Advances in Applied Mathematics 应用数学进展, 2026, 15(4), 271-285
Published Online April 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.154156

Hans汉斯

基于图正则化非负矩阵分解的scATAC-seq数据分析

张焱杰

青岛大学数学与统计学院，山东 青岛

收稿日期：2026年3月8日；录用日期：2026年4月2日；发布日期：2026年4月9日

摘　要

针对scATAC-seq数据的稀疏性与噪声问题，本研究提出一种融合图正则化与核范数约束的非负矩阵分解模型。该模型通过矩阵分解恢复全局信息，并利用细胞相似性图保持局部流形结构，从而在低维空间中实现生物一致性表示。实验表明，本方法能有效恢复缺失值，并在聚类与可视化中揭示更清晰的细胞亚群，提升下游分析性能。

关键词

数据稀疏性，非负矩阵分解，图拉普拉斯正则化，聚类分析

Analyzing Single-Cell ATAC-seq Data with Graph Regularized Non-Negative Matrix Factorization

Yanjie Zhang

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: March 8, 2026; accepted: April 2, 2026; published: April 9, 2026

Abstract

To address the challenges of sparsity and technical noise in single-cell ATAC sequencing data, this study proposes a non-negative matrix factorization method integrated with graph regularization and nuclear norm constraint. This approach restores global information through low-rank decomposition and preserves the local manifold structure by leveraging a cell similarity graph, thereby

文章引用: 张焱杰. 基于图正则化非负矩阵分解的 scATAC-seq 数据分析[J]. 应用数学进展, 2026, 15(4): 271-285.
DOI: 10.12677/aam.2026.154156

achieving biologically consistent representations in the low-dimensional space. Experiments demonstrate that our method effectively recovers missing values, reveals clearer cell subpopulations in clustering and visualization tasks, and enhances the performance of downstream scATAC-seq data analysis.

Keywords

Data Sparsity, Non-Negative Matrix Factorization, Graph Laplacian Regularization, Cluster Analysis

1. 引言

染色质的开放与闭合状态，即染色质可及性，调控着基因表达的时空特异性，是决定细胞功能的核心机制。在全基因组范围内绘制染色质开放图谱，对于解析细胞命运、发育过程及疾病发生具有重要意义。传统 DNase-seq [1]等技术依赖大量细胞，难以应用于稀有样本。而 ATAC-seq [2]技术利用 Tn5 转座酶直接标记开放染色质区域，以高效率、低样本量的优势实现了全基因组可及性检测，并能同时推断核小体位置[3]与转录因子[4]结合信息，现已成为该领域的主流方法。

单细胞 ATAC-seq (scATAC-seq) [5]结合单细胞条形码技术，实现对成千至数万个细胞的染色质可及性测量，相比群体水平实验能够揭示细胞异质性、推断分化轨迹并发现稀有细胞群体。然而，scATAC-seq 数据高维、极度稀疏，细胞–峰矩阵中仅约 1%~5%的条目为非零，使下游分析任务如细胞聚类(scABC [6])、调控特定细胞状态的转录因子鉴定(chromVAR [7])以及顺式调控共可及区域的预测(Cicero [8])面临巨大挑战。为提升数据完整性和分析精度，数据插补方法被提出，通过区分技术零值与生物零值，对未被检测到的真实信号进行概率性恢复，从而改善聚类、差异峰分析和调控网络推断的可靠性。

为解决上述问题，本研究提出了一种基于图正则化非负矩阵分解的 scATAC-seq 数据插补方法。该方法的核心在于将稀疏的细胞–峰矩阵分解为低维特征矩阵与细胞系数矩阵，以揭示其内在的低维结构。非负矩阵分解的“加性”约束有利于学习具有生物学可解释性的部分特征表示。通过引入图正则化项，模型能够同时捕捉数据的局部细胞相似性，从而实现对缺失值的更合理估计。在实现上，我们采用膝点检测法[9]自适应确定分解秩，以权衡模型复杂度与拟合精度，并利用循环坐标下降算法[10]进行高效优化，确保了方法的可扩展性与稳定性。本研究旨在为 scATAC-seq 数据分析提供一个更加强大、可靠的工具，以助力于更精细的基因调控机制解析。

2. 预备知识

2.1. 非负矩阵分解

定义 2.1：设一个数据矩阵：

$$\boldsymbol{X}=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\right]\in\mathbb{R}^{M\times N},$$

其中每一列 $\boldsymbol{x}_j$ 是一个样本向量。非负矩阵分解的目标是找到两个非负矩阵：

$$\boldsymbol{W}\in\mathbb{R}_+^{M\times K},\boldsymbol{H}\in\mathbb{R}_+^{N\times K},$$

使他们的乘积可以近似原始矩阵 $\boldsymbol{X}$ ：

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}.$$

非负矩阵分解是一种将非负矩阵分解为低维非负因子矩阵的方法，主要用于提取数据的潜在结构。这里$\boldsymbol{W}$的列向量可以看作基向量，捕捉数据的基本结构或特征模式。$\boldsymbol{H}$的行向量可以看作每个样本在这些基向量上的系数表示。因此，非负矩阵分解方法相当于将每个样本向量表示为基向量的加权线性组合，即：

$$\boldsymbol{x}_j \approx \sum_{k=1}^{K} \boldsymbol{w}_k h_{jk},$$

其中$\boldsymbol{w}_k$是$\boldsymbol{W}$的第k列，h_{jk}是对应的系数。

在非负矩阵分解中，通常用两个矩阵之间欧式距离的平方(即矩阵差的 Frobenius 范数的平方)来衡量近似程度，其目标函数定义如下：

$$O_1 = \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} \right\|^2 = \sum_{i,j} \left(x_{ij} - \sum_{k=1}^{K} w_{ik} h_{jk} \right)^2. \tag{1}$$

由于该目标函数对$\boldsymbol{W}$和$\boldsymbol{H}$单独是凸函数，但对两者的联合变量是非凸的，因此通常采用如下迭代更新算法求解局部最优解：

$$w_{ik} \leftarrow w_{ik} \frac{\left(\boldsymbol{X}\boldsymbol{H}\right)_{ik}}{\left(\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)_{ik}}, h_{\mathrm{jk}} \leftarrow h_{ik} \frac{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\right)_{jk}}{\left(\boldsymbol{H}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\right)_{jk}}. \tag{2}$$

该算法在每次迭代后保持$\boldsymbol{W}$，$\boldsymbol{H}$的非负性，并保证收敛至局部最优解。

2.2. 图拉普拉斯矩阵

在高维数据分析与流形学习中，数据常被认为分布在一个潜在的低维流形上。为捕捉并利用这一结构的局部几何特性，通常首先将数据点表示为图的节点，并根据其相似性构建邻接图。在此基础上，图拉普拉斯矩阵作为该图的数学表征，能够有效地将“局部平滑性”这一几何先验转化为可计算的约束项，从而在降维或表示学习过程中保持邻近样本在低维空间中的相似性。其构建过程如下：

1) 构建邻接图$G = (V, E)$：

给定包含N个样本(细胞)点的数据集$\{x_1, x_2, \cdots, x_N\}$，其中每个样本点对应图中的一个节点。首先，通过$k$-近邻法建立邻接关系，即对于每一个样本$x_j$，找到其$k$个最相似的样本点，并在图中对应的节点间建立连接。

2) 定义权重矩阵$\boldsymbol{S}$：

邻接图构建完成后，为每条边赋予权重，以量化样本间的相似性，形成权重矩阵$\boldsymbol{S} \in \mathbb{R}^{N \times N}$。常用的权重$\boldsymbol{S} = \left[\boldsymbol{S}_{jl} \right]$定义方式包括：

- 0~1 权重矩阵：若节点j与l相连，则$\boldsymbol{S}_{jl} = 1$；否则$\boldsymbol{S}_{jl} = 0$。
- 热核(Heat Kernel)权重：若节点j与l相连，则

$$\boldsymbol{S}_{jl} = \exp\left(\frac{-\left\| x_j - x_l \right\|^2}{\sigma} \right),$$

其中$\sigma > 0$为尺度参数。热核权重与流形上的 Laplace - Beltrami 算子具有内在联系，能够较好地捕捉局部几何结构。

- 点积(Dot-Product)权重：若节点j与l相连，则

$$\boldsymbol{S}_{jl} = \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{x}_l,$$

若样本向量已归一化为单位长度，则点积等价于余弦相似度。点积权重在文本或信息检索领域非常常用，而对于图像数据，热核权重可能更适合。

3) 构建图拉普拉斯矩阵 $\boldsymbol{L}$：

首先，根据权重矩阵 $\boldsymbol{S}$ 定义度矩阵 $\boldsymbol{D}\in\mathbb{R}^{N\times N}$，它是一个对角矩阵，其对角元素 $\boldsymbol{D}_{ii}=\sum_j \boldsymbol{S}_{ij}$ 表示与节点 i 相连的所有边的权重之和。

图拉普拉斯矩阵 $\boldsymbol{L}$ 可定义为：

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}.$$

在实际应用中，对称标准化拉普拉斯矩阵 $\boldsymbol{L}_{\text{sym}}$ 更为常用，其定义为：

$$\boldsymbol{L}_{sym}=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{\frac{1}{2}}.$$

该形式具有更好的谱性质，并能确保缩放不变性。

2.3. 核范数

在矩阵分析与优化中，核范数是一种常用的矩阵范数，用于度量矩阵的“秩”特性。它在低秩矩阵分解、矩阵补全、信号恢复以及特征提取等领域中具有广泛应用。核范数的引入主要是为了在保持优化问题凸性的同时，对矩阵的秩进行有效约束。

定义 2.2：设矩阵 $\boldsymbol{X}\in\mathbb{R}^{M\times N}$，其奇异值分解(SVD)为：

$$\boldsymbol{X}=\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^{\mathrm{T}},$$

其中，$\boldsymbol{P}\in\mathbb{R}^{M\times M}$ 和 $\boldsymbol{Q}\in\mathbb{R}^{N\times N}$ 是正交矩阵，$\boldsymbol{\Sigma}=diag\left(\sigma_1,\sigma_2,\cdots,\sigma_r\right)$ 为奇异值对角矩阵，且满足 $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r\geq 0$，则矩阵 $\boldsymbol{X}$ 的核范数定义为其所有奇异值的和：

$$\left\|\boldsymbol{X}\right\|_*=\sum_{i=1}^{r}\sigma_i.$$

矩阵的秩是表示其有效维度的重要指标，在矩阵分解与恢复问题中，我们希望寻找一个低秩近似以捕捉数据的主要结构，但直接以矩阵秩作为约束会导致优化问题成为非凸且在计算上是 NP 难(NP-hard)的问题。因此，核范数常被视为矩阵秩的最紧凸近似，即：

$$\operatorname{rank}\left(\boldsymbol{X}\right)\approx\left\|\boldsymbol{X}\right\|_*.$$

因此，优化问题 $\min\limits_{\boldsymbol{X}}\operatorname{rank}\left(\boldsymbol{X}\right)$ 通常会替换为其凸替代的形式：

$$\min_{\boldsymbol{x}}\left\|\boldsymbol{X}\right\|_*.$$

这种替代的方法在显性地保持矩阵分解低秩特性的同时，使得问题变得可求解。

性质 2.1：对任意矩阵 $\boldsymbol{X}\in\mathbb{R}^{\mathrm{M}\times N}$，其核范数可以表示为：

$$\left\|\boldsymbol{X}\right\|_*=\min_{\boldsymbol{X}=\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}}\frac{1}{2}\left(\left\|\boldsymbol{W}\right\|^2+\left\|\boldsymbol{H}\right\|^2\right),\tag{3}$$

其中最小值在 $\boldsymbol{W}=\boldsymbol{P}\boldsymbol{\Sigma}^{(1/2)}$，$\boldsymbol{H}=\boldsymbol{Q}\boldsymbol{\Sigma}^{(1/2)}$ 时取得，$\boldsymbol{X}=\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^{\mathrm{T}}$ 是矩阵的奇异值分解。

3. 方法模型

本章系统阐述针对 scATAC-seq 数据极端稀疏性问题所提出的图正则化非负矩阵分解模型。模型整体分为数据预处理、图拉普拉斯矩阵构建、模型构建与优化求解等环节。其核心在于将核范数约束与图拉普拉斯正则化同时引入非负矩阵分解框架(如图 1)中，从而兼顾数据的全局低秩结构恢复与局部流形保

持，实现对单细胞染色质开放信号的去噪与插补。

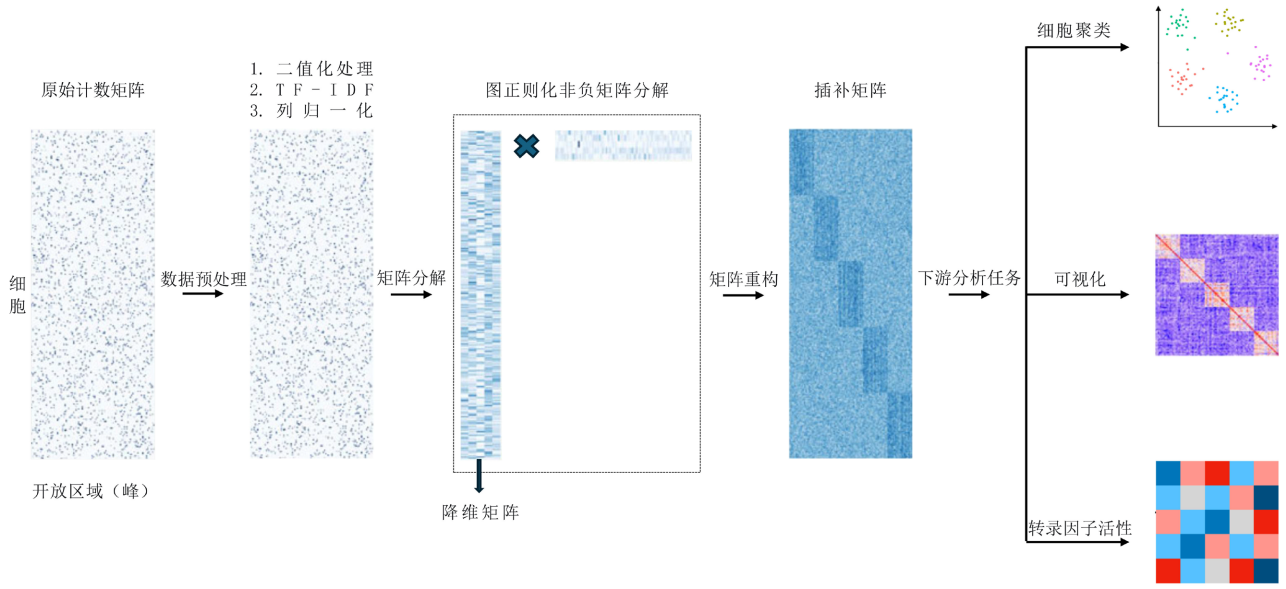

Figure 1. Graph regularized non-negative matrix factorization model framework
图 1. 图正则化非负矩阵分解模型框架

3.1. 数据预处理

scATAC-seq 原始数据经质控筛选后，可表示为峰–细胞计数矩阵 $\boldsymbol{X}_{raw} \in \mathbb{R}^{M\times N}$，其中 M 表示峰值的数量，N 表示单细胞数量，矩阵元素 x_{ij}^{raw} 表示第 i 个峰在第 j 个细胞中的测序片段计数。由于该矩阵通常存在文库大小差异、技术噪声及极端稀疏性(>95%的零值)，直接建模易引入偏差。为此，我们设计了以下三步预处理流程，旨在保留真实生物学信号的同时抑制技术变异。

1) 二值化处理

scATAC-seq 数据的本质是反映染色质区域开放与否，而非开放强度。因此，我们首先对计数矩阵进行二值化处理：

$$x'_{ij} = \begin{cases} 1, & x_{ij}^{raw} > 0 \\ 0, & x_{ij}^{raw} = 0 \end{cases},$$

得到二值矩阵 $\boldsymbol{X}' \in \{0,1\}^{M\times N}$。该转化使得每一个元素仅反应染色质是否处于开放状态，从而消除测序深度差异带来的影响，强调了染色质的“开/关”本质特征。

2) TF-IDF 变换

为进一步校正细胞间测序深度差异并增强特异性峰值的鉴别力，我们借鉴信息检索的思想，对二值矩阵 $\boldsymbol{X}'$ 应用词频–逆文档频率(TF-IDF)变换：

$$x''_{ij} = \frac{x'_{ij}}{\sum_{p=1}^{m} x'_{pj}} \times \log\left(\frac{n}{\sum_{q=1}^{n} x'_{iq}}\right),$$

其中，第一部分(TF)为细胞内峰开放频率，实现细胞层面的归一化，以校正不同细胞的测序深度；第二部分(IDF)抑制了在多数细胞中普遍开放的非特异性峰值(如启动子区域)，同时提升了细胞特异性峰值的相对权重，从而强化了细胞亚群间的差异信号。

3) 列归一化

最后，我们对矩阵 $\boldsymbol{X}''$ 的每一列(对应单个细胞)进行 L2 范数归一化

$$x_{ij} = \frac{x''_{ij}}{\sqrt{\sum_{k=1}^{m}\left(x''_{kj}\right)^2}},$$

得到最终输入矩阵 $\boldsymbol{X} \in \mathbb{R}^{M\times N}$。该步骤将所有细胞的特征向量映射到高维的单位球面上，使细胞间的余弦相似度计算更加稳定合理，减少因尺度差异造成的伪差异。

3.2. 图拉普拉斯矩阵构建

为了在低维嵌入空间中保持细胞间的局部流形结构，本模型引入了图拉普拉斯正则化项。该正则化项的核心思想是：若两个细胞在原始特征空间中具有高度的相似性，则它们在降维后的低维表示中也应彼此接近。

图拉普拉斯矩阵的构建直接在预处理后的稀疏矩阵上进行，本文采用的预处理流程为图构建奠定了坚实基础。首先，二值化处理使点积能够衡量细胞间共同开放的峰数量，TF-IDF 变换通过差异化加权增强了细胞特异性信号的贡献，而 L2 范数归一化则使点积等价于余弦相似度，稳定地反映了细胞开放谱的方向一致性。针对高维稀疏空间直接建图可能遭遇的“维度灾难”，本文采用余弦相似度，它关注向量方向的一致性而非绝对距离，在高维空间中仍能有效区分细胞的开放模式，保证了所建图的可靠性。

其次，0~1 权重仅表示连接有无，丢失相似度的连续信息，在图正则化中无法区分“高度相似”与“边缘相似”的细胞对；而热核权重需调节尺度参数 σ，在高维稀疏空间中距离分布集中，σ 难以客观确定。点积权重无需额外超参数，且更符合生物学实际。

基于此相似性度量，且需满足图拉普拉斯矩阵满足半正定性。本文采用对称 k-近邻法策略构建权重矩阵 S_{raw}。即对于任意两个细胞 i 和 j，只要满足 j 是 i 的 k 近邻或 i 是 j 的 k 近邻，则在二者之间建立一条无向边，边的权重即为两点间的点积相似度。实际操作中，我们首先为每个细胞独立保留其 k 近邻的相似度，再通过对称化操作

$$S = \frac{S_{raw} + S_{raw}^{T}}{2}$$

获得最终的对称权重矩阵。

基于对称权重矩阵 S，我们构建度矩阵 $D = diag\left(\sum_j S_{ij}\right)$。在本文的图正则化非负矩阵分解模型中，我们采用对称标准化拉普拉斯矩阵，该形式的拉普拉斯矩阵具有半正定性，有利于算法的收敛稳定性。

3.3. 模型构建

将预处理后的矩阵进行图正则化非负矩阵分解建模。该模型假设染色质开放模式可由低秩结构近似，并通过引入核范数约束与图拉普拉斯正则化项，同时实现全局低秩恢复与局部流形保持。

首先我们定义目标函数，令

$$\boldsymbol{X} \approx \hat{\boldsymbol{M}} = \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}.$$

其中 $\boldsymbol{W} \in \mathbb{R}^{M\times K}$ 为峰基矩阵，$\boldsymbol{H} \in \mathbb{R}^{N\times K}$ 为细胞低维表示矩阵，$\mathrm{K} \ll \min(M,N)$ 为矩阵分解的秩，模型的目标函数可定义如下：

$$\hat{\boldsymbol{M}} = \underset{M\geq 0}{\operatorname{argmin}} \sum_{i,j}\left(m_{ij} - x_{ij}\right)^2 + \lambda_1 \left\|\boldsymbol{M}\right\|_* + \lambda_2 tr\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right). \tag{4}$$

其中，$\left\|\boldsymbol{M}\right\|_* = \sum_i \sigma_i(\boldsymbol{M})$ 为核范数，用来约束矩阵 $\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}$ 的低秩性，以恢复全局结构并防止过拟合；$\boldsymbol{L}$ 为

对称标准化图拉普拉斯矩阵，其作用是在低维空间上施加平滑性约束，保证相似细胞在低维空间中具有相近的表示。λ_1 和 λ_2 分别为核范数项和图正则化项的平衡参数。

直接优化核范数项 $\|\boldsymbol{M}\|_*$ 涉及对矩阵乘积的奇异值分解，计算代价较高。根据性质 2.1，核范数具有如公式(3)的变分形式，将其带入公式(4)中，可得目标函数的等价形式：

$$\min_{\boldsymbol{W},\boldsymbol{H}} f(\boldsymbol{W},\boldsymbol{H}) = \sum\left(\left(\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\right)_{ij} - x_{ij}\right)^2 + \frac{\lambda_1}{2}\left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{H}\|_F^2\right) + \lambda_2 tr\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right). \tag{5}$$

该转化将隐式的核范数约束显式地表示为 $\boldsymbol{W}$ 与 $\boldsymbol{H}$ 的 Frobenius 范数正则化项，显著降低了计算复杂度。

在上述矩阵分解的框架下，我们希望通过优化目标函数 $f(\boldsymbol{W},\boldsymbol{H})$ 来得到最优的低秩分解矩阵 $\boldsymbol{W}$ 与 $\boldsymbol{H}$。由于目标函数关于 $\boldsymbol{W}$ 与 $\boldsymbol{H}$ 的联合变量是非凸的，因此直接求解全局最优解是困难的。然而，该目标函数具有分块凸性。即当固定其中一个矩阵时，目标函数关于另一个矩阵的优化问题是凸的。这种分块变量凸性为交替优化策略提供了理论基础，即通过将原问题分解为两个凸子问题并迭代求解，可有效逼近稳定解。

3.4. 基于循环坐标下降的优化算法

基于上述目标函数分块凸性，本文采用循环坐标下降法对 $\boldsymbol{W}$ 与 $\boldsymbol{H}$ 进行交替优化。循环坐标下降法的核心思想是：在固定其他所有变量的前提下，每次仅优化单个矩阵元素，将高维非负约束优化问题分解为一系列一元凸子问题。该方法具有以下优势：每次更新中，通过 $\max(0,\cdot)$ 直接投影，保证其非负性；每个元素更新都是凸优化问题，因此每个元素更新均可通过闭式解直接计算；每次更新仅涉及局部行列，计算复杂度低，适合大规模稀疏 scATAC-seq 数据。

具体地，定义 $\boldsymbol{E}_{it}$ 为仅在 (i,t) 处为 1 的矩阵，则对于 $\boldsymbol{W}$ 中的元素 w_{it}，考虑更新 $w_{it} \leftarrow w_{it} + s$，子问题为：

$$\min_{s:\, w_{it}+s\geq 0} g_{it}^{\boldsymbol{W}}(s) = f(\boldsymbol{W} + s\boldsymbol{E}_{it}, \boldsymbol{H}).$$

将 $g_{it}^{\boldsymbol{W}}(s)$ 在 $s=0$ 处二阶泰勒展开：

$$g_{it}^{\boldsymbol{W}}(s) = g_{it}^{\boldsymbol{W}}(0) + \left(g_{it}^{\boldsymbol{W}}\right)'(0)s + \frac{1}{2}\left(g_{it}^{\boldsymbol{W}}\right)''(0)s^2,$$

其中，一阶偏导和二阶偏导分别为：

$$\left(g_{it}^{\boldsymbol{W}}\right)'(0) = \frac{\partial f}{\partial w_{it}} = 2\sum_j\left(\left(\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\right)_{ij} - x_{ij}\right)h_{jt} + \lambda_1 w_{it},$$

$$\left(g_{it}^{\boldsymbol{W}}\right)''(0) = \frac{\partial^2 f}{\partial w_{it}^2} = 2\sum_j h_{jt}^2 + \lambda_1.$$

可得二次函数的最优无约束步长为 $s^* = -\left(g_{it}^{\boldsymbol{W}}\right)'(0) \Big/ \left(g_{it}^{\boldsymbol{W}}\right)''(0)$，结合非负约束可得

$$s^* = \max\left(-w_{it}, -\frac{\left(g_{it}^{\boldsymbol{W}}\right)'(0)}{\left(g_{it}^{\boldsymbol{W}}\right)''(0)}\right).$$

从而得到 w_{it} 的更新规则：

$$w_{it}^{new} = w_{it} + s^* = \max\left(0, w_{it} - \frac{\left(g_{it}^{\boldsymbol{W}}\right)'(0)}{\left(g_{it}^{\boldsymbol{W}}\right)''(0)}\right) = \max\left(0, \frac{2\sum_j h_{jt}\left(x_{ij} - \sum_{s\neq t} w_{is}h_{js}\right)}{2\sum_j h_{jt}^2 + \lambda_1}\right), \tag{6}$$

类似地，对于 $\boldsymbol{H}$ 中的元素 h_{jt}，定义子问题

$$\frac{\min}{s:h_{jt}+s\geq 0} g_{jt}^{\boldsymbol{H}}(s)=f\left(\boldsymbol{W},\boldsymbol{H}+s\boldsymbol{E}_{jt}\right).$$

其中，一阶偏导和二阶偏导分别为：

$$\left(g_{jt}^{\boldsymbol{H}}\right)'(0)=\frac{\partial f}{\partial h_{jt}}=2\sum_i\left(\left(\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\right)_{ij}-x_{ij}\right)w_{it}+\lambda_1 h_{jt}+2\lambda_2\left(\boldsymbol{L}\boldsymbol{H}\right)_{jt},$$

$$\left(g_{jt}^{\boldsymbol{H}}\right)''(0)=\frac{\partial^2 f}{\partial h_{jt}^2}=2\sum_i w_{it}^2+\lambda_1+2\lambda_2 L_{jj},$$

其中 $\left(\boldsymbol{L}\boldsymbol{H}\right)_{jt}=\sum_q L_{jq}h_{qt}$。可得 h_{jt} 更新规则：

$$h_{jt}^{new}=\max\left(0,h_{jt}-\frac{\left(g_{jt}^{\boldsymbol{H}}\right)'(0)}{\left(g_{jt}^{\boldsymbol{H}}\right)''(0)}\right)=\max\left(0,\frac{2\sum_i w_{it}\left(x_{ij}-\sum_{s\neq t}w_{is}h_{js}\right)-2\lambda_2\sum_{q\neq j}L_{jq}h_{qt}}{2\sum_{i=1}^m w_{it}^2+\lambda_1+2\lambda_2 L_{jj}}\right). \tag{7}$$

在坐标下降法中，迭代停止条件基于投影梯度的相对变化量，该条件确保非负约束下 KKT 条件的违反程度被准确量化：若变量已为 0 且梯度为负(即元素可以进一步减小)，则投影梯度为 0；否则投影梯度等于梯度本身。

每次迭代计算所有变量的投影梯度绝对值之和，记为 $V^{(k)}$，设初始迭代的 $V^{(0)}$ 为基准，矩阵分解模型迭代停止条件为：

$$\frac{V^{(k)}}{V^{(0)}}\leq \mathrm{tol}.$$

其中 tol (本模型 $\mathrm{tol}=10^{-4}$)是预设的容忍度。该条件表明，第 k 步解的一阶最优性违反程度已降至初始值的万分之一以下，即优化过程已充分收敛。若 $V^{(0)}=0$，则初始点已满足最优性条件，算法直接终止。此外，考虑到计算资源，当迭代次数大于 T (本模型 $T=500$)时，停止迭代。

Table 1. Algorithmic framework for graph regularized non-negative matrix factorization model
表 1. 图正则化非负矩阵分解模型算法框架

算法 1：图正则化非负矩阵分解模型算法框架
参数限定：λ_1，λ_2，K，k，T，$K_{\max}$，$K_{\min}$，tol。
目标：找出最优秩 K 及对应的非负矩阵分解 $\boldsymbol{W}$ 和 $\boldsymbol{H}$，并生成插补矩阵 $\hat{\boldsymbol{M}}=\boldsymbol{W}\boldsymbol{H}^T$。
输入：$\boldsymbol{L}$，$\boldsymbol{X}_{\mathrm{raw}}$。
步骤一、数据预处理：0~1 二值变换，TF-IDF 变换，列归一化变换。
步骤二、秩估计与分解，对于每个 $K\in(K_{\min},K_{\max})$ 并行执行：
1. 初始化：使用 NNDSVD 方法生成 $\boldsymbol{W}$，$\boldsymbol{H}$。
2. 坐标下降迭代，从 $t=0$ 到 $t=T-1$，执行：
(a) 更新 $\boldsymbol{W}$：按照公式(6)更新元素 w_{it}；
(b) 更新 $\boldsymbol{H}$：按照公式(7)更新元素 h_{jt}；
(c) 若 $\frac{V^{(k)}}{V^{(0)}}\leq \mathrm{tol}$，迭代则提前结束。
3. 计算重构误差 $e_K\leq\left\|\boldsymbol{X}-\boldsymbol{W}_K\boldsymbol{H}_K\right\|^2$，并存储 $\boldsymbol{W}_K$，$\boldsymbol{H}_K$。
步骤三、最优秩选择：
使用肘部法则在 (K,e_K) 曲线上确定拐点 $K=k^*$，提取 $\boldsymbol{W}=\boldsymbol{W}_K$，$\boldsymbol{H}=\boldsymbol{H}_K$。
步骤四、返回 $\boldsymbol{W}$，$\boldsymbol{H}$ 及插补矩阵 $\hat{\boldsymbol{M}}=\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}$。

本模型涉及四个主要超参数：矩阵分解秩 K 正则化参数 λ_1 和 λ_2，以及构建图拉普拉斯矩阵 L 时的近邻数 k。其中，秩 K 决定了潜在空间的维度，直接影响数据的低维表示能力；λ_1 控制矩阵分解的稀疏性与模型复杂度；λ_2 则用于平衡局部图结构保持项，以维持相似细胞在潜空间中的流形一致性；而 k 决定图结构的局部连通性强度。在超参数选择上，我们采用膝点检测法对 K 进行自动筛选。具体地，我们在预设区间内计算不同秩下的残差平方和，并通过膝点检测法检测确定误差下降趋缓的拐点，作为最优的秩选择结果。该方法能够有效平衡模型复杂度与拟合精度，从而自动确定合适的潜在维度。其余超参数我们分别设置：核范数正则化系数 $\lambda_1=1$，图正则化系数 $\lambda_2=0.5$，图近邻数 $k=30$。

基于上述迭代规则，可以把本模型的算法归纳如表 1。

3.5. 模型性能评估指标

为系统评估所提出模型在 scATAC-seq 数据上的插补效果，我们从细胞缺失信息的恢复能力、聚类紧凑性及聚类一致性三个方面进行了综合分析。

3.5.1. 真实信号恢复能力评估

首先，我们检验模型插补后的矩阵是否能够有效恢复单细胞层面的真实染色质开放信号。我们进行了如下分析：首先，将 scATAC-seq 数据按细胞类型进行聚合后进行峰调用(peak calling)，以识别各细胞类型特异的开放染色质区域。这些区域被定义为该细胞类型的“真实开放区域”。随后，在每个单细胞中，我们将插补得到的染色质可及性分数与对应细胞类型的真实开放区域标签进行比较。具体而言，我们将每个基因组区域归类为真阳性(TP)、假阳性(FP)、真阴性(TN)和假阴性(FN)。基于这些统计量，我们计算每个细胞的精确率–召回率曲线下面积(AUPR)，以衡量模型在单细胞水平上恢复真实可及性信号的能力。

3.5.2. 聚类紧致性评估

其次，为评估插补结果在细胞聚类中的区分度与聚合度，我们计算了所有细胞的平均轮廓系数(Silhouette Score)。对于任意一个细胞 x，其轮廓系数定义为：

$$\text{silhouette}(x)=\frac{b(x)-a(x)}{\max\left(a(x),b(x)\right)},$$

其中，$a(x)$ 表示细胞 x 与同一聚类中其他细胞之间的平均距离，$b(x)$ 表示细胞 x 与最近的不同聚类中所有细胞之间的平均距离。

3.5.3. 聚类一致性评估

为了进一步验证插补矩阵对细胞聚类结构的改进效果，我们对插补矩阵进行 T-SNE 方法进行降维，在此基础上，分别使用 k-中心点聚类(k-medoids)和层次聚类(hierarchical clustering)方法对细胞进行聚类。

聚类准确性指标采用调整兰德指数(Adjusted Rand Index, ARI)进行衡量。假设对于包含 n 个细胞的数据集 D，其真实标签划分为 $U=\{U_1,U_2,\cdots,U_r\}$，模型预测的聚类结果为 $V=\{V_1,V_2,\cdots,V_s\}$，定义每个聚类交集的细胞数为：

$$c_{ij}=\left|U_i\cap V_j\right|,$$

其中 $i\in\{1,2,\cdots,r\}$, $j\in\{1,2,\cdots,s\}$。

则 ARI 的计算公式为：

$$\text{ARI}=\frac{\sum_{ij}\binom{c_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}},$$

其中 $a_i = \sum_{j=1}^{s} c_{ij}, b_j = \sum_{i=1}^{r} c_{ij}$ 。

4. 实验与结果

4.1. 数据集与预处理

本实验使用来源于 GEO 数据库(编号 GSE107816)人类 T 细胞 scATAC-seq 数据集，包括四种类型的 T 细胞，分别为 Jurkat T 细胞、memory T 细胞、naive T 细胞以及 Th17 T 细胞。

针对原始测序数据，我们进行了严格的数据处理。首先，我们对数据进行接头序列修剪和低质量端去除。接着，比对参考人类基因组(hg19)，采用 Bowtie2 排除不匹配配对，同时去除 PCR 重复片段，仅保留比对质量高且配对正确的读段，并过滤掉线粒体、Y 染色体及未装配序列的读段。

为保证数据质量，本研究仅保留具有至少 500 个唯一片段的单细胞。随后，基于保留的 scATAC-seq 信号构建文库，并使用 MACS2 进行峰识别，将峰值区域从峰中心向两侧各延伸 250 bp，同时剔除与 ENCODE 黑名单区域重叠的峰值。最终我们保留了 765 个细胞以及 49,344 个染色质区域。最后将每个细胞在各峰值区间内的有效片段计数整合为峰值 - 细胞计数矩阵。

4.2. 插补结果评价

为直观评估不同插补方法对单细胞染色质可及性数据结构的还原效果，本文采用均匀流形近似与投影(UMAP)对原始计数矩阵及本模型得到的插补矩阵进行降维可视化，结果如图 2 所示。

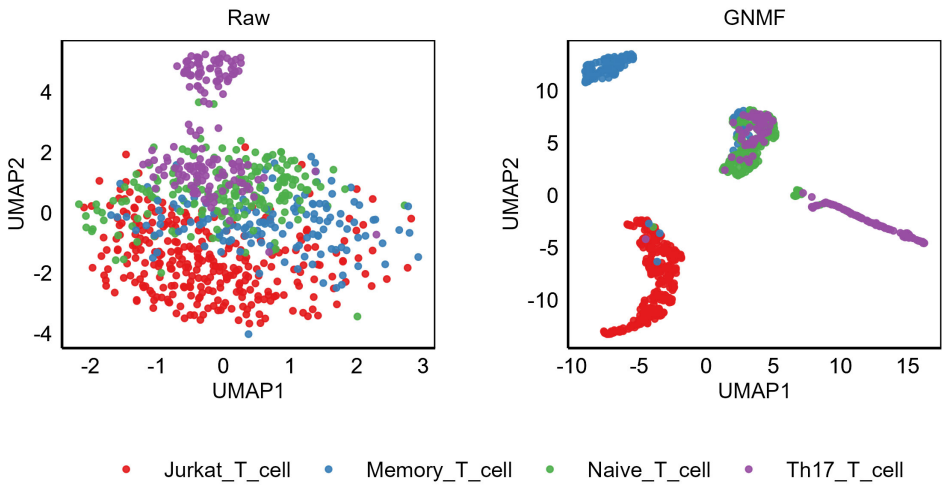

Figure 2. Analysis of UMAP visualizations before and after imputation
图 2. 插补前后 UMAP 可视化分析

可视化结果表明，基于图正则化非负矩阵分解的插补结果对于细胞聚类结构的区分度有明显提升。为了定量评估所提出模型的表现，我们使用第三章中描述的评价指标对插补矩阵进行分析。

首先，为评估不同插补方法对 scATAC-seq 数据中真实生物学信号的恢复能力，本文采用精确率 - 召回率曲线下面积(AUPR)作为评价指标。图 3 展示了原始矩阵以及各方法在所有细胞上的 AUPR 分布，其中黑色横线和红色横线分别表示中位数与均值，方法按平均 AUPR 降序排列。

实验结果表明，本文方法在平均 AUPR 与中位数 AUPR 上均显著优于所有对比方法，说明方法能够有效恢复 scATAC-seq 数据中缺失的真实生物学信息，在保留稀有开放事件的同时抑制技术噪声，展现出卓越的信号复原能力。

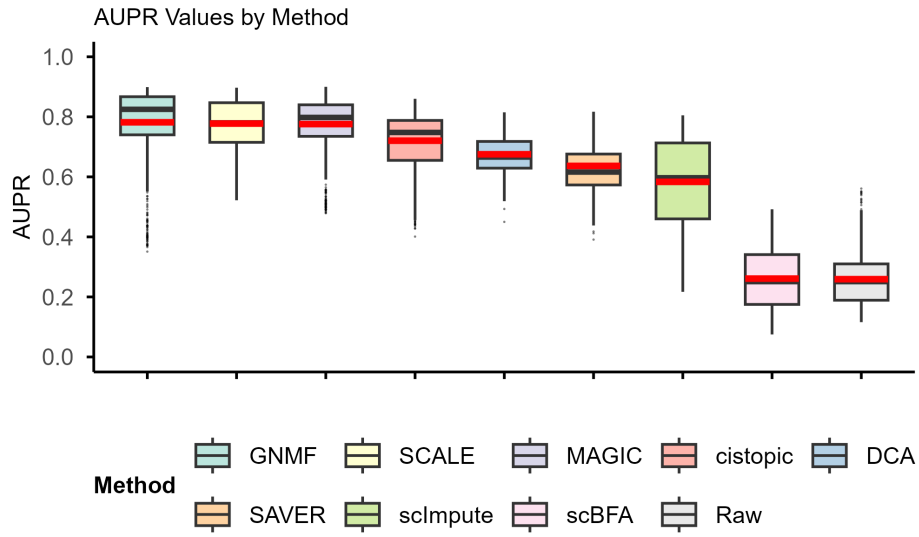

Figure 3. Comparative analysis of AUPR values across different methods
图 3. 不同方法的 AUPR 指标对比分析

其次，为量化评估不同插补方法对细胞聚类性能的影响，本文进一步引入轮廓系数作为评价指标。图 4 展示了原始矩阵以及 8 种不同插补方法的平均轮廓系数对比结果比较结果。

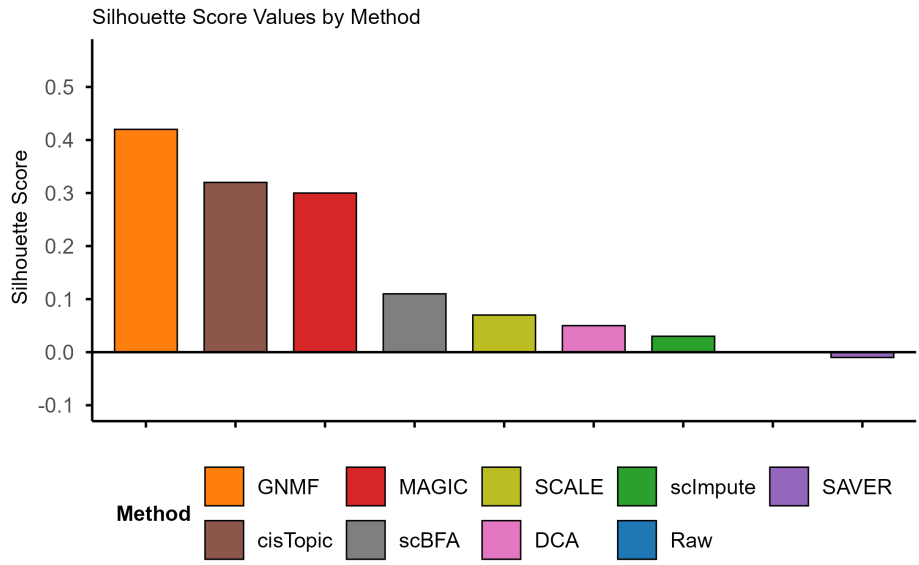

Figure 4. Comparative analysis of average silhouette coefficients among various methods
图 4. 不同方法的平均轮廓系数对比分析

结果表明，本模型在数据处理中展现出显著优势，其平均轮廓系数明显高于其他插补方法，说明本方法能够有效增强同类细胞染色质可及性模式的相似性，提升不同细胞类型之间的可区分性。

最后，为评估不同插补方法对细胞聚类性能的影响，本文分别在原始数据及各插补后数据上进行聚类一致性分析。首先采用 T-SNE 降维方法对数据进行降维，随后分别使用 k-中心点聚类(k-medoids)和层次聚类(hierarchical clustering)对细胞进行聚类。聚类结果如图 5 所示。

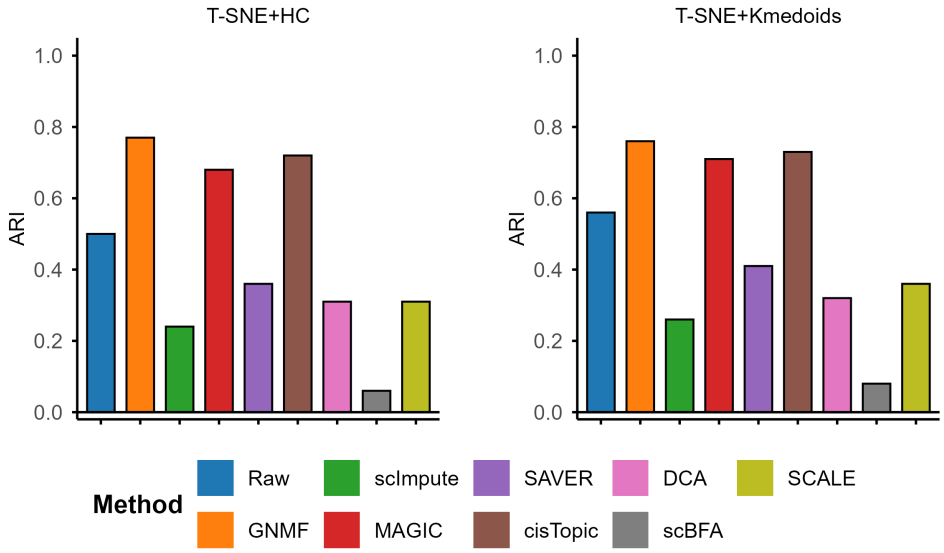

Figure 5. Comparative analysis of ARI values across different methods
图 5. 不同方法的 ARI 指标对比分析

实验结果显示，在两种聚类算法下，本文所提出的图正则化非负矩阵分解插补方法均取得最高的调整兰德指数(ARI)，显著优于原始数据及其他对比方法。这些结果表明，基于非负矩阵分解方法插补方案能够更准确地恢复细胞间的真实相似性结构，从而提升聚类结果与真实细胞类型之间的一致性。

4.3. 下游任务分析

为验证插补数据对 scATAC-seq 下游分析的有效性，本文选取三项关键任务进行对比评估：识别与单细胞相关的调控特征(ChromVAR)、估计基因活性得分和 DNA 相互作用(Cicero)及针对 scATAC-seq 数据专门设计的聚类方法(scABC)。所有任务均在原始矩阵与插补矩阵上分别执行，以比较插补带来的性能提升。其中，chromVAR 方法和 Cicero 方法首先分别将 scATAC-seq 矩阵转换至转录因子和基因特征空间，随后使用各流程的标准方法进行聚类分析。

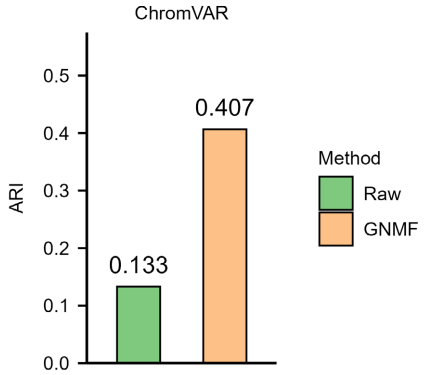

Figure 6. Clustering comparison analysis of the ChromVAR method before and after imputation
图 6. 插补前后 ChromVAR 方法聚类对比分析

首先，基于 ChromVAR 方法分别计算原始数据与插补数据中每个细胞的 motif 可及性偏差得分，进而推断转录因子活性。基于该得分矩阵进行层次聚类，以调整兰德指数(ARI)评估聚类与细胞类型标签的

一致性；同时绘制各细胞类型中关键转录因子的活性分布小提琴图，结果如图 6、图 7 所示。实验表明，相较于原始稀疏矩阵，插补矩阵显著提升了转录因子活性的细胞类型特异性，使亚群间差异更为清晰，说明基于本模型得到的插补矩阵推断的转录因子活性在细胞类型区分度上较原始数据有显著提升，表明插补过程有效恢复了染色质可及性数据中真实的调控信号。

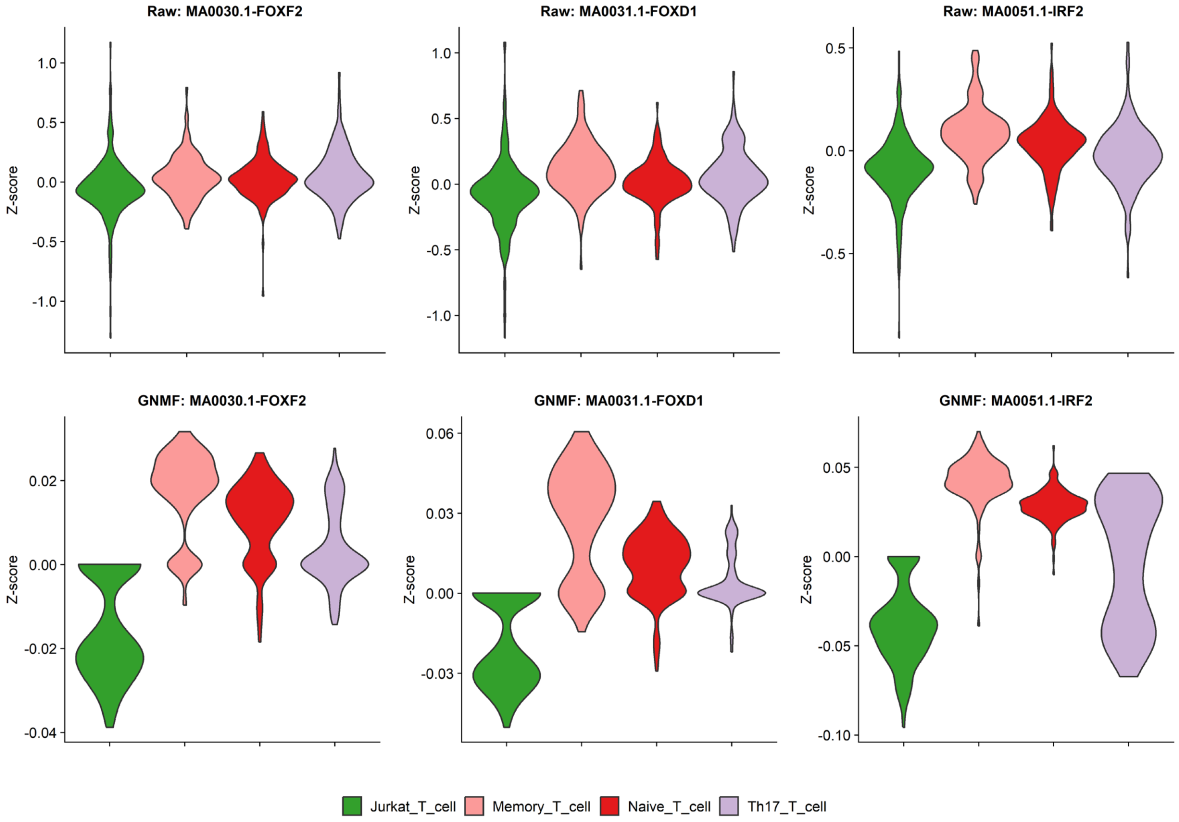

Figure 7. Comparative analysis of transcription factor activity distribution before and after imputation
图 7. 插补前后转录因子的活性分布对比分析

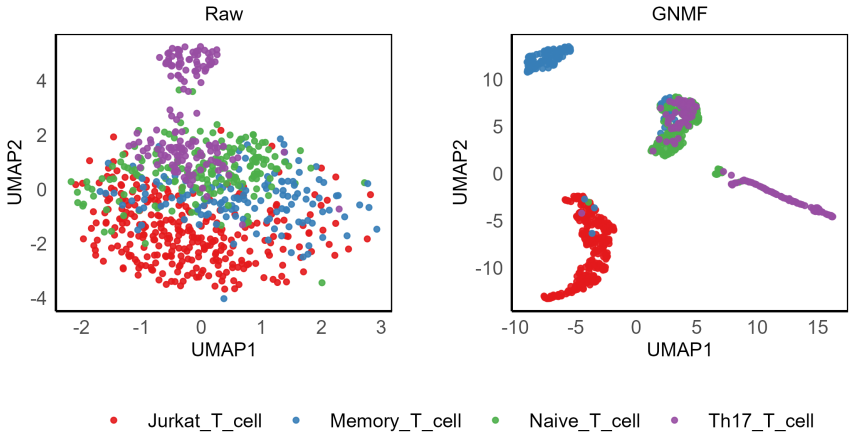

Figure 8. Comparative visualization analysis using the Cicero method before and after imputation
图 8. 插补前后 Cicero 方法可视化对比分析

接下来，我们对原始数据及插补数据进行 Cicero 比对分析。Cicero 方法首先构建细胞类型特异的基因活性矩阵，继而通过 PCA 与 t-SNE 降维并进行无监督聚类。其可视化结果与聚类结果如图 8、图 9 所示。实验结果表明，基于插补矩阵推断的顺式调控网络显著优于原始数据，说明数据插补有效恢复了细胞类型特异的顺式调控架构，增强了基因调控网络推断的可靠性。

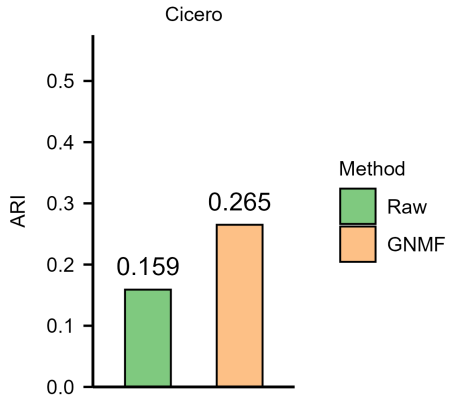

Figure 9. Comparative analysis of clustering using the Cicero method before and after imputation
图 9. 插补前后 Cicero 方法聚类对比分析

进一步探究其在完全无监督的染色质开放模式分析中的应用潜力。为此，我们引入 scABC 算法，该方法专为单细胞 ATAC-seq 数据设计，能够在不依赖任何先验注释信息的条件下，直接基于染色质可及性矩阵进行细胞聚类与差异开放区域识别。通过比较原始数据与插补数据的聚类结果与真实细胞类型标签的一致性，我们揭示了数据插补对恢复染色质开放模式固有结构的作用。对比结果如图 10 所示。结果表明，插补后聚类结果与真实细胞类型标签的一致性有明显的提升。

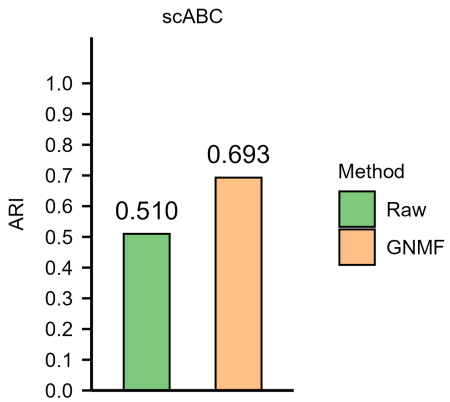

Figure 10. Comparative analysis of clustering with the scABC method before and after imputation
图 10. 插补前后 scABC 方法聚类对比分析

三项下游任务分析结果共同表明：基于图正则化非负矩阵分解模型的插补的 scATAC-seq 数据，能够有效提升数据在调控信号恢复、网络重建及细胞分群等多个维度的分析性能。

5. 总结

本文提出了一种融合图正则化与核范数约束的非负矩阵分解模型，用于解决 scATAC-seq 数据的高

维稀疏性问题。该模型通过核范数约束恢复全局低维结构，同时利用图拉普拉斯正则化保持细胞间的局部流形几何，在低维空间中实现生物学一致的表示。在人类 T 细胞数据集上的实验表明：本方法能有效恢复真实的染色质开放信号，提升细胞聚类紧致性与准确性，并有效地改善了下游任务分析(如 chromVAR、Cicero、scABC)。

参考文献

[1] Song, L. and Crawford, G.E. (2010) DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, **2010**, pdb.prot5384. https://doi.org/10.1101/pdb.prot5384

[2] Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nature Methods*, **10**, 1213-1218. https://doi.org/10.1038/nmeth.2688

[3] Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G. and Greenleaf, W.J. (2015) Structured Nucleosome Fingerprints Enable High-Resolution Mapping of Chromatin Architecture within Regulatory Regions. *Genome Research*, **25**, 1757-1770. https://doi.org/10.1101/gr.192294.115

[4] Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M. and Costa, I.G. (2019) Identification of Transcription Factor Binding Sites Using ATAC-seq. *Genome Biology*, **20**, Article No. 45. https://doi.org/10.1186/s13059-019-1642-2

[5] Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., *et al*. (2015) Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation. *Nature*, **523**, 486-490. https://doi.org/10.1038/nature14590

[6] Zamanighomi, M., Lin, Z., Daley, T., Chen, X., Duren, Z., Schep, A., *et al*. (2018) Unsupervised Clustering and Epigenetic Classification of Single Cells. *Nature Communications*, **9**, Article No. 2410. https://doi.org/10.1038/s41467-018-04629-3

[7] Schep, A.N., Wu, B., Buenrostro, J.D. and Greenleaf, W.J. (2017) chromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data. *Nature Methods*, **14**, 975-978. https://doi.org/10.1038/nmeth.4401

[8] Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., *et al*. (2018) Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, **71**, 858-871.e8. https://doi.org/10.1016/j.molcel.2018.06.044

[9] Satopaa, V., Albrecht, J., Irwin, D. and Raghavan, B. (2011) Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. 2011 31*st International Conference on Distributed Computing Systems Workshops*, Minneapolis, 20-24 June 2011, 166-171. https://doi.org/10.1109/icdcsw.2011.20

[10] Hsieh, C. and Dhillon, I.S. (2011) Fast Coordinate Descent Methods with Variable Selection for Non-Negative Matrix Factorization. *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, 21-24 August 2011, 1064-1072. https://doi.org/10.1145/2020408.2020577