AI大模型基于病理切片预测空间蛋白

CELL 和 Nature Medicine 文章复现及原理全解析课程目录

【模块一：课前预热（4节）】

第1节：Python环境搭建

1. Spyder 和 Anaconda 软件安装（Windows、Linux、Mac）
2. Conda 环境管理和镜像设置
3. 使用 conda 和 pip 安装 Python 包
4. Jupyter Lab 安装和使用

第2节：Python编程语言入门

1. Python 缩进与命名规范
2. 包和模块的基本概念，import 的三种写法
3. 对象属性与方法调用
4. 自定义函数 def：参数、返回值、位置参数与关键字参数
5. 条件语句和循环语句

第3节：Python数据结构进阶

1. 列表、元组、字典、集合等基本数据结构
2. 生成和索引、增删改查、排序、统计与去重
3. 矩阵新建和行列取子集、布尔索引（numpy）
4. 数据框新建、行列选择、数据类型转换（pandas）

第4节：seaborn 和 matplotlib 绘图

1. matplotlib 基本绘图流程
2. seaborn 常用图：histplot、boxplot、violinplot、barplot、heatmap
3. 使用 plotnine（ggplot 风格）绘图
4. 自定义颜色、配色、拼图和图片保存

【模块二：Nature Medicine 复现（10节）】

第1讲：病理图像与 CODEX 超大图像读取

1. H&E 图像文件结构与基础读取，理解 OME-TIFF 底层结构

2. 病理大图多级分辨率金字塔与按需读取
3. CODEX 多通道图像高维结构与 XML 通道解析
4. 空间多通道融合、共定位、单通道增强与伪彩融合

第2讲：数字病理切片图像和 CODEX 多模态图像对齐

1. 对齐前基线建立，明确 H&E 与 CODEX/IF 在配准中的角色
2. 理解 Palom 的核心数据流、参数体系与 Aligner 构建
3. 缩略图级粗对齐与仿射矩阵中的缩放和平移
4. 分块精对齐与局部位移场优化
5. 配准结果输出、质量控制与效果评估

第3讲：H&E 与 DAPI 双模态细胞核分割

1. StarDist 预训练模型选择与双模态细胞核分割
2. 细胞核形态特征提取
3. CODEX 质心映射到 H&E 坐标系
4. 最近邻距离分析与亚细胞级配准精度评估

第4讲：配准后单细胞表达矩阵构建与空间分布模式

1. 双模态图像读取，从细胞核到细胞区域的单细胞边界划分
2. 多通道 CODEX 强度转化为 cell × gene 表达矩阵
3. 单细胞矩阵质控、标准化、降维与聚类分析
4. 局部空间回投与 Marker 基因空间展示
5. 全切片图像预处理与智能切块全流程

第5讲：复现 Nature Medicine 论文 WSI 与 CODEX 处理流程

1. 双模态数据读取与 H&E 全切片预处理
2. 伪影多层质控与高质量 Tile 筛选
3. H&E 染色标准化与 Tile 级特征提取
4. CODEX 通道归一化与 Tile 表达矩阵构建

第6讲：HEX 模型训练

1. HEX 大模型训练数据格式转换与标准化
2. 配对数据质控与训练前完整性验证
3. HEX 训练流程与输入数据对接
4. 模型训练参数微调与 Checkpoint 保存

5. 结合 Pearson 相关系数、MSE 等指标做性能评估

第7讲：HEX 大模型推理

1. 理解 H&E Patch 输入形式与蛋白预测任务
2. 掌握 HEX 网络结构与 MUSK 视觉编码器作用
3. 推理前预处理与标准化
4. 从单张或多张 H&E Patch 到 40 维蛋白表达向量的批量推理
5. 整理预测矩阵并进行结果展示与生物学解读

第8讲：HEX 的 WSI 高分辨率虚拟蛋白生成与空间可视化

1. 全切片推理任务设计与关键参数配置
2. 掌握全切片虚拟蛋白预测结果存储架构
3. AI 大模型在 WSI 上的高分辨率推理流程
4. 生成论文级虚拟蛋白空间表达图谱
5. 比较不同标记物共定位并理解全切片空间微环境

第9讲：从论文复现到搭建自己的 AI 大模型

1. 理解 AI 大模型项目从脚本层面到项目层面的完整搭建流程
2. 掌握图像格式理解、多模态配准、空间映射、Patch 生成、模型训练与推理的整体骨架
3. 学习 AI 大模型搭建八步法：任务定义、数据配对、标准化、空间对齐、特征构建、模型训练、性能评估、多模态融合
4. 理解输入、标签、训练策略与评估体系的设计逻辑
5. 建立从小样本跑通到正式训练和微调的渐进式训练思路

第10讲：思路迁移到自己的课题

1. HE 切片预测 IHC
2. 病理切片预测空间转录组
3. 病理图像结合临床做预后预测
4. 借助单细胞图谱辅助构建图像标签

【模块三：CELL 复现（4节）】

第11讲：多级精配准

1. 双模态全图读取与高质量 ROI 自动筛选
2. H&E RGB ROI 和 CODEX DAPI ROI 裁切与配准输入构建

3. 三级渐进式配准策略：Rigid、Non-Rigid 与 Micro
4. 配准质量控制、多层次配准结果可视化与形变场解析

第12讲：CELL 主刊 GigaTIME 大模型复现与自定义 WSI 局部推理

1. 样例数据预处理与二值掩码解包
2. 非细胞区域过滤与激活密度定量评估
3. 自定义 WSI 读取与组织区域自动定位
4. ROI Patch 网格提取与第一轮全通道推理
5. 通道空间热图重建与局部空间模式解析

第13讲：GigaTIME 大模型全面解析

1. 任务目标与整体框架
2. 模型结构解析：从编码到解码
3. UNet 在空间预测中的作用
4. 损失函数与模型优化

第14讲：最小可运行测试与训练

1. 从预训练权重到推理评估，验证模型环境与数据接口
2. 理解 Dice 系数与 Pearson 相关系数
3. 通过最小训练实验走通训练、验证和指标记录流程
4. 训练结果输出与历史曲线分析

【模块四：TCGA 公共数据库病理切片挖掘（4节）】

第15讲：TCGA 队列构建、临床数据解析与 WSI 预处理

1. TCGA 临床数据和病理切片标准化下载流程
2. 提取生存信息与关键临床变量
3. 建立 slide_id 与 case_id 的准确映射关系
4. LUAD/LUSC 队列结构检查与数据完整性质控
5. 从全切片生成 20x 与 40x 两套坐标流程

第16讲：TCGA 病理切片特征提取与训练数据准备

1. 从坐标读取、Patch 提取到双流特征生成的完整流程
2. WSI 形态特征与虚拟蛋白特征结构解析
3. 双流特征规模对比与多模态信息互补

4. 训练数据整理与按 case_id 分组的 K-Fold 划分

第17讲：生存预测模型训练与 Kaplan-Meier 评估

1. 双流生存模型训练与跨队列验证
2. 风险分数与生存时间、删失状态和临床分期整合分析
3. C-index 与早期分期亚组性能评估
4. Kaplan-Meier 生存曲线与风险分层验证

第18讲：虚拟蛋白下游分析

1. 下游分析数据体系与整体框架搭建
2. 蛋白与生物标志物关联分析
3. 虚拟蛋白与病理分期关联分析
4. 单个蛋白生存分析
5. 多蛋白 Signature 构建与生存分析