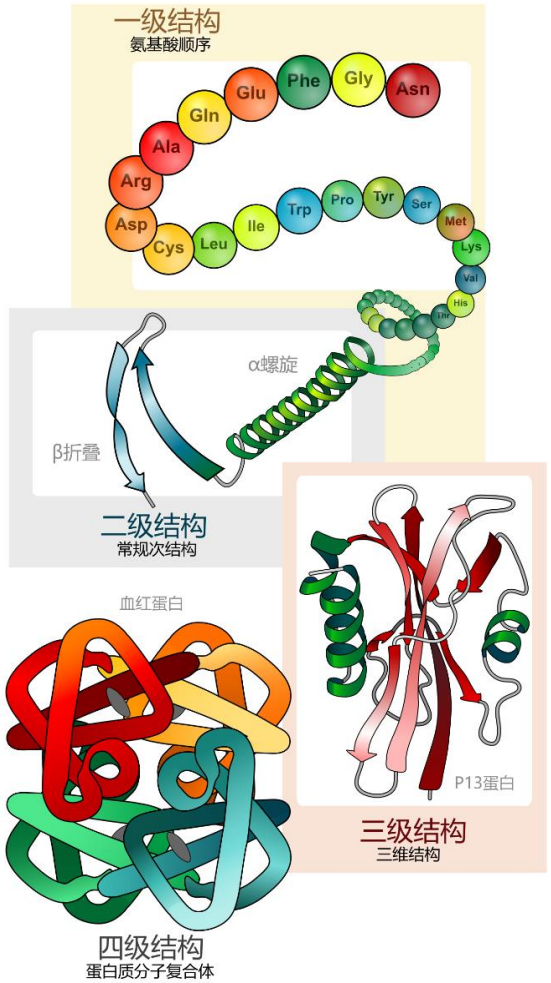
基于多源知识融合的高精度蛋白质功能预测研究

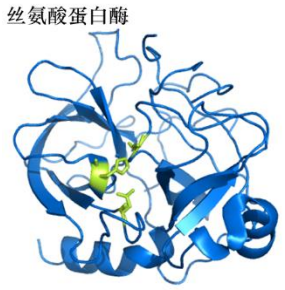
南京农业大学 智慧农业学院

汇报人：朱一亨

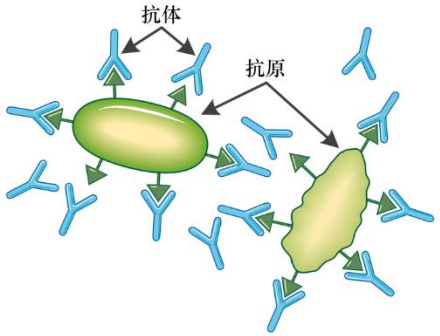
2025年7月5日

01 蛋白质的生物功能



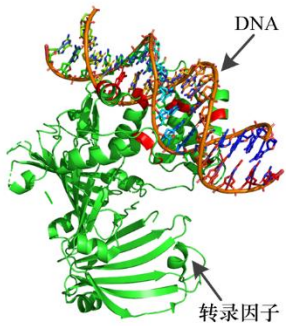


(a) 催化反应



(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)

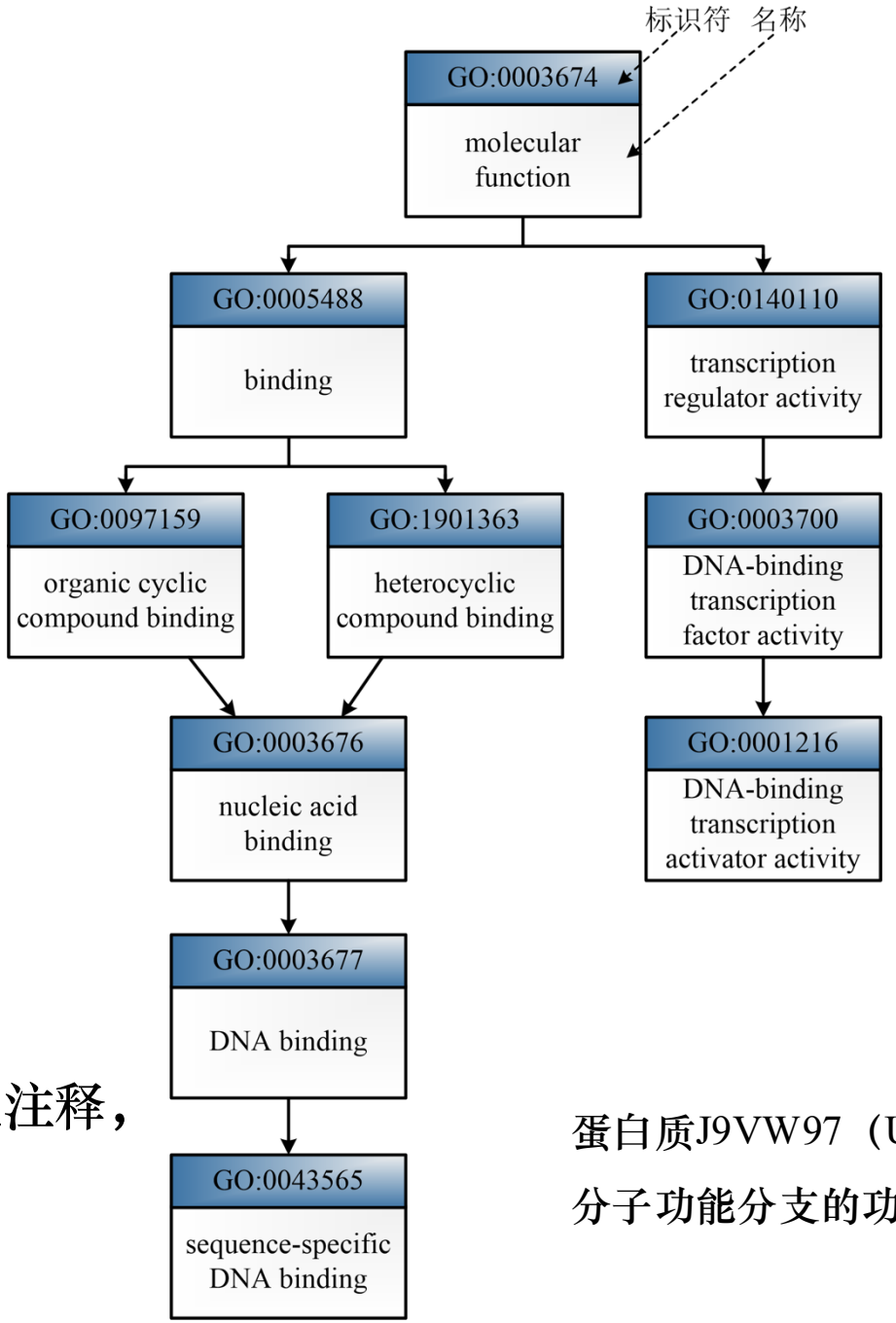
  分子功能 (Molecular Function, MF)

  生物过程 (Biological Process, BP)

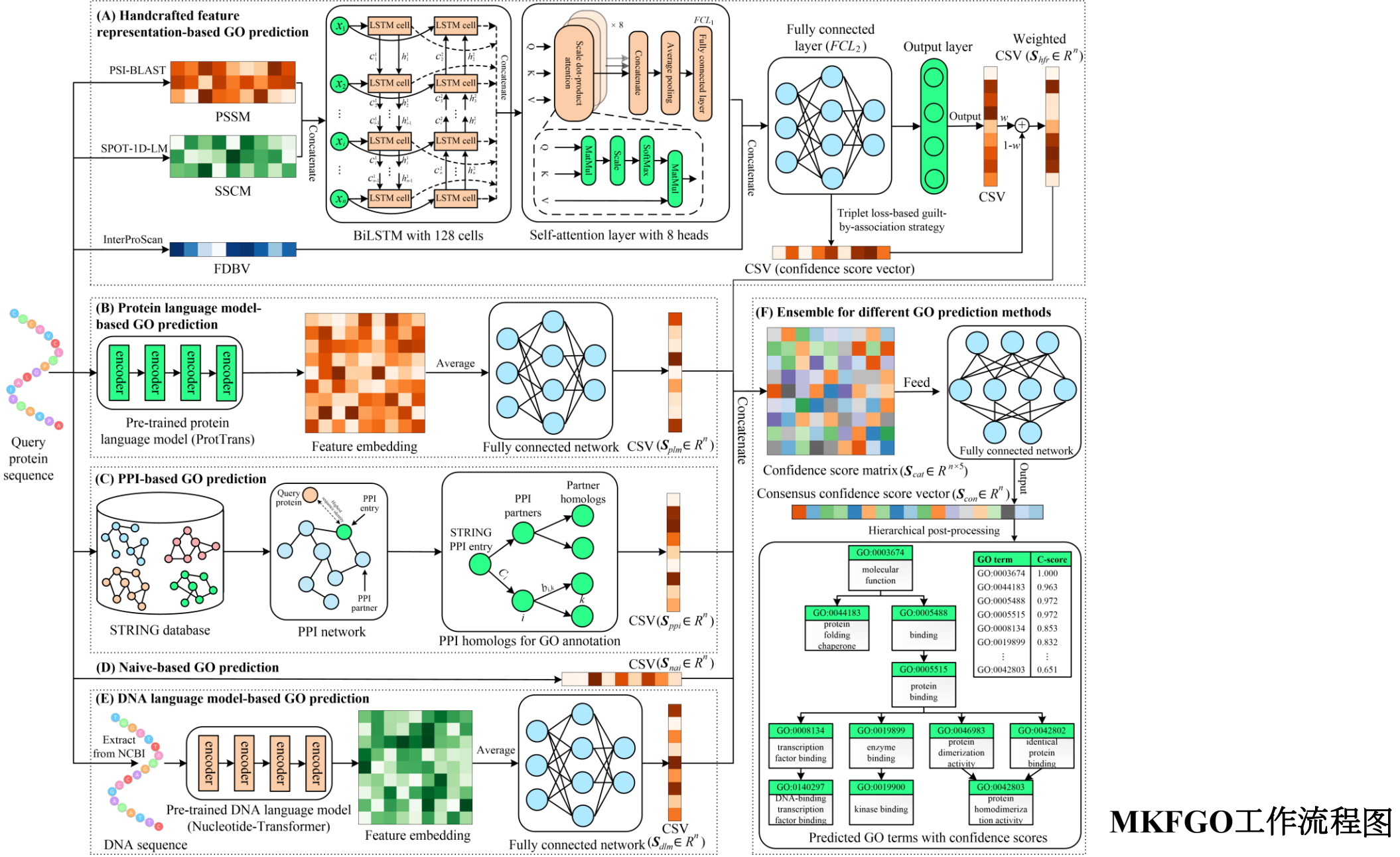
  细胞组件 (Cellular Component, CC)

- 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。



蛋白质J9VW97（UniProt ID）在分子功能分支的功能注释图



MKFGO工作流程图

MKFGO: Integrating Multi-Source Knowledge Fusion with Pre-Trained Language Model for High-Accuracy Protein Function Prediction. **Briefings in Bioinformatics**. (Revision and Under Review)

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

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.

Contact @Yi-Heng Zhu

MKFGO的输入页面

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
TGGCCCAGGGTGGTATGTAATTTTCGCTCGGCCTGGGACGGGGCCCAGGCCGGGCCCAGCCTGGTGGAGCGTCCAGGTCTGG
GTGCGAAGCCAGGCCCCTGGGCGGAGGTGAGGGGTGGTCTGAGG
```

MKFGO Prediction Results (Model I)

[Download result.zip for all prediction results]

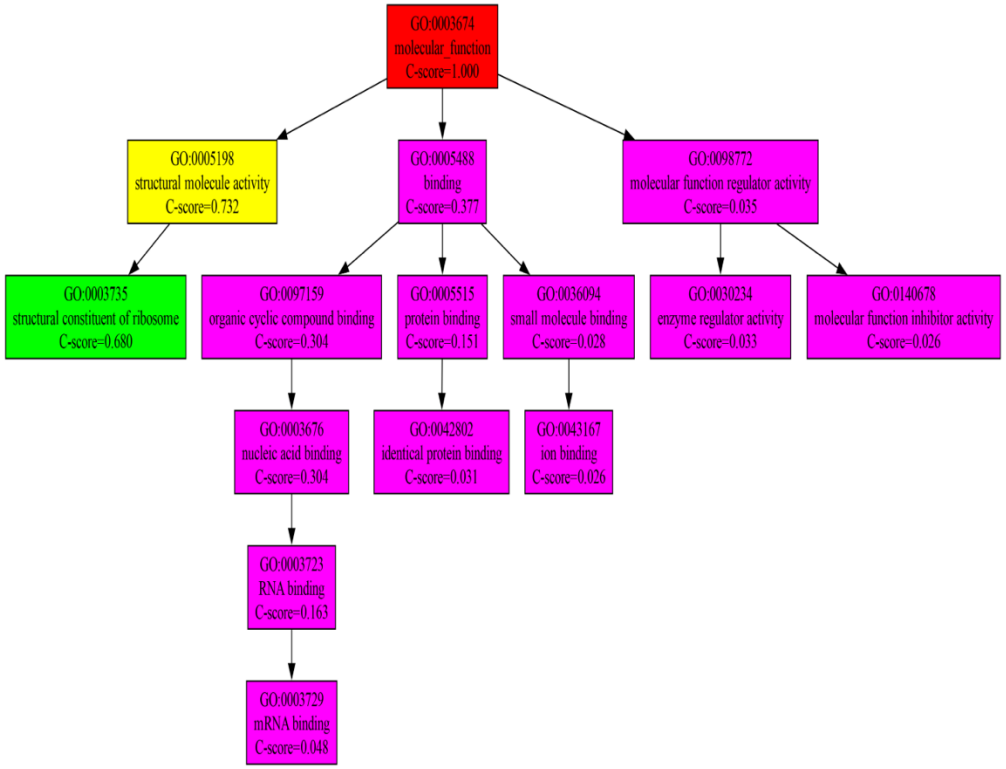
A0A1D8PPE0

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]

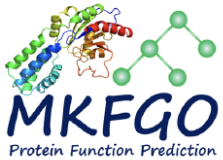
MKFGO的预测结果页面

已开发的生物信息学工具:

- 蛋白质功能预测
- 非编码基因功能预测
- 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

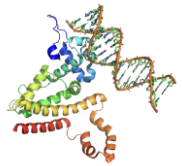


MKFGO

Protein Function Prediction

Integrating Multi-Source Knowledge Fusion with Pre-Trained Language Model for High-Accuracy Protein Function Prediction

bioRxiv (2025) Access Tool

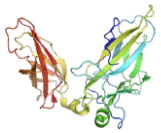


ULDNA

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

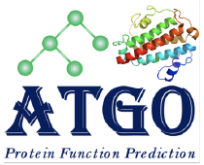


ICCPred

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

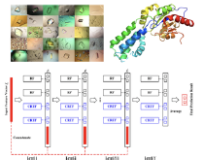
PLOS Comp. Biol. (2022) Access Tool

TripletGO

Protein function prediction

Integrating Transcript Expression Profiles with Protein Homology Inferences for Gene Function Prediction

GPB (2022) Access Tool

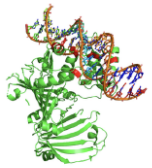


DCFCrystal

Protein crystallization prediction

Accurate Multi-Stage Prediction of Protein Crystallization Propensity Using Deep-Cascade Forest with Sequence-Based Features

Brief. Binform. (2021) Access Tool

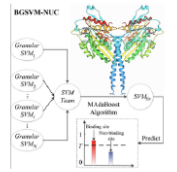


DNAPred

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

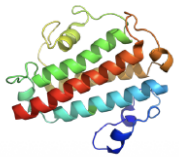


BGSVM-NUC

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

Publications

[1] **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**. (Revision and Under Review).

[2] Rui Ye, Yu Ding, Jing Zhang *, **Yi-Heng Zhu** *. A Transformer-Based Transfer Learning Algorithm for Time Series Imputation and Forecasting with Data Scarcity. **Expert Systems With Applications**. (Under Review).

[3] 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: 115862.

[4] **Yi-Heng Zhu**, Zi Liu, Yu Ding, Zhiwei Ji*, and Dong-Jun Yu*. Machine Learning for Protein Function Prediction, Chapter in the book of "Protein Function Prediction" [M], **Elsevier**, 2025, DOI: 10.1007/978-1-0716-4662-5.

主要研究方向

(1) 生物大分子的功能注释

- 蛋白质功能预测 （Protein Function Prediction）
- 酶功能预测 (Enzyme Function Prediction)
- 非编码基因功能预测 (Non-Coding RNA Function Prediction)

(2) 生物大分子的相互作用预测

- 蛋白质-配体的相互作用/亲和力预测 (Protein-Ligand Interaction/Binding Affinity Prediction)
- 药物-靶标相互作用/亲和力预测 (Drug-Target Interaction/Binding Affinity Prediction)

(3) 蛋白质设计（Protein Design: Backbone-to-Sequence Reconstruction）

谢谢各位专家观看

请各位专家批评指正！