataset using the author's source codes. [e]The prediction models are directly downloaded from the author's web platforms. [f]Coverage is the proportion of the number of test proteins with available prediction scores divided by the total number of test proteins.

MKFGO和SOTA方法在1522个测试蛋白质上的性能比较

07 基于多源知识融合与无监督语言模型的蛋白质功能预测方法MKFGO

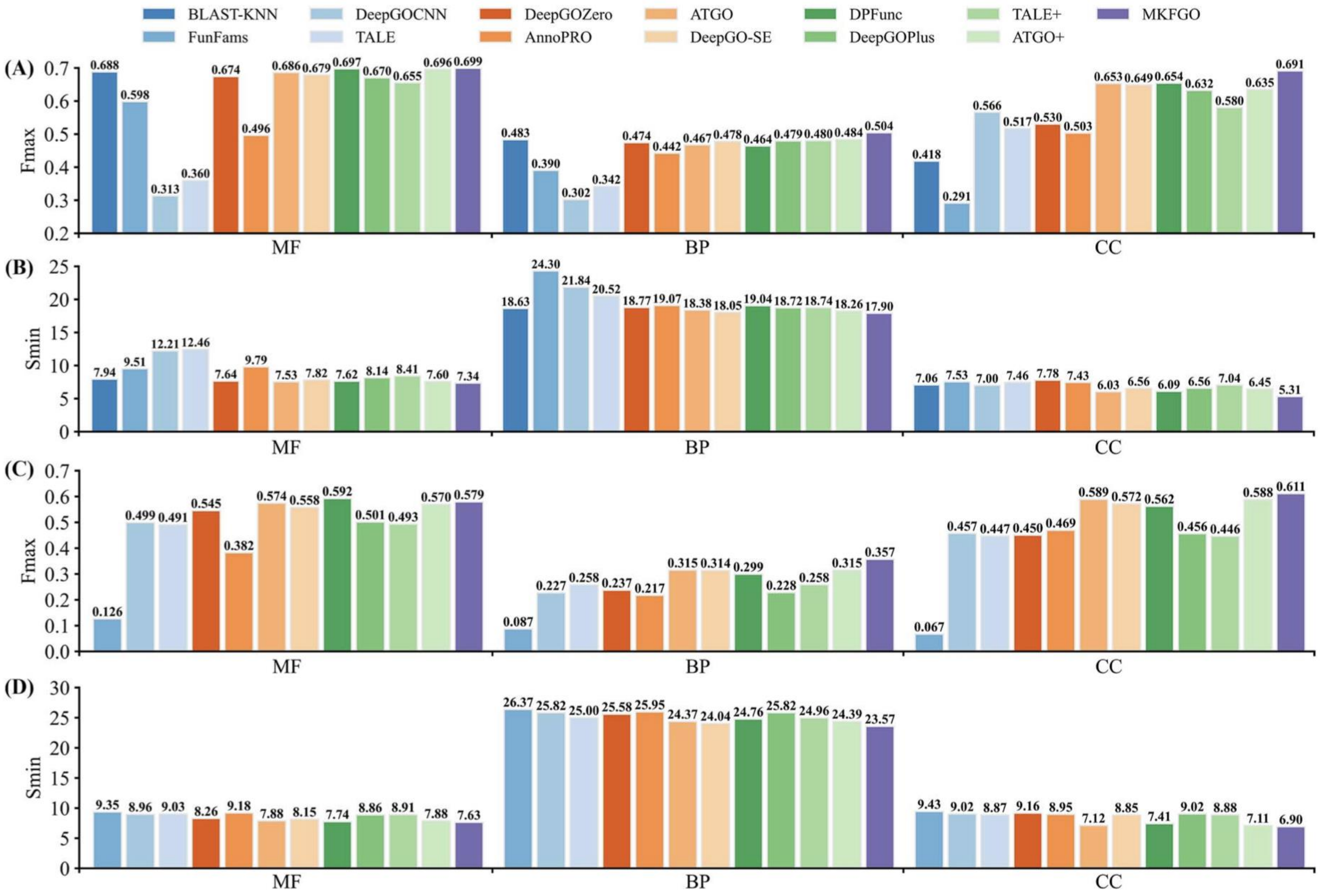
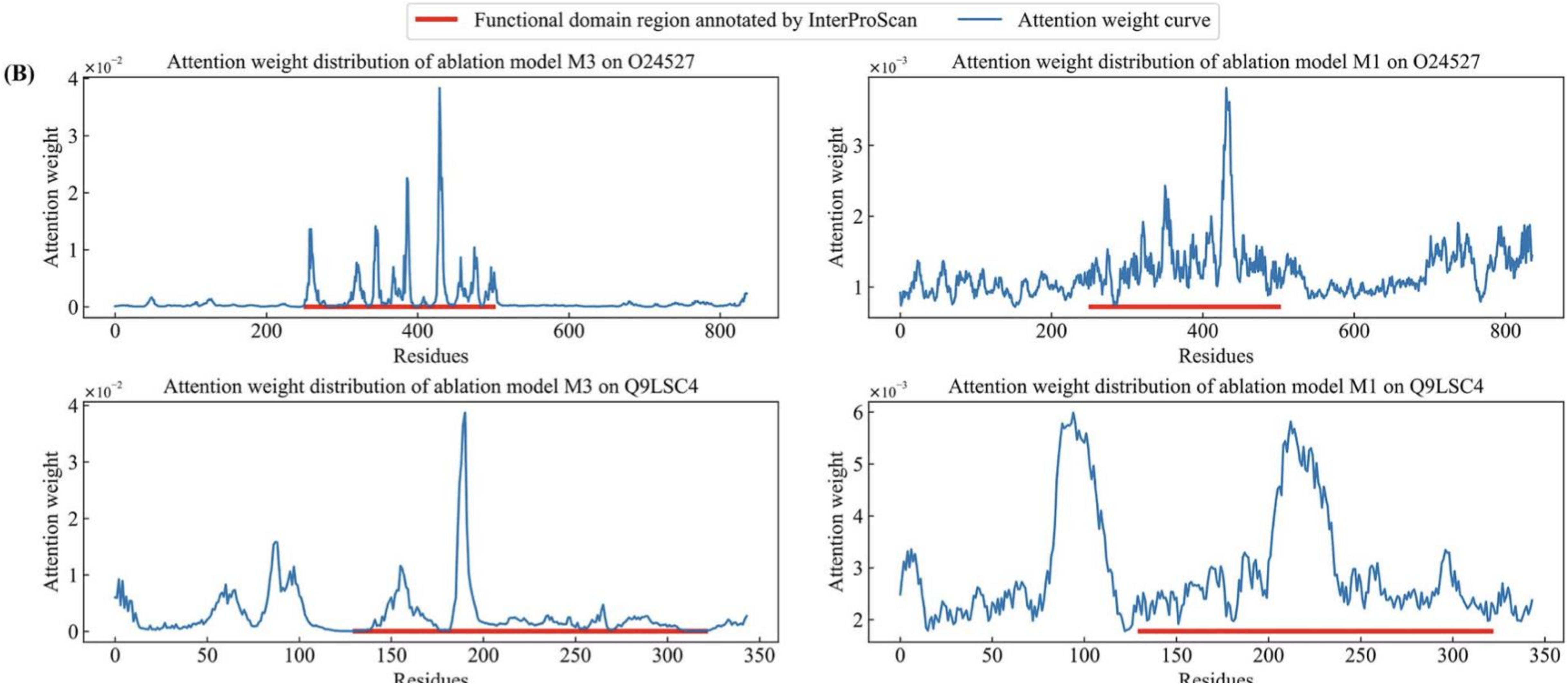


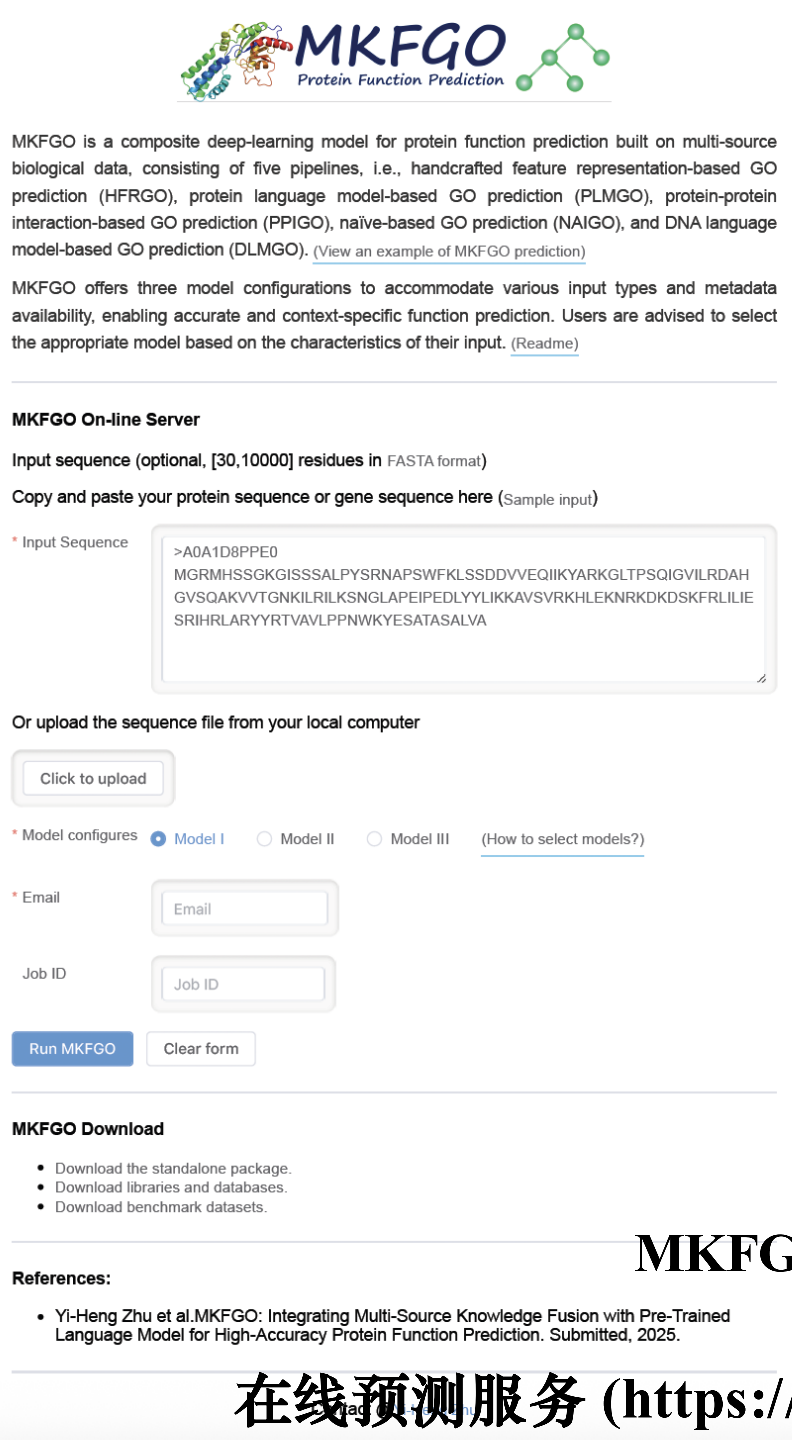
Figure 2. The performance comparison among 13 function prediction methods on the new species and nonhomology proteins across three GO aspects. (A) The F_{max} values on the 300 test proteins from 158 new species. (B) The S_{min} values on the 300 new species proteins. (C) The F_{max} values on the 305 non-homologous test proteins. (D) The S_{min} values on the 305 non-homologous proteins.

MKFGO和SOTA方法在新物种蛋白质和非同源蛋白质上的性能比较

07 基于多源知识融合与无监督语言模型的蛋白质功能预测方法MKFGO



自注意力机制的生物可解释性分析



MKFGO is a composite deep-learning model for protein function prediction built on multi-source biological data, consisting of five pipelines, i.e., handcrafted feature representation-based GO prediction (HFRGO), protein language model-based GO prediction (PLMGO), protein-protein interaction-based GO prediction (PPIGO), naïve-based GO prediction (NAIGO), and DNA language model-based GO prediction (DLMGO). (View an example of MKFGO prediction)

MKFGO offers three model configurations to accommodate various input types and metadata availability, enabling accurate and context-specific function prediction. Users are advised to select the appropriate model based on the characteristics of their input. (Readme)

MKFGO On-line Server

Input sequence (optional, [30,10000] residues in FASTA format)

Copy and paste your protein sequence or gene sequence here (Sample input)

* Input Sequence

```
>A0A1D8PPE0
MGRMHSSGKGISSSALPYSRNAPSWFKLSSDDVVEQIIKYARKGLTPSQIGVILRDAH
GVSQAKVVTGNKILRILKSNGLAPEIPEDLYYLIKKAVSVRKHLEKNRKDKDSKFRLILIE
SRIHRLARYYRTVAVLPPNWKYESATASALVA
```

Or upload the sequence file from your local computer

Click to upload

* Model configures ● Model I ○ Model II ○ Model III (How to select models?)

* Email: Email

Job ID: Job ID

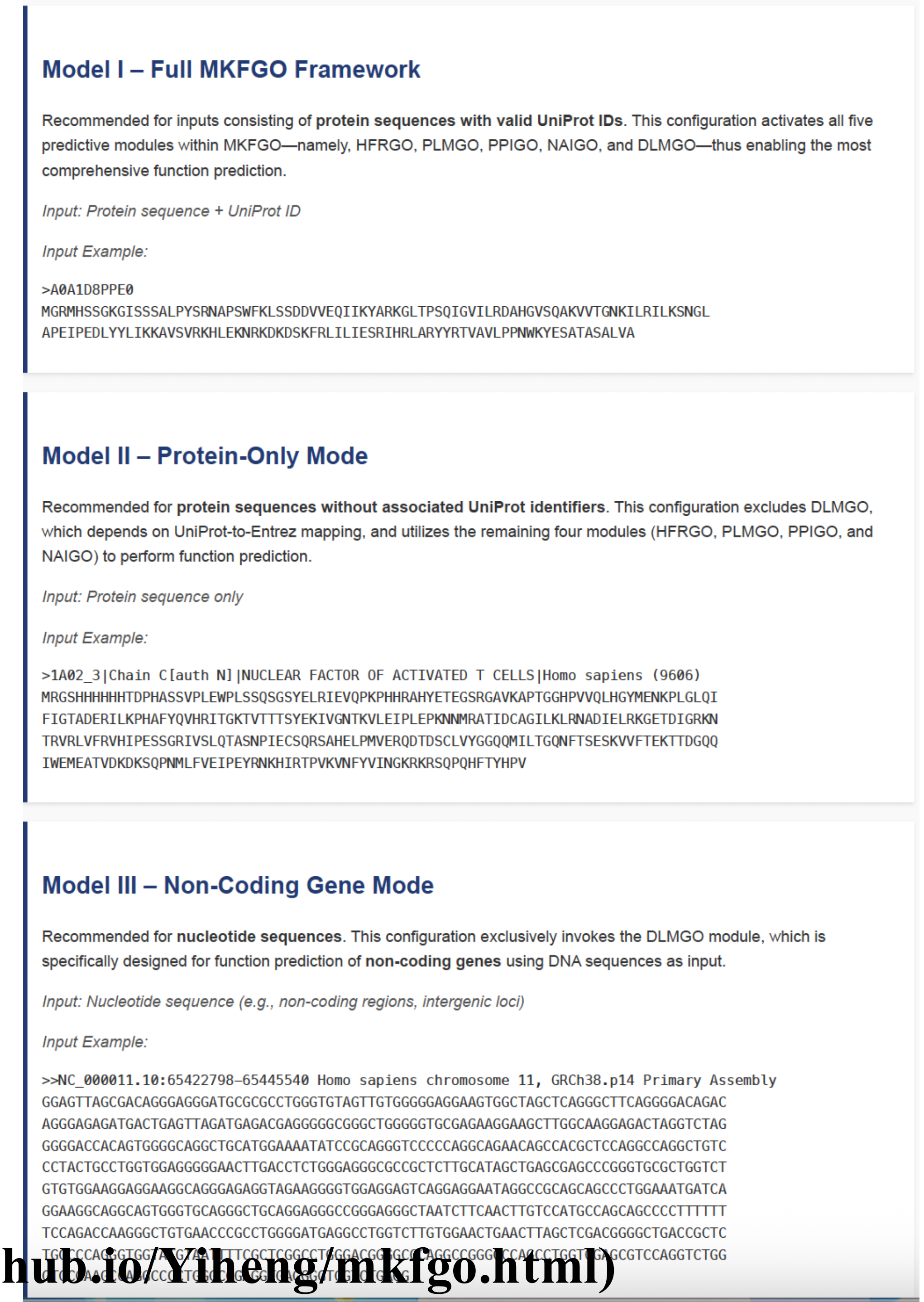
Run MKFGO | Clear form

MKFGO Download

- Download the standalone package.
- Download libraries and databases.
- Download benchmark datasets.

References:

- Yi-Heng Zhu et al.MKFGO: Integrating Multi-Source Knowledge Fusion with Pre-Trained Language Model for High-Accuracy Protein Function Prediction. Submitted, 2025.



Model I – Full MKFGO Framework

Recommended for inputs consisting of **protein sequences with valid UniProt IDs**. This configuration activates all five predictive modules within MKFGO—namely, HFRGO, PLMGO, PPIGO, NAIGO, and DLMGO—thus enabling the most comprehensive function prediction.

Input: Protein sequence + UniProt ID

Input Example:

```
>A0A1D8PPE0
MGRMHSSGKGISSSALPYSRNAPSWFKLSSDDVVEQIIKYARKGLTPSQIGVILRDAHGVSQAKVVTGNKILRILKSNGL
APEIPEDLYYLIKKAVSVRKHLEKNRKDKDSKFRLILIESRIHRLARYYRTVAVLPPNWKYESATASALVA
```

Model II – Protein-Only Mode

Recommended for **protein sequences without associated UniProt identifiers**. This configuration excludes DLMGO, which depends on UniProt-to-Entrez mapping, and utilizes the remaining four modules (HFRGO, PLMGO, PPIGO, and NAIGO) to perform function prediction.

Input: Protein sequence only

Input Example:

```
>1A02_3|Chain C[auth N]|NUCLEAR FACTOR OF ACTIVATED T CELLS|Homo sapiens (9606)
MRGSHHHHHHTDPHASSVPLEWPLSSQSGSYELRIEVQPKPHHRAHYETEGSRGAVKAPTGGHPVVQLHGYMENKPLGLQI
FIGTADERILKPHAFYQVHRITGKTVTTTSYEKIVGNTKVLEIPLEPKNNMRATIDCAGILKLRNADIELRKGETDIGRKN
TRVRLVFRVHIPESSGRIVSLQTASNPIECSQRSAHELPMVERQDTDSCLVYGGQQMILTGQNFTSESKVVFTEKTTDGQQ
IWEMEATVDKDKSQPNMLFVEIPEYRNKHIRTPVKVNFYVINGKRKRSQPQHFTYHPV
```

Model III – Non-Coding Gene Mode

Recommended for **nucleotide sequences**. This configuration exclusively invokes the DLMGO module, which is specifically designed for function prediction of **non-coding genes** using DNA sequences as input.

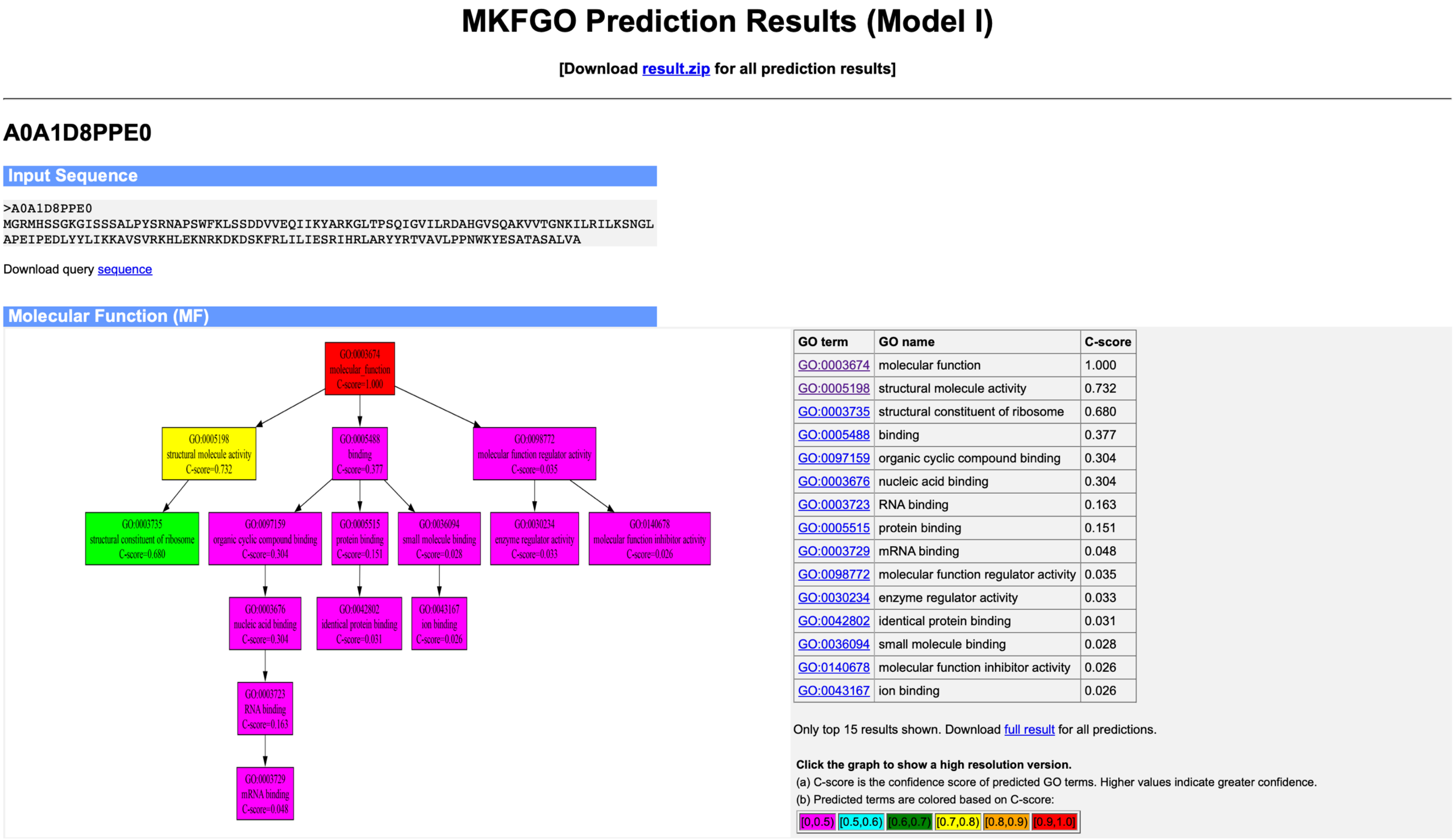
Input: Nucleotide sequence (e.g., non-coding regions, intergenic loci)

Input Example:

```
>>NC_000011.10:65422798-65445540 Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly
GGAGTTAGCGACAGGGAGGGATGCGCGCCTGGGTGTAGTTGTGGGGGAGGAAGTGGCTAGCTCAGGGCTTCAGGGGACAGAC
AGGGAGAGATGACTGAGTTAGATGAGACGAGGGGGCGGGCTGGGGGTGCGAGAAGGAAGCTTGGCAAGGAGACTAGGTCTAG
GGGGACCACAGTGGGGCAGGCTGCATGGAAAATATCCGCAGGGTCCCCCAGGCAGAACAGCCACGCTCCAGGCCAGGCTGTC
CCTACTGCCTGGTGGAGGGGGAACTTGACCTCTGGGAGGGCGCCGCTCTTGCATAGCTGAGCGAGCCCGGGTGCGCTGGTCT
GTGTGGAAGGAGGAAGGCAGGGAGAGGTAGAAGGGGTGGAGGAGTCAGGAGGAATAGGCCGCAGCAGCCCTGGAAATGATCA
GGAAGGCAGGCAGTGGGTGCAGGGCTGCAGGAGGGCCGGGAGGGCTAATCTTCAACTTGTCCATGCCAGCAGCCCCTTTTTT
TCCAGACCAAGGGCTGTGAACCCGCCTGGGGATGAGGCCTGGTCTTGTGGAACTGAACTTAGCTCGACGGGGCTGACCGCTC
TG[illegible]CCCAG[illegible]TGG[illegible]TTTCGCTCGGCCT[illegible]GGACGG[illegible]AGGCCGGG[illegible]CCA[illegible]CTGGT[illegible]GCGTCCAGGTCTGG
[illegible]
```

MKFGO的输入页面

在线预测服务 (https://yiheng-zhu.github.io/Yiheng/mkfgo.html)



MKFGO Prediction Results (Model I)

[Download result.zip for all prediction results]

A0A1D8PPE0

Input Sequence

```
>A0A1D8PPE0
MGRMHSSGKGISSSALPYSRNAPSWFKLSSDDVVEQIIKYARKGLTPSQIGVILRDAHGVSQAKVVTGNKILRILKSNGL
APEIPEDLYYLIKKAVSVRKHLEKNRKDKDSKFRLILIESRIHRLARYYRTVAVLPPNWKYESATASALVA
```

Download query sequence

Molecular Function (MF)

GO term	GO name	C-score
GO:0003674	molecular function	1.000
GO:0005198	structural molecule activity	0.732
GO:0003735	structural constituent of ribosome	0.680
GO:0005488	binding	0.377
GO:0097159	organic cyclic compound binding	0.304
GO:0003676	nucleic acid binding	0.304
GO:0003723	RNA binding	0.163
GO:0005515	protein binding	0.151
GO:0003729	mRNA binding	0.048
GO:0098772	molecular function regulator activity	0.035
GO:0030234	enzyme regulator activity	0.033
GO:0042802	identical protein binding	0.031
GO:0036094	small molecule binding	0.028
GO:0140678	molecular function inhibitor activity	0.026
GO:0043167	ion binding	0.026

Only top 15 results shown. Download full result for all predictions.

Click the graph to show a high resolution version.

(a) C-score is the confidence score of predicted GO terms. Higher values indicate greater confidence.

(b) Predicted terms are colored based on C-score:

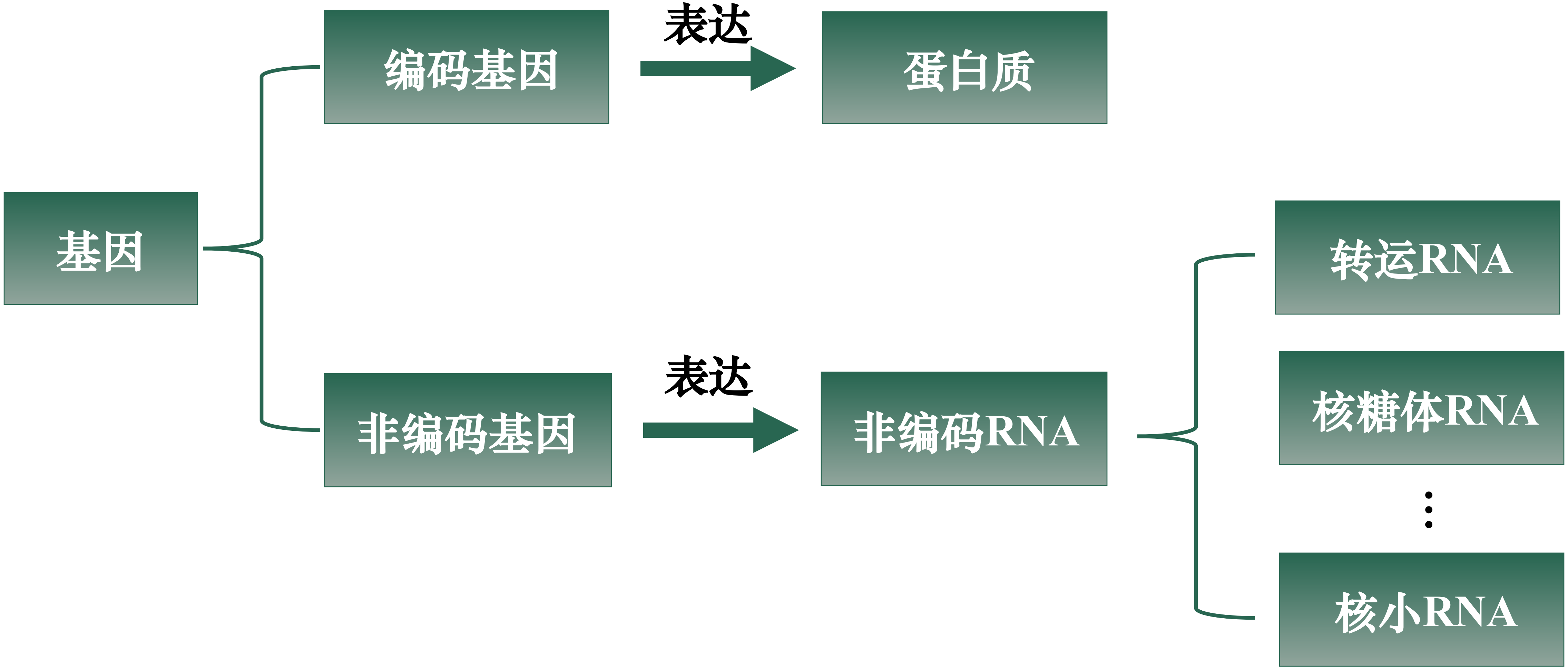
[0,0.5)	[0.5,0.6)	[0.6,0.7)	[0.7,0.8)	[0.8,0.9)	[0.9,1.0]

MKFGO的预测结果页面

02 Part two

基因功能预测

01 基因功能的执行者



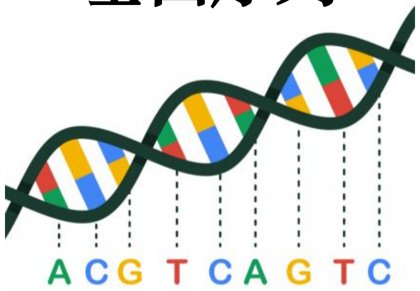
- 基因功能注释是将基因及其表达产物（蛋白质或RNA）的功能信息系统地归纳并标注到该基因上的过程。

02 基因功能预测的必要性

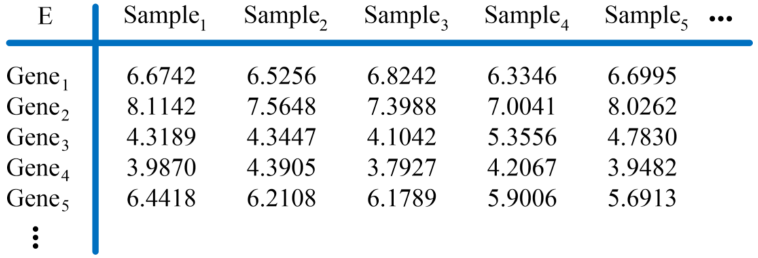
- 基因功能的注释通常依赖于对其表达产物（如蛋白质或RNA）的功能解析，这一过程往往繁琐且耗时。

- 直接从基因序列及其表达模式等原始数据出发，利用深度学习等算法对基因功能进行系统预测，从而减少对实验验证的依赖并提高注释效率。

基因序列

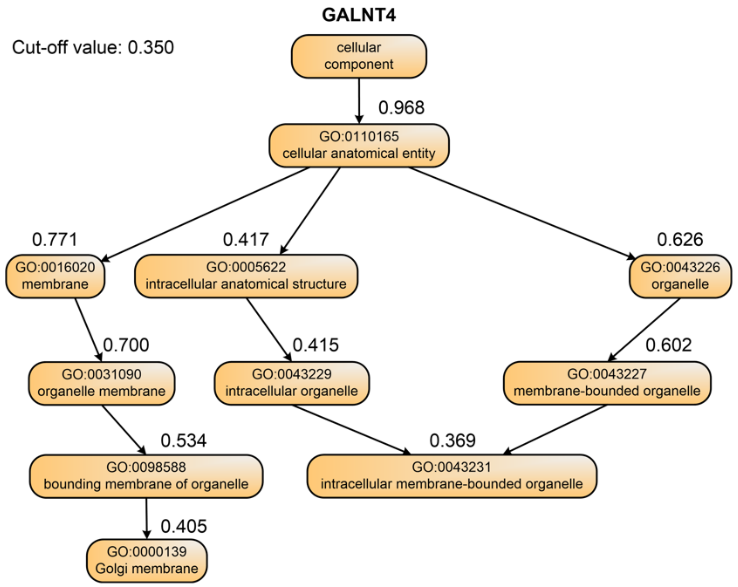


基因表达数据



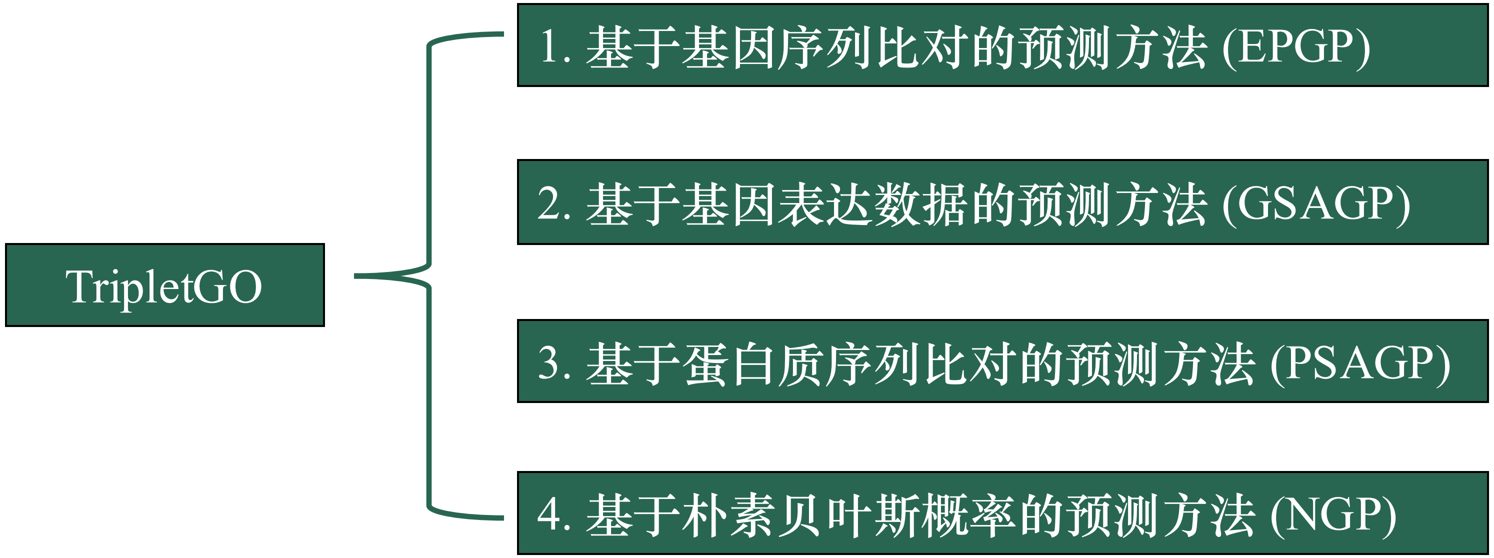
E	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	···
Gene_1	6.6742	6.5256	6.8242	6.3346	6.6995	
Gene_2	8.1142	7.5648	7.3988	7.0041	8.0262	
Gene_3	4.3189	4.3447	4.1042	5.3556	4.7830	
Gene_4	3.9870	4.3905	3.7927	4.2067	3.9482	
Gene_5	6.4418	6.2108	6.1789	5.9006	5.6913	
⋮						

深度学习算法


GALNT4
Cut-off value: 0.350
cellular component
0.968
GO:0110165 cellular anatomical entity
0.771
0.417
0.626
GO:0016020 membrane
GO:0005622 intracellular anatomical structure
GO:0043226 organelle
0.700
0.415
0.602
GO:0031090 organelle membrane
GO:0043229 intracellular organelle
GO:0043227 membrane-bounded organelle
0.534
0.369
GO:0098588 bounding membrane of organelle
GO:0043231 intracellular membrane-bounded organelle
0.405
GO:0000139 Golgi membrane

基因功能注释图 (GO注释)

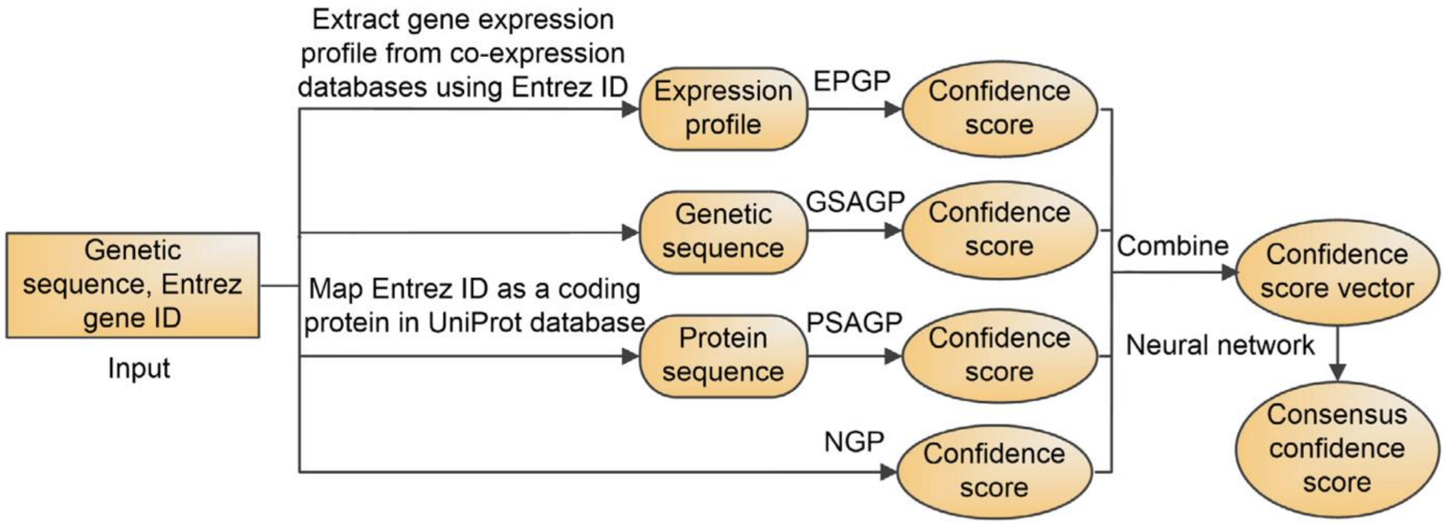
03 基于度量学习与多源信息融合的功能预测方法 TripletGO



TripletGO

1. 基于基因序列比对的预测方法 (EPGP)
2. 基于基因表达数据的预测方法 (GSAGP)
3. 基于蛋白质序列比对的预测方法 (PSAGP)
4. 基于朴素贝叶斯概率的预测方法 (NGP)

TripletGO的工作框架图


Extract gene expression profile from co-expression databases using Entrez ID
Expression profile
EPGP
Confidence score
Genetic sequence, Entrez gene ID
Input
Genetic sequence
GSAGP
Confidence score
Combine
Confidence score vector
Map Entrez ID as a coding protein in UniProt database
Protein sequence
PSAGP
Confidence score
Neural network
NGP
Confidence score
Consensus confidence score

编码基因：1, 2, 3, 4

非编码基因：1, 2, 4

Yi-Heng Zhu, Chengxin Zhang, Yan Liu, Gilbert Omenn, Peter Freddolino, Dong-Jun Yu, Yang Zhang. Integrating Transcript Expression Profiles with Protein Homology Inferences for Gene Function Prediction. **Genomics, Proteomics & Bioinformatics**. 2022.

03 基于度量学习与多源信息融合的功能预测方法 TripletGO

- 基因表达数据：基因表达过程中测量的基因转录产物 RNA在细胞中的表达量（丰度）。

- 核心思想：若两个基因的表达数据具有相似性，则它们的功能具有相似性。

- 需要解决的关键问题：如何度量两个基因的表达相似性？（从数学角度分析，如何度量两个基因表达向量的相关性？）

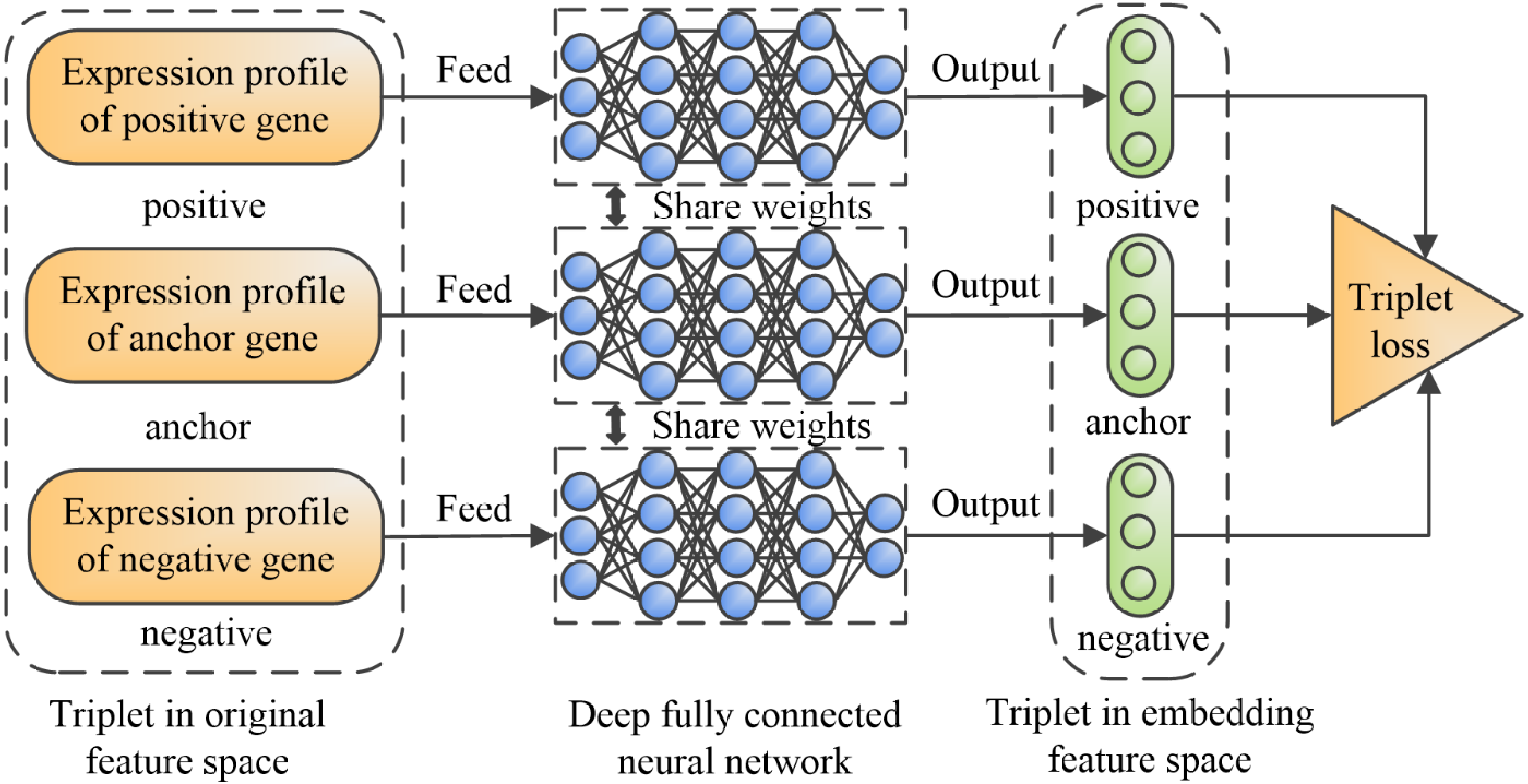
- 常用的表达相似性度量指标：皮尔逊相关系数（PCC）、互信息排序（MR）、斯皮尔曼相关系数（SRC）、欧氏距离（ED）和加权内积（WIP）

- 缺陷：无监督的方法无法关联表达相似性与功能相似性。

E	$Sample_1$	$Sample_2$	$Sample_3$	$Sample_4$	$Sample_5$	…
$Gene_1$	6.6742	6.5256	6.8242	6.3346	6.6995	
$Gene_2$	8.1142	7.5648	7.3988	7.0041	8.0262	
$Gene_3$	4.3189	4.3447	4.1042	5.3556	4.7830	
$Gene_4$	3.9870	4.3905	3.7927	4.2067	3.9482	
$Gene_5$	6.4418	6.2108	6.1789	5.9006	5.6913	
⋮						

基因表达数据

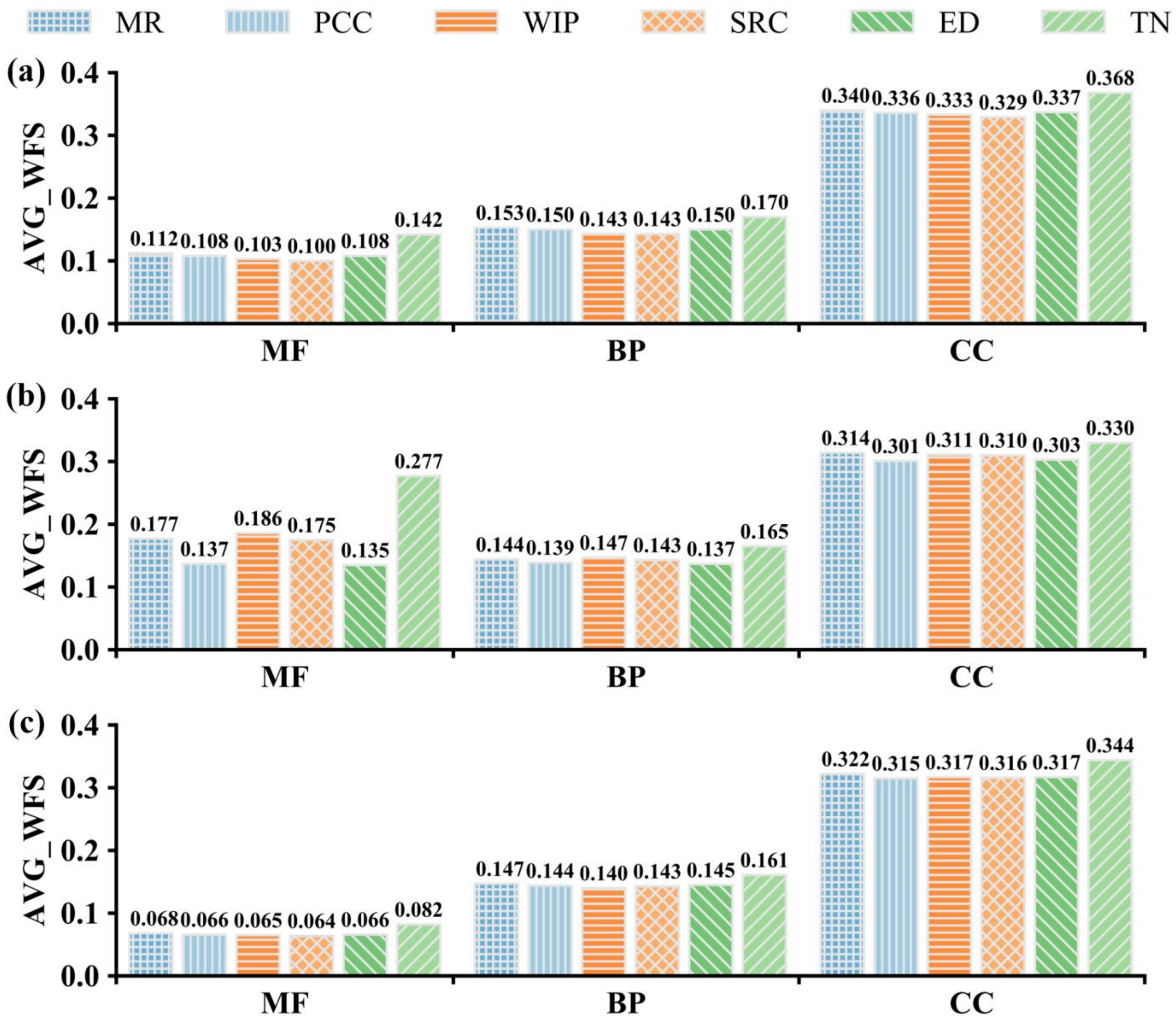
03 基于度量学习与多源信息融合的功能预测方法 TripletGO


Expression profile of positive gene
positive
Expression profile of anchor gene
anchor
Expression profile of negative gene
negative
Feed
Feed
Feed
Share weights
Share weights
Output
Output
Output
positive
anchor
negative
Triplet loss
Triplet in original feature space
Deep fully connected neural network
Triplet in embedding feature space

三元组网络度量基因表达相似性

$$Tripletloss =$$
$$max(d(anc, pos) + margin - d(anc, neg),\ 0)$$

03 基于度量学习与多源信息融合的功能预测方法 TripletGO



MR：互信息排序 PCC：皮尔逊相关系数

WIP：和加权内积

SRC：斯皮尔曼相关系数

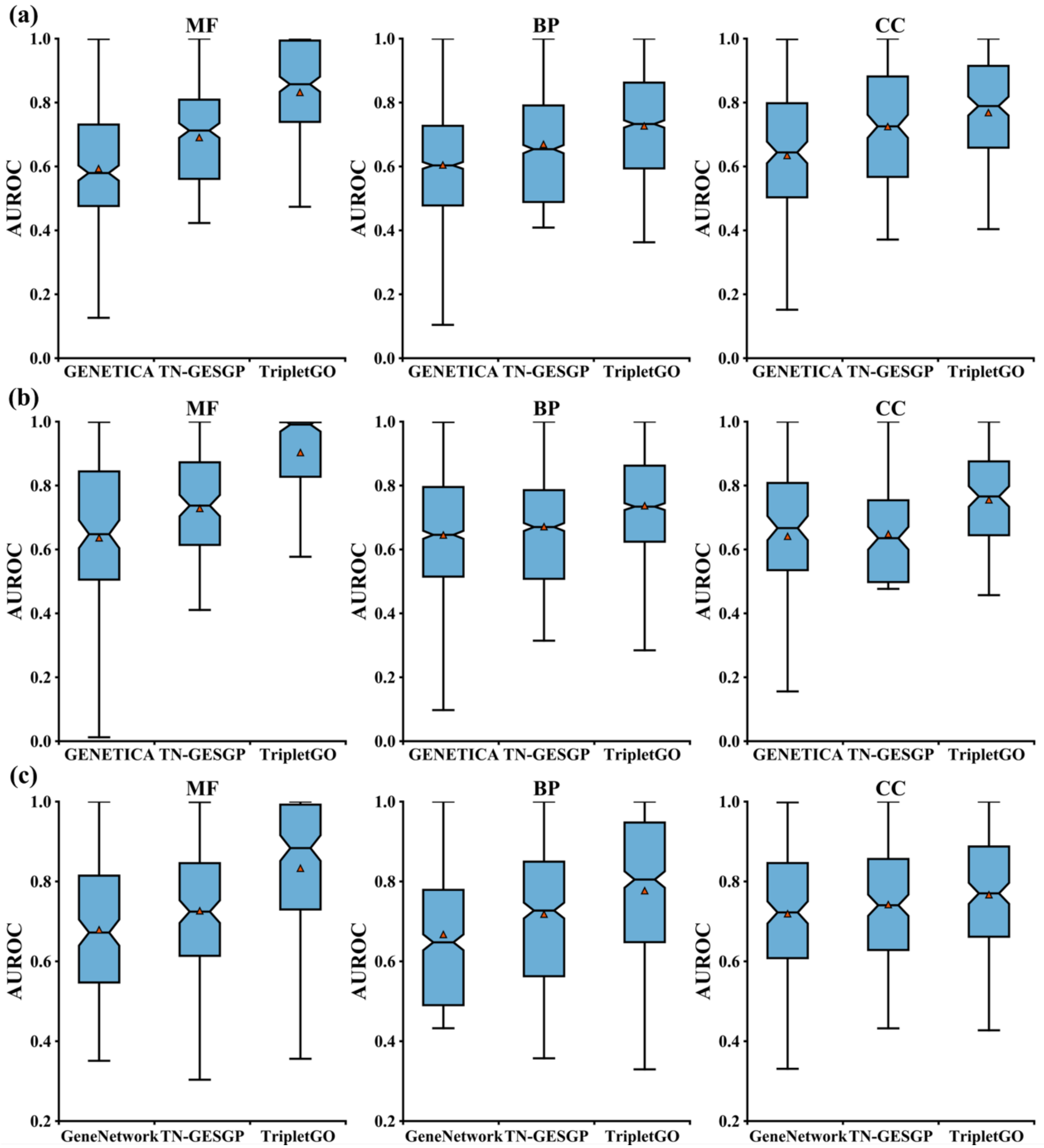
ED：欧氏距离

TN：三元组神经网络

不同基因表达相似性度量方法在三组测试集上的AVG_WFS值

（a）Gene Targets测试集中5656个编码基因；（b）Gene Targets测试集中98个非编码基因；（c）CAFA3 Protein Targets with Expression Data测试集中2433个蛋白质。

03 基于度量学习与多源信息融合的功能预测方法 TripletGO

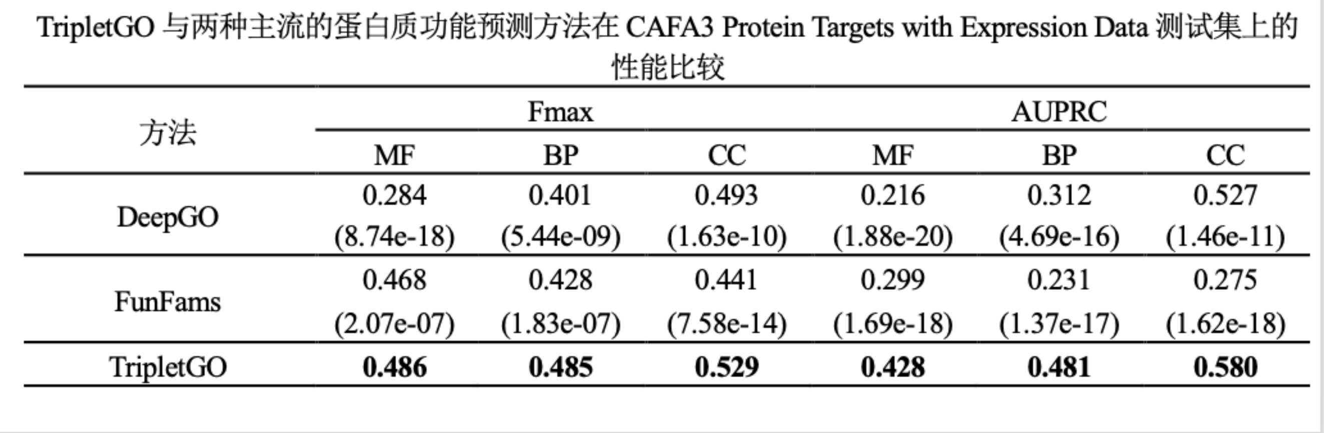


不同基因功能预测方法在三组公共测试集上面向GO术语的AUROC分布箱线图

（a）GENETICA、TN-GESGP和TripletGO在Human Test I上的AUROC分布图；

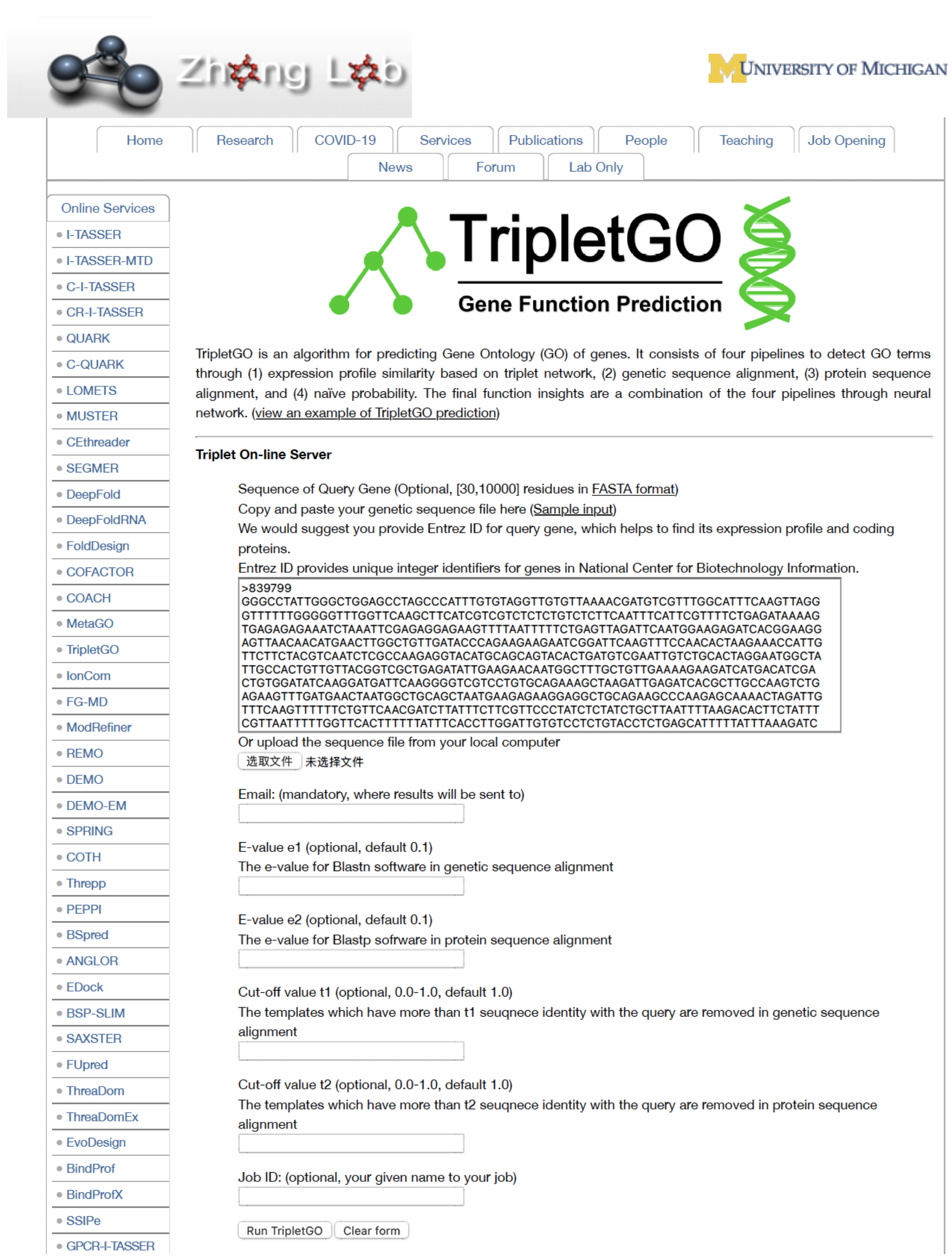
（b）GENETICA、TN-GESGP和TripletGO在Mouse Test I上的AUROC分布图；

（c）GeneNetwork、TN-GESGP和TripletGO在Human Test II上的AUROC分布图



TripletGO 与两种主流的蛋白质功能预测方法在 CAFA3 Protein Targets with Expression Data 测试集上的性能比较

方法	Fmax			AUPRC		
	MF	BP	CC	MF	BP	CC
DeepGO	0.284 (8.74e-18)	0.401 (5.44e-09)	0.493 (1.63e-10)	0.216 (1.88e-20)	0.312 (4.69e-16)	0.527 (1.46e-11)
FunFams	0.468 (2.07e-07)	0.428 (1.83e-07)	0.441 (7.58e-14)	0.299 (1.69e-18)	0.231 (1.37e-17)	0.275 (1.62e-18)
TripletGO	**0.486**	**0.485**	**0.529**	**0.428**	**0.481**	**0.580**



TripletGO is an algorithm for predicting Gene Ontology (GO) of genes. It consists of four pipelines to detect GO terms through (1) expression profile similarity based on triplet network, (2) genetic sequence alignment, (3) protein sequence alignment, and (4) naïve probability. The final function insights are a combination of the four pipelines through neural network. (view an example of TripletGO prediction)

Triplet On-line Server

Sequence of Query Gene (Optional, [30,10000] residues in FASTA format)
Copy and paste your genetic sequence file here (Sample input)
We would suggest you provide Entrez ID for query gene, which helps to find its expression profile and coding proteins.
Entrez ID provides unique integer identifiers for genes in National Center for Biotechnology Information.

```
>839799
GGGCCTATTGGGCTGGAGCCTAGCCCATTTGTGTAGGTTGTGTTAAAACGATGTCGTTTGGCATTTCAAGTTAGG
GTTTTTTGGGGGTTTGGTTCAAGCTTCATCGTCGTCTCTCTGTCTCTTCAATTTCATTCGTTTTCTGAGATAAAAG
TGAGAGAGAAATCTAAATTCGAGAGGAGAAGTTTTAATTTTTCTGAGTTAGATTCAATGGAAGAGATCACGGAAGG
AGTTAACAACATGAACTTGGCTGTTGATACCCAGAAGAAGAATCGGATTCAAGTTTCCAACACTAAGAAACCATTG
TTCTTCTACGTCAATCTCGCCAAGAGGTACATGCAGCAGTACACTGATGTCGAATTGTCTGCACTAGGAATGGCTA
TTGCCACTGTTGTTACGGTCGCTGAGATATTGAAGAACAATGGCTTTGCTGTTGAAAAGAAGATCATGACATCGA
CTGTGGATATCAAGGATGATTCAAGGGGTCGTCCTGTGCAGAAAGCTAAGATTGAGATCACGCTTGCCAAGTCTG
AGAAGTTTGATGAACTAATGGCTGCAGCTAATGAAGAGAAGGAGGCTGCAGAAGCCCAAGAGCAAAACTAGATTG
TTTCAAGTTTTTTCTGTTCAACGATCTTATTTCTTCGTTCCCTATCTCTATCTGCTTAATTTTAAGACACTTCTATTT
CGTTAATTTTTGGTTCACTTTTTTATTTCACCTTGGATTGTGTCCTCTGTACCTCTGAGCATTTTTATTTAAAGATC
```

Or upload the sequence file from your local computer
选取文件 未选择文件

Email: (mandatory, where results will be sent to)

E-value e1 (optional, default 0.1)
The e-value for Blastn software in genetic sequence alignment

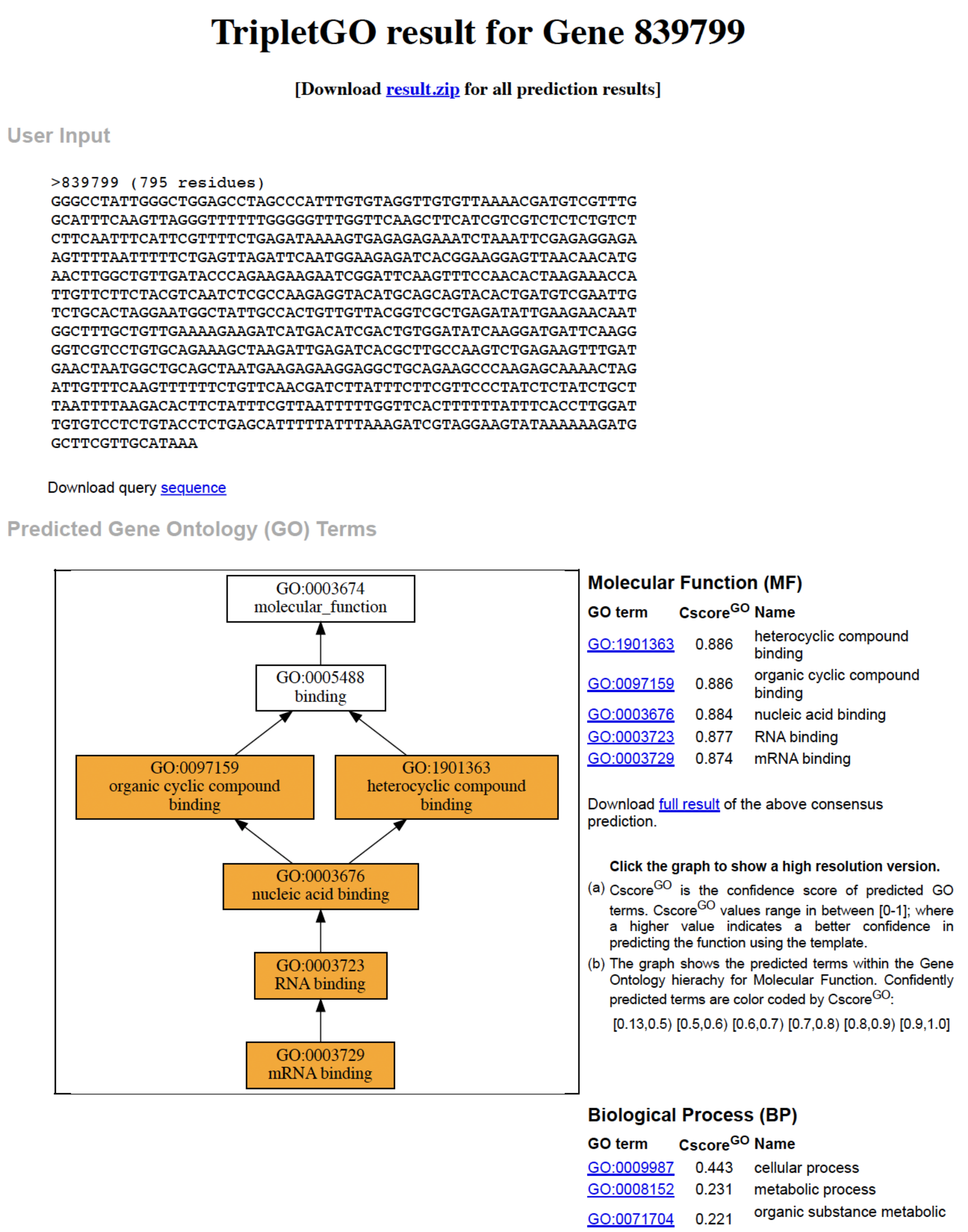
E-value e2 (optional, default 0.1)
The e-value for Blastp sofrware in protein sequence alignment

Cut-off value t1 (optional, 0.0-1.0, default 1.0)
The templates which have more than t1 seuqnece identity with the query are removed in genetic sequence alignment

Cut-off value t2 (optional, 0.0-1.0, default 1.0)
The templates which have more than t2 seuqnece identity with the query are removed in protein sequence alignment

Job ID: (optional, your given name to your job)

Run TripletGO Clear form



TripletGO result for Gene 839799

[Download result.zip for all prediction results]

User Input

```
>839799 (795 residues)
GGGCCTATTGGGCTGGAGCCTAGCCCATTTGTGTAGGTTGTGTTAAAACGATGTCGTTTG
GCATTTCAAGTTAGGGTTTTTTGGGGGTTTGGTTCAAGCTTCATCGTCGTCTCTCTGTCT
CTTCAATTTCATTCGTTTTCTGAGATAAAAGTGAGAGAGAAATCTAAATTCGAGAGGAGA
AGTTTTAATTTTTCTGAGTTAGATTCAATGGAAGAGATCACGGAAGGAGTTAACAACATG
AACTTGGCTGTTGATACCCAGAAGAAGAATCGGATTCAAGTTTCCAACACTAAGAAACCA
TTGTTCTTCTACGTCAATCTCGCCAAGAGGTACATGCAGCAGTACACTGATGTCGAATTG
TCTGCACTAGGAATGGCTATTGCCACTGTTGTTACGGTCGCTGAGATATTGAAGAACAAT
GGCTTTGCTGTTGAAAAGAAGATCATGACATCGACTGTGGATATCAAGGATGATTCAAGG
GGTCGTCCTGTGCAGAAAGCTAAGATTGAGATCACGCTTGCCAAGTCTGAGAAGTTTGAT
GAACTAATGGCTGCAGCTAATGAAGAGAAGGAGGCTGCAGAAGCCCAAGAGCAAAACTAG
ATTGTTTCAAGTTTTTTCTGTTCAACGATCTTATTTCTTCGTTCCCTATCTCTATCTGCT
TAATTTTAAGACACTTCTATTTCGTTAATTTTTGGTTCACTTTTTTATTTCACCTTGGAT
TGTGTCCTCTGTACCTCTGAGCATTTTTATTTAAAGATCGTAGGAAGTATAAAAAAGATG
GCTTCGTTGCATAAA
```

Download query sequence

Predicted Gene Ontology (GO) Terms

Molecular Function (MF)

GO term	Cscore^{GO}	Name
GO:1901363	0.886	heterocyclic compound binding
GO:0097159	0.886	organic cyclic compound binding
GO:0003676	0.884	nucleic acid binding
GO:0003723	0.877	RNA binding
GO:0003729	0.874	mRNA binding

Download full result of the above consensus prediction.

Click the graph to show a high resolution version.

(a) Cscore^{GO} is the confidence score of predicted GO terms. Cscore^{GO} values range in between [0-1]; where a higher value indicates a better confidence in predicting the function using the template.

(b) The graph shows the predicted terms within the Gene Ontology hierachy for Molecular Function. Confidently predicted terms are color coded by Cscore^{GO}:

[0.13,0.5) [0.5,0.6) [0.6,0.7) [0.7,0.8) [0.8,0.9) [0.9,1.0]

Biological Process (BP)

GO term	Cscore^{GO}	Name
GO:0009987	0.443	cellular process
GO:0008152	0.231	metabolic process
GO:0071704	0.221	organic substance metabolic

在线预测服务 (https://aideepmed.com/TripletGO/)

03 Part three

蛋白质-配体相互作用预测

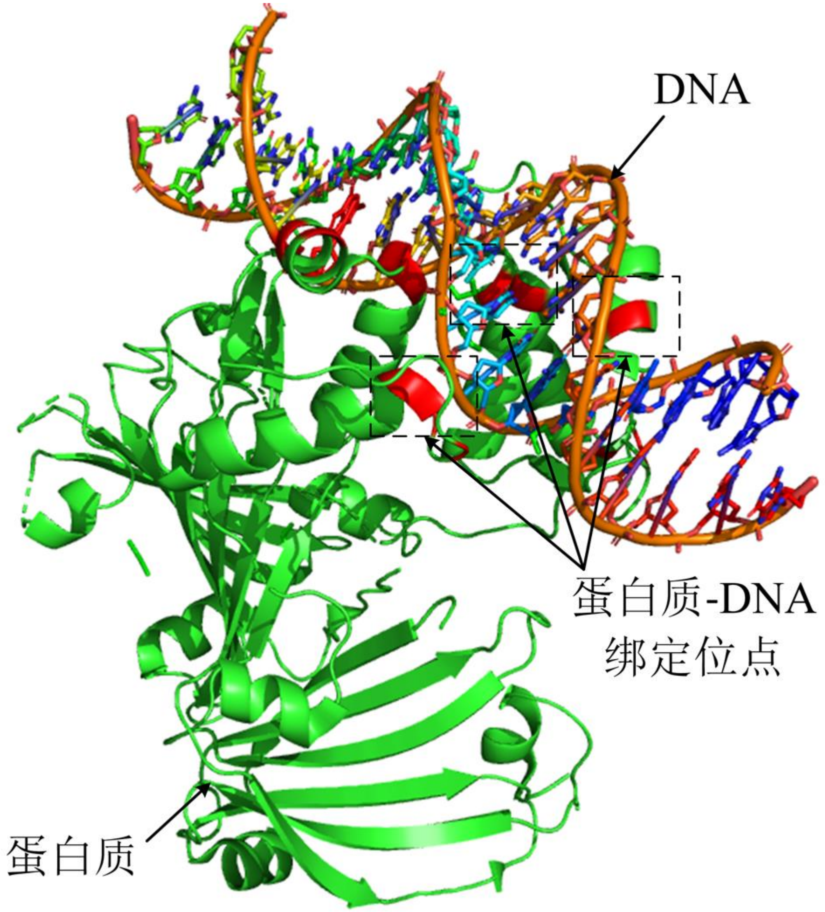
01 蛋白质-配体相互作用

- 蛋白质在发挥其生物学功能时往往并非孤立运作，而是需要与配体（如 DNA、RNA、金属离子等）发生特异性的物理相互作用，从而共同完成其功能活动。

- 相互作用预测主要涵盖两个层级：

（1）蛋白质层级：识别是否发生相互作用并估计结合强度；

（2）残基层级：定位具体的结合位点或功能性关键残基。



蛋白质-DNA复合物三维结构图

02 蛋白质-配体相互作用预测研究进展

[1] Yi-Heng Zhu, Jun Hu, Yong Qi, Xiao-Ning Song, Dong-Jun Yu. *Boosting Granular Support Vector Machines for the Accurate Prediction of Protein-Nucleotide Binding Sites*. **Combinatorial Chemistry & High Throughput Screening,** 2019.

[2] Yi-Heng Zhu, Jun Hu, Xiao-Ning Song, Dong-Jun Yu*. *DNAPred: Accurate Identification of DNA-binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines*. **Journal of Chemical Information and Modeling,** 2019.

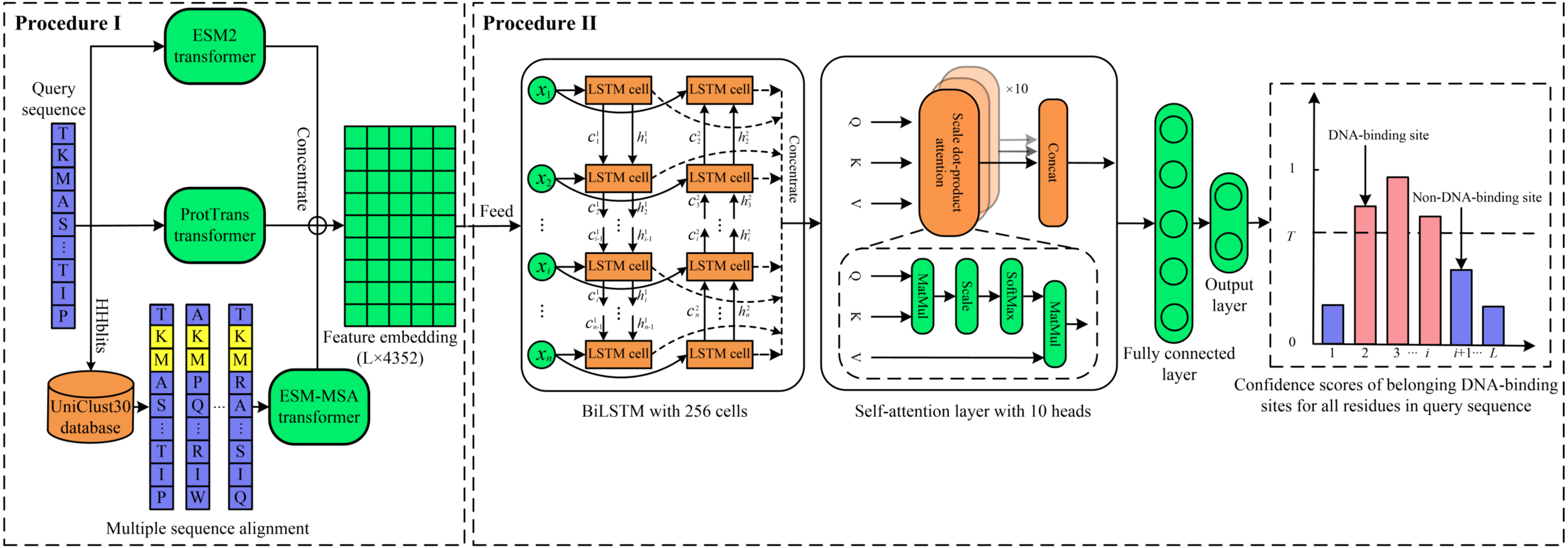
[3] Zi Liu#, Yi-Heng Zhu#, Long-Chen Shen, Xuan Xiao, Wang-Ren Qiu, Dong-Jun Yu*. *Integrating Unsupervised Language Model with Multi-View Multiple Sequence Alignments for High-Accuracy Inter-Chain Contact Prediction*. **Computers in Biology and Medicine**, 2023. (*)

[4] Yi-Heng Zhu, Zi Liu, Zhiwei Ji*, Dong-Jun Yu*. *ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for High-Accuracy Protein-DNA Binding Site Prediction*. **Briefings in Bioinformatics**, 2024. (*)

[5] Zi Liu, Wang-Ren Qiu, Yan Liu, He Yan, Wenyi Pei*, Yi-Heng Zhu*, and Jing Qiu*. *A Comprehensive Review of Computational Methods for Protein-DNA Binding Site Prediction*. **Analytical Biochemistry**, 2025.

02 基于无监督语言模型与多源信息融合的蛋白质-DNA绑定位点预测方法 ULDNA

- 主要贡献：融合多种蛋白质大语言模型，显著地提升了蛋白质-DNA绑定位点预测精度。



ULDNA工作框架图

Yi-Heng Zhu, Zi Liu, Zhiwei Ji*, Dong-Jun Yu*. ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for High-Accuracy Protein-DNA Binding Site Prediction. Briefings in Bioinformatics, 2024.

02 基于无监督语言模型与多源信息融合的蛋白质-DNA绑定位点预测方法 ULDNA

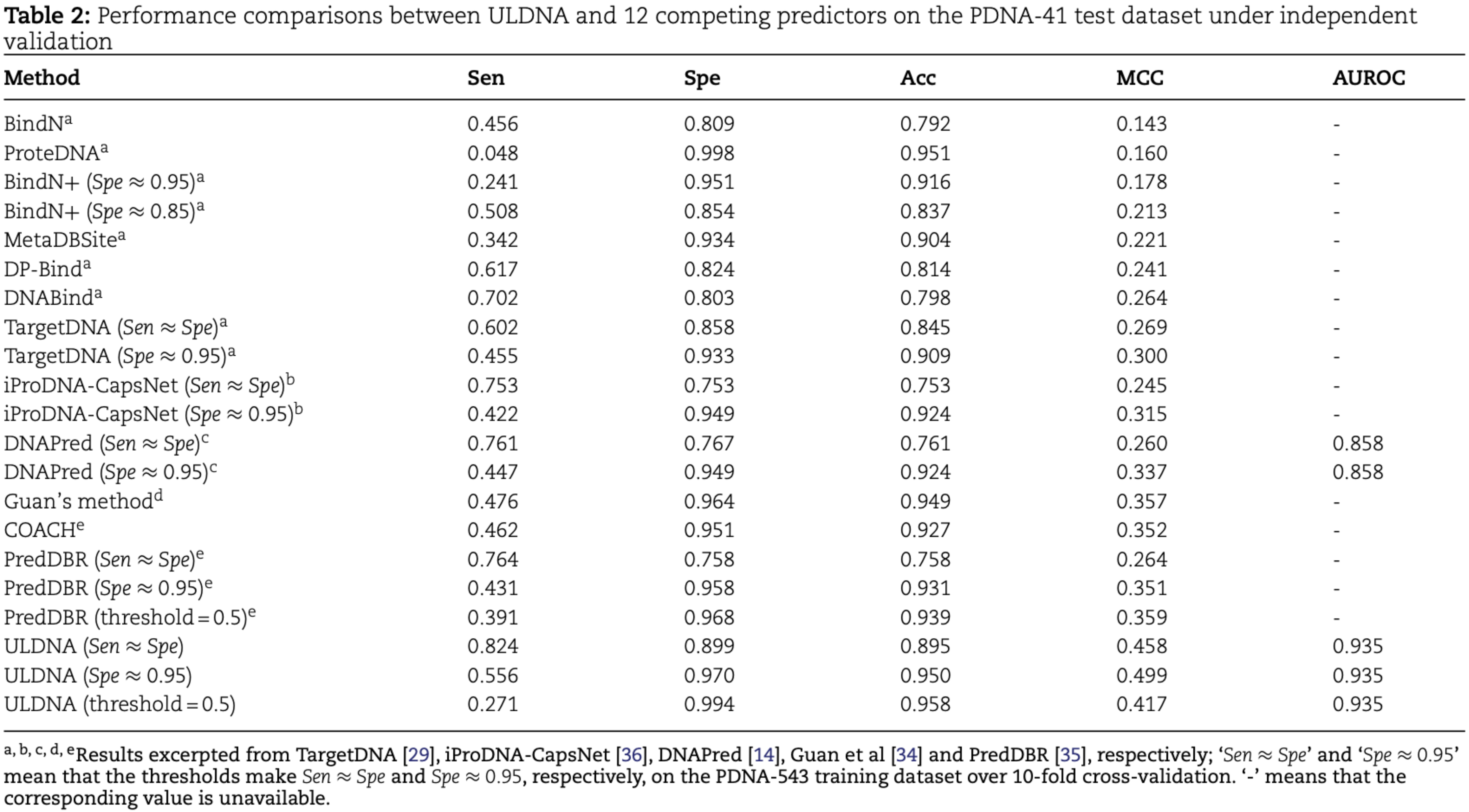


Table 2: Performance comparisons between ULDNA and 12 competing predictors on the PDNA-41 test dataset under independent validation

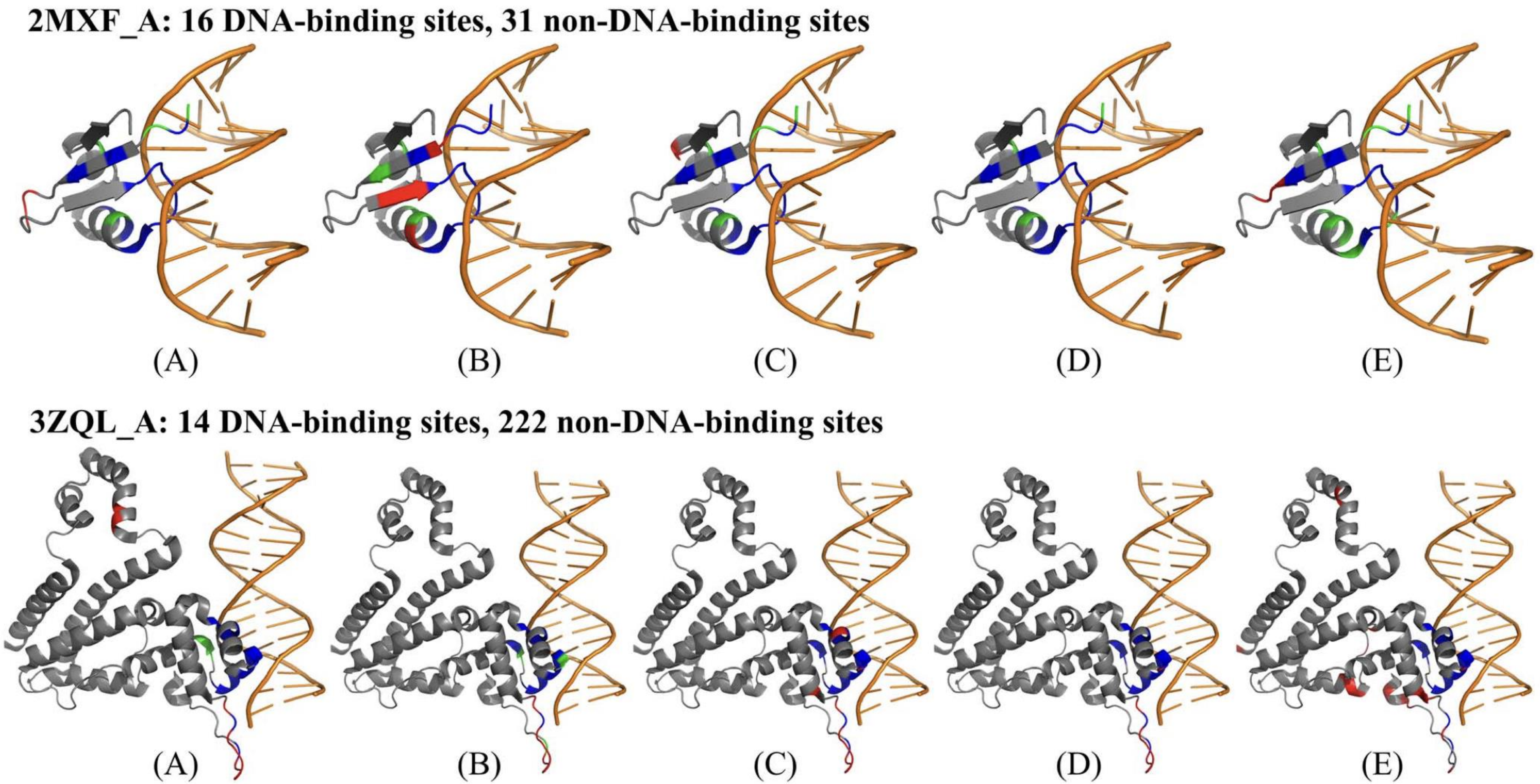
Method	Sen	Spe	Acc	MCC	AUROC
BindN[a]	0.456	0.809	0.792	0.143	-
ProteDNA[a]	0.048	0.998	0.951	0.160	-
BindN+ (*Spe* ≈ 0.95)[a]	0.241	0.951	0.916	0.178	-
BindN+ (*Spe* ≈ 0.85)[a]	0.508	0.854	0.837	0.213	-
MetaDBSite[a]	0.342	0.934	0.904	0.221	-
DP-Bind[a]	0.617	0.824	0.814	0.241	-
DNABind[a]	0.702	0.803	0.798	0.264	-
TargetDNA (*Sen* ≈ *Spe*)[a]	0.602	0.858	0.845	0.269	-
TargetDNA (*Spe* ≈ 0.95)[a]	0.455	0.933	0.909	0.300	-
iProDNA-CapsNet (*Sen* ≈ *Spe*)[b]	0.753	0.753	0.753	0.245	-
iProDNA-CapsNet (*Spe* ≈ 0.95)[b]	0.422	0.949	0.924	0.315	-
DNAPred (*Sen* ≈ *Spe*)[c]	0.761	0.767	0.761	0.260	0.858
DNAPred (*Spe* ≈ 0.95)[c]	0.447	0.949	0.924	0.337	0.858
Guan's method[d]	0.476	0.964	0.949	0.357	-
COACH[e]	0.462	0.951	0.927	0.352	-
PredDBR (*Sen* ≈ *Spe*)[e]	0.764	0.758	0.758	0.264	-
PredDBR (*Spe* ≈ 0.95)[e]	0.431	0.958	0.931	0.351	-
PredDBR (threshold = 0.5)[e]	0.391	0.968	0.939	0.359	-
ULDNA (*Sen* ≈ *Spe*)	0.824	0.899	0.895	0.458	0.935
ULDNA (*Spe* ≈ 0.95)	0.556	0.970	0.950	0.499	0.935
ULDNA (threshold = 0.5)	0.271	0.994	0.958	0.417	0.935

[a, b, c, d, e] Results excerpted from TargetDNA [29], iProDNA-CapsNet [36], DNAPred [14], Guan et al [34] and PredDBR [35], respectively; '*Sen* ≈ *Spe*' and '*Spe* ≈ 0.95' mean that the thresholds make *Sen* ≈ *Spe* and *Spe* ≈ 0.95, respectively, on the PDNA-543 training dataset over 10-fold cross-validation. '-' means that the corresponding value is unavailable.

ULDNA和SOTA方法在41个DNA结合蛋白质上的性能比较

02 基于无监督语言模型与多源信息融合的蛋白质-DNA绑定位点预测方法 ULDNA

案例分析

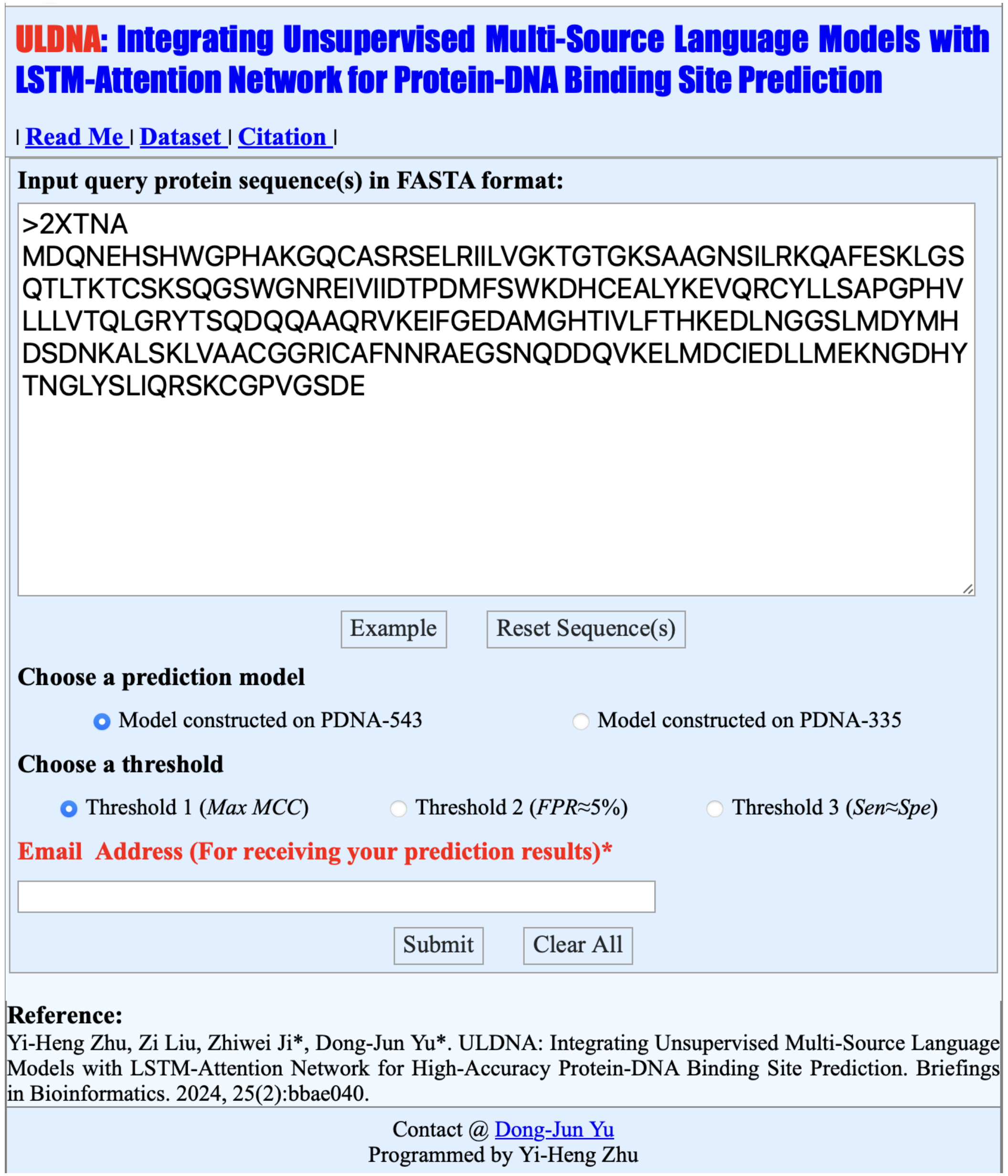


2MXF_A: 16 DNA-binding sites, 31 non-DNA-binding sites

3ZQL_A: 14 DNA-binding sites, 222 non-DNA-binding sites

Figure 4. Visualization of prediction results for two proteins (2MXF_A and 3ZQL_A) using five DNA-binding site prediction models: (A) LA-ESM2, (B) LA-ProtTrans, (C) LA-ESM-MSA, (D) ULDNA, (E) PredDBR. The atomic-level native structure of each protein is downloaded from the PDB database and then plotted as the cartoon picture using PyMOL software [70]. The color scheme is used as follows: DNA in orange, true positives in blue, false positives in red and false negatives in green.

02 基于无监督语言模型与多源信息融合的蛋白质-DNA绑定位点预测方法 ULDNA



ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for Protein-DNA Binding Site Prediction

| Read Me | Dataset | Citation |

Input query protein sequence(s) in FASTA format:

```
>2XTNA
MDQNEHSHWGPHAKGQCASRSELRIILVGKTGTGKSAAGNSILRKQAFESKLGS
QTLTKTCSKSQGSWGNREIVIIDTPDMFSWKDHCEALYKEVQRCYLLSAPGPHV
LLLVTQLGRYTSQDQQAAQRVKEIFGEDAMGHTIVLFTHKEDLNGGSLMDYMH
DSDNKALSKLVAACGGRICAFNNRAEGSNQDDQVKELMDCIEDLLMEKNGDHY
TNGLYSLIQRSKCGPVGSDE
```

Choose a prediction model

● Model constructed on PDNA-543 ○ Model constructed on PDNA-335

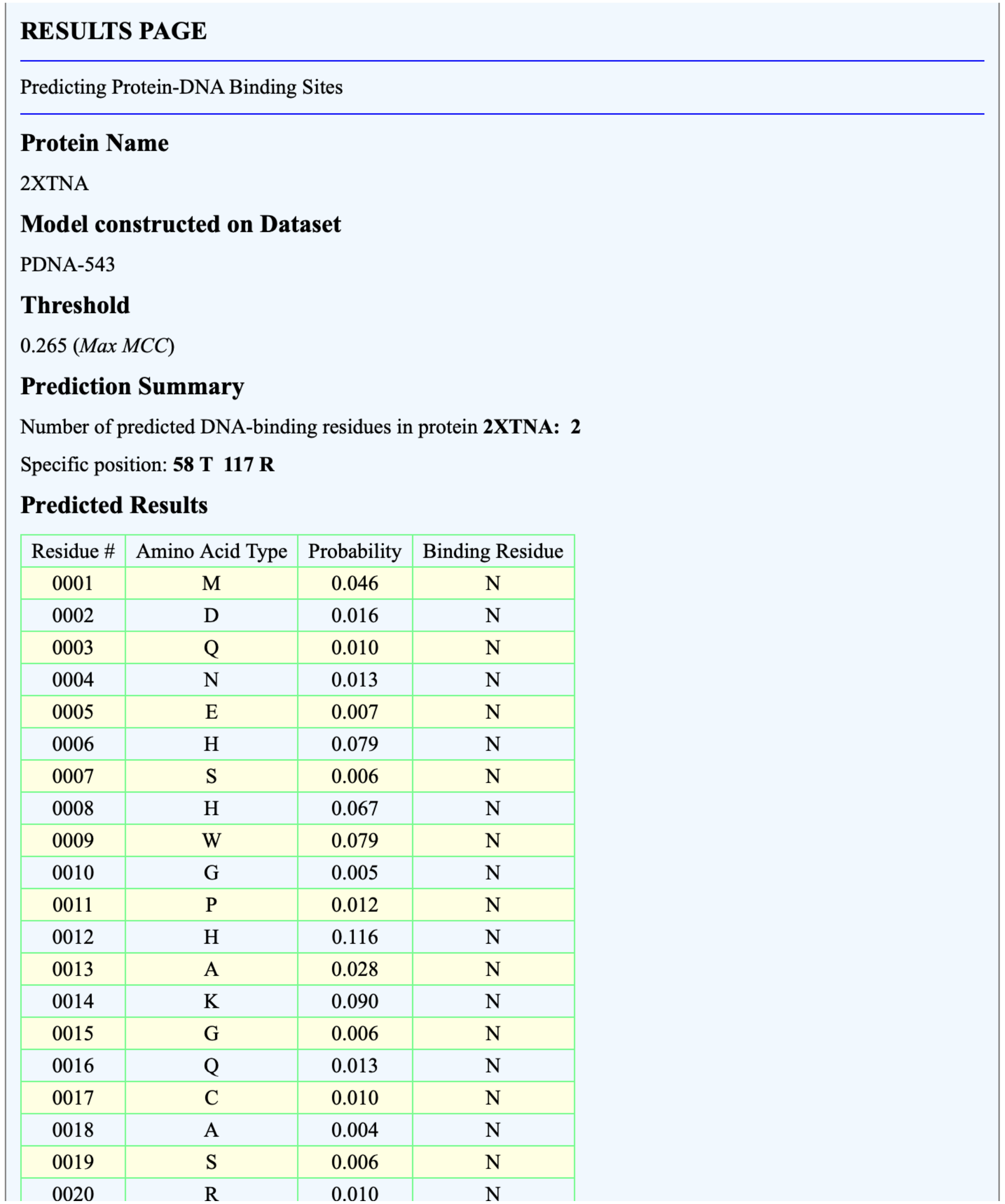
Choose a threshold

● Threshold 1 (*Max MCC*) ○ Threshold 2 (*FPR*≈5%) ○ Threshold 3 (*Sen*≈*Spe*)

Email Address (For receiving your prediction results)*

Reference:
Yi-Heng Zhu, Zi Liu, Zhiwei Ji*, Dong-Jun Yu*. ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for High-Accuracy Protein-DNA Binding Site Prediction. Briefings in Bioinformatics. 2024, 25(2):bbae040.

Contact @ Dong-Jun Yu
Programmed by Yi-Heng Zhu



RESULTS PAGE

Predicting Protein-DNA Binding Sites

Protein Name

2XTNA

Model constructed on Dataset

PDNA-543

Threshold

0.265 (*Max MCC*)

Prediction Summary

Number of predicted DNA-binding residues in protein **2XTNA: 2**

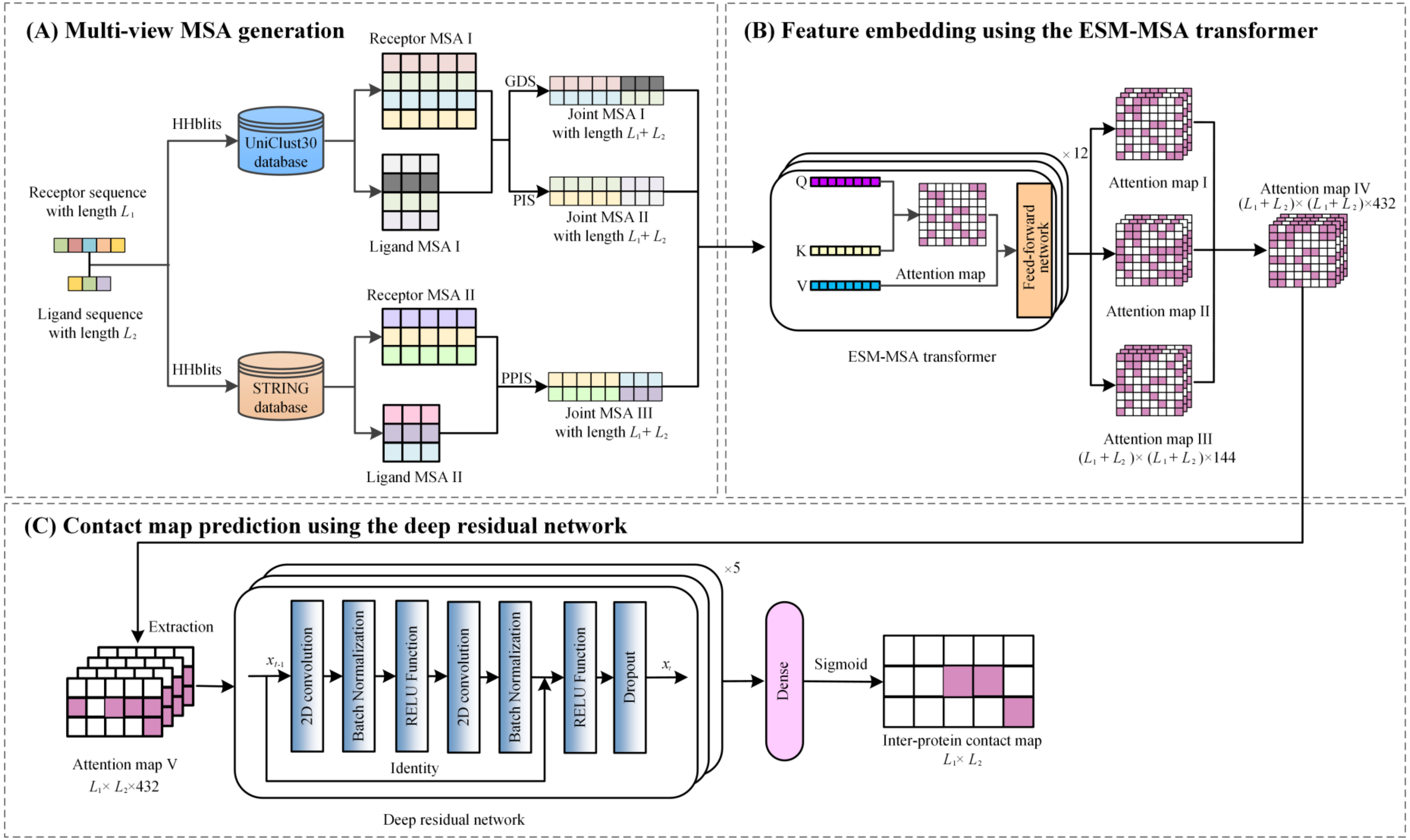
Specific position: **58 T 117 R**

Predicted Results

Residue #	Amino Acid Type	Probability	Binding Residue
0001	M	0.046	N
0002	D	0.016	N
0003	Q	0.010	N
0004	N	0.013	N
0005	E	0.007	N
0006	H	0.079	N
0007	S	0.006	N
0008	H	0.067	N
0009	W	0.079	N
0010	G	0.005	N
0011	P	0.012	N
0012	H	0.116	N
0013	A	0.028	N
0014	K	0.090	N
0015	G	0.006	N
0016	Q	0.013	N
0017	C	0.010	N
0018	A	0.004	N
0019	S	0.006	N
0020	R	0.010	N

在线预测服务 (http://csbio.njust.edu.cn/bioinf/dnapred/)

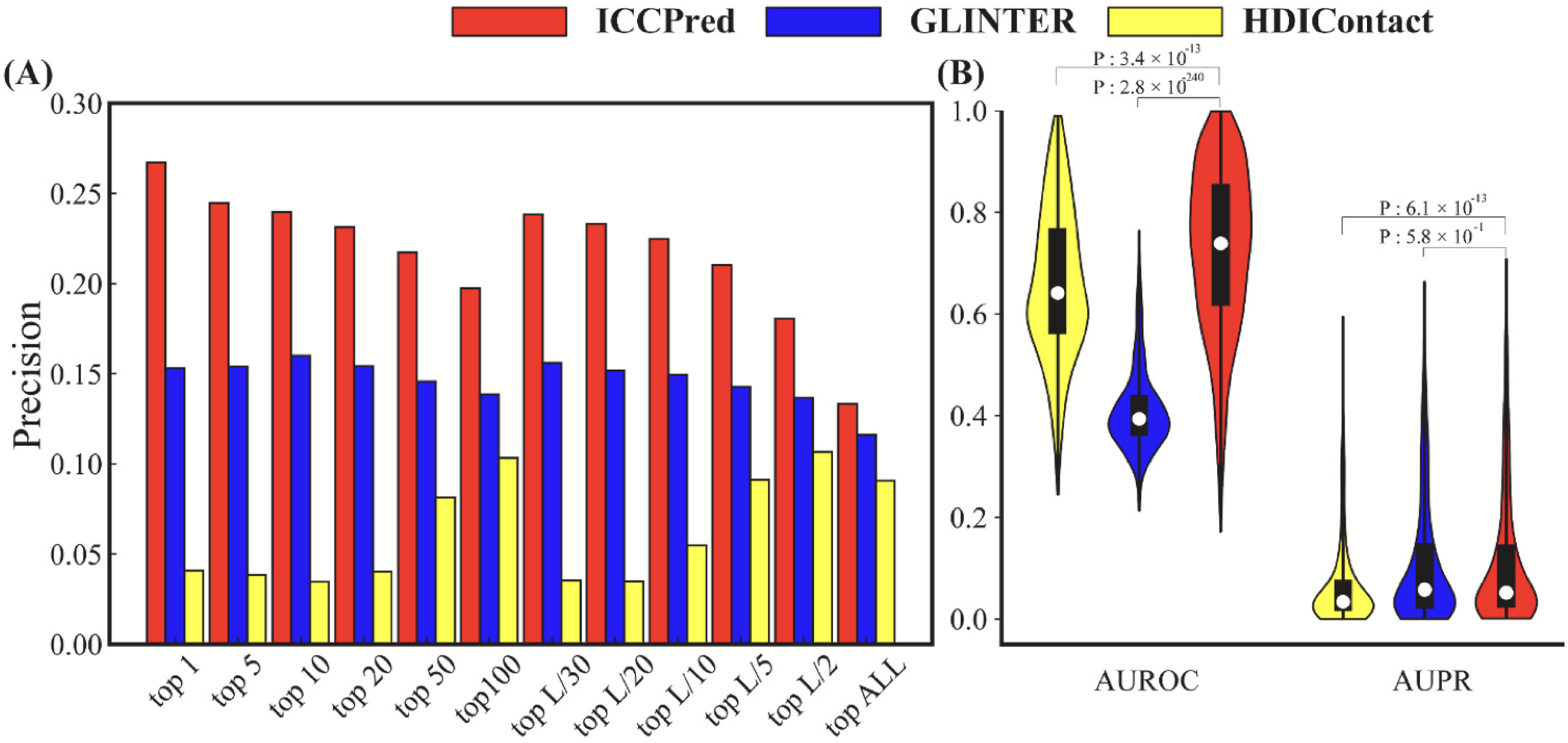
03 基于无监督语言模型与深度残差神经网络的蛋白质链接接触图预测方法ICCPred



ICCPred 工作框架图

Zi Liu#, Yi-Heng Zhu#, Long-Chen Shen, Xuan Xiao, Wang-Ren Qiu, Dong-Jun Yu. Integrating Unsupervised Language Model with Multi-View Multiple Sequence Alignments for High-Accuracy Inter-Chain Contact Prediction. **Computers in Biology and Medicine**. 2023.

03 基于无监督语言模型与深度残差神经网络的蛋白质链接接触图预测方法ICCPred



ICCPred与SOTA方法在630个测试蛋白质上的性能比较

03 基于无监督语言模型与深度残差神经网络的蛋白质链接接触图预测方法ICCPred

案例分析

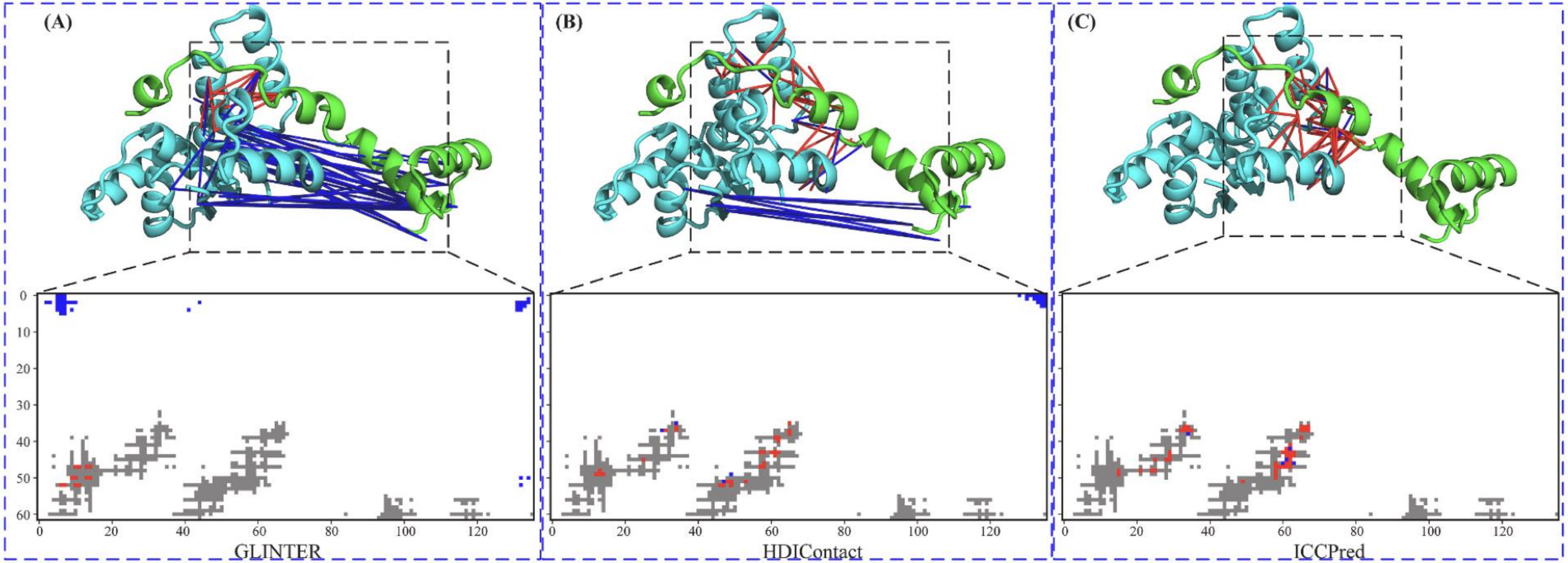


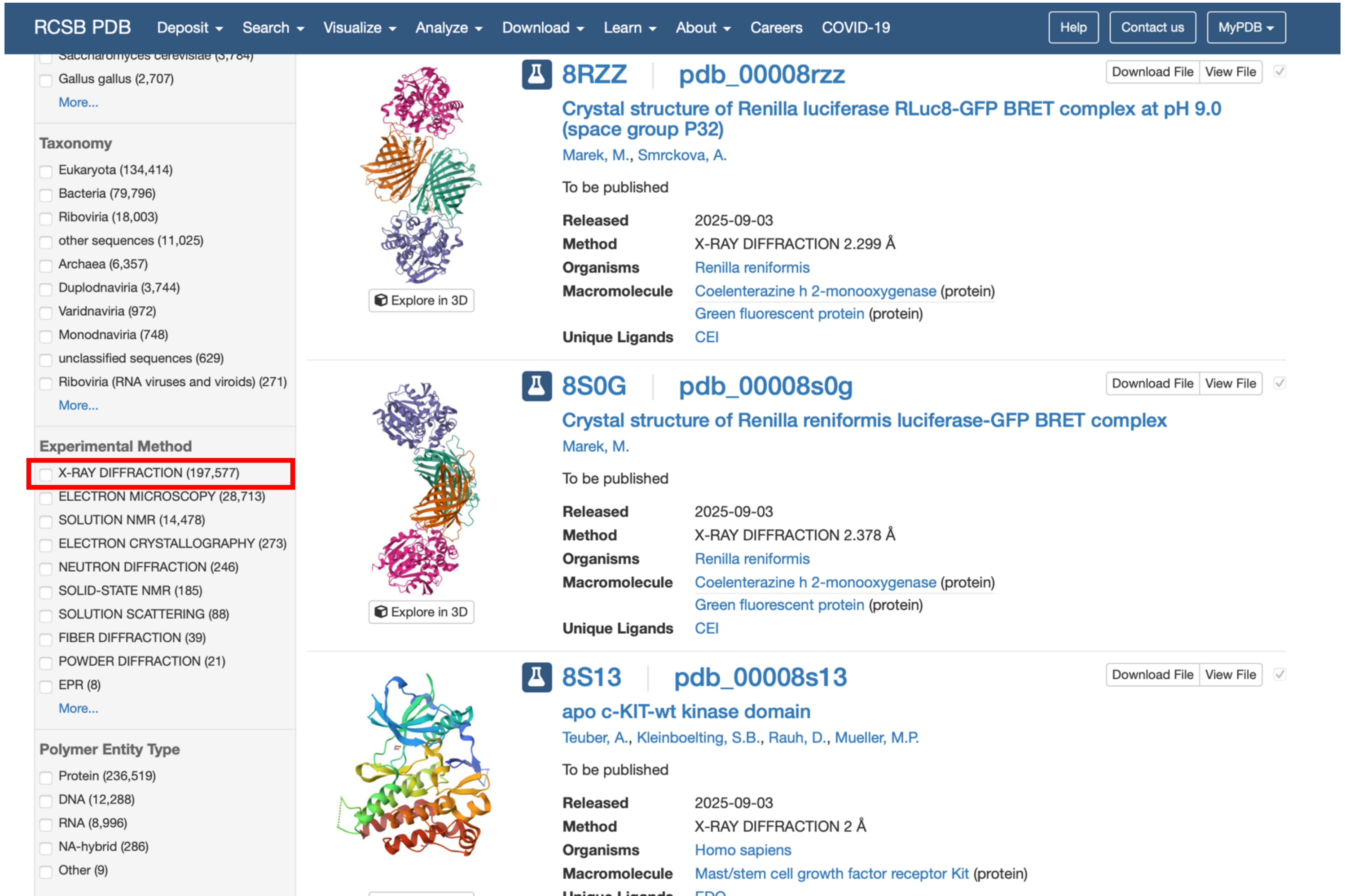
Fig. 3. Illustrative examples for GLINTER (A), HDIContact (B), and ICCPred (C) on protein complex 6A7V at the top 50 predicted contacts. Native structures of two monomers are shown in green and cyan, respectively. True positives are depicted in red, while solid blue lines represent false positives. On the bottom of each panel, grey dots indicate naive contacts, red dots represent the true positives in the top 50 predicted contacts, and blue dots are false positives.

04 Part two

蛋白质结晶倾向性预测

01 X射线晶体衍射技术

- ➢ X射线晶体衍射技术（X-ray crystallography）是解析蛋白质三维结构最主要的手段。
- ➢ 据统计，在PDB数据库中，大约有81.7%的蛋白质结构是通过X射线晶体衍射技术解析的。



□ Saccharomyces cerevisiae (3,784)
□ Gallus gallus (2,707)
More...

Taxonomy
□ Eukaryota (134,414)
□ Bacteria (79,796)
□ Riboviria (18,003)
□ other sequences (11,025)
□ Archaea (6,357)
□ Duplodnaviria (3,744)
□ Varidnaviria (972)
□ Monodnaviria (748)
□ unclassified sequences (629)
□ Riboviria (RNA viruses and viroids) (271)
More...

Experimental Method
□ X-RAY DIFFRACTION (197,577)
□ ELECTRON MICROSCOPY (28,713)
□ SOLUTION NMR (14,478)
□ ELECTRON CRYSTALLOGRAPHY (273)
□ NEUTRON DIFFRACTION (246)
□ SOLID-STATE NMR (185)
□ SOLUTION SCATTERING (88)
□ FIBER DIFFRACTION (39)
□ POWDER DIFFRACTION (21)
□ EPR (8)
More...

Polymer Entity Type
□ Protein (236,519)
□ DNA (12,288)
□ RNA (8,996)
□ NA-hybrid (286)
□ Other (9)

8RZZ | pdb_00008rzz
Download File | View File
Crystal structure of Renilla luciferase RLuc8-GFP BRET complex at pH 9.0 (space group P32)
Marek, M., Smrckova, A.
To be published
Explore in 3D

Released	2025-09-03
Method	X-RAY DIFFRACTION 2.299 Å
Organisms	Renilla reniformis
Macromolecule	Coelenterazine h 2-monooxygenase (protein) Green fluorescent protein (protein)
Unique Ligands	CEI

8S0G | pdb_00008s0g
Download File | View File
Crystal structure of Renilla reniformis luciferase-GFP BRET complex
Marek, M.
To be published
Explore in 3D

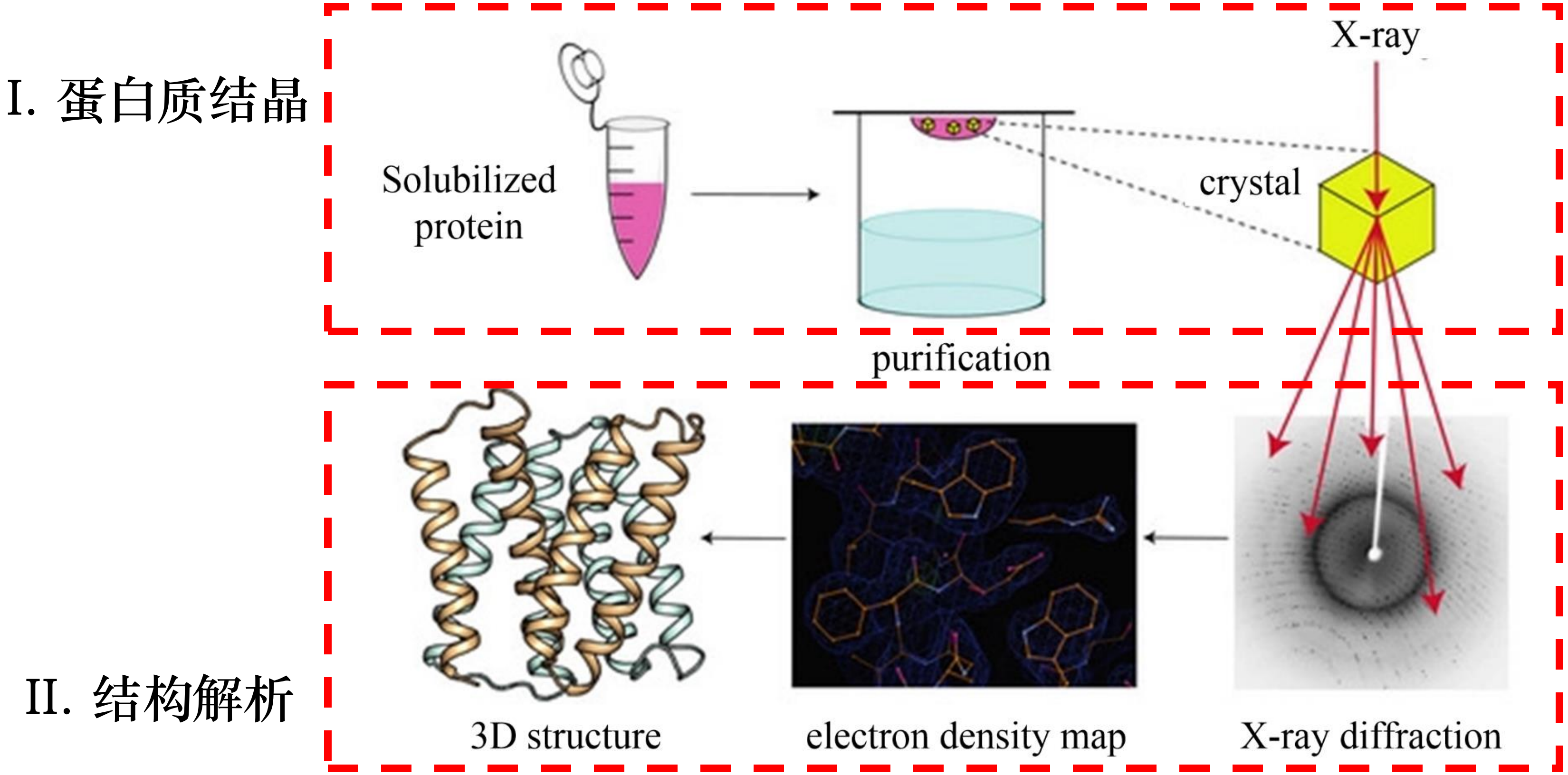
Released	2025-09-03
Method	X-RAY DIFFRACTION 2.378 Å
Organisms	Renilla reniformis
Macromolecule	Coelenterazine h 2-monooxygenase (protein) Green fluorescent protein (protein)
Unique Ligands	CEI

8S13 | pdb_00008s13
Download File | View File
apo c-KIT-wt kinase domain
Teuber, A., Kleinboelting, S.B., Rauh, D., Mueller, M.P.
To be published

Released	2025-09-03
Method	X-RAY DIFFRACTION 2 Å
Organisms	Homo sapiens
Macromolecule	Mast/stem cell growth factor receptor Kit (protein)

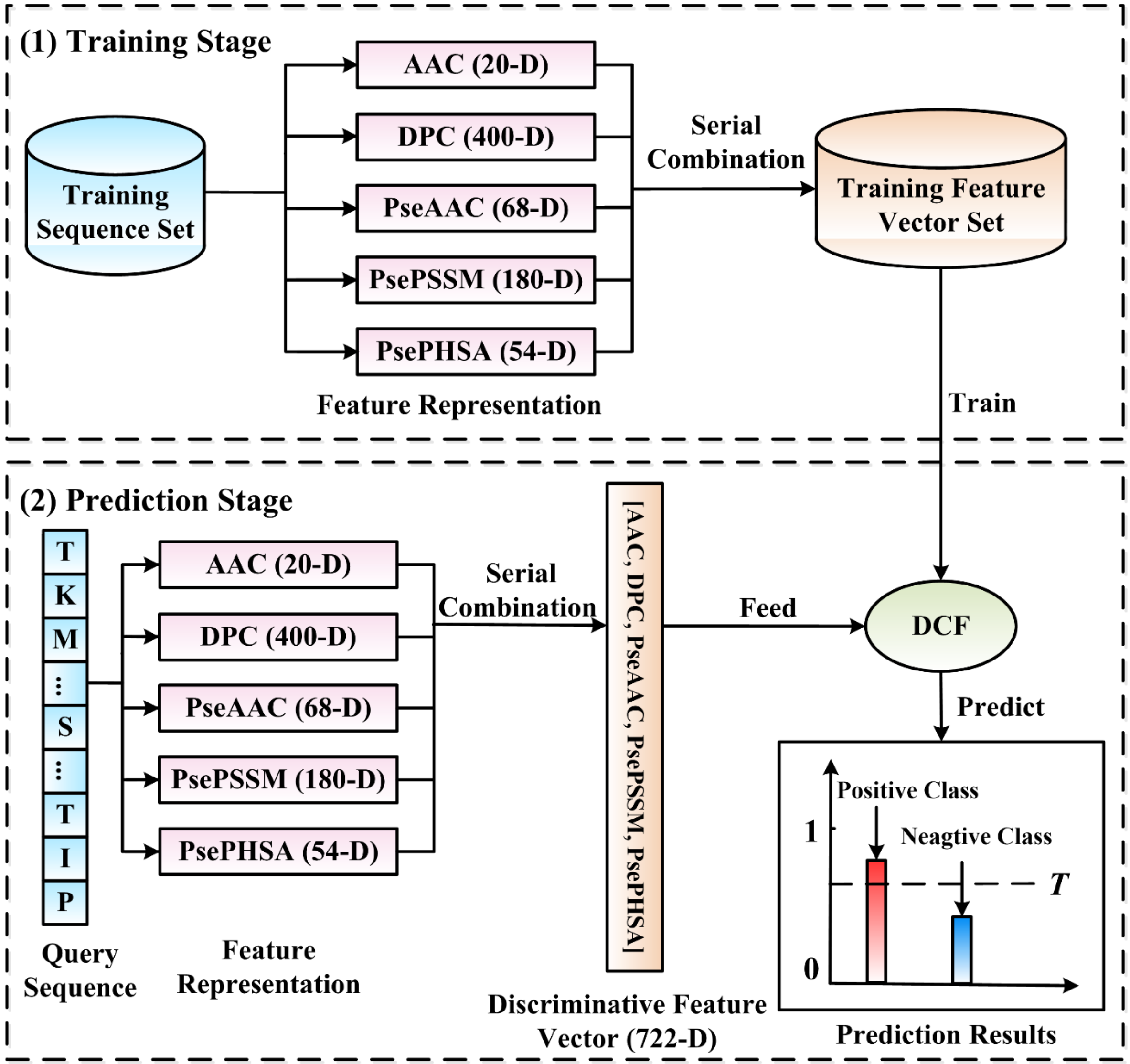
02 蛋白质结晶倾向性预测研究的必要性

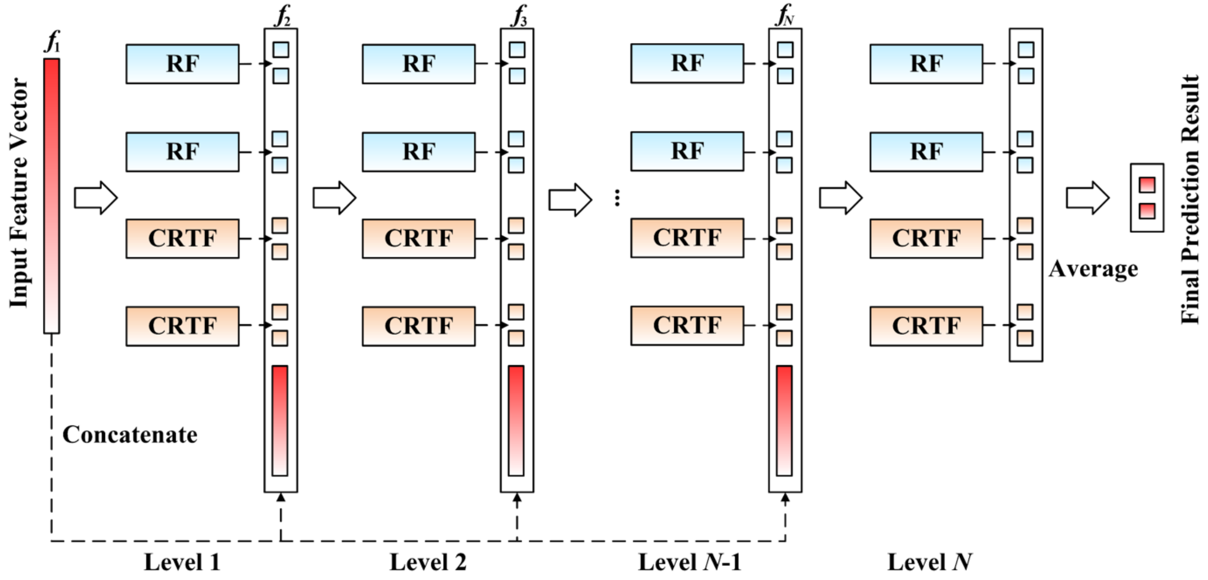
- 在实际的蛋白质结构解析过程中， X射线晶体衍射技术的成功率只有10%左右。其主要原因是大量实验蛋白无法得到可供衍射的晶体。



X射线晶体衍射测定蛋白质三维结构的步骤

03 基于多视角特征融合与深度级联随机森林的预测方法 DCFCrystal

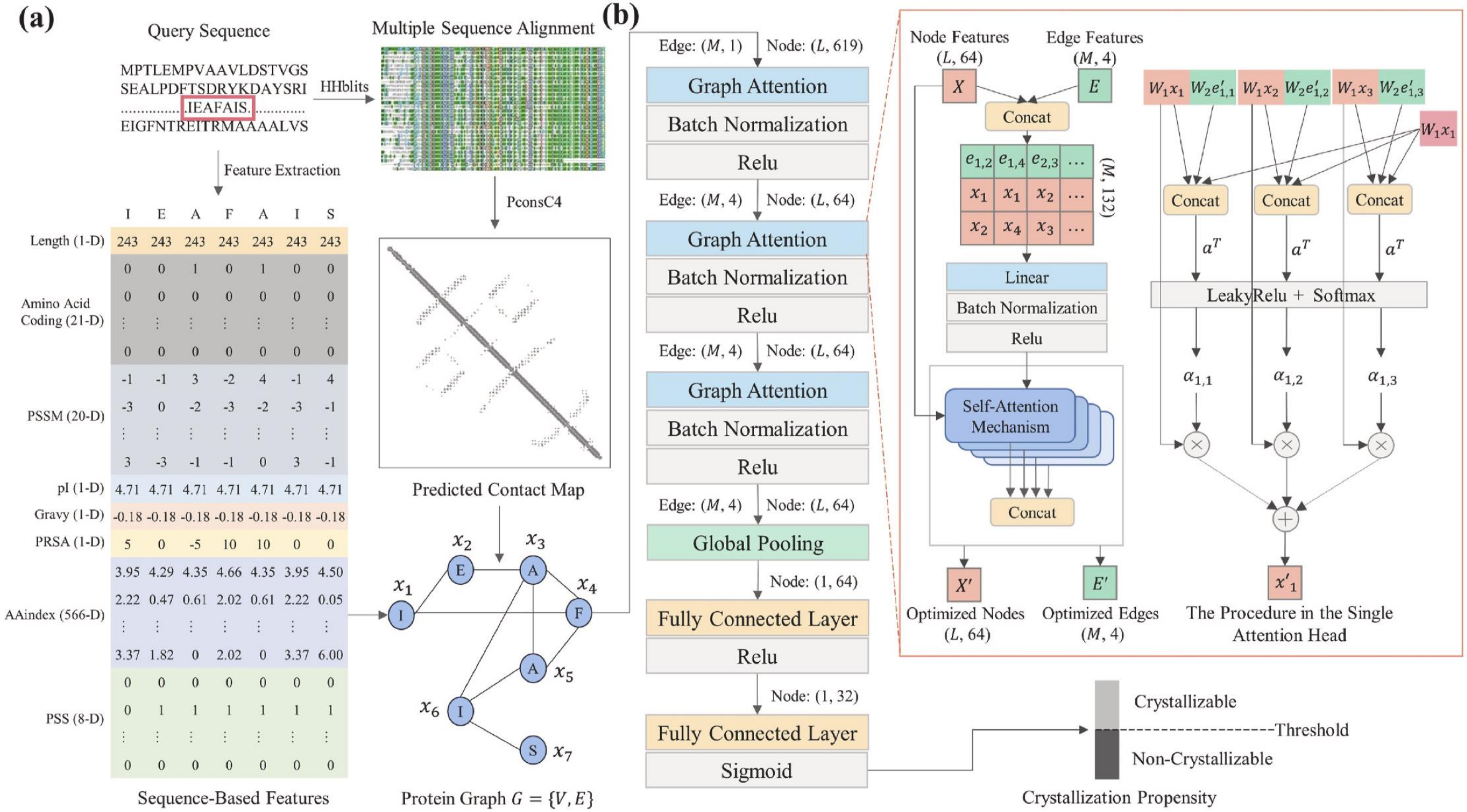




DCFCrystal 工作框架图

Yi-Heng Zhu, Jun Hu, Fang Ge, Fuyi Li, Jiangning Song*, Yang Zhang*, Dong-Jun Yu*. Accurate Multi-Stage Prediction of Protein Crystallization Propensity Using Deep-Cascade Forest with Sequence-Based Features. Briefings in Bioinformatics, 2021.

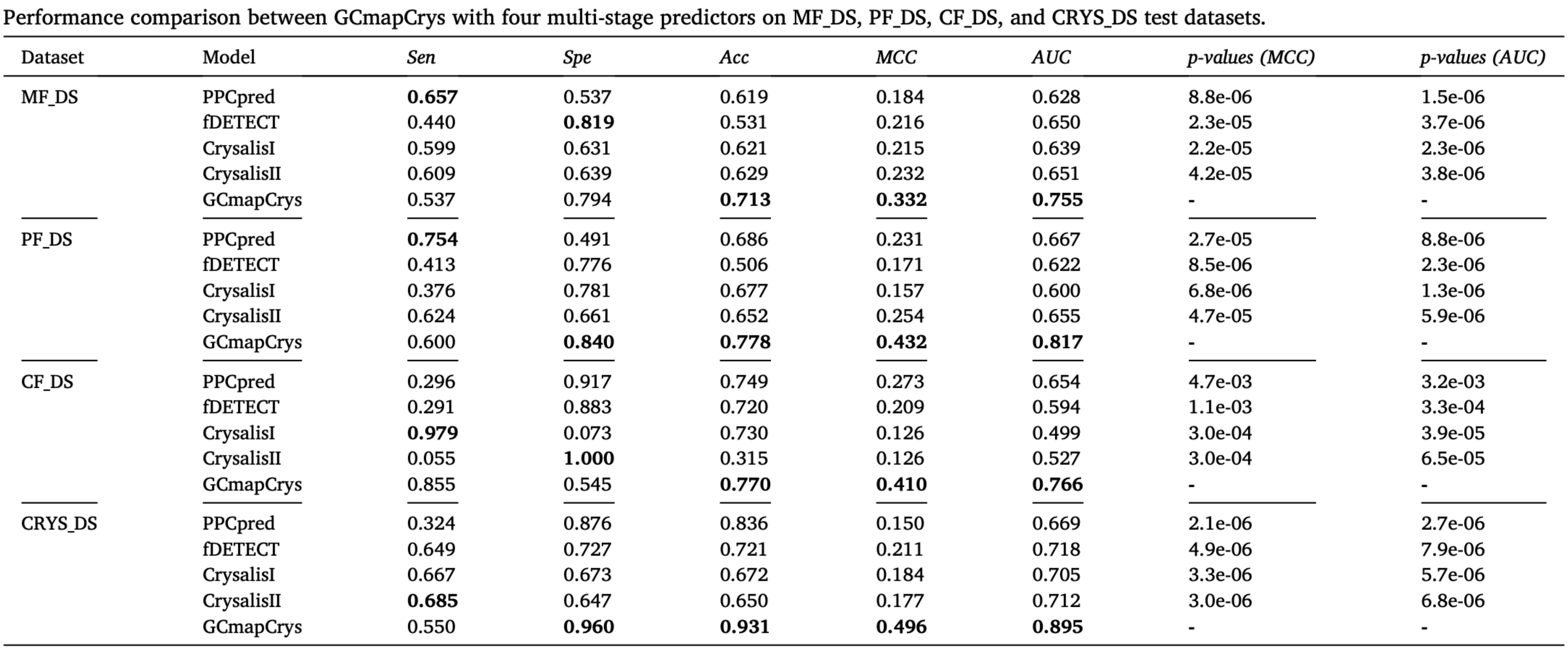
04 基于图注意力网络与残基接触图的预测方法 GCmapCrys



	I	E	A	F	A	I	S
Length (1-D)	243	243	243	243	243	243	243
Amino Acid Coding (21-D)	0	0	1	0	1	0	0
	0	0	0	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	0	0	0
PSSM (20-D)	-1	-1	3	-2	4	-1	4
	-3	0	-2	-3	-2	-3	-1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	3	-3	-1	-1	0	3	-1
pI (1-D)	4.71	4.71	4.71	4.71	4.71	4.71	4.71
Gravy (1-D)	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18
PRSA (1-D)	5	0	-5	10	10	0	0
AAindex (566-D)	3.95	4.29	4.35	4.66	4.35	3.95	4.50
	2.22	0.47	0.61	2.02	0.61	2.22	0.05
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	3.37	1.82	0	2.02	0	3.37	6.00
PSS (8-D)	0	0	0	0	0	0	0
	0	1	1	1	1	1	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	0	0	0

Peng-Hao Wang#, Yi-Heng Zhu#, Xibei Yang, Dong-Jun Yu. GCmapCrys: Integrating Graph Attention Network with Predicted Contact Map for Multi-Stage Protein Crystallization Propensity Prediction. Analytical Biochemistry, 2023.

04 基于图注意力网络与残基接触图的预测方法 GCmapCrys



Performance comparison between GCmapCrys with four multi-stage predictors on MF_DS, PF_DS, CF_DS, and CRYS_DS test datasets.

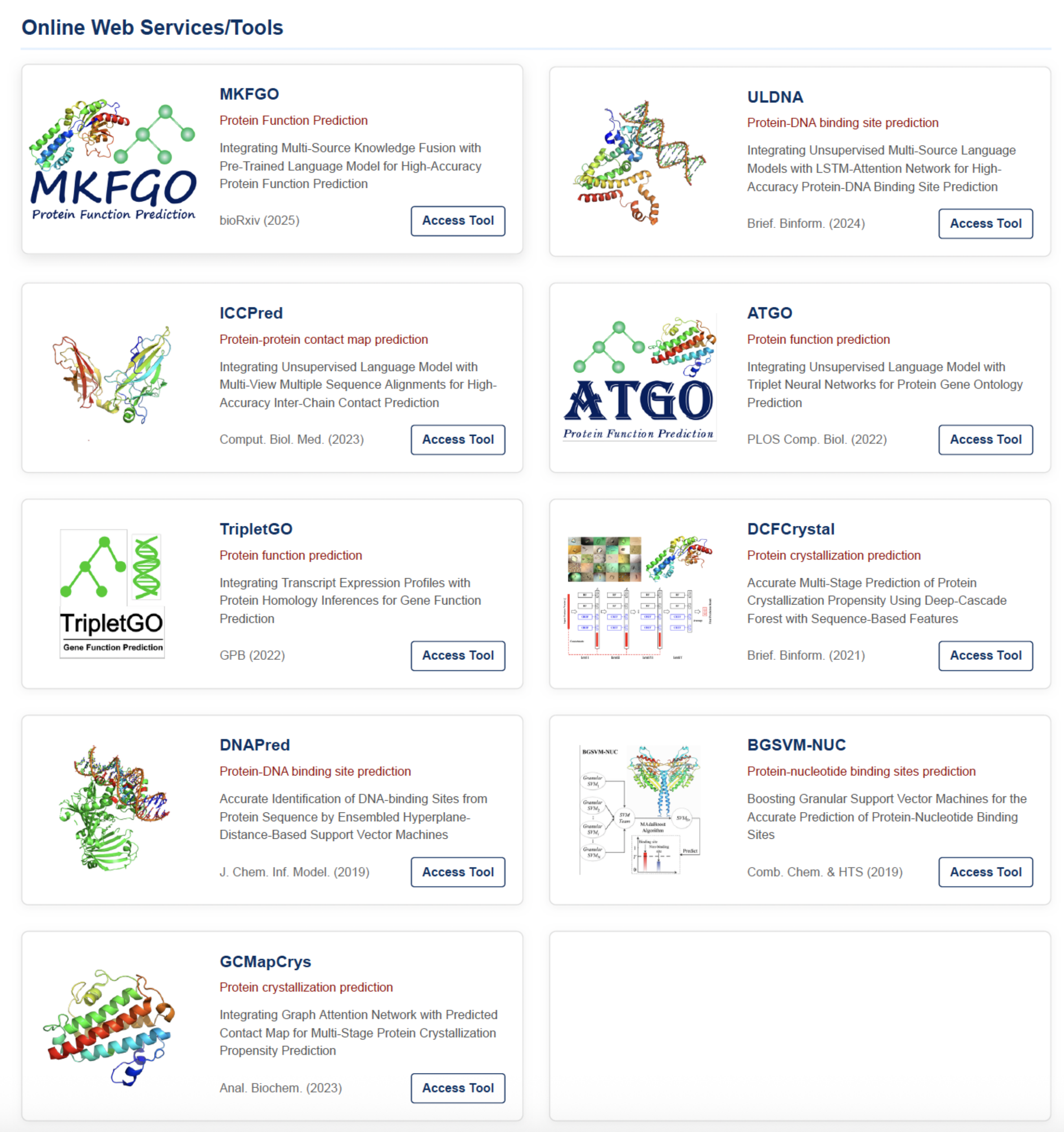
Dataset	Model	*Sen*	*Spe*	*Acc*	*MCC*	*AUC*	*p-values (MCC)*	*p-values (AUC)*
MF_DS	PPCpred	**0.657**	0.537	0.619	0.184	0.628	8.8e-06	1.5e-06
	fDETECT	0.440	**0.819**	0.531	0.216	0.650	2.3e-05	3.7e-06
	CrysalisI	0.599	0.631	0.621	0.215	0.639	2.2e-05	2.3e-06
	CrysalisII	0.609	0.639	0.629	0.232	0.651	4.2e-05	3.8e-06
	GCmapCrys	0.537	0.794	**0.713**	**0.332**	**0.755**	-	-
PF_DS	PPCpred	**0.754**	0.491	0.686	0.231	0.667	2.7e-05	8.8e-06
	fDETECT	0.413	0.776	0.506	0.171	0.622	8.5e-06	2.3e-06
	CrysalisI	0.376	0.781	0.677	0.157	0.600	6.8e-06	1.3e-06
	CrysalisII	0.624	0.661	0.652	0.254	0.655	4.7e-05	5.9e-06
	GCmapCrys	0.600	**0.840**	**0.778**	**0.432**	**0.817**	-	-
CF_DS	PPCpred	0.296	0.917	0.749	0.273	0.654	4.7e-03	3.2e-03
	fDETECT	0.291	0.883	0.720	0.209	0.594	1.1e-03	3.3e-04
	CrysalisI	**0.979**	0.073	0.730	0.126	0.499	3.0e-04	3.9e-05
	CrysalisII	0.055	**1.000**	0.315	0.126	0.527	3.0e-04	6.5e-05
	GCmapCrys	0.855	0.545	**0.770**	**0.410**	**0.766**	-	-
CRYS_DS	PPCpred	0.324	0.876	0.836	0.150	0.669	2.1e-06	2.7e-06
	fDETECT	0.649	0.727	0.721	0.211	0.718	4.9e-06	7.9e-06
	CrysalisI	0.667	0.673	0.672	0.184	0.705	3.3e-06	5.7e-06
	CrysalisII	**0.685**	0.647	0.650	0.177	0.712	3.0e-06	6.8e-06
	GCmapCrys	0.550	**0.960**	**0.931**	**0.496**	**0.895**	-	-

GCmapCrys 与SOTA在不同基准数据集上的性能比较

已开发的生物信息学工具:

- 蛋白质功能预测
- 基因功能预测
- 蛋白质-配体相互作用预测
- 蛋白质链间接触图预测
- 蛋白质结晶倾向性预测

https://yiheng-zhu.github.io/Yiheng/index.html#services



Online Web Services/Tools

MKFGO
Protein Function Prediction
Integrating Multi-Source Knowledge Fusion with Pre-Trained Language Model for High-Accuracy Protein Function Prediction
bioRxiv (2025)
Access Tool

ULDNA
Protein-DNA binding site prediction
Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for High-Accuracy Protein-DNA Binding Site Prediction
Brief. Binform. (2024)
Access Tool

ICCPred
Protein-protein contact map prediction
Integrating Unsupervised Language Model with Multi-View Multiple Sequence Alignments for High-Accuracy Inter-Chain Contact Prediction
Comput. Biol. Med. (2023)
Access Tool

ATGO
Protein function prediction
Integrating Unsupervised Language Model with Triplet Neural Networks for Protein Gene Ontology Prediction
PLOS Comp. Biol. (2022)
Access Tool

TripletGO
Protein function prediction
Integrating Transcript Expression Profiles with Protein Homology Inferences for Gene Function Prediction
GPB (2022)
Access Tool

DCFCrystal
Protein crystallization prediction
Accurate Multi-Stage Prediction of Protein Crystallization Propensity Using Deep-Cascade Forest with Sequence-Based Features
Brief. Binform. (2021)
Access Tool

DNAPred
Protein-DNA binding site prediction
Accurate Identification of DNA-binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines
J. Chem. Inf. Model. (2019)
Access Tool

BGSVM-NUC
Protein-nucleotide binding sites prediction
Boosting Granular Support Vector Machines for the Accurate Prediction of Protein-Nucleotide Binding Sites
Comb. Chem. & HTS (2019)
Access Tool

GCMapCrys
Protein crystallization prediction
Integrating Graph Attention Network with Predicted Contact Map for Multi-Stage Protein Crystallization Propensity Prediction
Anal. Biochem. (2023)
Access Tool

科研团队信息

- 教授1人
- 副教授1人
- 讲师2人
- 博士研究生3人
- 硕士研究生9人

团队主页：cdsic.njau.edu.cn

研究方向

人工智能与模式识别

人工智能的理论及应用
大数据计算与模式识别

生物信息与系统生物学

多组学数据整合分析与计算
复杂生物系统的数学建模与预测

致谢主要合作者

於东军, 教授,
南京理工大学

Yang Zhang, 教授
新加坡国立大学

Jiangning Song，教授
莫纳什大学

郑伟，教授
南开大学

张成辛，教授
中国科学院

邱望仁，教授
景德镇陶瓷大学

刘岩，教授
扬州大学

葛芳，副教授
南京邮电大学

闫贺，副教授
南京林业大学

刘子，讲师
景德镇陶瓷大学

谢谢各位专家观看

请各位专家批评指正！