# 姚培根

**AI 制药 / 生物医药数据平台开发**

2027 届硕士 · 生物信息学　18162654635　bacontesla314@gmail.com

研究聚焦微生物功能多肽挖掘与生物医学知识智能化，具备大规模数据处理、模型训练评估及在线平台开发经验。

## 教育背景

| | | | |
|---|---|---|---|
| **同济大学** | 985 / 双一流 | 生物信息学 · 硕士 | 2024 — 2027 |
| **华中农业大学** | 211 / 双一流 | 生物信息学 · 学士 | 2020 — 2024 |

## 专业技能

**生物信息分析：**微生物基因组与多肽数据整合、序列质控及聚类、功能注释与富集分析；具备 Prodigal、MMseqs2、HMMER、eggNOG-mapper 实践，使用 GO、KEGG、NCBI Taxonomy 等生物知识资源。

**AI 建模与知识挖掘：**Python、PyTorch；蛋白质语言模型 ESM2 / ProtT5 特征提取、多任务学习与知识蒸馏；LLM 指令微调（LoRA）、实体关系抽取、本体归一化、知识图谱与 RAG。

**数据平台与工程：**FastAPI、PostgreSQL、HDF5、Neo4j、FAISS、React / TypeScript；具备数据流水线、RESTful API 与可视化页面开发经验，使用 Linux、Git、Docker 管理开发与运行环境。

**英语能力：**大学英语六级（CET-6）526 分。

## 项目经历

### miProbe | 微生物多肽数据与 AI 分析平台

2025.08 — 2026.08

面向功能多肽与潜在药物分子挖掘，承担数据处理、嵌入计算、模型训练评估及平台开发；数据库已部署并提供公开 API。

- **多源数据与特征资源：**整合 5 个公共数据库，完成质控、去重、聚类与注释，构建约 **10.8 亿条**非冗余多肽资源；计算 ESM2 / ProtT5 序列向量，形成 **8 TB**预计算嵌入，以 PostgreSQL 管理元数据、HDF5 分片存储特征，支持批量访问与下游复用。
- **功能预测与模型评估：**在 **42 个**实验表征数据集上比较蛋白质语言模型嵌入与 one-hot、k-mer 等传统特征，嵌入模型在 **30 项任务**中获得最高 AUC；构建多任务 MLP，同时预测 7 项理化性质。
- **抗菌肽候选挖掘：**基于约 190 万条序列及 4 个教师模型训练序数回归学生模型，教师共识评分预测准确率 **82.0%**、Kappa **0.689**；利用预计算嵌入，双 RTX 4090 在 **16 小时**内完成全库推理，筛得约 **24.4 万条**高共识评分候选抗菌肽。
- **在线平台与研发接口：**开发多属性检索、序列详情、功能与理化性质展示及基因组邻域可视化；以 FastAPI、React 实现交互平台，通过公开 API 与 Python 工具包为外部分析流程提供序列、注释和嵌入访问。
- **项目成果认可：**“面向生物医药智能研发的微生物短肽高质量数据集”**拟入选 2026 年上海市行业高质量数据集案例**（单位：上海市第一妇婴保健院；推荐单位：静安区数据局）。

平台：biosino.org/iMAC/miProbe　工具包：github.com/dfwlab/miProbe

### 知识驱动的微生物组学分析 AI 智能体构建

2026.09 — 至今

围绕微生物—疾病—代谢物知识整合，开发文献抽取、图谱与证据问答原型，并通过领域实体识别微调优化知识获取能力。

- **文献知识抽取与治理：**开发 PubMed / PMC 文献筛选、XML 解析与段落处理流水线，使用 LLM 抽取 4 类实体、7 类关系；结合结构化输出校验、原文证据核验及 MeSH / NCBI Taxonomy / ChEBI / GO 本体归一化，组织可溯源的 Neo4j 知识图谱。
- **图谱与文献协同智能体：**实现“问题解析—图谱召回—文献检索—证据汇总”的受控流程，结合 FAISS 向量检索生成带文献出处的回答；开发疾病知识挖掘模块，由文献证据判定关联方向并导出先验知识集，探索 TTD 药物—靶点—疾病数据整合。
- **领域微调与效果验证：**整合 CRAFT、Bacteria Biotope、BioRED 人工标注语料，完成 Llama 1B / 3B LoRA 微调，并与 DeepSeek Flash API few-shot 方案比较；在同一 **298 条验证样本**上，**3B 微调模型**相比于 DeepSeek API 少样本学习的 F1 score **63.41%**，提高了 **17.29 个百分点**，达到 **80.70%**。

## 学术成果

**共同第一作者（排名第一）| Nature Communications**　已接收，待正式发表

**miProbe: Microbial Peptide Repository of Biological Embeddings**

**2026 年影响因子 / JCR 分区：**18.1 / Q1

主要负责数据收集与分析、论文初稿撰写与修稿阶段结果补充、图表制作及 Web 平台开发。论文代码：github.com/dfwlab/miProbe-paper

网页版简历