基于对比学习增强自编码器的单细胞转录组聚类方法

1 引言

单细胞转录组测序技术能够在单个细胞水平上解析基因表达模式，在发育生物学[1]、自身免疫[2]及癌症研究[3]等领域展现出巨大的应用价值。然而，由于单细胞转录组数据普遍具有高度稀疏性、噪声较大以及维度高等特点，对其进行直接分析面临较大挑战。聚类分析能够在无监督条件下根据基因表达特征对细胞进行分组，从而揭示细胞间潜在的异质性与功能差异，因此已成为单细胞转录组数据分析中最为核心的方法之一。通过聚类分析，研究人员可以识别具有相似表达模式的细胞群体，进而探究细胞在组织结构中的组成关系，以及其在生理过程或疾病发生发展中的动态变化[4]。

传统的单细胞转录组聚类方法通常采用主成分分析[5]（Principal Component Analysis, PCA）对高维表达数据进行降维，并结合聚类算法（如 K-均值算法[6]、Leiden 算法[7]等）对细胞进行分组。然而，PCA 本质上是一种线性降维方法，假设数据分布于线性子空间中，在当前单细胞数据规模和复杂性不断提升的背景下[8]，其对复杂非线性细胞结构的刻画能力受到明显限制。此外，PCA 对异常值和噪声较为敏感，在处理具有高噪声、高稀疏性特征的单细胞转录组数据时稳定性较差，容易将技术噪声误识为主要成分，从而影响特征表示质量。

随着机器学习方法的发展，越来越多的研究将深度学习技术引入单细胞转录组聚类分析，以应对更加复杂的数据分布。其中，自编码器（Autoencoder）作为一种有效的表示学习模型，能够通过重构输入数据的方式学习其潜在低维表示，在细胞特征提取方面得到了广泛应用[9-11]。然而，传统自编码器的训练目标主要聚焦于最小化重构误差，缺乏对隐空间中样本判别性结构的显式约束，导致其学习到的特征在聚类等下游任务中的表现仍存在一定局限。

为提升自编码器学习到的潜在表示在聚类任务中的判别能力，本文在传统自编码器框架基础上引入对比学习机制，通过联合优化重构损失与对比损失，引导模型在保留表达信息的同时，学习更加紧凑且具有良好可分性的聚类友好特征表示。具体而言，对同一细胞的表达数

据施加不同的数据增强操作以构造正样本对，并通过编码器映射至隐空间；对比损失鼓励同一细胞不同视图的表示保持相似，同时拉开不同细胞之间的距离。模型训练过程中联合优化重构损失与对比损失，从而在保留原始表达信息的同时，增强潜在特征的判别性。实验结果表明，引入对比学习的自编码器在聚类任务中能够获得更加紧凑且分离良好的细胞簇。实验代码见 github：https://github.com/BlueberryOreo/Machine-Learning-suda-2025-fall-Course-Report

2 方法

2.1 自编码器设计

自编码器由编码器（Encoder）和解码器（Decoder）两部分组成。其中，编码器用于将输入数据映射到低维隐空间，以提取其潜在特征；解码器则根据隐空间表示对数据进行重构，尽可能还原原始输入。本文采用神经网络作为编码器和解码器的基本结构，其整体架构如图 1 所示。其中 $\mathcal{E}$ 和 $\mathcal{D}$ 分别表示编码器和解码器，$\mathbf{x}$ 表示原始输入数据，$\mathbf{z}$ 表示编码器学习得到的隐空间特征，$\hat{\mathbf{x}}$ 表示解码器重构得到的输出数据。

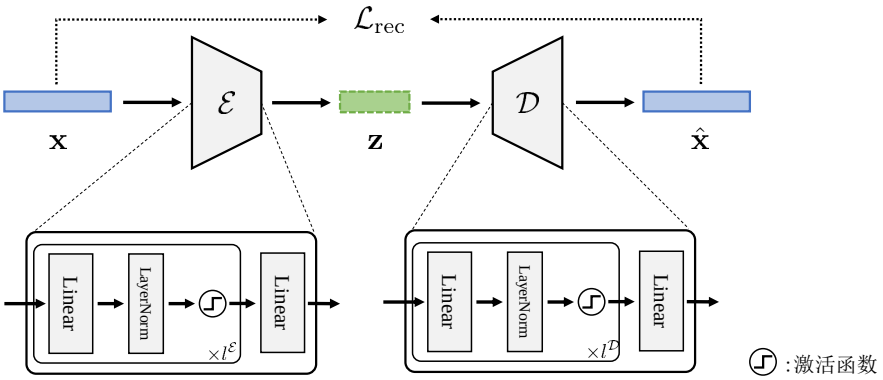


图 1 自编码器架构

自编码器的训练过程以重构误差最小化为目标，本文采用基于均方误差（Mean Squared Error, MSE）的重构损失函数，用于衡量原始输入数据与重构数据之间的差异，其定义如下：

$$\mathcal{L}_{\text{rec}} = \frac{1}{d}\sum_{i=1}^{d}(x_i - \hat{x}_i)^2 \tag{1}$$

其中 d 表示原始数据的维度，x_i 和 $\hat{x}_i$ 分别表示原始数据与重构数据在第 i 个维度上的取值。

2.2 对比学习

仅依赖重构损失对自编码器进行训练，缺乏对隐空间中样本判别性结构的显式约束，容易导致模型学习到的特征在聚类任务中的区分能力受限。为此，本文在重构损失的基础上引入对比学习机制，通过联合优化重构目标与对比目标，引导模型在保持数据表达能力的同时，学习更加紧凑且具有良好可分性的聚类友好特征表示。

为构建对比学习所需的正负样本对，本文采用“同一细胞不同视图”的数据增强策略，对原始细胞数据进行多视图变换。具体而言，采用基因随机丢弃、微量高斯噪声注入以及文库大小缩放三种数据增强方式。记增强后的数据为 $\tilde{\mathbf{x}}$，通过编码器对其进行特征提取得到对应的隐空间表示 $\tilde{\mathbf{z}}$。在一个批次内，同一细胞的不同视图所对应的特征表示构成正样本对 $(\mathbf{z}, \tilde{\mathbf{z}}^{+})$，而不同细胞的不同视图之间则构成负样本对 $(\mathbf{z}, \tilde{\mathbf{z}}^{-})$。

参考 Chen 等[12]，本文采用归一化温度尺度交叉熵（Normalized Temperature-scaled Cross Entropy, NT-Xent）损失函数对正负样本对进行对比学习，其定义如下：

$$\mathcal{L}_{\text{con}} = \frac{1}{2B} \sum_{i=1}^{2B} -\log \frac{\exp\left(s(\mathbf{z}_i, \mathbf{z}_i)/\tau\right)}{\sum_{k=1}^{2B} \mathbb{1}_{k \neq i} \exp\left(s(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{2}$$

其中，B 表示批次大小，$s(\cdot,\cdot)$ 表示样本间的相似度度量。本文采用余弦相似度来衡量隐空间表示之间的相似性，τ 为温度参数，用于调节相似度分布的平滑程度；$\mathbb{1}_{k \neq i}$ 为指示函数，当 $k \neq i$ 时取值为 1，否则取值为 0，以排除样本自身在分母中的贡献。

2.3 联合优化

基于式 (1) 和式 (2)，模型最终的训练损失函数可定义为：

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{con}} \tag{3}$$

其中 λ 为对比学习损失所占权重。

3 实验

3.1 实验设置

本文在 Tosches 等[13] 和 Schaum 等[14] 提供的单细胞转录组数据集上对所提出的方法进行聚类分析评估。Tosches 等[13] 对红耳龟（Trachemys scripta elegans）雌性个体端脑背侧皮层区域进行采样，共获得 18664 个单细胞样本，覆盖 23500 个基因。Schaum 等[14] 构建了涵盖小鼠 20 种器官和组织的单细胞转录组参考图谱，总计包含超过 100000 个单细胞样本。本文从该数据集中选取膈肌（Diaphragm）和肺（Lung）两个组织的细胞数据进行实验，分别包含 870 个细胞和 1676 个细胞，均测量了 23341 个基因。

在模型设置中，自编码器的输入维度设为 2000，中间隐藏层维度为 1024，隐空间特征维度为 128。编码器与解码器的网络层数 $l^{\mathcal{E}}$ 和 $l^{\mathcal{D}}$ 均设置为 2 层，采用 GELU 函数作为激活函数，并在网络中引入 dropout 正则化策略，dropout 率设为 0.1。模型训练过程中，批次大小设置为 512，学习率为 0.001，训练轮数为 100 轮。对比学习损失的权重系数 λ 设为 1.0，温度参数 τ 设为 0.05。在聚类阶段，本文采用单细胞分析中常用的基于 kNN 图的 Leiden 聚类算法[7]对隐空间特征进行聚类分析。具体而言，借助 Scanpy 库[15]构建 k 近邻图，设置邻居数为 15，并采用余弦相似度作为样本间相似度度量。对于分辨率，在红耳龟数据集（以下简称 Turtle）上设置为 0.6，在小鼠肺细胞数据集（以下简称 Lung）上设置为 0.7，在小鼠膈肌细胞数据集（以下简称 Diaphragm）上设置为 0.5。

实验在一台 Linux 服务器上完成。服务器采用双路 AMD EPYC 7402 处理器（每颗 24 核、支持超线程，总计 48 物理核心 / 96 逻辑线程），CPU 主频 2.8 GHz，并具备多 NUMA 节点拓扑结构。训练使用一块 NVIDIA GeForce RTX 3090 GPU 进行加速。软件环境方面，实验使用 Python 3.11，并基于 PyTorch 2.5.0 搭建模型，以保证实验的可复现性与一致性。

3.2 数据预处理

为保证数据质量并提高模型训练的稳定性，在模型训练之前对原始细胞数据进行了系统的预处理。本文采用单细胞分析中常用的预处理流程，主要包括质量控制（Quality Control, QC）、高变基因筛选、数据标准化以及对数化处理。所有预处理步骤均基于 Scanpy 库[15]实现。为保证不同数据集之间的可比性，本文在三个数据集中均统一选取 2000 个高变基因作为后续分析和模型训练的输入特征。

3.3 实验结果

表 1 实验结果，粗体表示最好结果

方法	Turtle			Diaphragm			Lung		
	NMI ↑	ARI ↑	ACC ↑	NMI ↑	ARI ↑	ACC ↑	NMI ↑	ARI ↑	ACC ↑
PCA	**0.868**	0.847	0.908	0.893	0.850	0.911	0.775	0.481	0.653
基准自编码器	0.860	**0.913**	0.910	0.912	0.901	0.939	0.746	0.466	0.575
自编码器 + 对比学习	0.866	**0.913**	**0.935**	**0.958**	**0.979**	**0.991**	**0.823**	**0.722**	**0.786**

实验结果如表 1 所示。为充分验证所提方法在聚类任务中的有效性，实验中额外引入了传统非神经网络方法作为对比，即基于 PCA 进行特征提取后再进行聚类。从结果可以观察到，相较于基准自编码器方法，引入对比学习机制后，模型在所有数据集上的聚类性能均得到进一步提升，其中在 Diaphragm 和 Lung 数据集上的提升尤为显著。这一现象表明，对比学习能够有效增强潜在特征的判别性，从而有助于在聚类过程中形成更加紧凑且分离良好的细胞簇。

此外，与基于 PCA 的特征提取方法相比，自编码器提取的特征在 ARI 指标上表现出明显优势，说明 PCA 所学习的线性表示在样本对关系建模方面存在一定局限性。相比之下，自编码器能够通过非线性映射学习更加符合数据分布的潜在表示，而进一步引入对比学习后，类内一致性和类间区分性均得到显著改善，从而在聚类结构质量上取得更优的表现。

3.4 可视化分析

附录图 A-1、A-2 和 A-3 展示了基于标准自编码器及引入对比学习的自编码器所提取特征，在细胞数据上通过 UMAP 算法[16] 进行可视化的结果。可以观察到，引入对比学习后，不同类别细胞在特征空间中的分离程度明显增强，而同一类别细胞的特征分布更加紧凑。这一现象与表 1 中 NMI 和 ARI 指标的整体提升高度一致，表明对比学习有效增强了潜在特征的全局结构一致性以及样本对层面的类内一致性。尤其在 Lung 数据集上（图 A-3），标准自编码器提取的特征中同类细胞分布较为分散，导致其在 ARI 和 ACC 指标上表现受限；而在引入对比学习后，同类细胞特征显著收敛、类间边界更加清晰，对应地在定量结果中表现为 ARI 和 ACC 的大幅提升。该结果进一步说明，对比学习通过同时压缩类内距离并拉开类间距离，为聚类算法提供了更加判别性的特征表示，从而促成更加紧凑且分离良好的细胞簇。

4 总结

本文围绕单细胞转录组数据高维、稀疏和噪声大的特点，提出了一种基于对比学习增强自编码器的单细胞聚类方法。针对传统 PCA 等线性降维方法以及仅以重构误差为目标的自编码器在刻画复杂非线性结构和聚类判别性方面的不足，本文在自编码器框架中引入对比学习机制，通过联合优化重构损失与对比损失，引导模型学习更加紧凑且具有良好可分性的潜在特征表示。

在多个真实单细胞转录组数据集（Turtle、Diaphragm 和 Lung）上的实验结果表明，与 PCA 以及基准自编码器方法相比，引入对比学习后的自编码器在 NMI、ARI 和 ACC 等聚类评价指标上均取得了更优的性能，尤其在结构复杂的数据集上表现出显著优势。UMAP 可视化结果进一步验证了该方法能够有效增强类内一致性、拉大类间距离，从而形成更加清晰、稳定的细胞簇结构。

综上所述，本文方法通过对比学习与自编码器的有机结合，在保持表达信息重构能力的同时显著提升了特征表示的判别性，为单细胞转录组数据的无监督聚类分析提供了一种有效且具有推广潜力的解决方案。

参考文献

[1] Semrau S, Goldmann J E, Soumillon M, Mikkelsen T S, Jaenisch R, Van Oudenaarden A. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells[J]. Nature communications, 2017, 8(1): 1096.

[2] Gaublomme J T, Yosef N, Lee Y, Gertner R S, Yang L V, Wu C, Pandolfi P P, Mak T, Satija R, Shalek A K, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity [J]. Cell, 2015, 163(6): 1400-1412.

[3] Patel A P, Tirosh I, Trombetta J J, Shalek A K, Gillespie S M, Wakimoto H, Cahill D P, Nahed B V, Curry W T, Martuza R L, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma[J]. Science, 2014, 344(6190): 1396-1401.

[4] Jiang S, Wang C, Sun Q, Zhang Z. A robust multi-scale clustering framework for single-cell RNA-seq data analysis[J]. Scientific Reports, 2025, 15(1): 18543.

[5] Pearson K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 1901, 2(11): 559-572.

[6] McQueen J B. Some methods of classification and analysis of multivariate observations[C]//Proc. of 5th Berkeley Symposium on Math. Stat. and Prob. 1967: 281-297.

[7] Traag V A, Waltman L, Van Eck N J. From Louvain to Leiden: guaranteeing well-connected communities[J]. Scientific reports, 2019, 9(1): 1-12.

[8] Lopez R, Regier J, Cole M B, Jordan M I, Yosef N. Deep generative modeling for single-cell transcriptomics[J]. Nature methods, 2018, 15(12): 1053-1058.

[9] Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach[J]. Nature Machine Intelligence, 2019, 1(4): 191-198.

[10] Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly M P, Hu G, Li M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis[J]. Nature communications, 2020, 11(1): 2338.

[11] Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data[J]. Nature communications, 2021, 12(1): 1873.

[12] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. 2020: 1597-1607.

[13] Tosches M A, Yamawaki T M, Naumann R K, Jacobi A A, Tushev G, Laurent G. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles[J]. Science, 2018, 360(6391): 881-888.

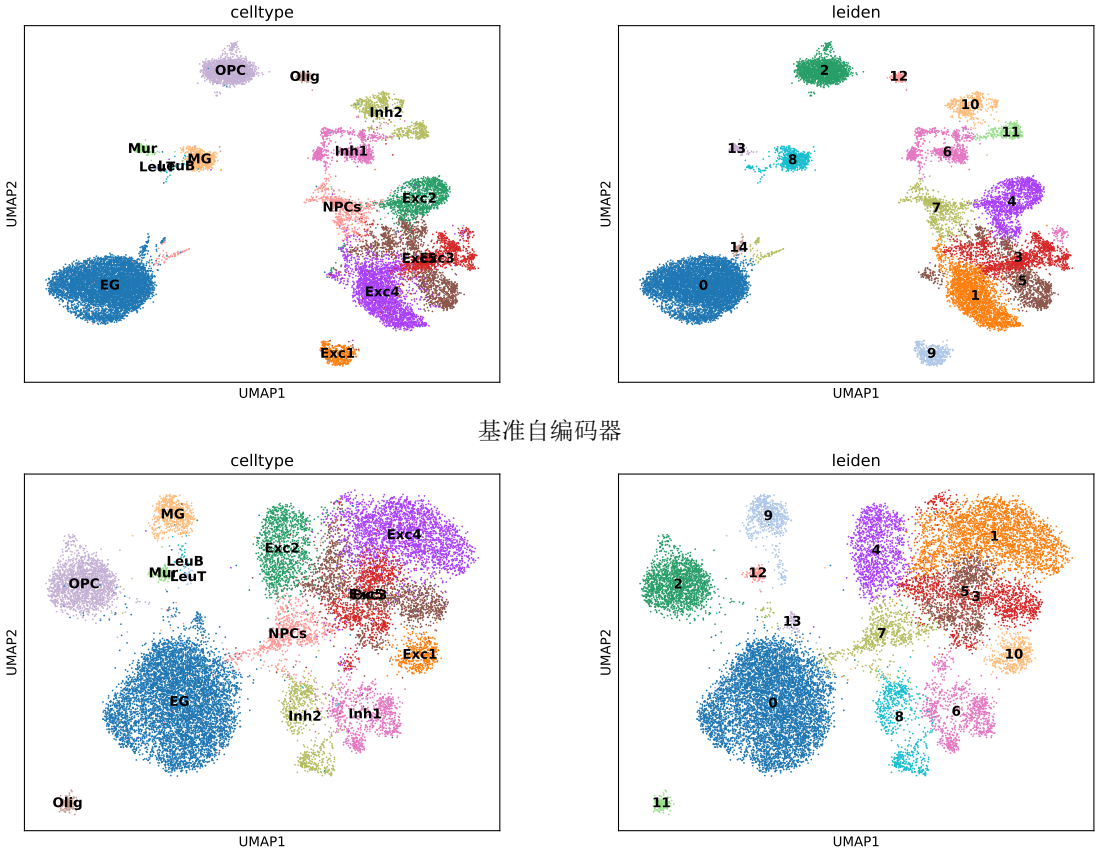
[14] Schaum N, Karkanias J, Neff N F, May A P, Quake S R, Wyss-Coray T, Darmanis S, Batson J, Botvinnik O, Chen M B, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium[J]. Nature, 2018, 562(7727): 367.

[15] Wolf F A, Angerer P, Theis F J. SCANPY: large-scale single-cell gene expression data analysis[J]. Genome biology, 2018, 19(1): 15.

[16] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint arXiv:1802.03426, 2018.

附录

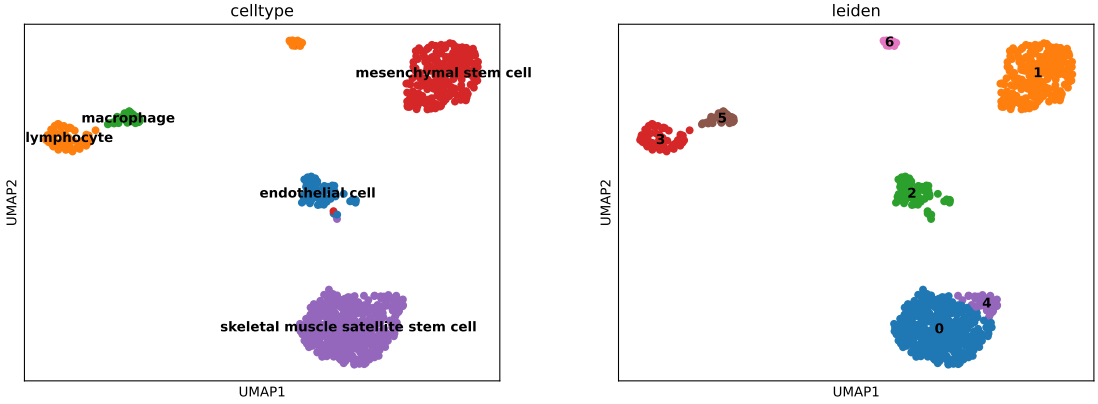
A　可视化结果



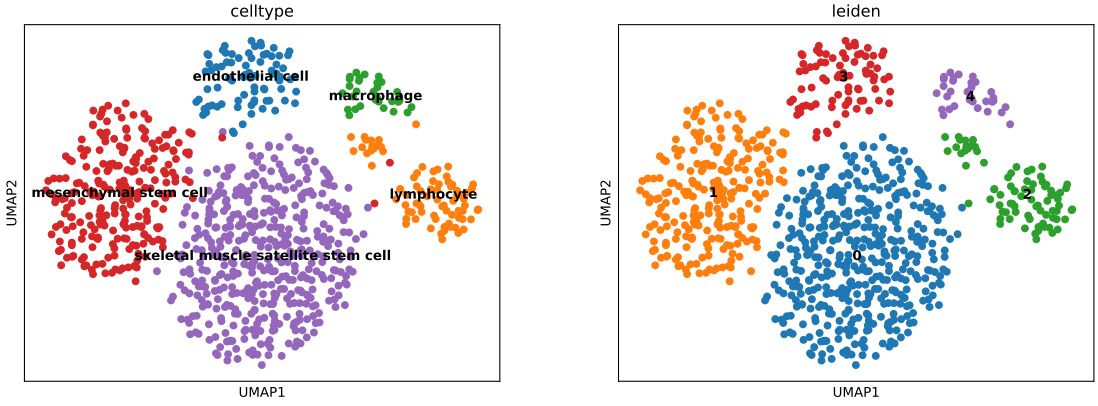
基准自编码器

自编码器 + 对比学习

图 **A-1**　Turtle 数据集 UMAP 可视化结果

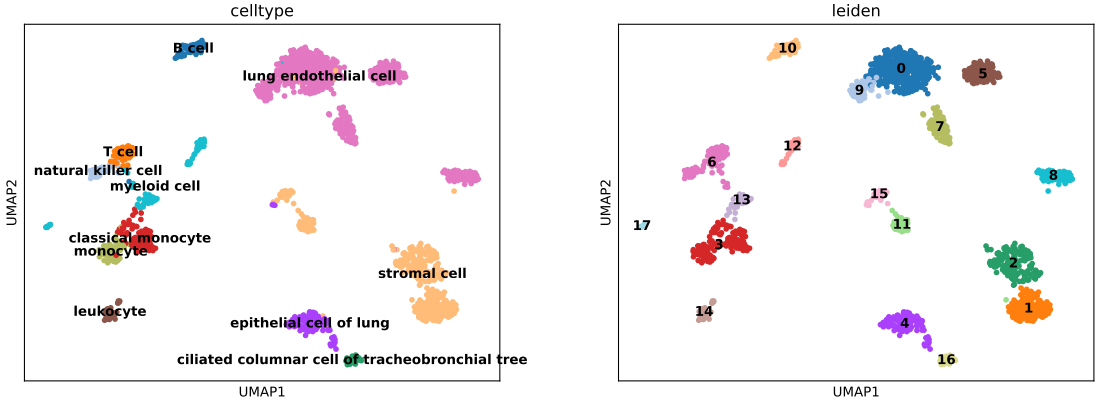


基准自编码器

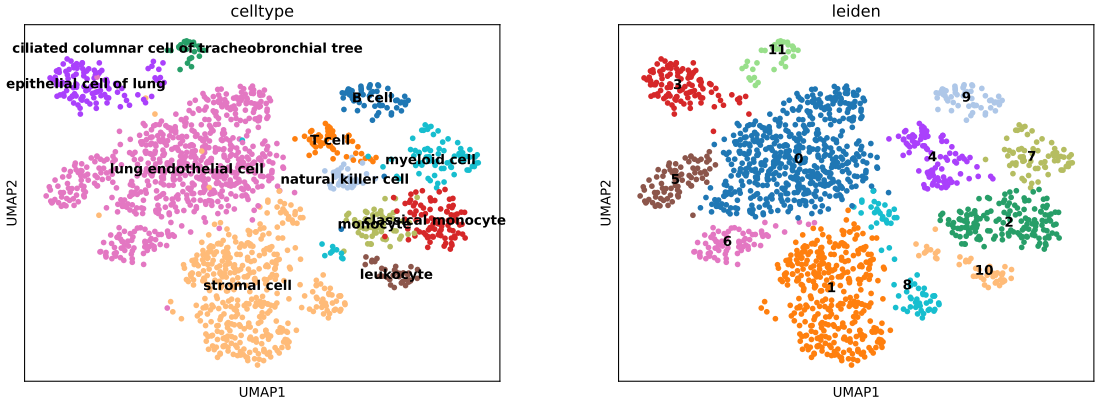


自编码器 + 对比学习

图 A-2　Diaphragm 数据集 UMAP 可视化结果



基准自编码器



自编码器 + 对比学习

图 A-3　Lung 数据集 UMAP 可视化结果