图神经网络 对单细胞转录组数据分类和缺失值填补的提升

杨锟　鲁文哲　王诗艺

南开大学数学科学学院

2025 年 4 月 2 日

背景

1. 什么是单细胞 RNA 测序(scRNA-seq)?
   - 高通量技术，测量单细胞分辨率的基因表达
   - 提供海量数据，但有缺失值、噪声和高维问题
2. 为何使用图神经网络(GNNs)?
   - GNN 能自然建模关系（如细胞-细胞或细胞-基因）为图结构
   - 在分类和填补任务中灵活高效
3. 研究目标
   - 提高细胞类型分类准确性
   - 有效填补缺失基因表达数据

scRNA-seq 的流程

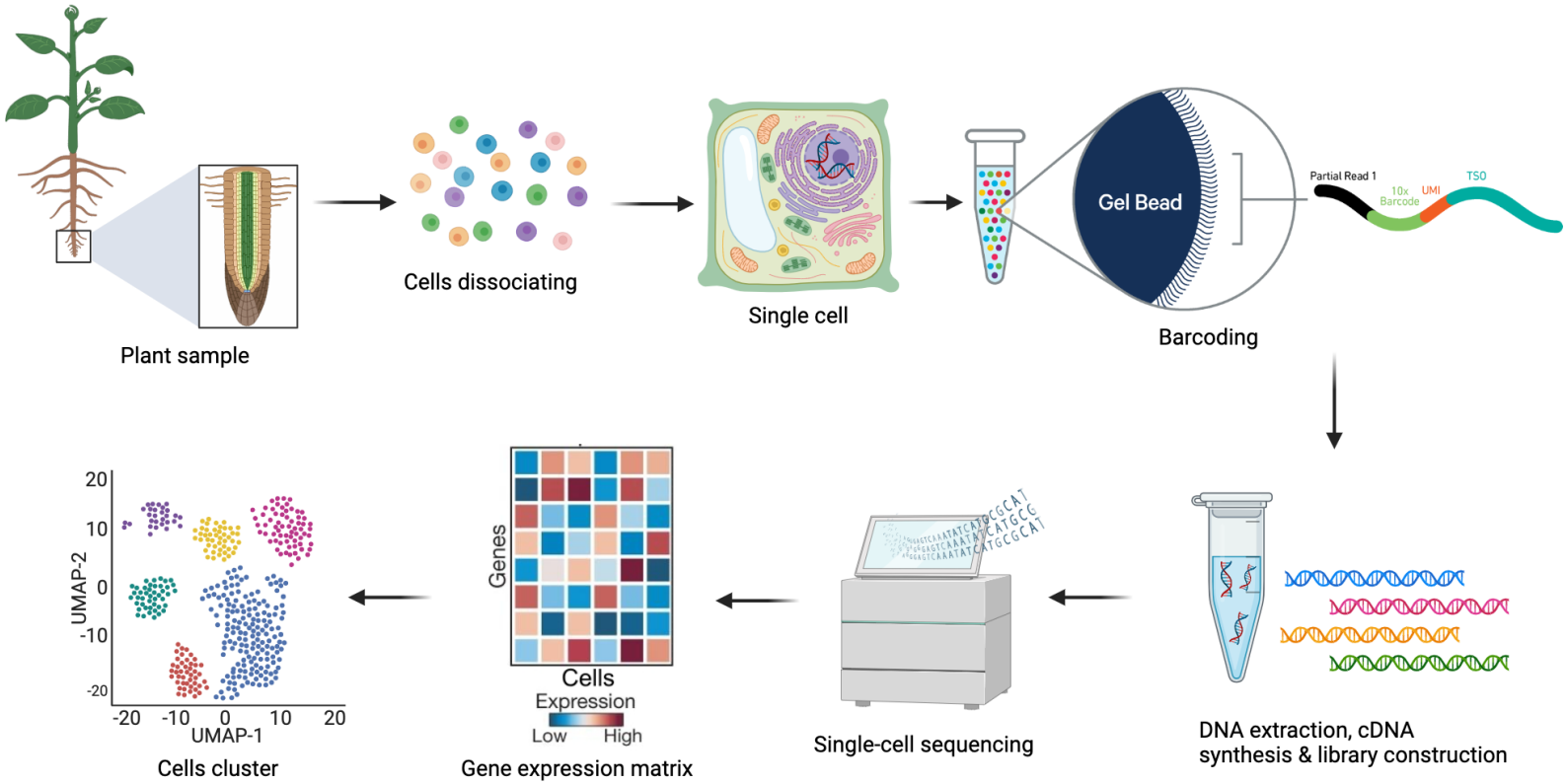

图 1: scRNA-seq Process[1]

[1]图片来源：Single cell RNA sequencing; scRNA-seq

GNN 基础

1. 什么是 GNNs?
   - 在图结构上运行的神经网络：节点（如细胞、基因）、边（节点之间的关系）
   - 沿邻居传播和聚合特征
2. 主要 GNN 类型[2]
   - **GCN**：简化谱卷积，聚合邻居特征
   - **DGCNN**：用于细胞分类任务，动态计算图连接
   - **GraphSAGE**：通过采样和聚合进行归纳学习
   - **GAT**：基于注意力的邻居加权
   - **GIN**：适用于图同构任务
3. 与 scRNA-seq 的关联
   - 图结构捕捉细胞间相似性或细胞-基因交互

[2]关于 GNNs 各种模型的详细介绍可访问：Best Graph Neural Network architectures

GNNs 的不同运用

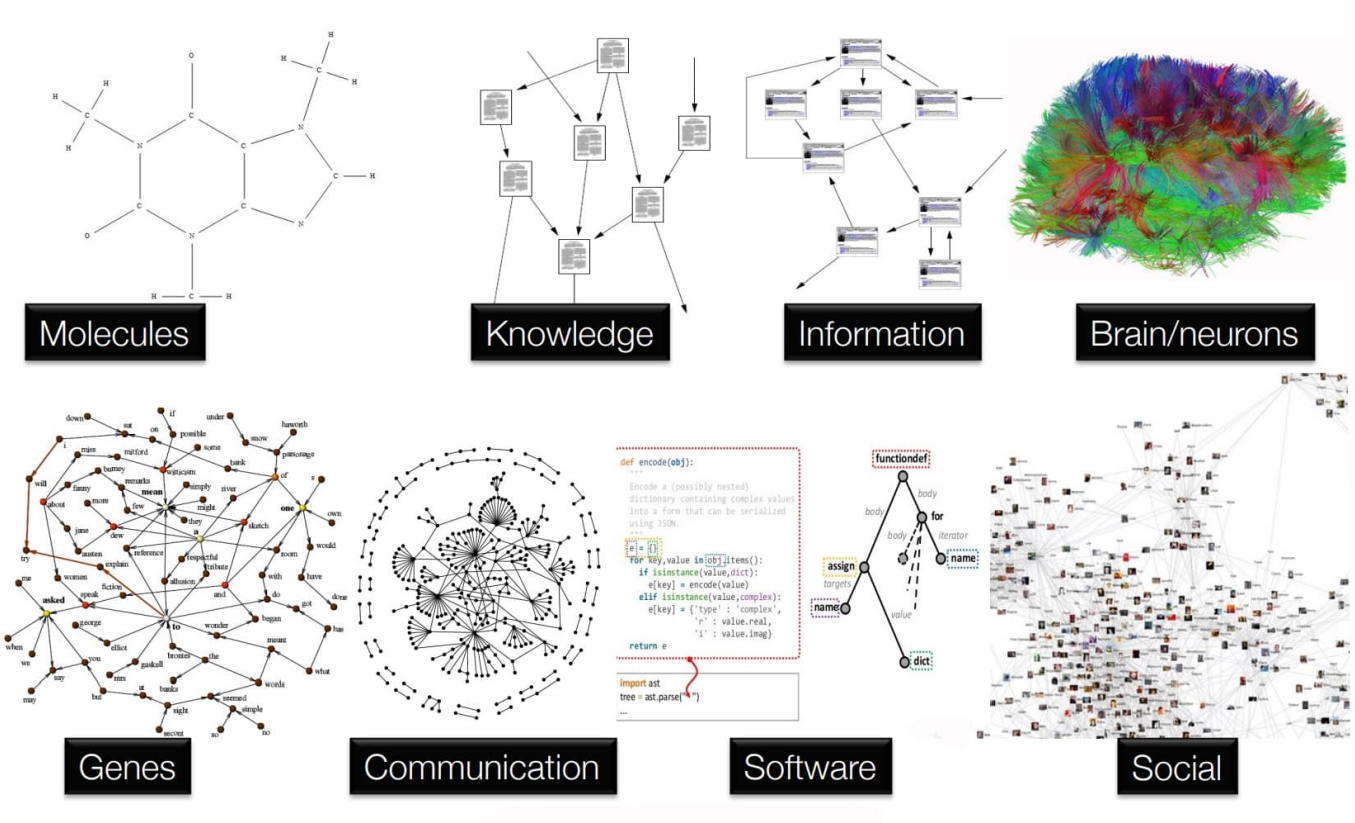

图 2: GNNs Application[3]

[3]图片来源: What Are Graph Neural Networks?

问题陈述

1. scRNA-seq 分析的挑战
   - 细胞分类：根据基因表达分配细胞类型
   - 数据填补：填补缺失或噪声的基因表达值
2. 现有方法
   - SVM（支持向量机）和 MLP（多层感知器）：传统基线，处理非结构化数据能力有限
   - 先前 GNN 研究（如 Buterez 等、Wang 等）：显示潜力但需改进
3. 论文中的突破
   - 基准测试 GNN 在分类和填补中的表现
   - 将GRAPE 框架[4]新颖应用于 scRNA-seq

[4]关于 GRAPE 框架的详细介绍可访问：Graph Representation Learning

方法-机器学习 (i)

PCA[5] & KNN[6]

- 主成分分析法(PCA)
  - 降维技术，用于减少数据维度，保留主要信息；
  - 通过线性变换将原始特征投影到新的坐标系；
  - 用于数据可视化、去噪和特征提取。
- 邻近算法(KNN)
  - 基于实例的分类和回归算法；
  - 找到距离最近的 K 个邻居，根据多数投票（分类）或加权平均（回归）确定结果；
  - 简单但对噪声敏感，计算成本随数据量增加而升高。

[5] 关于 PCA 的详细介绍可访问：Principal Component Analysis

[6] 关于 KNN 的详细介绍可访问：K-Nearest Neighbor(KNN) Algorithm

方法-机器学习 (ii)

MLP[7] & SVM[8]

- 多层感知器(MLP)
  - 前馈神经网络，用于解决非线性问题；
  - 通过反向传播算法训练，调整权重以最小化损失函数；
  - 适合复杂任务，但需要大量数据和计算资源。
- 支持向量机(SVM)
  - 监督学习方法，通过寻找最大间隔超平面来分类数据；
  - 使用核技巧（如线性核、RBF 核）处理非线性可分数据，将数据映射到高维空间；
  - 对小数据集效果好，但对参数敏感。

[7] 关于 MLP 的详细介绍可访问：Multilayer Perceptrons in Machine Learning

[8] 关于 SVM 的详细介绍可访问：Support Vector Machine (SVM) Algorithm

机器学习方法

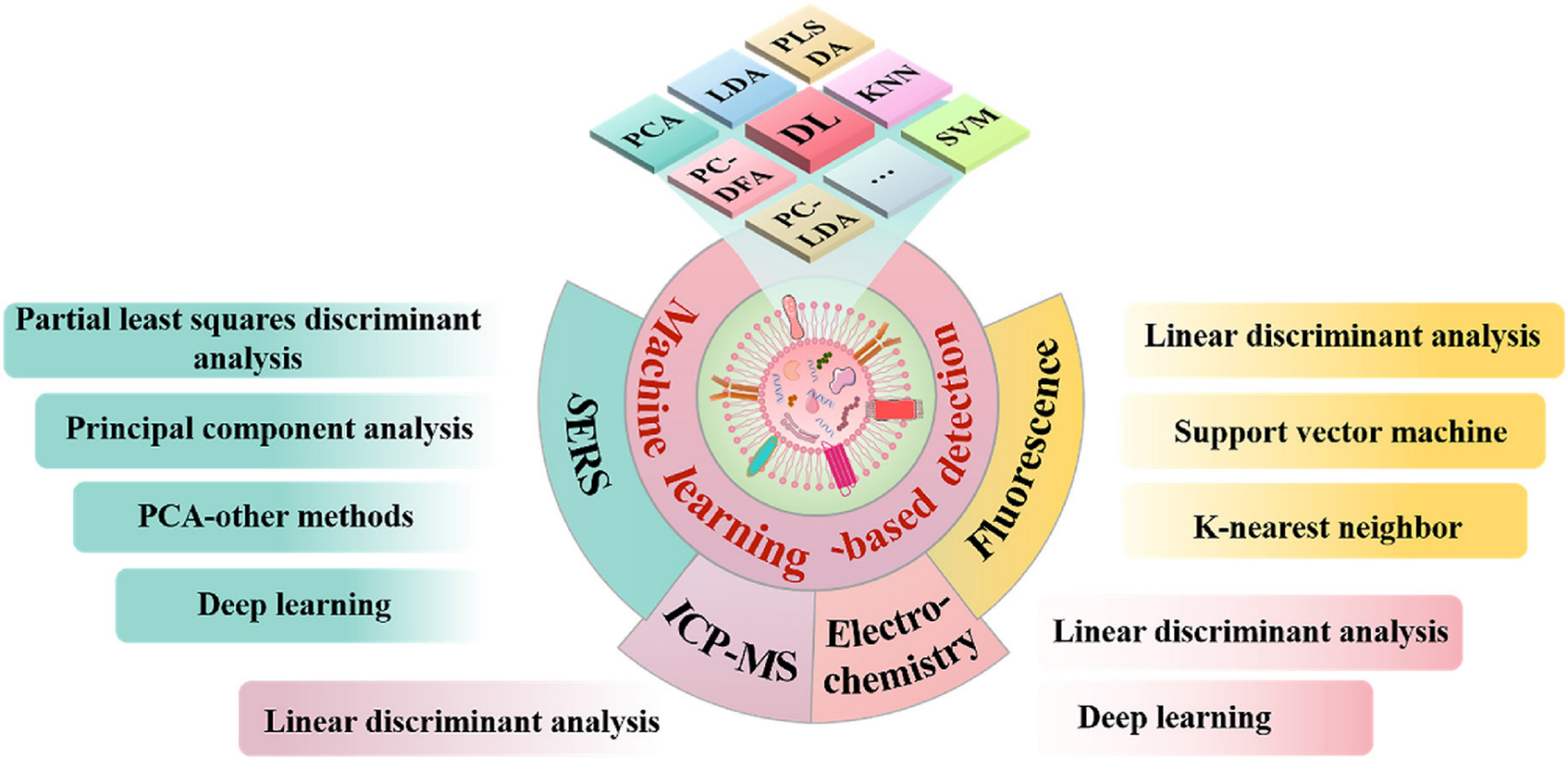

图 3: Machine Learning Methods[9]

[9]图片来源：Advances of machine learning-assisted small extracellular vesicles detection strategy

方法-机器学习 (iii)

DGN(Differentiable Group Normalisation)[10]

- DGN——可微组归一化，是用于解决图神经网络(GNN)中过度平滑问题的一项技术。
- DGN 通过对同一组内的节点进行归一化处理，将不同组的节点分布区分开来。其中，组的数量由用户自定义。如此一来，它能在确保相似标签节点的表示保持平滑的同时，让不同标签节点的表示具有明显差异。

分组归一化 $\Rightarrow$ 增加组间差异 $\Rightarrow$ 引入可学习参数

[10] 关于 DGN 和过度平滑问题的详细介绍可访问：
Towards Deeper Graph Neural Networks with Differentiable Group Normalization

DGN

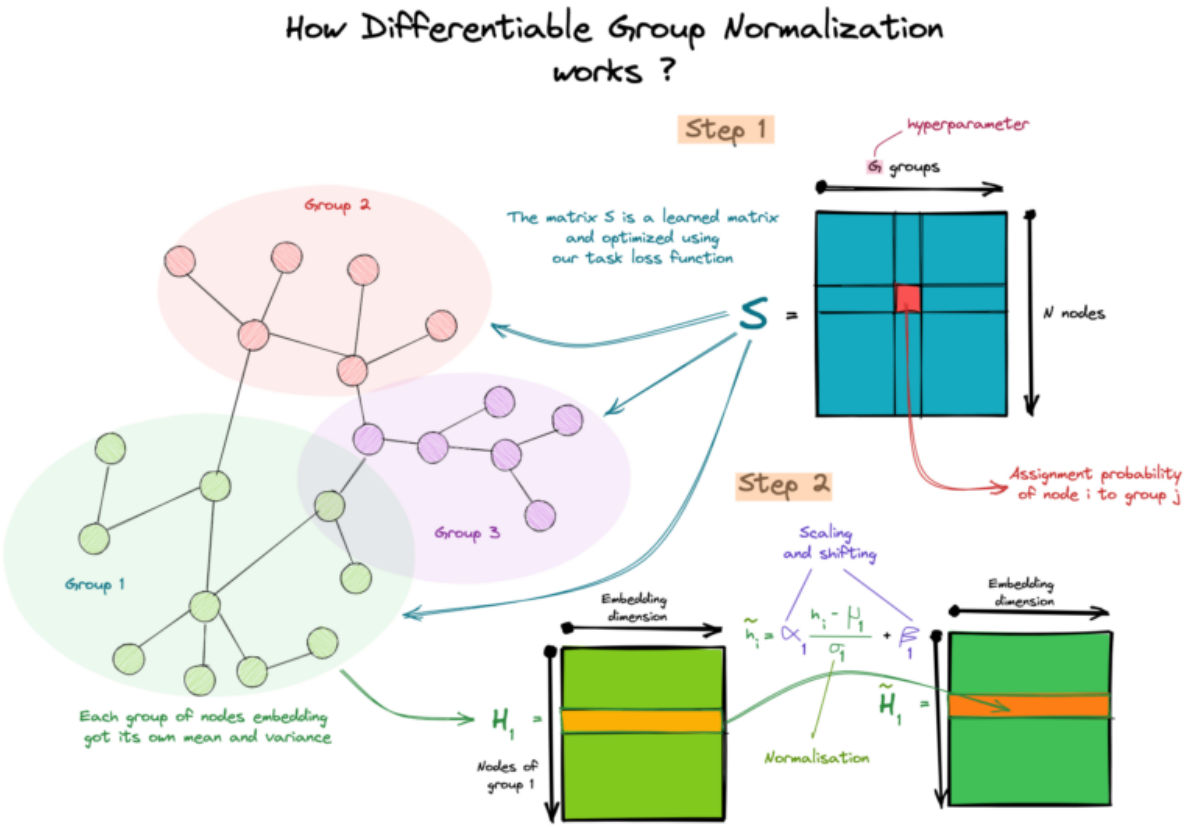

图 4: How DGN works[11]

[11]图片来源：Over smoothing issue in graph neural network

方法 - 细胞分类

1. 方法
   - 将 scRNA-seq 数据表示为细胞-细胞图
   - 邻接矩阵：PCA+KNN
   - 模型：带可微群归一化(DGN)的 GCN，缓解过平滑
2. 过平滑问题
   - 层数增加使节点表示过于相似
   - DGN 分离节点群，保持区分度
3. 基准
   - SVM（线性核）、MLP（2 层，64 隐藏单元）

方法 - 数据填补

1. 方法
   - 二部图：细胞节点和基因节点，边为基因表达值
   - GRAPE 框架：
     - 3 层 GraphSAGE 作为数据模型
     - MLP 用于填补，GCN 用于分类
   - 修改：
     - 可学习基因嵌入（非one-hot encoding[12]）
     - PCA+KNN 邻接用于分类
     - 加入重构损失来正则化
2. 目标
   - 填补缺失数据并提升分类效果

[12] 关于独热编码的详细介绍可访问：One Hot Encoding in Machine Learning

独热编码

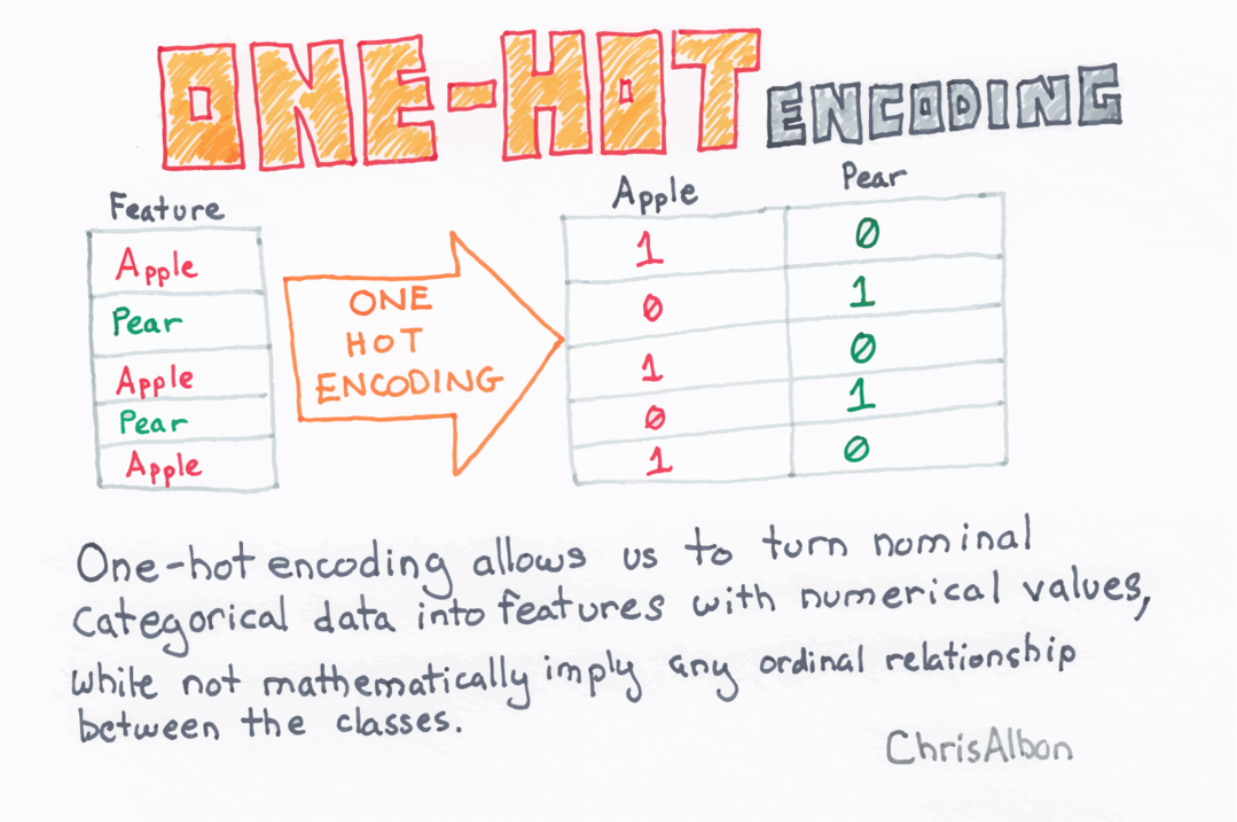

图 5: One-hot Encoding[13]

[13]图片来源：Label Encoder and One Hot Encoding

实验 - 数据集

1. Paul15 数据集(低维多类)
   - 髓系祖细胞发育
   - 685 个基因，19 种细胞类型
2. PBMC3K 数据集(高维多类)
   - 外周血单核细胞
   - 1838 个基因，8 种细胞类型
3. 设置
   - 70-30% 训练-测试划分
   - GNN 训练 2000 轮，学习率 0.001

实验结果

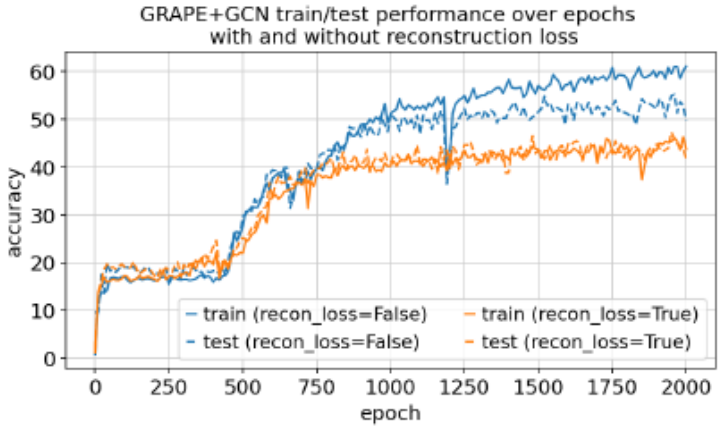

Figure 7: Performance over epochs for GRAPE+GCN model with gene embedding dimension of 64, with and without reconstruction loss penalty. The top 400 genes are used.

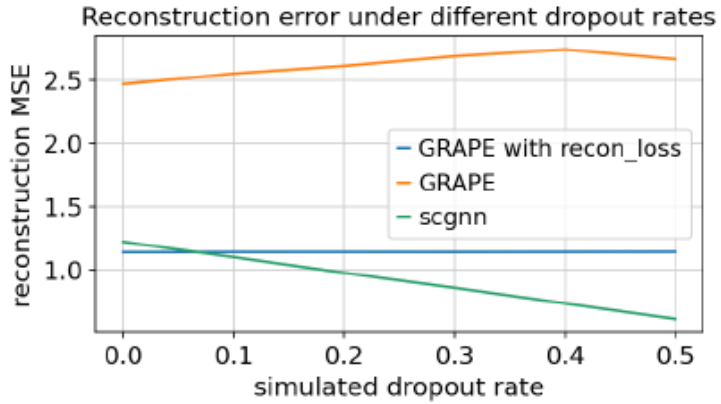

Figure 8: Reconstruction loss on simulated data dropout using GRAPE+GCN model with embedding size of 64 and trained on 400 genes.

图 6: Result[14]

[14] 图片来源:
Improving Classification and Data Imputation for Single-Cell Transcriptomics with Graph Neural Networks

实验 - 分类结果

1. Paul15 结果
   - GCN+DGN：59.4% 准确率（SVM：58.6%，MLP：52.8%）
   - 最佳邻接：PCA+KNN
2. PBMC3K 结果
   - GCN（1、2、4 层）：95.2-95.6%（SVM：94.2%，MLP：90.3%）
   - 随基因维度增加表现更优
3. 关键见解
   - GCN+DGN 缓解过平滑，超越基准

实验 - 填补结果

1. GRAPE 表现
   - 匹配 SVM/GCN 准确率（Paul15 约 58%），泛化能力更佳
   - 随基因数和嵌入大小（如 64）增加而提升
2. 重构
   - 带重构损失：训练更稳定，泛化差距小
   - 模拟缺失：低缺失率下优于 VAE scGNN

讨论与洞见

1. 分类
   - GCN+DGN 在 PCA+KNN 邻接下表现最佳
   - 性能随基因维度增加而提升
2. 填补
   - GRAPE 正则化分类，有效处理缺失数据
3. 挑战
   - 邻接选择关键但优化困难
   - 无 DGN 的深层模型有过平滑问题

研究展望 (i)

异构图学习

- 当前 GRAPE 框架采用的二分图数据表示本质上是同构图，将所有节点同等对待。然而，细胞和基因的生物学意义不同，在同构图设置下，它们的表示可能会趋同，影响下游任务性能。未来可引入异构图学习方法[15]，通过引入依赖于节点和边类型的参数，显式地对节点和边的异质性进行建模，使不同类型的节点和边有独立的表示，进而提升单细胞转录组学分析的效果。

[15] 如：Heterogeneous Graph Transformer

研究展望 (ii)

改进插补模型和重构损失

- 现有插补模型和重构损失的设计仍有优化空间。可尝试采用VAE 图模型[16]替换当前的插补模型，并将基因调控网络纳入重构损失的计算中。这样能更准确地捕捉基因之间的调控关系，提升基因表达数据插补的准确性，为后续细胞分类和其他分析任务提供更可靠的数据基础。

[16] scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses

研究展望 (iii)

细胞聚类应用

- GNN 在细胞聚类任务中已有成功应用案例[17]。未来可将细胞聚类层集成到 GRAPE 框架中，开展无监督细胞聚类研究，并与标准聚类方法[18]进行比较。这有助于发现细胞的潜在亚群结构，深入理解单细胞数据中的细胞异质性，为生物学研究提供更深入的见解。

[17] 如：Deep Modularity Networks(DMoN)

[18] 如：Leiden algorithm

Thank You for Your Listening!