
# 科研经验交流分享

南京农业大学 人工智能学院
汇报人：朱一亨
2024年10月09日

# 目 录

01 Part one

# 前言和思考

# 为什么要读研/读博?

- 就业
- 深厚的知识储备
- 独立分析/解决问题的能力
- 精神财富（面对挫折的勇气，抗压能力等）

# 科研进展不顺时，如何调整心态?

- 找到释放压力的途径（运动，游戏，旅行）
- 多跟师兄、老师、亲友交流
- 科研从来都不是一帆风顺，需要不断试错

02 Part two

# 如何独立开展科研工作

# 如何独立的开展科研工作

- 选题
- 科研方法和计划安排
- 实验结果分析
- 定期交流
- 论文写作

# 选题

- 导师建议
- 自我兴趣

一个好的课题需满足以下特点：

（1）科学价值

（2）与课题组/学校的优势相结合

（3）可行性

（4）关注前沿热点

（5）持续性

# 科研方法和计划安排

- 文献调研

（1）本研究领域有哪些好期刊?

（2）在谷歌学术/固定期刊上根据(研究课题)关键字搜索

（3）选择20-30篇（一区以上）期刊泛读

（读摘要，解决了什么具体问题，采用了什么方法）

（4）选1-2篇论文精读，复现论文（跑代码，实验结果和分析）

- 文献调研的预期效果：

（1）明确课题的研究意义和重要性

（2）该领域具体有哪几类方法？Baseline是什么？

（3）近3年提出了哪些重要的方法？

（4）该领域还存在哪些不足和挑战？

（5）养成整理文献的好习惯

## Protein Function Prediction

### Overview

1. 深度学习在蛋白质功能预测中的应用. 合成生物学, 2023. [PDF]
2. A comprehensive computational benchmark for evaluating deep learning-based protein function prediction approaches. **Briefings in Bioinformatics**, 2024. [PDF]
3. Protein function prediction with gene ontology: from traditional to deep learning models. **PeerJ**, 2021. [PDF]
4. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. **Genome Biology**, 2019. [PDF]
5. Protein function annotation using protein domain family resources. **Methods**, 2016. [PDF]

### Evaluation Metrics

1. A large-scale assessment of sequence database search tools for homology-based protein function prediction. **bioRxiv**, 2023. [PDF]
2. Evaluation: A large-scale evaluation of computational protein function prediction. **Nature Methods**, 2013. [PDF]

### Toolbars

1. InterPro in 2022. **Nucleic Acids Research**, 2022. [PDF] [Web Server]

### Public Database

1. Protein sequence database: UniProt.
2. Protein structure database: PDB, AlphaFold database.
3. Protein function database: Gene Ontology, GOA.
4. Protein-ligand strcture database: BioLip.
5. Gene co-expression database: COXPRESdb, ATTED-II.

### Template-Based Methods

1. QAUST: Protein function prediction using structure similarity, protein interaction, and functional motifs. **Genomics Proteomics Bioinformatics**, 2021. **Source: structure, protein-protein network, and functional motifs.** [PDF]
2. MLC: Metric learning on expression data for gene function prediction. **Bioinformatics**, 2020. **Source: gene expression.** [PDF] [Code]
3. INGA 2.0: Improving protein function prediction for the dark proteome. **Nucleic Acids Research**, 2019. **Source: sequenc,protein-protein network, domain.** [PDF] [Web Server]
4. MetaGO: Predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. **Journal of Molecular Biology**, 2018. **Source: sequence, structure, and protein-protein network.** [PDF] [Web Server]

### Pre-Trained Model-Based Methods

1. AnnoPRO: A strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. **Genome Biology**, 2024. [PDF] [Code]
2. DualNetGO: A dual network model for protein function prediction via effective feature selection. **Bioinformatics**, 2024. [PDF] [Code]
3. Domain-PFP: Protein function prediction using function-aware domain embedding representations. **Communications Biology**, 2023. [PDF] [Code]
4. CFAGO: Cross-fusion of network and attributes based on attention mechanism for protein function prediction. **Bioinformatics**, 2023. [PDF] [Code]
5. MELISSA: Semi-supervised embedding for protein function prediction across multiple networks **bioRxiv**, 2023. [PDF] [Code]
6. HiFun: Homology independent protein function prediction by a novel protein-language self-attention model. **Briefings in Bioinformatics**, 2023. [PDF] [Code]
7. PredGO: Large-scale predicting protein functions through heterogeneous feature fusion. **Briefings in Bioinformatics**, 2023. [PDF] [Code]
8. MGEGFP: A multi-view graph embedding method for gene function prediction based on adaptive estimation with GCN. **Briefings in Bioinformatics**, 2023. [PDF] [Code]
9. MMSMAPlus: A multi-view multi-scale multi-attention embedding model for protein function prediction. **Briefings in Bioinformatics**, 2023. [PDF] [Code]
10. PFmulDL: A novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. **Computers in Biology and Medicine**, 2022. [PDF] [Code]
11. DeepFRI: Structure-based protein function prediction using graph convolutional networks. **Nature Communications**, 2021. [PDF] [Web Server] [Code]
12. deepNF: Deep network fusion for protein function prediction. **Bioinformatics**, 2018. [PDF] [Code]

### Large Language Model-Based Methods

1. DeepGO-SE: Protein function prediction as approximate semantic entailment. **Nature Machine Intelligence**, 2024. **Model: ESM2.** [PDF] [Code]
2. PhiGnet: Accurate prediction of protein function using statistics-informed graph networks. **Nature Communications**, 2024. **Model: ESM-1b.** [PDF] [Code]
3. GPSFun: Geometry-aware protein sequence function predictions with language models. **Nucleic Acids Research**, 2024. **Model: ESM2.** [PDF] [Web Server]
4. DeepGOMeta: Predicting functions for microbes. **bioRxiv**, 2024. [PDF] [Code]
5. GNNGO3D: Protein function prediction based on 3D structure and functional hierarchy learning. **IEEE Transactions on Knowledge and Data Engineering**, 2023. **Model: ESM-1b.** [PDF]
6. Struct2GO: Protein function prediction based on graph pooling algorithm and AlphaFold2 structure information. **Bioinformatics**, 2023. **Model: SeqVec.** [PDF] [Code]
7. SPROF-GO: Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. **Briefings in Bioinformatics**, 2023.

# 科研方法和计划安排

- 科研方法

（1）循序渐进（在现有的模型上逐步改进，不要期望一步到位）

（2）先构建简单模型（baseline），再逐步增加新模块（便于观察模块的效果）

（3）设计实验方案/模型时，不要机械搬运，要思考合理性和生物学意义

（4）数据集的构建（挑选有潜力的数据集）

（5）常用的代码/模块要写注释，并上传到Github

（6）重要的实验结果和数据，要及时记录和保留

# 科研方法和计划安排

- 计划安排

（1）时间保证 (珍惜时间，有效时间，一周最低40小时)

（2）每天做好计划安排（写清单）

（3）有好的idea，要随时记录下来

# 实验结果分析

（1）选择对比方法时，选近3-5年高水平期刊的论文方法

（2）消融实验要充分

（3）要考虑实验数据的随机性

（4）负面的实验数据，要多分析效果不好的原因

# 定期交流

（1）及时找导师交流（准备好PPT，方法和模型细节，实验数据）

（2）多跟同门师兄弟姐妹交流（口头）

（3）组会时多提问

# 论文写作

- 常用的软件

（1）Word/LaTeX

（2）Endnote

（3）PPT/Visio

## ➢ 论文架构

（1）Abstract

（2）Introduction

（3）Method

（4）Result and Discussion

（5）Conclusion

写作顺序: (3)(4)(2)(5)(1)

每写完一段，回过去仔细读一遍，检查每一句描述是否准确，逻辑是否通顺

全部写完以后，再通读1-2遍，看段落间过度是否自然通顺

- Abstract

(1) 非常重要，编辑/审稿人最先看的内容

(2) 体现出研究问题的科学意义、研究方法创新性和研究内容的贡献

(3) 凝练、反复修改、逻辑顺畅

- Introduction

(1) 研究背景（课题的科学意义及重要性）

(2) 文献综述（现有的方法有哪几类，每类有哪些代表性的方法）

(3) 现有方法的不足和挑战

(4) 本文方法针对(3)解决了什么问题，有什么创新性，贡献在哪里，取得了哪些显著的效果?

- Method

(1) 先画流程图，务必做到步骤层次分明、逻辑顺畅

(2) 根据流程图，将自己的方法逐步拆解，每一步取一个标题

(3) 根据每一步的标题，具体写方法细节

- Result

(1) 与现有的方法比较: 8-10个，包含近三年的方法

(2) 消融实验: 分析每个模块的贡献度

(3) Case study：选择具有重要生物学意义的Case分析，将所提方法的贡献进一步升华

- Discussion

（1）所提的方法相比于现有的方法在精度上提升了多少

（2）我们的方法为什么好？别人的方法为什么不好？（结合算法和生物学背景解释）

## ➢ Conclusion

(1) 总结我们方法取得的成果，并简要说明原因 （与Introduction首尾呼应）

(2) 我们的方法还存在哪些不足或提升空间

(3) 指出未来的研究方向

03 Part three

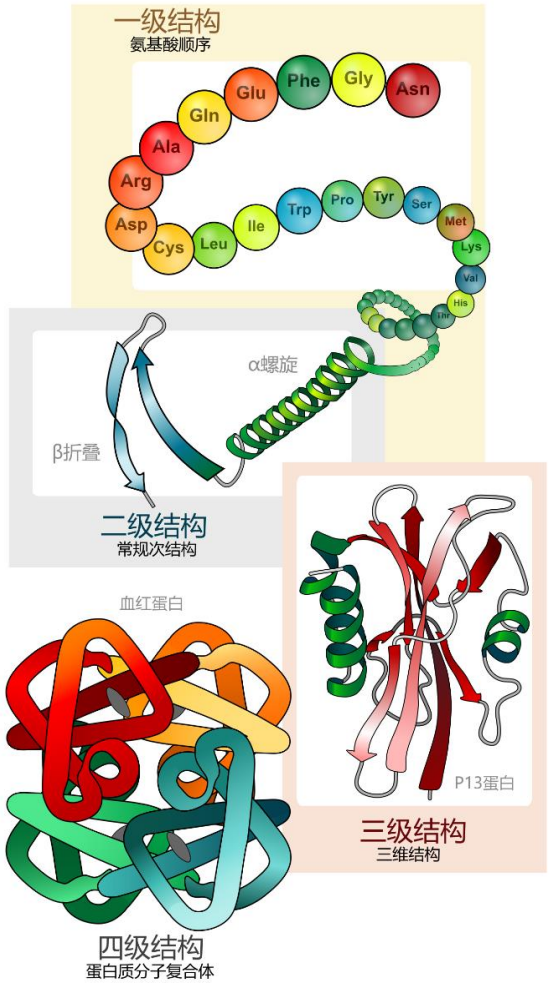
# 我的研究课题

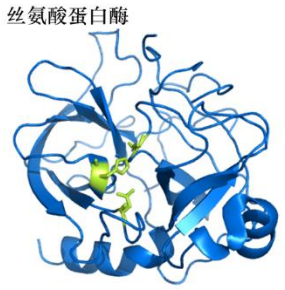
## ➢ 我的研究课题

（1）蛋白质功能预测

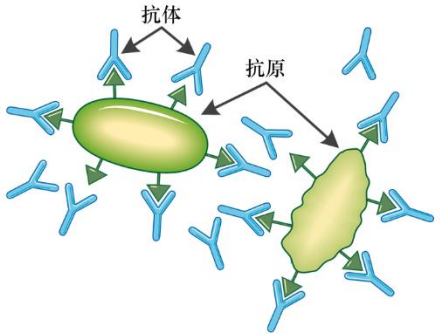
（2）蛋白质-配体绑定位点预测

# 蛋白质的生物功能



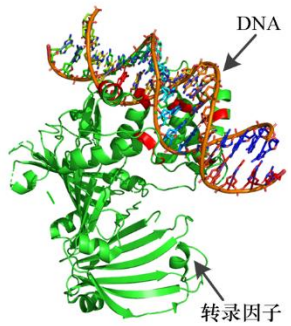




(a) 催化反应




(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。

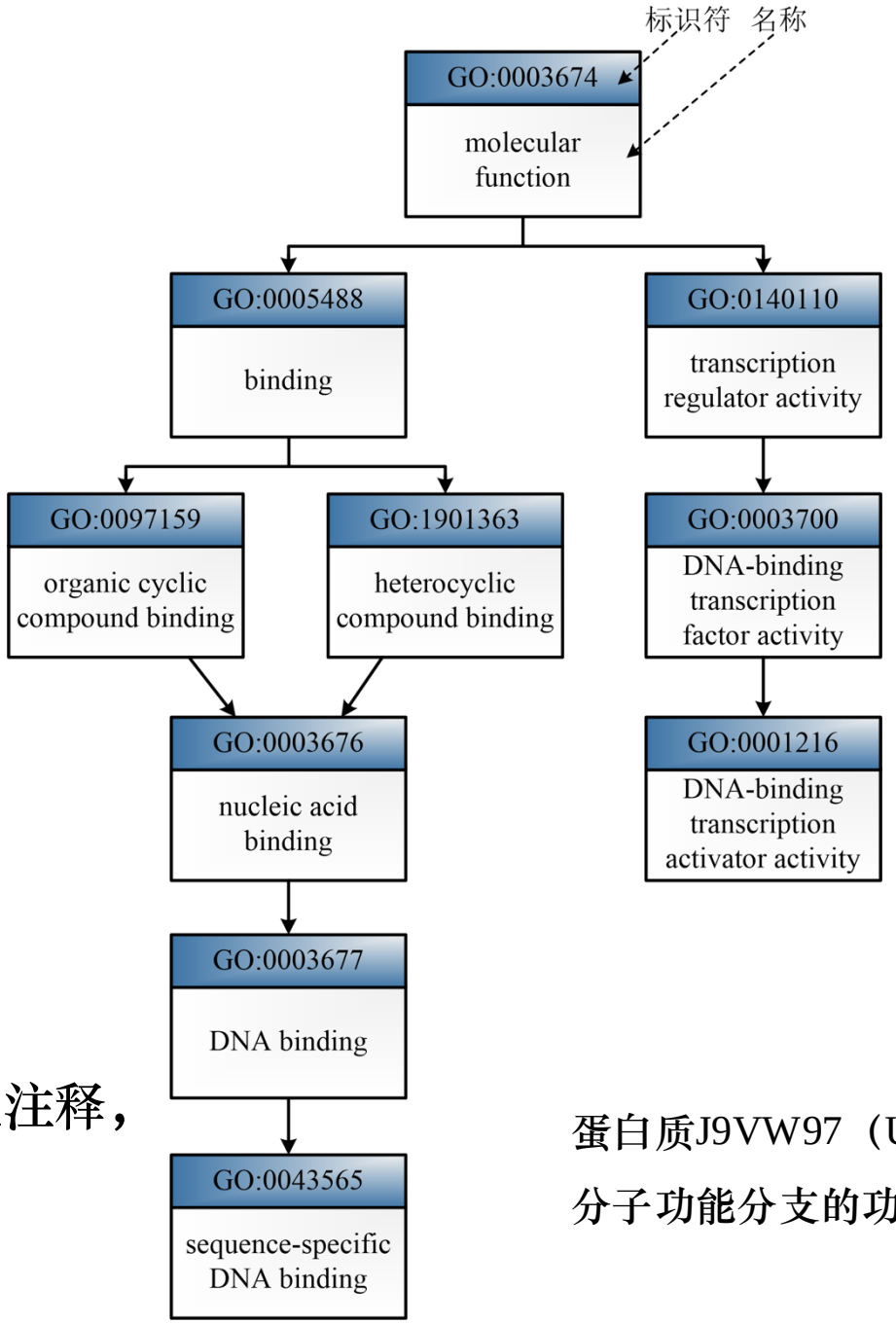
# 蛋白质的功能注释方法

- 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)

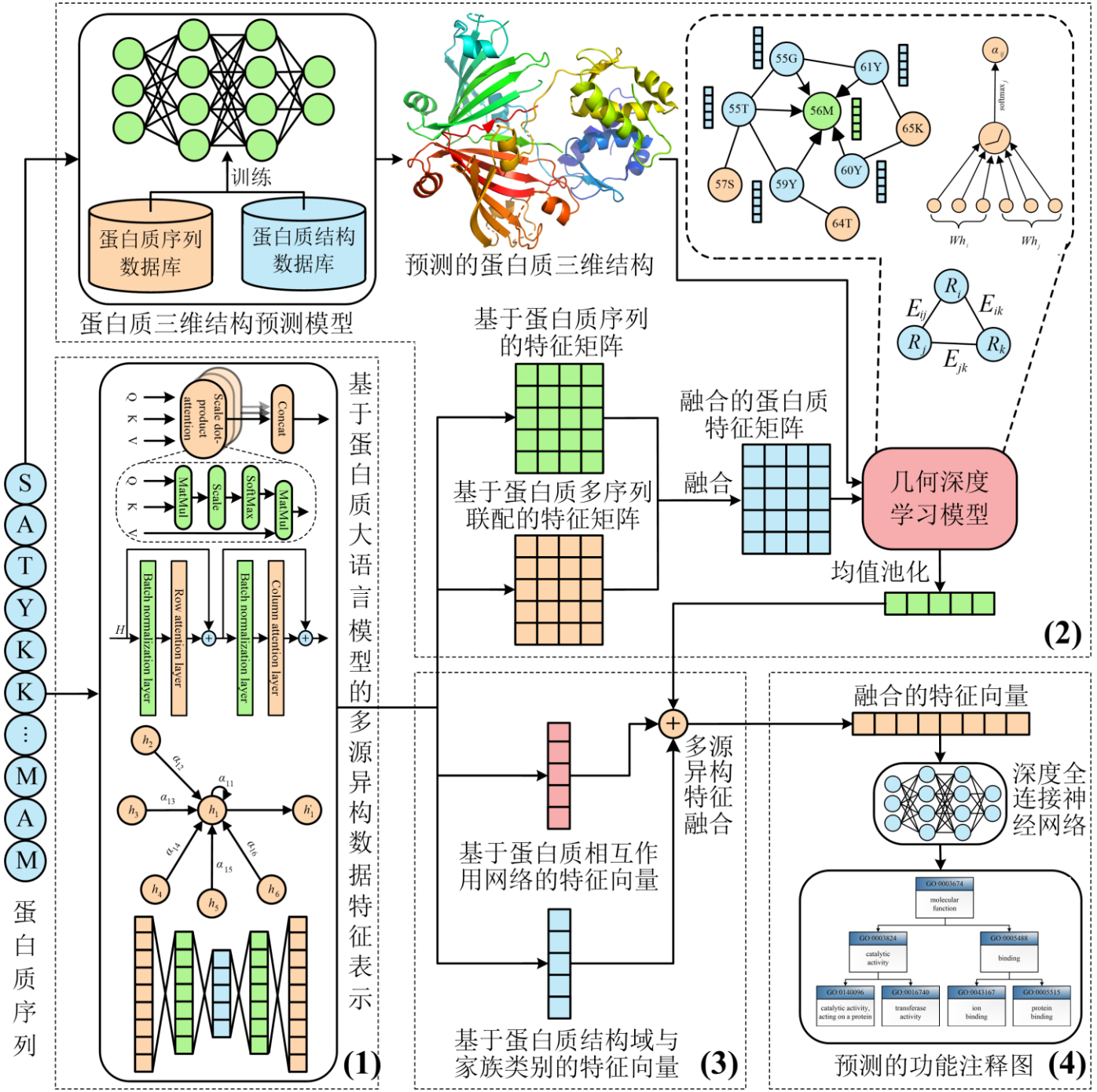



- 蛋白质功能注释目标

  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。

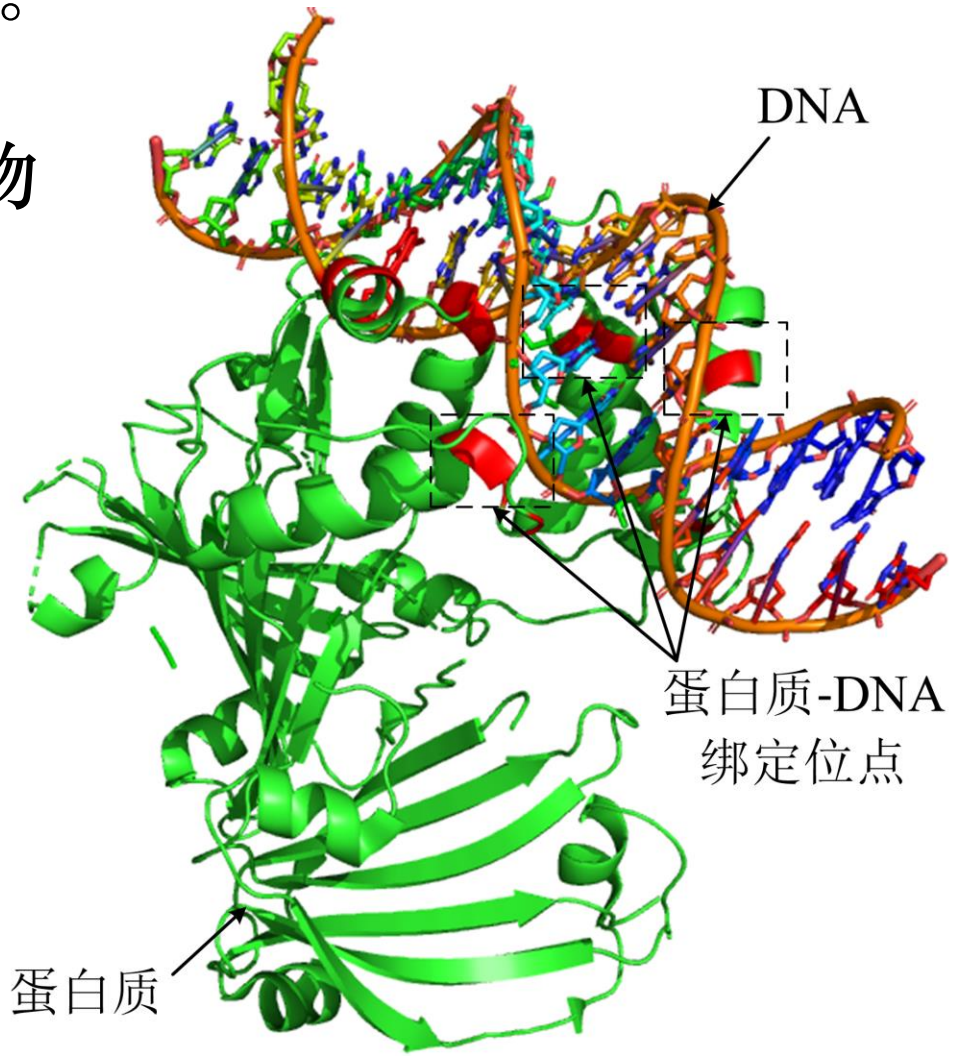
蛋白质J9VW97（UniProt ID）在分子功能分支的功能注释图

# 蛋白质功能预测模型流程图

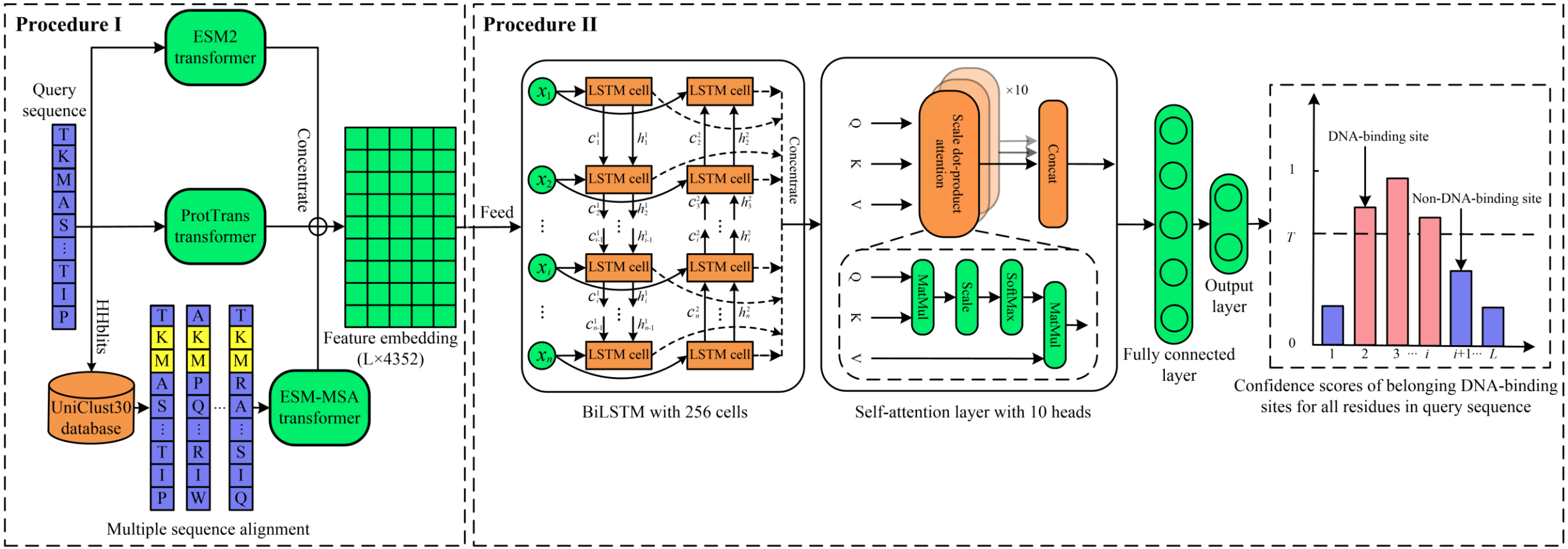


# 蛋白质-配体绑定位点

- 蛋白质在发挥功能时，并不是孤立存在的。
- 蛋白质 + 配体 ⟶ 蛋白质-配体复合物



# 蛋白质-配体绑定位点预测流程图



## ➢ 其他关注的课题

（1）蛋白质-配体亲和力预测

（2）药物-靶标相互作用预测

（3）RNA-蛋白质相互作用预测

（4）蛋白质结晶倾向性预测

（5）蛋白质序列设计

（6）转录因子结合位点预测

（7）RNA甲基化位点预测

谢谢各位老师和同学观看

请批评指正！