Modeling and Simulation 建模与仿真, 2025, 14(5), 818-828
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.145436

Hans汉斯

SCINNO：一种基于深度学习的自编码器聚类分析方法

齐紫瑶

大连交通大学理学院，辽宁 大连

收稿日期：2025年4月27日；录用日期：2025年5月20日；发布日期：2025年5月28日

摘　要

单细胞测序技术的快速发展为解析细胞异质性提供了前所未有的分辨率，但高噪声数据与复杂细胞亚群的精准聚类仍是重大挑战。文章提出了一种融合深度去噪网络与多头自注意力机制的深度聚类框架，旨在提升单细胞数据的特征表示能力与聚类鲁棒性。首先，设计基于深度去噪自编码器的深度去噪网络(DN)，通过引入InfoNCE对比损失函数增强特征解耦能力，有效抑制数据噪声并提取低维干净特征；随后，提出一种结合多头自注意力机制的深度聚类网络(CN)，利用注意力权重捕捉特征间全局关联性，并通过模糊K-means算法动态优化隶属度矩阵U与聚类中心。在多个公开单细胞数据集上的实验表明，本方法较其他聚类算法具有更好的聚类效果，为单细胞数据的高效解析提供了新的理论支持与技术工具。

关键词

单细胞测序，自编码器，自注意力机制，聚类

SCINNO: A Deep Learning-Based Autoencoder Clustering Analysis Method

Ziyao Qi

School of Science, Dalian Jiaotong University, Dalian Liaoning

Received: Apr. 27th, 2025; accepted: May 20th, 2025; published: May 28th, 2025

Abstract

The rapid development of single-cell sequencing technology has provided unprecedented resolution

文章引用: 齐紫瑶. SCINNO: 一种基于深度学习的自编码器聚类分析方法[J]. 建模与仿真, 2025, 14(5): 818-828.
DOI: 10.12677/mos.2025.145436

for resolving cellular heterogeneity, but accurate clustering of noisy data and complex cell subpopulations remains a major challenge. This paper proposes a deep clustering framework that integrates a deep denoising network and a multi-head self-attention mechanism, aiming to improve the feature representation ability and clustering robustness of single-cell data. Firstly, a deep denoising network (DN) based on a deep denoising autoencoder was designed, and the feature decoupling ability was enhanced by introducing the InfoNCE contrast loss function, which effectively suppressed the data noise and extracted low-dimensional clean features. Subsequently, a deep clustering network (CN) combined with a multi-head self-attention mechanism was proposed, which used attention weights to capture the global correlation between features and dynamically optimized the membership matrix U and the clustering center through the fuzzy K-means algorithm. Experiments on multiple public single-cell datasets show that the proposed method has a better clustering effect than other clustering algorithms and provides new theoretical support and technical tools for the efficient analysis of single-cell data.

Keywords

Single-Cell Sequencing, Autoencoder, Self-Attention Mechanism, Clustering

1. 引言

单细胞 RNA 测序(scRNA-seq)技术的革命性突破为生命科学研究开辟了新纪元。该技术通过对单个细胞转录组的高通量测序，实现了对复杂生物系统中细胞异质性的精细化解析[1]。在发育生物学领域，该技术已成功绘制出斑马鱼胚胎细胞分化的动态图谱；在肿瘤研究中，其揭示了肿瘤微环境中免疫细胞的多样性特征。然而，该技术产生的数据具有显著的高维特性(通常包含数千个基因表达特征)、极端的稀疏性(零值比例超过 80%)以及复杂的技术噪声(批次效应、捕获效率差异等)，这些特性给后续的生物信息学分析带来了严峻挑战。传统的聚类分析方法在处理此类数据时暴露出明显局限性。以 K-means [2]为代表的线性聚类算法，依赖于研究者对特征空间的先验假设，难以捕捉基因表达网络的非线性关系。层次聚类虽然能够揭示数据的层次结构，但计算复杂度高且容易受噪声干扰。近年来兴起的深度学习方法为解决该问题提供了新路径。基于自编码器的深度嵌入聚类模型(DEC)通过构建端到端的特征学习框架，在多个数据集上展现出优于传统方法的性能。

2. 研究现状

单细胞 RNA 测序(scRNA-seq)技术的突破性进展使得在单个细胞水平解析基因表达谱成为可能，为揭示细胞异质性、发育轨迹及疾病机制提供了关键手段。然而，scRNA-seq 数据固有的高维度、稀疏性及技术噪声严重制约了细胞亚群的精准识别。传统聚类方法(如 K-means、层次聚类)依赖人工特征工程与先验假设，难以捕捉复杂的非线性特征关系；而基于深度学习的聚类模型(如深度嵌入聚类，DEC)虽能自动提取特征，却常因噪声干扰导致潜在空间重叠，影响聚类效果。因此，如何构建兼具去噪能力与特征关联性建模的深度聚类框架，成为当前研究的核心挑战。

近年来，自编码器(AE)与变分自编码器(VAE)因其强大的特征重构能力被广泛应用于单细胞数据分析。然而，现有方法在两方面存在局限：其一，去噪过程多依赖简单重构损失，难以区分噪声与真实生物变异；其二，特征交互建模多采用全连接层，缺乏对全局关联的显式建模。针对上述问题，本文提出

一种创新性深度聚类框架，其核心贡献如下：

1) 深度去噪网络与 InfoNCE 对比损失的融合：在深度去噪自编码器(DAE)中引入 InfoNCE 对比损失，通过最大化干净特征与去噪特征的互信息，增强模型对噪声的鲁棒性。相较于传统均方误差(MSE)损失，InfoNCE 通过正负样本对比学习，有效保留数据内在结构，抑制过平滑问题。

2) 多头自注意力机制与模糊聚类的协同优化：在深度聚类网络(CN)中嵌入多头自注意力层，通过并行注意力头捕获特征间多尺度依赖关系，并结合模糊 K-means 算法动态更新隶属度矩阵。该设计不仅提升了特征表示的判别性，还缓解了聚类中心初始化敏感性问题。

本文在多个单细胞数据集上进行了广泛验证。结果表明，相较于其他模型，本文模型在聚类效果上均显著优于现有方法。本研究不仅为单细胞数据分析提供了新的技术路径，其模块化设计思路还可扩展至多组学数据整合领域，具有重要的理论价值与应用前景。

3. 材料和方法

3.1. 数据集

在本文中，我们考虑在常用的数据集上进行比较，然后通过实验来证明本研究模型的有效性和价值。总共在五个包含不同人类和小鼠组织的数据集上评估了该模型，其中 Lawlor [3]和 Muraro [4]是两个包含未知细胞类型的数据集，Bmcite [5]和 Bhattacherjee [6]是两个包含超过 10,000 个细胞的数据集。这些数据集的详细信息如表 1 所示。

Table 1. Details of the scRNA-seq dataset
表 1. scRNA-seq 数据集的详细信息

No.	Dataset	Cell Source	Cells	Original	Cell Types
1	Kolodziejski	Mouse	704	38,653	3
2	Lawlor	Human	638	26,616	8
3	Muraro	Human	3072	19,059	11
4	Bhattacherjee	Mouse	24,822	21,000	8
5	Bmcite	Mouse	30,672	17,009	5

3.2. 预处理

我们借助 Python 包 SCANPY [7]对原始的单细胞 RNA 测序(scRNA-seq)数据开展预处理工作。scRNA-seq 数据呈现为矩阵形式，矩阵的每一行对应一个细胞，每一列对应一个基因，并且每个细胞所对应的基因数量是相同的。为了减少无用基因对模型计算过程以及聚类准确性的不利影响，我们会针对每个数据集，去除那些在超过 95%的细胞中计数都为零的基因。之后，对数据进行归一化处理，使其平均值为 0 且方差为 1。归一化操作能够让数据处于统一的尺度上，有利于后续的分析。完成归一化后，再对数据进行对数变换，对数变换可以有效压缩数据的动态范围，使数据分布更加符合分析要求。经过上述预处理步骤后，我们会挑选出前 2500 个高度可变的基因，将这些基因的数据作为 SCINNO 模型的输入数据。这些高度可变的基因往往蕴含着更多关于细胞状态和功能的关键信息，有助于提升模型的性能和分析的准确性。

3.3. 模型结构

SCINNO 的整体流程如图 1 所示。SCINNO 模型由加入了 InfoNCE 对比损失函数的深度去噪网络和加入了多头自注意力机制的深度聚类网络组成。

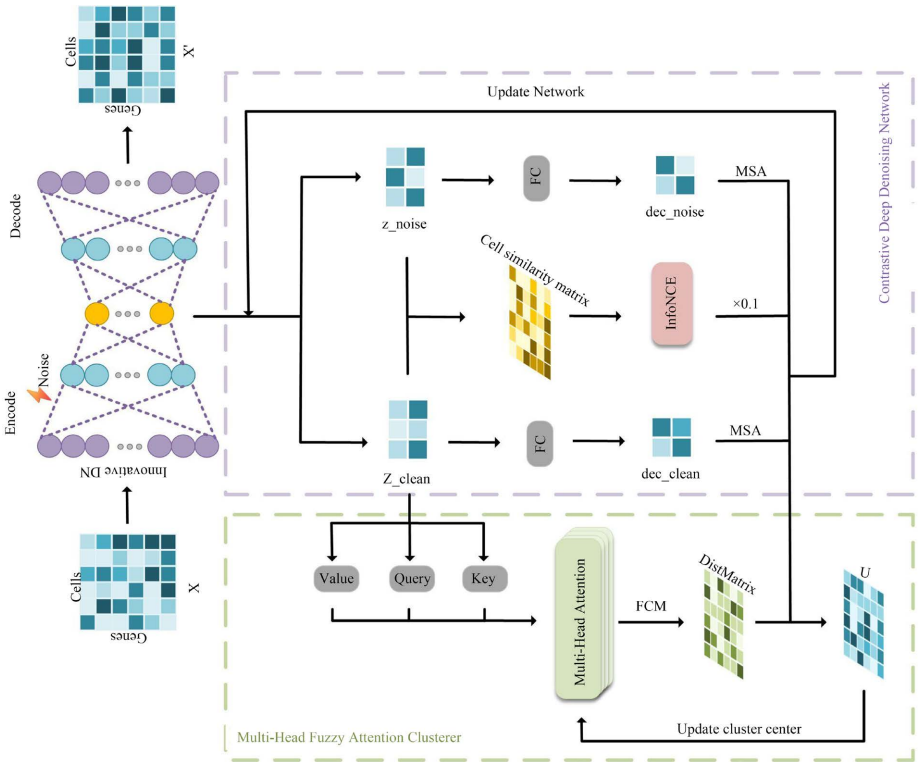

Figure 1. Diagram of the model structure of SCINNO
图 1. SCINNO 的模型结构图

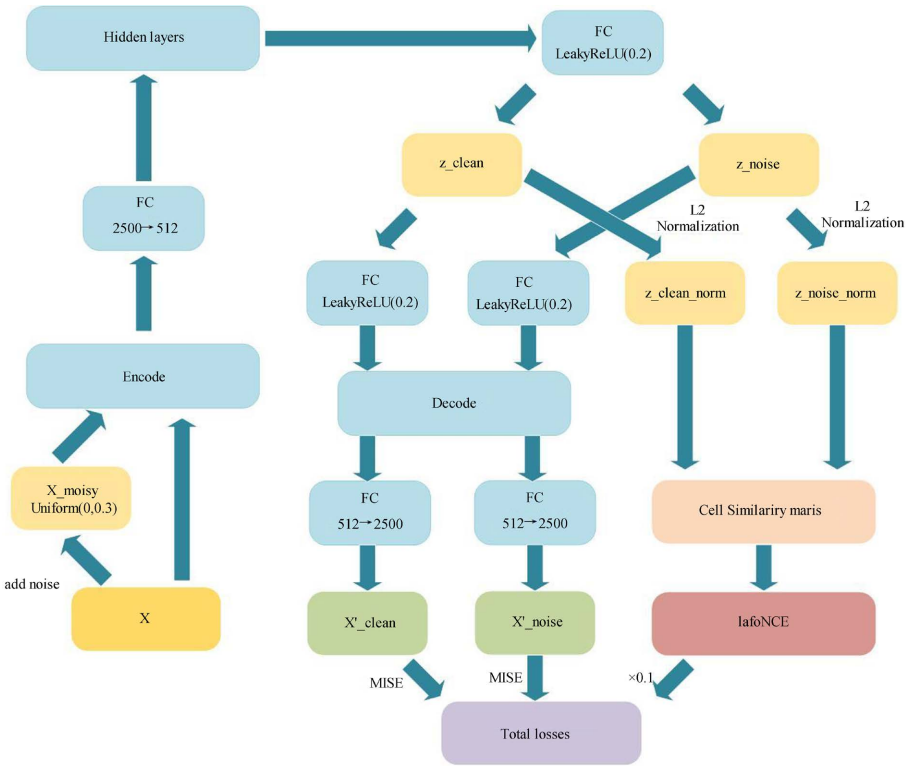

Figure 2. Flowchart of DN
图 2. DN 的流程图

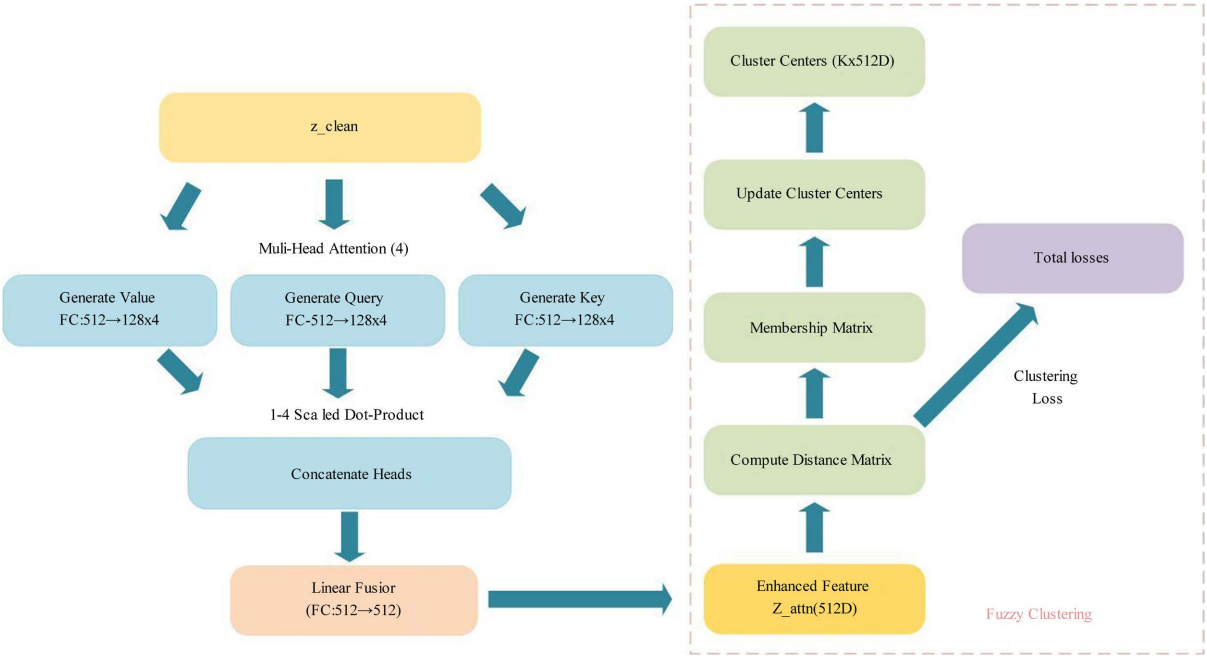

Figure 3. Flowchart of CN
图 3. CN 的流程图

深度去噪网络的具体结构如图 2 所示。深度聚类网络的具体结构如图 3 所示。

SCINNO 模型通过深度去噪网络(DN_Innovative)与深度聚类网络(CN_Innovative)的协同设计，实现了单细胞数据从噪声抑制到精准聚类的端到端优化。DN_Innovative 模块采用双路径编码策略，在原始基因表达矩阵中注入均匀噪声(幅度 0.3 的随机扰动)生成噪声数据，分别通过共享权重的编码器提取干净特征与噪声特征。其中，干净路径直接编码原始数据，噪声路径学习含噪输入的鲁棒表示，二者通过重构损失(MSE)强制解码器恢复原始表达谱，同时引入对比损失(InfoNCE)最大化干净与噪声特征间的互信息，迫使模型忽略技术噪声、聚焦生物学信号。CN_Innovative 模块则基于多头自注意力机制建模基因间的全局依赖关系，将去噪后的特征映射到多头子空间(4 头)，通过并行计算注意力权重捕捉不同基因组合的共表达模式(如细胞周期相关基因簇)，生成更具判别性的融合特征。在此基础上，动态模糊聚类算法为每个细胞计算与各簇中心的软隶属度，采用指数衰减函数(模糊因子 $\gamma = 2.0$)优化概率分配，并通过加权平均迭代更新簇中心位置，有效处理细胞状态的连续过渡与边界模糊问题。端到端训练将重构误差、对比对齐与聚类紧密度目标统一为联合损失函数，通过反向传播同步优化编码器、注意力层与聚类中心参数，避免了传统分阶段流程中特征学习与聚类目标的割裂，从而在抑制噪声传播的同时增强聚类判别性能，最终实现高噪声单细胞数据的高精度亚群解析。

3.4. 数学原理和公式

3.4.1. 噪声输入机制和双分支编码器

给定原始基因表达向量 $x \in R^{2500}$ (维度由高可变基因筛选确定)，通过均匀分布噪声生成增强视图：

$$\tilde{x} = x + \delta, \delta \sim U(-\epsilon, \epsilon) \odot 1 \tag{1}$$

$\epsilon = 0.3$ 控制噪声幅度，通过网格搜索验证其对模型鲁棒性的平衡效果(过大破坏语义，过小降低抗噪性)；噪声项 δ 的每个元素独立服从均匀分布，即每个维度的噪声值在 $[-0.3, 0.3]$ 内随机生成，$1 \in \mathbb{R}^{2500}$ 是全 1 向量，确保噪声均匀施加到所有基因维度；$U(-\epsilon, \epsilon)$ 表示均匀分布，生成与原始数据匹配的噪声。

共享权重的编码器 f_e 分别处理干净数据 x 和噪声数据 $\tilde{x}$，生成潜在特征：

$$\mathbf{z}_{\text{clean}} = \text{LeakyReLU}\left(\mathbf{W}_2 \cdot \text{LeakyReLU}\left(\mathbf{W}_1\mathbf{x}\right)\right),\ \mathbf{z}_{\text{noise}} = \text{LeakyReLU}\left(\mathbf{W}_2 \cdot \text{LeakyReLU}\left(\mathbf{W}_1\tilde{\mathbf{x}}\right)\right) \tag{2}$$

$\mathbf{W}_1 \in \mathbb{R}^{512\times512}, \mathbf{W}_2 \in \mathbb{R}^{2500\times512}$ 是编码器的权重矩阵，将输入从高维基因空间映射到 512 维潜在空间；LeakyRelu 激活函数缓解梯度消失问题，允许少量负值信息通过。

3.4.2. 信息对比损失(InfoNCE Loss)

对批次内样本计算归一化相似度，最大化正样本对 $\left(\mathbf{z}_{\text{clean}}^{(i)}, \mathbf{z}_{\text{noise}}^{(i)}\right)$ 的互信息：

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s_{ii}/\tau\right)}{\sum_{j=1}^{B}\exp\left(s_{ij}/\tau\right)},\ s_{ij} = \frac{\left\langle \hat{\mathbf{z}}_{\text{clean}}^{(i)}, \hat{\mathbf{z}}_{\text{noise}}^{(j)} \right\rangle}{\left\|\hat{\mathbf{z}}_{\text{clean}}^{(i)}\right\|_2 \left\|\hat{\mathbf{z}}_{\text{noise}}^{(j)}\right\|_2} \tag{3}$$

$B=128$ 是批次大小，平衡计算效率与负样本数量；$\tau=0.5$ 是温度参数，调节相似度分布尖锐度(值越小，相似度越集中，实验验证其最优性)；$\hat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$ 表示 L2 归一化后的特征，消除幅值对相似度的影响。

3.4.3. 双路径重构损失

强制编码器保留原始数据的可逆性：

$$\mathcal{L}_{\text{recon}} = \frac{1}{B}\sum_{i=1}^{B}\left(\left\|\mathbf{x}^{(i)} - f_d\left(\mathbf{z}_{\text{clean}}^{(i)}\right)\right\|_2^2 + \left\|\mathbf{x}^{(i)} - f_d\left(\mathbf{z}_{\text{noise}}^{(i)}\right)\right\|_2^2\right) \tag{4}$$

解码器结构：

$$f_d\left(\mathbf{z}\right) = \mathbf{W}_4 \cdot \text{LeakyReLU}\left(\mathbf{W}_3\mathbf{z}\right),\quad \mathbf{W}_3 \in \mathbb{R}^{512\times512}, \mathbf{W}_4 \in \mathbb{R}^{2500\times512} \tag{5}$$

重构损失确保潜在特征 z_{clean} 和 z_{niose} 保留原始数据的全局结构；共享解码器权重 f_d 迫使双路径特征空间对齐。

3.4.4. 多头自注意力机制

将潜在特征 $z_{\text{clean}} \in \mathbb{R}^{512}$ 分解为 $H=4$ 头，生成注意力增强特征 z_{attn}：

$$Q_h = W_h^Q Z_{\text{clean}},\quad K_h = W_h^K Z_{\text{clean}},\quad V_h = W_h^V Z_{\text{clean}} \tag{6}$$

$$\text{Attn}_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right)V_h \tag{7}$$

$$\mathbf{z}_{\text{attn}} = \mathbf{W}^O\left[\text{Attn}_1 \| \cdots \| \text{Attn}_H\right],\quad d_h = \frac{512}{H} = 128 \tag{8}$$

$W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{512\times128}$ 是各头的投影矩阵，学习不同子空间的关联模式；$\sqrt{d_h} = \sqrt{128}$ 缩放点积结果，防止梯度爆炸；$\mathbf{W}^O \in \mathbb{R}^{512\times512}$ 将拼接后的多头输出映射回原维度。

3.4.5. 模糊聚类损失

基于注意力增强特征 z_{attn}，计算软隶属度和聚类损失：

$$u_{ik} = \frac{\exp\left(-\gamma\left\|\mathbf{z}_{\text{attn}}^{(i)} - \mathbf{c}_k\right\|^2\right)}{\sum_{j=1}^{K}\exp\left(-\gamma\left\|\mathbf{z}_{\text{attn}}^{(i)} - \mathbf{c}_j\right\|^2\right)} \tag{9}$$

$$\mathcal{L}_{\text{cluster}} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}u_{ik}^m\left\|\mathbf{z}_{\text{attn}}^{(i)} - \mathbf{c}_k\right\|^2 + \alpha\sum_{i,k}u_{ik}\log u_{ik} \tag{10}$$

$\gamma = 2.0$ 制隶属度衰减速率(值越大，样本越倾向于单一类别)； $m = 2.0$ 是模糊指数，调节隶属度软硬程度($m \to \mathbf{1}$ 时退化为硬聚类)；$\alpha = 0.1$ 是熵正则项系数，防止所有样本隶属同一类别的退化解；$\mathbf{c}_k \in R^{512}$ 是可学习的聚类中心，表示细胞类群的原型特征。

3.4.6. 总优化目标

联合对比学习、重构与聚类损失，动态平衡优化方向：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{contrast}} + \lambda_2 \mathcal{L}_{\text{cluster}}, \quad \lambda_1 = 0.1, \lambda_2 = 0.5 \tag{11}$$

$\lambda_1 = 0.1$ 控制对比损失权重，避免过度扭曲特征空间； $\lambda_2 = 0.5$ 强化聚类损失的主导作用；训练初期以重构损失为主，后期逐步增加聚类损失权重。

3.5. 评价指标

3.5.1. NMI

归一化互信息(Normalized Mutual Information, NMI) [8]：

$$\text{NMI}(U,V) = \frac{2 \cdot MI(U,V)}{H(U) + H(V)}$$

$\text{MI}(U,V)$ 是互信息(Mutual Information)， $H(U)$ 和 $H(V)$ 分别是聚类结果 U 和真实标签 V 的熵。

互信息(MI)是衡量两个聚类结果之间的依赖关系，计算公式为：

$$\text{MI}(U,V) = \sum_{i=1}^{R} \sum_{j=1}^{C} P(i,j) \log \left(\frac{P(i,j)}{P_U(i) P_V(j)} \right)$$

$P(i,j)$：样本同时属于聚类 U 的第 i 类和真实标签 V 的第 j 类的概率； $P_U(i)$ 和 $P_V(j)$：样本分别属于聚类 U 的第 i 类和真实标签 V 的第 j 类的概率。

熵是衡量聚类结果或真实标签的不确定性：

$$H(U) = -\sum_{i=1}^{R} P_U(i) \log P_U(i), H(V) = -\sum_{j=1}^{C} P_V(j) \log P_V(j)$$

归一化：将互信息除以两个熵的算术平均，使结果范围在[0, 1]之间。值越接近 1，表示聚类结果与真实标签的一致性越高。

3.5.2. ARI

调整兰德指数(Adjusted Rand Index, ARI) [9]：

$$\text{ARI} = \frac{\text{RI} - \text{E}[\text{RI}]}{\max(\text{RI}) - \text{E}[\text{RI}]}$$

其中： RI 是兰德指数(Rand Index)； $\text{E}[\text{RI}]$ 是兰德指数的期望值。

兰德指数(RI)：计算样本对在聚类结果和真实标签中的一致性比例：

$$\text{RI} = \frac{a + d}{a + b + c + d}$$

其中， a：在聚类和真实标签中均被分到同一类的样本对数； b：在聚类中被分到同一类，但在真实标签中被分到不同类的样本对数； c：在聚类中被分到不同类，但在真实标签中被分到同一类的样本对数； d：在聚类和真实标签中均被分到不同类的样本对数。

调整：通过减去随机情况下的期望值 $\text{E}[\text{RI}]$ 并进行归一化，使结果范围在[0, 1]之间； $\text{ARI} = 1$：完全

一致； $ARI = 0$：随机水平； $ARI < 0$：比随机更差。

列联表法简化计算：

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}$$

其中：n_{ij} 是聚类类别 i 和真实标签 j 的共同样本数，a_i 是聚类类别 i 的样本数，b_j 是真实标签 j 的样本数，n 是总样本数。

4. 结果与讨论

4.1. 聚类效果比较

为了评估 SCINNO (结果图中名为 Innovative)的聚类性能，我们在五个真实的 scRNA-seq 数据集上运行 SCINNO 模型来获得聚类效果。这些数据集的标签信息都是通过大量的生物实验获得，同时与多种具有默认参数的单细胞聚类方法进行比较，包括 Celltree [10]、SHARP [11]、SIMLR [12]、Seurat [13]、scENA [14]和 SCGAE [15]。此外，我们通过两个常用的聚类指标(NMI, ARI)来评估每个聚类模型，来展示我们的模型性能。详细的指标值如下图 4 所示，从图中我们可以看到，SCINNO 模型对每个数据集的整体效果优于其他对比方法。具体来说，SCINNO 在每个数据集上实现了最佳的 NMI 和 ARI 值。

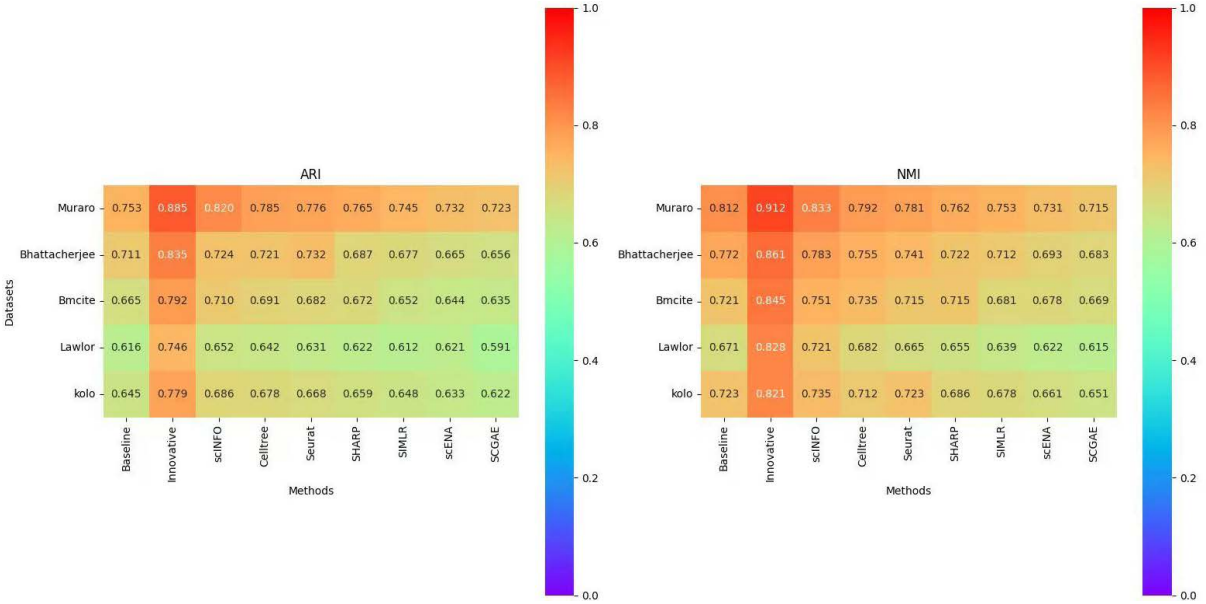

Figure 4. Clustering performance of SCINNO and other clustering methods
图 4. SCINNO 和其他聚类方法的聚类性能

4.2. 聚类效果可视化

为了更清楚地显示 SCINNO 模型的聚类效应，我们使用 T-SNE 将真实数据集投影到 2D 空间中，更直观地展示预测效果。如图 5 所示，一个点代表一个细胞，不同的颜色代表不同的细胞类型。从图中可以看出，在 Muraro、Bmcite 和 Bhattacherjee 数据集上，SCINNO 预测的不同颜色簇具有明显的边界，可

以清晰地区分类别。

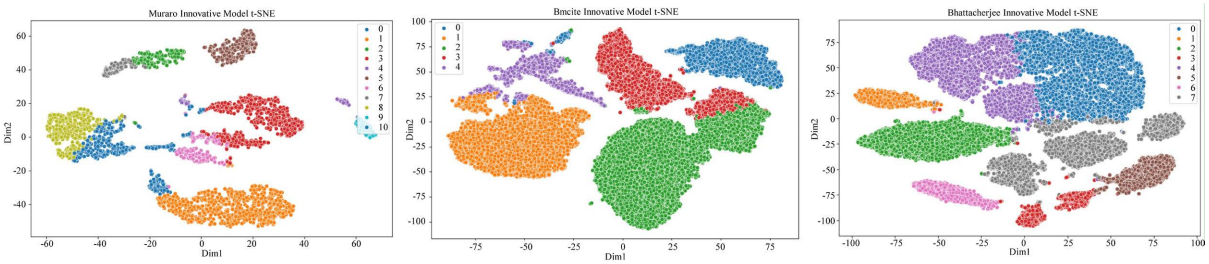

Figure 5. Visualization of the clustering effect of the SCINO model on the datasets
图 5. SCINNO 模型在数据集上的聚类效果可视化

4.3. 消融实验

为了验证 InfoNCE 对比损失函数和多头自注意力机制对模型性能具有贡献，我们分别关闭 InfoNCE 对比损失函数和多头自注意力机制，来和 SCINNO 模型进行模型效果对比。从图 6 可以看出，当分别关闭 InfoNCE 对比损失函数和多头自注意力机制时，模型性能都有所下降，这也说明了这两者对模型效果的提升是有重要贡献的。

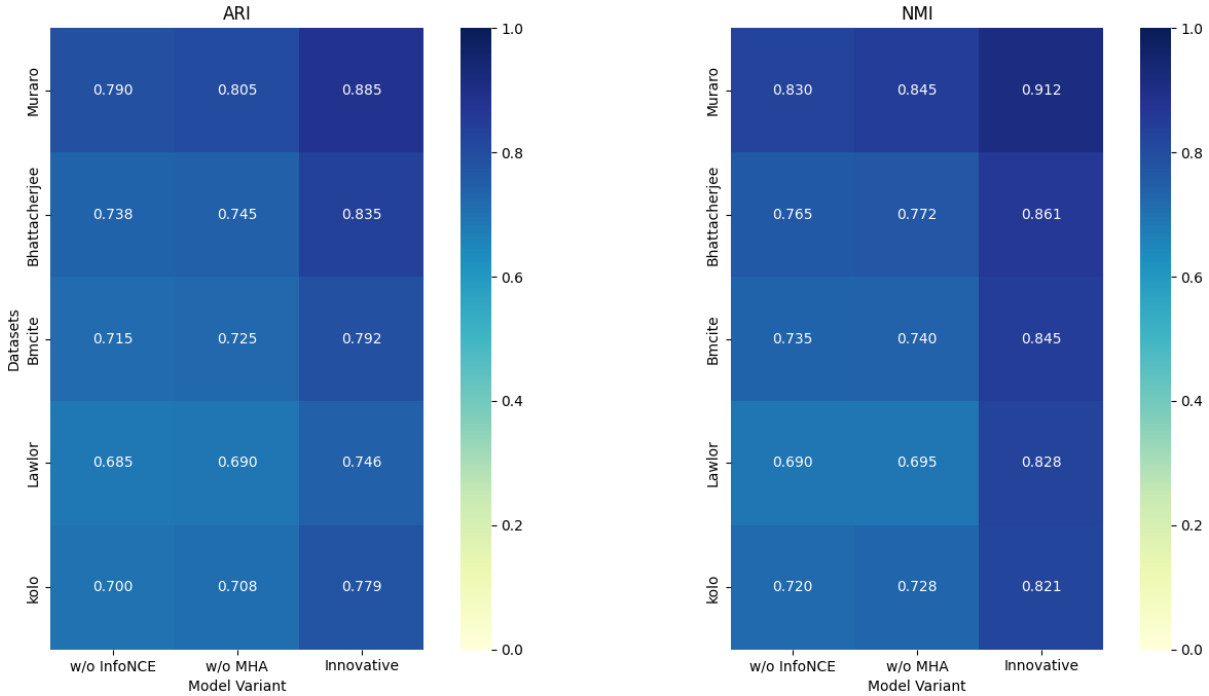

Figure 6. Ablation experiments of the SCINNO model
图 6. SCINNO 模型的消融实验

4.4. 讨论

在本文中我们提出了一种基于深度模型的聚类方法——SCINNO，该方法的核心架构包含两个关键模块：DN_Innovative 模块通过噪声注入模拟技术噪声，构建干净与噪声数据的双路径编码网络，结合重构损失(MSE Loss)和对比损失(InfoNCE Loss)实现去噪特征学习；CN_Innovative 模块则通过多头自注意

力机制增强特征交互，结合动态模糊聚类优化样本与聚类中心的软分配关系。模型通过端到端联合训练，将特征学习、去噪和聚类目标统一优化，总损失函数包含重构误差、对比对齐和聚类目标的综合约束。我们的模型具有这几个优势：一是通过噪声对比学习对齐干净与噪声数据的潜在空间(z_clean vs z_noise)，提升模型对测序误差等技术噪声的鲁棒性；二是引入多头自注意力捕捉基因调控网络的非线性关系，突破传统线性降维方法的局限性；三是采用动态模糊聚类策略优化样本隶属度矩阵，显著提高对重叠簇和稀有细胞类型的识别能力；四是通过端到端训练避免分步方法的误差累积，确保特征学习与聚类目标的深度适配。我们的模型的聚类效果相比于其他模型取得了显著的提升，为单细胞数据的聚类分析提供一种强有力的工具。

尽管 SCINNO 模型已取得显著进展，但其架构与算法仍存在可优化空间，具体方向包括：当前 DN_Innovative 模块采用固定幅度的均匀噪声(如 0.3)模拟技术噪声，然而实际单细胞数据中噪声类型复杂多样(如 dropout、批次效应、测序深度差异)。改进方向：基于细胞表达特征(如稀疏性、基因表达均值)自适应调整噪声幅度。例如，高稀疏细胞(dropout 较多)可注入更高强度噪声，迫使模型学习更强的抗噪能力或者是结合均匀噪声、高斯噪声与掩码噪声(模拟 dropout)，通过多模态噪声注入提升泛化性。

5. 结语

未来的研究我们可以进行多模态的数据融合，结合空间转录组、单细胞 ATAC-seq 等多组学数据，构建多维度特征空间。也可以考虑引入时序分析技术追踪细胞状态转变，揭示发育过程或疾病进展中的关键调控节点以及开发自适应超参数优化工具，减少人工干预。同时，建立统一的质量控制标准和基准测试平台，提升算法可比性。总之，我们后续仍会在单细胞测序方面继续努力，开发出更优秀、更好的模型来应用于生物医学的研究之中。

参考文献

[1] AlJanahi, A.A., Danielsen, M. and Dunbar, C.E. (2018) An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy—Methods & Clinical Development*, **10**, 189-196. https://doi.org/10.1016/j.omtm.2018.07.003

[2] Likas, A., Vlassis, N. and J. Verbeek, J. (2003) The Global K-Means Clustering Algorithm. *Pattern Recognition*, **36**, 451-461. https://doi.org/10.1016/s0031-3203(02)00060-2

[3] Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., *et al*. (2016) Single-Cell Transcriptomes Identify Human Islet Cell Signatures and Reveal Cell-Type-Specific Expression Changes in Type 2 Diabetes. *Genome Research*, **27**, 208-222. https://doi.org/10.1101/gr.212720.116

[4] Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., *et al*. (2016) A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, **3**, 385-394. https://doi.org/10.1016/j.cels.2016.09.002

[5] Bhattacherjee, A., Djekidel, M.N., Chen, R., Chen, W., Tuesta, L.M. and Zhang, Y. (2019) Cell Type-Specific Transcriptional Programs in Mouse Prefrontal Cortex during Adolescence and Addiction. *Nature Communications*, **10**, Article No. 4169. https://doi.org/10.1038/s41467-019-12054-3

[6] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., *et al*. (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902. https://doi.org/10.1016/j.cell.2019.05.031

[7] Wolf, F.A., Angerer, P. and Theis, F.J. (2018) SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biology*, **19**, Article No. 15. https://doi.org/10.1186/s13059-017-1382-0

[8] Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., *et al*. (2020) Single-Cell RNA-Seq Clustering: Datasets, Models, and Algorithms. *RNA Biology*, **17**, 765-783. https://doi.org/10.1080/15476286.2020.1728961

[9] Petegrosso, R., Li, Z. and Kuang, R. (2019) Machine Learning and Statistical Methods for Clustering Single-Cell RNA-Sequencing Data. *Briefings in Bioinformatics*, **21**, 1209-1223. https://doi.org/10.1093/bib/bbz063

[10] duVerle, D.A., Yotsukura, S., Nomura, S., Aburatani, H. and Tsuda, K. (2016) Celltree: An R/Bioconductor Package to Infer the Hierarchical Structure of Cell Populations from Single-Cell RNA-Seq Data. *BMC Bioinformatics*, **17**, Article No. 363. https://doi.org/10.1186/s12859-016-1175-6

[11] Wan, S., Kim, J. and Won, K.J. (2020) SHARP: Hyperfast and Accurate Processing of Single-Cell RNA-Seq Data via

Ensemble Random Projection. *Genome Research*, **30**, 205-213. https://doi.org/10.1101/gr.254557.119

[12] Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. and Batzoglou, S. (2017) Visualization and Analysis of Single-Cell RNA-Seq Data by Kernel-Based Similarity Learning. *Nature Methods*, **14**, 414-416. https://doi.org/10.1038/nmeth.4207

[13] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) Spatial Reconstruction of Single-Cell Gene Expression Data. *Nature Biotechnology*, **33**, 495-502. https://doi.org/10.1038/nbt.3192

[14] Cui, Y., Zhang, S., Liang, Y., Wang, X., Ferraro, T.N. and Chen, Y. (2021) Consensus Clustering of Single-Cell RNA-Seq Data by Enhancing Network Affinity. *Briefings in Bioinformatics*, **22**, bbab236. https://doi.org/10.1093/bib/bbab236

[15] Luo, Z., Xu, C., Zhang, Z. and Jin, W. (2021) A Topology-Preserving Dimensionality Reduction Method for Single-Cell RNA-Seq Data Using Graph Autoencoder. *Scientific Reports*, **11**, Article No. 20028. https://doi.org/10.1038/s41598-021-99003-7