

附件 1:

新工科毕业设计研究与实践项目与选题征集表

项目名称	基于深度学习的蛋白质复合物识别系统设计与实现		
负责人	张宇	联系方式	13612174418; zhangyuai@163.com
所在学院	人工智能		
学院联系人		联系电话	
拟设置指导教师数	1	指导教师学科背景	人工智能
拟指导学生数	张宇	拟招募选题学生的学院(专业)	
项目目的及意义(200字内): 本项目旨在开发机器学习算法,自动识别冷冻电子断层扫描(CryoET)图像中五类蛋白质复合物,解决细胞内超微结构解析的关键瓶颈。通过挖掘 cryoET 数据门户中数万张未标注的断层图,揭示细胞“暗物质”,为疾病机制研究和药物靶点发现提供技术基础。项目采用 F_B=4 评估指标,强化召回率以契合科学发现需求,优先避免漏检真实粒子。该研究将推动计算细胞生物学发展,加速病毒-宿主互作、蛋白质异常聚集等领域突破,构建从成像数据到生物学洞见的自动化分析管道,具有重要的生物医学应用价值与 AI for Science 示范意义。			
项目已具备的条件(200字内): 项目已具备三大核心实施条件:开源数据与算力资源完备。数据层面,开源数据提供 7 个训练样本(630×630×184 体素)及标注坐标,涵盖 6 类蛋白质颗粒;cryoET 数据门户公开数万张标准化断层图,可扩展训练数据。算力层面,实验室有多张 Nvidia4090GPU。技术基础上,开源人工智能算法众多,数据接口库均为开源,工具链完善,可直接复现与优化。			
项目期限与目标(300字内): 本项目旨在开发基于深度学习的冷冻电子断层扫描(CryoET)蛋白质复合物自动识别系统,建立从原始断层图像到生物结构解析的高效分析管道。技术目标为:构建 3D 热图回归模型,采用 SegResNet 或 2.5D U-Net 架构,在 F_B=4 评估指标下达到 0.75 以上分数,其中 hard 类颗粒召回率≥0.85,以优先保障科学发现完整性;设计抗噪声数据增强与各向异性卷积模块,解决信噪比低、Z 轴分辨率差与缺失楔效应等成像难题;实现滑动窗口推理与 TensorRT 加速,将单样本处理时间压缩至 2 分钟内。应用目标为:准确识别 ribosome、virus-like particle、apo-ferritin、thyroglobulin、β-galactosidase 五类蛋白质,支持 cryoET 数据门户异构数据的领域自适应,实现跨实验条件稳定泛化。最终交付开源工具包、预训练模型及标准化接口,为计算细胞生物学提供可扩展的 AI 框架,推动病毒-宿主互作、蛋白质聚集疾病等机制研究的突破,构建 AI for Science 在结构生物学中的示范应用。			

项目主要研究方向（选题）及任务（1000字内，应至少设计2个研究方向）：

本项目聚焦冷冻电子断层扫描（CryoET）蛋白质自动识别，围绕人工智能模型研发与工程化部署两大核心方向，构建从算法创新到实用化平台的完整技术体系，着力解决低信噪比三维目标检测的科学难题与计算效率瓶颈，推动AI技术在结构生物学领域的规模化应用。

方向一：高性能人工智能模型研发与系统测试

本方向致力于突破CryoET极端成像条件下的蛋白质颗粒检测精度与泛化能力瓶颈，开发面向稀疏目标的三维热图回归模型，并建立覆盖全生命周期的自动化测试体系，确保模型在真实科研场景中的鲁棒性与可信度。

任务1：鲁棒3D检测架构设计与优化

开发基于SegResNet与2.5D U-Net的混合架构模型，融合注意力机制捕获长程空间依赖，采用各向异性卷积核（XY平面 7×7 ，Z轴 $3\times 1\times 1$ ）适配分辨率差异，设计多尺度特征金字塔融合亚纳米级细节与数十纳米级上下文。针对标注稀缺问题，引入BYOL自监督预训练，利用无标注断层图学习噪声不变表征；针对类别不平衡，构建加权Focal Loss（ $\gamma=2$ ），对hard类颗粒（thyroglobulin、 β -galactosidase）施加倍率权重，并采用高斯掩码放大中心区域损失贡献3倍。通过梯度检查点与混合精度训练，实现 $630\times 630\times 184$ 体素数据在16GB显存下的端到端训练。

任务2：自动化测试与性能验证体系构建

搭建覆盖单元、集成与部署三层次的测试框架：单元层面对每个3D卷积层、上采样模块进行数值梯度检查与输出shape验证；集成层面在Kaggle官方7样本训练集、自建50张多物种测试集（含EMPIAR-10045公开数据）上执行交叉验证，监控FB、召回率、定位误差（ <2 体素）等8项指标，实施Early Stopping与模型检查点策略；部署层面构建持续集成流水线，每次代码提交触发自动化测试，在Kaggle GPU环境中验证推理耗时（目标 <2 分钟/样本）与内存占用峰值。建立对抗鲁棒性测试集，通过傅里叶域楔形掩码模拟缺失楔效应、添加泊松-高斯噪声，确保模型在信噪比下降50%时召回率衰减不超过10%。

任务3：领域自适应与少样本迁移机制研究

针对cryoET数据门户的异构数据（不同电镜品牌、样品厚度、细胞类型），开发基于MAML的元学习框架，在6个公开数据集上学习通用初始化参数，使新样本微调步数从1000步降至100步；设计风格归一化层（AdaIN+InstanceNorm），消除灰度分布差异。构建主动学习模块，对低置信度预测区域自动推荐人工复核，迭代扩充标注池。目标实现跨数据集FB下降 $<5\%$ ，在 β -amylase等零权重类别上伪标签准确率达75%以上，支撑模型向用户私有数据的低成本迁移。

方向二：易用软件平台开发与高效部署

本方向聚焦算法工程化转化，开发覆盖数据管理、模型训练、推理分析、结果可视化的全流程工具链，降低结构生物学家使用AI的技术门槛，实现CryoET智能分析的云边端协同部署。

任务1：模块化工具链与图形化界面开发

封装cryoet-ml-toolkit核心库，提供pip安装包，集成cryoet-dataportal数据接口、Patch滑动窗口采样、TensorRT加速推理、FB评估等API，支持Python脚本一键调用。基于PyQt开发跨平台图形界面，实现拖拽式数据加载、超参数配置（学习率、patch size）、实时监控训练曲线、交互式结果审查（3D热图叠加显示、

颗粒坐标编辑)。内置 napari-cryoet 插件, 支持在 napari 中直接运行 AI 预测并可视化颗粒置信度云图, 提供坐标导出为 ChimeraX 兼容格式。编写详尽的 Sphinx 文档与 Jupyter 教程, 覆盖从新手入门到高级调优全流程。

任务 2: 云端弹性推理服务与性能极致优化

在阿里云 PAI 平台部署模型即服务 (MaaS), 用户上传 tomogram 文件后, 系统自动分块、调度 GPU 资源, 90 秒内返回 JSON 格式颗粒列表 (含坐标、类别、置信度)。采用 Kubernetes 自动扩缩容, 支撑 100 用户并发请求。优化推理引擎: 将 PyTorch 模型导出为 ONNX 静态图, 经 TensorRT 算子融合与 FP16 量化, 在 A100 显卡上实现推理速度提升 8-10 倍 (单样本从 10 小时降至 1.2 小时); 开发多进程并行流水线, 数据加载、预处理、推理、后处理异步执行, CPU 利用率提升至 90%。构建 KDTree 空间索引加速 NMS, 后处理时间从 120 分钟压缩至 8 分钟。

通过两大方向的协同攻关, 项目将形成“算法创新—工程化—应用验证”的完整闭环, 既产出性能领先的 AI 模型, 又交付结构生物学家“开箱即用”的软件平台, 最终实现 CryoET 数据挖掘效率百倍提升, 加速疾病机制发现与药物靶点研究。