虚拟细胞

跟着二十篇顶刊文章学虚拟细胞课程目录

【模块一：虚拟细胞基础大模型】

第1讲：GET Foundation

模型实现的功能：

1. PBMC 微调与 GRN 推断
2. ATAC 评估：预测 ATAC-seq 可及性信号
3. 星形胶质细胞推断：跨细胞类型的基因表达预测
4. Motif-ATAC 预测：转录因子结合 motif 的可及性预测

课程内容：

1. 模型微调与 GRN：加载 GET 预训练模型、PBMC 微调 checkpoint，推理基因表达，提取注意力矩阵，构建 GRN（含 R pcalg 因果分析）
2. 模型 ATAC 预测评估：加载 ATAC 预测模型，预测 TSS 可及性，评估 Pearson 相关，可视化预测与真实结果
3. 模型细胞推断：加载星形胶质细胞数据，进行跨细胞类型表达推断，比较不同细胞类型表达模式
4. 模型 Motif-ATAC 预测：进行 TF 结合 motif 分析、Motif-ATAC 可及性预测与已知 TF motif 验证

第2讲：scGPT

模型实现的功能：

1. 细胞类型注释
2. 批次整合
3. 基因扰动预测
4. 基因调控网络推断

课程内容：

1. 细胞注释：加载预训练 scGPT 模型和注释 checkpoint，对 PBMC 10K 数据推理，预测细胞类型标签，生成混淆矩阵并计算准确率
2. 批次整合：加载 Immune_ALL_human 多批次数据和整合模型，生成细胞嵌入，UMAP 可视化批次效应消除，并计算 batch mixing 指标
3. 扰动预测：加载 Adamson Perturb-seq 数据和扰动微调 checkpoint，预测基因敲除后表达变化，计算 Pearson 相关并比较预测与真实表达

4. GRN 嵌入：加载基因嵌入矩阵，计算基因间余弦相似度，构建并可视化基因调控网络，对比已知 TF-target 关系

【模块二：网络生物学预测 -- 转录调控网络】

第3讲：AlphaGenome

模型实现的功能：

1. 多模态预测：从 DNA 序列预测 7000+ 个基因组 track
2. 变异效应评估：评估 SNP 对基因组特征的影响

课程内容：

1. AlphaGenome 快速入门与多模态预测：加载 AlphaGenome 模型，输入 DNA 序列，预测多种基因组 track，可视化预测结果，复现 5 个关键论文图
2. AlphaGenome 变异效应评估基准：加载 19 个预计算评估基准，进行 AUROC、AUPRC、Spearman 评估，并与其他模型对比

第4讲：scPRINT

模型实现的功能：

1. 细胞嵌入与零样本分类
2. GRN 推断
3. 去噪与表达恢复

课程内容：

1. 细胞嵌入与零样本预测：加载 scPRINT 模型，做零样本细胞嵌入，UMAP 可视化，计算 kNN 分类准确率并输出混淆矩阵
2. GRN 推断：提取注意力矩阵，构建基因调控网络，绘制热力图和网络图，并与已知 TF 进行对比
3. 去噪与表达恢复：进行表达值去噪，绘制散点图、分布对比和 marker 基因热力图

【模块三：scRNA-seq 虚拟蛋白组】

第5讲：scTranslator

课程内容：

1. 蛋白质丰度预测：加载 scTranslator 模型，进行 RNA 到 Protein 翻译推理，对比预测与真实蛋白质丰度，并进行 Pearson 相关评估
2. 伪基因敲除与下游分析：模拟基因敲除，预测蛋白质组变化，进行差异蛋白分析和聚类可视化

【模块四：基因扰动预测：从单基因到组合 CRISPR 虚拟敲除】

第6讲：STATE

课程内容：

1. 数据加载与架构介绍：加载 K562 评估数据，做扰动分布可视化、表达对比和 PCA 可视化
2. 扰动响应预测：加载 STATE 模型，完成推理预测、DE 基因对比和 Pearson 评估
3. 结果分析与可视化：绘制热力图，进行通路聚类分析和预测方向分析

第7讲：dbDiffusion

课程内容：

1. 数据探索与扰动聚类：加载 Yao Perturb-seq 数据，进行 PCA、Leiden 聚类和 UMAP 可视化
2. 扩散模型采样与可视化：加载 VAE 和 Diffusion 模型，完成扩散采样、VAE 解码与 UMAP 对比
3. 去偏差推断与评估：进行 PPI 网络去偏、置信区间估计和火山图可视化

第8讲：GEARS

课程内容：

1. 数据加载与探索：加载 Norman K562 数据，统计扰动条件，UMAP 可视化并分析数据分割
2. 扰动预测推理：加载 GEARS 模型，完成单基因和组合扰动预测，评估 Top 20 DE 基因 MSE 和 GI 分析
3. 模型评估与可视化：进行全数据集评估，比较 Pearson 相关、MSE、方向错误率及不同子组表现

第9讲：PRnet

模型实现的功能：

1. LINCS L1000 数据探索与 SMILES 编码
2. 药物扰动响应预测
3. 潜在空间与连接性分析

课程内容：

1. 数据加载与 PRnet 架构介绍：加载 LINCS L1000 demo 数据，展示 SMILES 编码和 Morgan Fingerprint，讲解 PRnet CVAE 架构
2. 药物扰动响应预测：加载预训练 PRnet，完成推理预测、R^2 /Pearson 评估和散点图展示
3. 潜在空间与药物嵌入分析：进行潜在空间 t-SNE、余弦相似度热力图和连接性评分分析

第10讲：scTenifoldKnk

课程内容：

1. 通过 PC 回归和 tensor decomposition 构建去噪 GRN

2. 进行虚拟 KO，将目标基因的调控影响清零
3. 利用流形对齐揭示 KO 前后每个基因的扰动程度

第11讲：scGen

模型实现的功能：

1. 扰动响应预测
2. 跨细胞类型泛化
3. 跨研究预测
4. 跨物种预测

课程内容：

1. 模型原理：理解 VAE 编码器-解码器结构、扰动向量学习和对照细胞加扰动向量的预测思想
2. 扰动预测：加载单细胞数据，训练 scGen 模型，估计扰动差异向量，对未见细胞类型进行扰动状态预测，并计算 Pearson 或 R^2 评估效果
3. 感染响应预测：加载感染数据，学习感染前后状态变化，预测未见细胞群感染后的表达谱，并比较 top DEGs 和整体表达相关性
4. 跨研究迁移：用研究 A 训练模型，将刺激效应迁移到研究 B，仅有对照数据的情况下预测刺激后表达状态，并检查关键响应基因恢复情况

第12讲：GenKI

模型实现的功能：

1. 虚拟基因敲除推断
2. KO 响应基因识别
3. 基因调控网络辅助解析
4. 双基因敲除模拟

课程内容：

1. 虚拟敲除基础分析：加载 WT 单细胞表达矩阵，构建 scGRN，训练 GenKI 模型，设定目标基因虚拟敲除，计算 WT 与虚拟 KO 的潜在分布差异并输出 KO-responsive genes 排名
2. 模型原理与训练：理解表达矩阵和调控网络联合建模、VGAE 结构、KL divergence 的作用及 bagging 提高稳定性的思路
3. 功能富集与网络解释：对 KO-responsive genes 做 GO/通路富集，构建 STRING 网络，并与已知文献结论对比

【模块五：空间微环境与多尺度虚拟细胞 — 前沿与展望】

第13讲：Celcomen

模型实现的功能：

1. 预测基因或细胞扰动后的空间组织改变

课程内容：

1. 环境验证与空间数据探索：加载 10x Xenium GBM 数据，进行空间坐标和细胞类型分布可视化
2. 模拟数据自一致性验证：生成模拟数据，完成自一致性检验和因果发现验证
3. 空间因果推断与扰动分析：进行 CCE 因果推断，加载 SCE 权重，完成空间扰动预测和因果效应可视化

第14讲：Nicheformer

模型实现的功能：

1. 从基因表达预测细胞的空间位置

课程内容：

1. 环境验证与数据探索：介绍模型架构，加载 PBMC 样本数据并完成基因词表对齐
2. Embedding 生成与可视化：使用 from_pretrained 加载模型，提取细胞嵌入，进行 UMAP 可视化和多层嵌入对比
3. 空间标签预测与迁移：进行零样本标签预测、kNN 标签迁移，并评估准确率

第15讲：PULSAR

课程内容：

1. 环境验证与模型探索：加载 PULSAR 模型，进行架构探索和参数统计
2. 零样本年龄回归预测：使用 OneK1K 数据生成供体嵌入，完成年龄回归预测和 MAE/R^2 评估
3. 疾病分类与供体检索：使用 Lupus 数据完成 SLE 疾病分类、供体检索和 AUROC 评估

第16讲：Tahoe-100M

课程内容：

1. 数据集概览与实验设计：介绍实验设计、元数据统计以及药物和细胞系分布
2. 转录组景观与批次效应：分析 UMAP 嵌入、批次效应、E-distance 和细胞周期
3. 药物响应机制与前沿展望：进行 pseudobulk DE 基因分析、药物机制归类和 Tahoe-x1 模型介绍

【模块六：形态学扰动预测 — 从转录组到细胞图像】

第17讲：IMPA

模型实现的功能：

1. 从转录组到细胞图像

课程内容：

1. 环境验证与数据探索：探索 BBBC021 数据集、6 种药物处理和细胞形态统计
2. 药物扰动形态预测：加载 StarGANv2 模型，完成 6 种已知药物形态预测以及生成与真实结果对比
3. 未知药物预测与扰动空间：进行 zero-shot 未知药物预测和 PCA 形态学空间分析

【模块七：扰动动力学 — 从静态预测到时间轨迹模拟】

第18讲：Squidiff

课程内容：

1. 模拟数据与扩散过程：生成模拟数据，进行扩散过程可视化、DDIM 采样和模型训练
2. 基因扰动预测：加载扰动模型，完成基因敲除预测以及预测与真实对比
3. 药物组合效应预测：加载药物模型，进行单药和组合药效预测，并分析协同效应

第19讲：CellOracle

课程内容：

1. 环境验证与数据探索：验证 CellOracle 安装，加载 Paul 2015 造血数据，完成细胞分群可视化
2. GRN 构建与网络分析：基于 scATAC 构建 base GRN，进行线性回归 GRN 拟合和网络连接度分析
3. TF 扰动模拟与可视化：完成 Gata1 KO、Spi1 KO 模拟，分析扰动效应和细胞命运转换

【模块八：评估与基准测试 — 如何判断扰动预测靠不靠谱？】

第20讲：scPerturBench

课程内容：

1. 环境验证与数据探索：加载 Kang IFN-β 数据集，比较扰动前后，并概览 27 种方法和 29 个数据集
2. 评估指标详解与实现：从零实现 MSE、PCC-delta、E-distance、Wasserstein、KL divergence、Common-DEGs 六个指标
3. 基准结果可视化与方法选择：进行方法排名热力图、性能对比和决策树可视化