arXiv:2601.22331v1 [cs.LG] 29 Jan 2026

通过批次依赖核与自适应采样实现可扩展的细胞绘画批次校正

Aditya Narayan Ravi [1] **Snehal Vadvalkar** [2] **Abhishek Pandey** [2] **Ilan Shomorony** [1]

摘要

细胞绘画是一种基于显微成像的高内涵成像分析方法，可生成丰富的细胞形态学特征谱，并通过量化细胞对化学扰动的响应来助力药物发现。然而，在大规模应用中，细胞绘画数据会受到显著的批次效应影响，这些效应源于实验室、仪器和实验方案之间的差异，可能掩盖生物学信号。我们提出 BALANS（基于局部亲和性与子采样的批次对齐），这是一种可扩展的批次校正方法，通过根据成对距离构建平滑的亲和矩阵来对齐不同批次的样本。对于 n 个数据点，BALANS 运用两种思路构建稀疏亲和矩阵$A \in \mathbb{R}^{n\times n}$。(i) 对于点 i 和 j，它根据 j 所属批次内从 i 到其第 k 个最近邻的距离设定局部尺度，再利用经这些批次感知局部尺度校准的高斯核计算 A_{ij}。(ii) BALANS 不会构建所有 n^2 个条目，而是采用自适应采样过程，优先处理累积邻居覆盖率较低的行，并且仅保留每行中亲和度最高的条目，从而得到稀疏但信息丰富的 A 近似矩阵。我们证明，这种采样策略在样本复杂度方面是阶数最优的，并提供近似保证；此外，我们还表明，BALANS 的运行时间关于 n 近乎线性。在多种真实世界细胞绘画数据集和受控的大规模合成基准测试上的实验表明，BALANS 可扩展至大型数据集，同时相比广泛使用的批次校正方法的原生实现缩短运行时间，且不牺牲校正质量。

1. 引言

基于图像的表型分析已成为研究细胞如何响应不同处理的有力工具（Regev等，2017年）。借助高通量显微成像，研究人员能够检测由化学或遗传扰动引起的细胞形状和结构变化（Rohban等，2017年）。这使研究人员能够根据化合物产生的细胞效应识别有前景的化合物，从而加速药物发现。细胞绘画检测是一种广泛使用的基于图像的表型分析方法（Bray等，2016年；Caicedo等，2017年）。细胞绘画由Bray等（2016年）最初开发，使用六种荧光染料标记八种不同的细胞组分，并通过四或五个通道对其成像。该方法可提供高内涵、单细胞分辨率的数据，成本效益高（Cimini等，2023年），并能提供与（Cutiongco等，2020年；Wawer等，2014年）以及基于蛋白质的检测互补的信息（Dagher等，2023年）。细胞绘画已成功结合特征提取技术，以生成数值化的形态学特征谱。CellProfiler是一种广泛使用的工具（Carpenter等，2006），它基于人工设计特征提取特征谱，例如从分割后的细胞组分中提取纹理、强度和形状描述符。近年来，人们还提出了DeepProfiler等基于深度学习的方法（Moshkov等，2024），使用EfficientNet等架构（Tan和Le，2019）提取特征谱。

批次效应会掩盖真实的生物学信号：大规模实验通常涉及多个来源（实验室、实验方案、显微镜），会引入批次依赖性伪影，从而掩盖真实的生物学信号。我们关注的信号是扰动所产生的形态学效应，不受技术噪声（例如微孔板或染色伪影）或无关生物变异（例如细胞密度或非特异性浓度效应）的影响。这些混杂因素在数据经过标准预处理后依然存在。Arevalo等人对细胞绘画数据中的批次效应以及现有批次校正技术的性能进行了全面分析Arevalo等人（2024年）。

为说明批次效应，我们使用来自 JUMP-CP联盟的精选阳性对照化合物分析真实的《Cell》细胞绘画数据（Chandrasekaran等人，2023年），

[1]伊利诺伊大学厄巴纳-香槟分校电气与计算机工程系 [2]美国艾伯维。通讯请联系：Aditya Narayan Ravi <anravi2@illinois.edu>。

预印本。*2026*年*2*月*2*日。

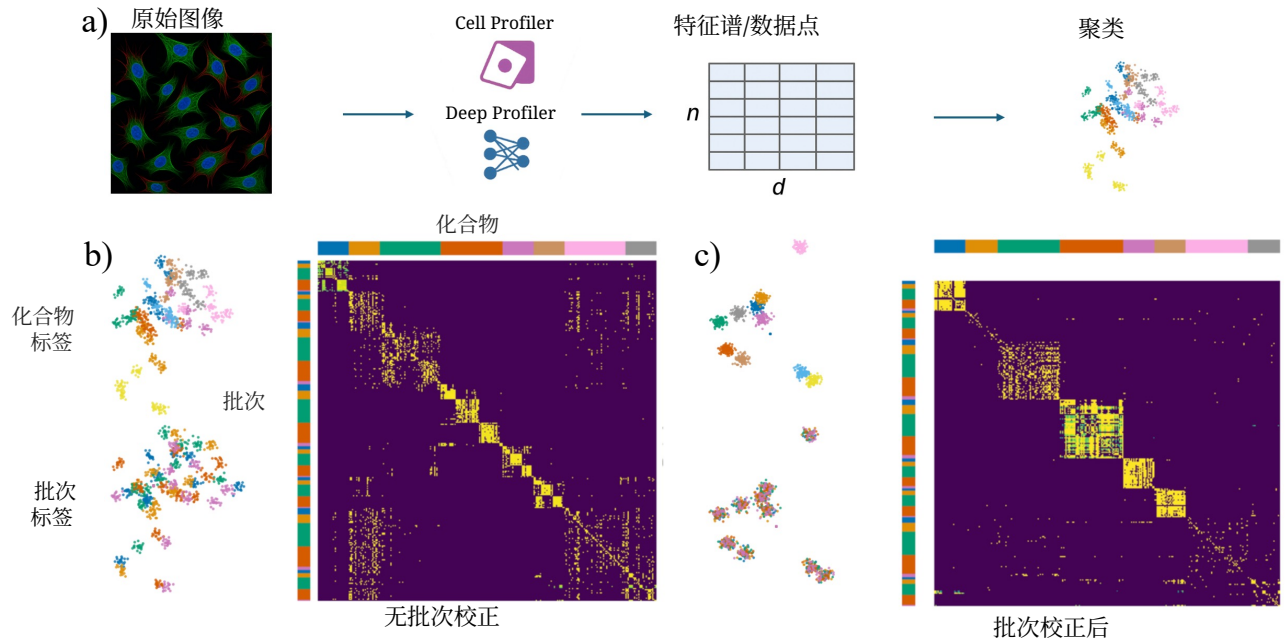


图 *1.* **(a)** 基于为每个 n 个细胞提取特征的标准细胞绘画分析流程。**(b)** 校正前：按化合物标签（上图）为 UMAP 图着色时，可以看到对应同一化合物的多个簇。按批次着色时，可以看到同一化合物对应的不同簇来自不同批次。亲和力热图显示出与批次对齐的块状模式。**(c)** BALANS 处理后：对应不同化合物的簇更加清晰，批次之间相互混合，亲和度与化合物标签相对应。热图的行按批次排序（左侧色条），列按标签排序（顶部色条）。

该数据涵盖四个数据集、13个来源和三个微观范围。这些化合物可诱导多种形态表型，因此非常适合用于可视化批次效应。我们使用通过 CellProfiler 提取的特征谱。所得亲和矩阵在进行任何批次校正之前，以热图形式显示于图1（b）中，呈现出与批次标签一致的块结构（热图左侧的颜色）。批次校正的目标是对细胞特征向量执行变换，使所得亲和矩阵呈现与化合物标签一致的块结构，如图1（c）所示。

基于亲和力的去噪：去噪方法广泛用于通过平滑相似的数据点来增强生物学信号（Van Dijk等人，2018年；Huang等人，2018年；Li和Li，2018年）。这些方法已成功应用于mRNA和蛋白质谱分析（Tran等人，2020）。例如，MAGIC（Van Dijk等人，2018年）构建了精心设计的亲和度矩阵，将每个点与其邻居取平均。我们从这种去噪技术中获得启发，尝试构建批次感知的亲和度矩阵，以用于去除批次效应。一项挑战在于，当数据点数量众多时，计算所有成对亲和度的计算成本可能很高；而细胞绘画实验通常会扩展到涵盖许多不同批次的数十万乃至数百万个单细胞谱（Arevalo等人，2024年）。因此，我们寻求设计一种批次感知的亲和度矩阵，以便以可扩展的方式用于批次校正，从而高效地校正大型数据集中的批次效应。

BALANS：在这项工作中，我们介绍了 BALANS——基于局部亲和性与子采样的批次对齐，这是一种新颖且可扩展的批次校正方法。与传统的基于亲和度的方法不同，BALANS 在构建平滑亲和力矩阵时纳入批次标签信息。该方法仅计算完整亲和矩阵的稀疏版本，从而实现可扩展性；因此，每个数据点都可以根据少量邻居进行平滑处理。

BALANS的第一个关键思想是使“接近程度”这一概念依赖于批次。对于对应于唯一特征谱的任意一对数据点 i 与 j，BALANS通过计算依赖于 j所在批次的局部尺度来调整它们的相似度。更确切地说，若 b_j 是与 j相关联的批次标签，它会计算从 i到其在 j所在批次内第 k 个最近邻的距离，记为$\sigma_{i,j} := d_i^{b_j}{}_{,(k)}$，并将其用作高斯核中的缩放因子。随后，$i$ 与 j 之间的亲和度定义为 $A_{ij} = \exp\left(-\|x_i - x_j\|^2/\sigma_{i,j}^2\right)$ · 这种缩放具有两种效应。它会增强来自较远或聚类较不紧密的批次中相容点的相对影响，并抑制来自较近或聚类更紧密的批次中具有误导性的亲和度。由此，BALANS有选择地识别出真正相似的跨批次邻居，尽管存在由批次引起的

失真。

BALANS 的第二个思路是避免计算完整的$n \times n$亲和矩阵，因为在大规模情况下这难以实现。相反，BALANS 会选取一部分 $S \subseteq [n]$ 个 $|S| \ll n$ 数据点并计算其亲和度行。BALANS 不会进行均匀采样，而是采用基于覆盖率的自适应策略：在每个步骤中 t，下一个点 i_t 被选中的概率与其相对于此前采样点的累积亲和度的倒数成正比。这样会优先选择其邻域尚未在先前样本中得到充分表示的数据点。每个计算出的亲和度行都通过肘部检测方法进行稀疏化，仅保留信息量最高的条目。由此得到的矩阵 $A_S \in \mathbb{R}^{|S| \times n}$ 呈现了完整亲和矩阵的部分视图 $A \in \mathbb{R}^{n \times n}$。其余条目则通过基于低秩近似的方法推断，该方法基于 A_S（Williams 和 Seeger，2000年）。BALANS 的概述见算法1。

我们从理论上证明，该采样策略仅使用 $|S| = O(K \log K)$ 个样本，便能以高概率保证每个簇的行数相近。这确保了对底层结构的覆盖度，从而能够对 A 进行低秩近似。我们证明，重构误差 $\|A - \widehat{A}\|_{\text{op}}$ 以高概率有界，其中 $\widehat{A}$ 由采样行构建。BALANS 相对于数据点数量以近线性时间 运行，因而可扩展到大型数据集。重要的是，完整亲和矩阵 A 从未被显式构建。相反，BALANS 将一个基于 A_S 定义的稀疏平滑算子应用于数据矩阵 X。

大型数据集上的实验： 我们在来自 JUMP 细胞绘画联盟（Chandrasekaran 等人，2023年）和 BBBC（Ljosa 等人，2012年）的真实世界数据集上，针对使用 CellProfiler 或 DeepProfiler 特征的不同场景进行评估（表 1）。在八个基准场景中，BALANS 的表现与最先进的批次校正方法相当；在规模最大、多样性最高的设置（表1）中，相对于未校正基线，其平均评估分数提升超过 30%。我们还报告了运行时间比较，表明 BALANS 速度快，并与现有方法具有竞争力。除准确性外，BALANS 还极为高效。尽管以 Python 实现，其运行速度仍快于许多经过优化的原生实现。在旨在模拟真实批次结构的合成基准测试上，它在保持高校正质量的同时，可在不到一小时内扩展到多达 500 万个数据点的数据集。这些结果表明，BALANS 是高内涵成像中可扩展且有效的批次校正解决方案。

相关工作： 细胞绘画已被广泛探索，所采用的是超越 CellProfiler 和 DeepProfiler 的基于深度学习的特征表示（Doron 等人，2023；Wong 等人，2023；Kraus 等人，2024）。批次校正已在多种生物数据模态中得到广泛研究（Ando 等人，2017年；Cuklina 等人，2021年；Yu 等人，2023年；Tran 等人，2020）。局部带宽估计（Zelnik-Manor 和 Perona，2004年；Herrmann，1997年；Knutsson 等人，1994年）和低秩近似（Williams 和 Seeger，2000；Drineas等人，2005年；Kumar等人，2012；Musco和 Musco，2017年）也具有成熟的理论基础。相关工作的综合综述见附录 K。

与 BBKNN 的比较： BBKNN（Polański 等人，2020）是一种用于构建批次感知 KNN 图的工具，与 BALANS 有关联。但其功能与 BALANS 存在根本差异。BBKNN 只生成图结构，不会生成校正后的特征谱或嵌入。尽管 BBKNN 定义了亲和矩阵，却没有提供将该矩阵高效应用于数据的方法。相比之下，BALANS 专门用于输出校正后的特征谱，而无需构建完整矩阵。由于 BBKNN 无法生成校正后的特征谱，因此比较仅限于基于图的指标。更详细的讨论及定量结果见附录I。

2. 关键概念

在 Cell Painting 实验中，细胞会受到各种扰动（例如化学或遗传扰动），随后通过荧光显微成像进行成像，再经过特征提取流程，获得代表细胞形态的特征谱。考虑以这种方式获得的 n 数据点 $x_1, \ldots\ldots, x_n \in \mathbb{R}^d$。每个数据点都标注有一个已知的批次标签 $b_i \in \{1, \ldots, B\}$，$i = 1, \ldots\ldots, n$，反映实验室、实验方案或所用显微镜等方面的差异。此外，每个数据点都属于一个未知的生物学簇。真正的生物学簇是一组具有共同潜在表型或相似表型的数据点，不受其处理标签或批次分配的影响。因此，每个数据点都有一个（未知的）簇标签 $c_i \in \{\mathcal{C}_1, \mathcal{C}_2 \ldots, \mathcal{C}_K\}$，$i = 1, \ldots\ldots, n$。我们的目标是为每个数据点找出一小组邻居，以反映其真实的生物学相似性。我们将利用这些邻居构建一个平滑算子，对该数据点进行校正。

为了确定用于平滑的邻居，需要衡量样本之间的相似度。这些相似度通常以亲和矩阵的形式编码，亲和矩阵可以通过多种方式构建（核函数（Hofmann 等人，2008）、最近邻图（Abbasifard 等人，2014）或学习得到的距离度量（Kulis 等人，2013））。然而，若直接根据数据计算亲和度而不进行校正，所得结果通常既反映簇归属，也反映外部变异来源，包括批次依赖效应和测量噪声。为处理成对相似性中的批次依赖失真，我们定义一种

批次依赖的局部尺度，用于每个点，以刻画从该点视角看各批次的距离。这些尺度随后用于通过局部尺度高斯核计算亲和度（Zelnik-Manor 和 Perona，2004年），其中两点之间的相似度取决于该尺度。

2.1. 批次依赖的局部尺度

令 $d_{ij} = \|x_i - x_j\|_2$ 表示数据点 x_i 与 x_j之间的欧几里得距离。根据成对距离定义亲和度的一种标准方法是使用高斯核$A_{ij} = \exp\left(-d_{ij}^2/\sigma^2\right)$，其中 $\sigma > 0$ 是用于控制局部性尺度的全局尺度。该核会为距离较近的点赋予较高的亲和度，广泛用于谱聚类（Zelnik-Manor 和 Perona，2004年）。然而，当不同批次可能呈现不同的密度模式、噪声水平或距离时，固定全局尺度 σ 这一假设并不适用。因此，亲和度估计可能较差。

在真实数据中可以直接观察到这一问题。请看图1(b)，该图展示了按上述方法构建的相关矩阵，其中包含对应于阳性对照化合物的数据点。尽管已知这些化合物会诱导强烈且多样的表型，但所得结构却表现出明显的批次内聚类，这并不理想。因此，为解决这一问题，我们定义批次依赖的局部尺度及其对应的亲和度。

定义 2.1（批次依赖的局部尺度；亲和度）。令 $d_{ij} = \|x_i - x_j\|_2$ 表示欧几里得距离，令 b_j 表示 x_j 所属的批次。令 $\sigma_{ij}^2 := \left(d_{i,}^{b_j}{}_{(k)}\right)^2$，其中$d_i^{b_j}{}_{,(k)}$ 表示 x_i 到其在批次 b_j内第 k近邻的距离。那么，批次依赖亲和度为$A_{ij} := \exp\left(-d_{ij}^2/\sigma_{ij}^2\right)$.

为展示其在真实数据中的效果，我们将批次依赖的亲和度应用于对应阳性对照的数据点，并构建相关矩阵。如图1(c)所示，所得热图呈现出更清晰的结构，在化合物层面实现了更清晰的跨批次对齐。通过在不同批次间调整局部尺度来计算批次依赖的亲和度，可以在批次效应下找回真实邻居，但计算开销较大。它需要 $O(n^2)$ 次距离计算，并反复进行 k-NN 搜索，以计算所有成对的σ_{ij}。为确保可扩展性，我们通过对部分行进行采样来估计亲和矩阵，每一行都能提供完整邻域的信息。我们从每一行获取该数据点的批次依赖局部尺度。这种基于行的采样也称为地标采样（Kumar 等人，2012年）。与随机采样 A中的元素不同，这种方法保留了计算批次依赖尺度所需的信息。在下一节中，我们将展示，仅使用少量行即可实现对 A 的准确近似。

2.2. 通过自适应地标采样进行低秩矩阵近似

在继续之前，我们首先需要为亲和矩阵的结构建立一个合适的模型。回到阳性对照实验，我们注意到，应用批次依赖的尺度后，亲和矩阵中呈现出块结构（图1(c)），反映了叠加噪声的块对角矩阵。这启发我们将其自然地分解为块对角矩阵分量与噪声之和。为此，我们将观测到的亲和矩阵 $\sim$ $\sim A \in \mathbb{R}^{n\times n}$ 建模为 $A = A_0 + E$，其中 其中 A_0 是编码聚类结构的块对角矩阵，而 E 表示与 A_0无关的残差噪声。我们假设 A_0 是对称的低秩矩阵，并具有以下结构。

假设 2.2（块结构亲和性）。索引集 $[n]$ 存在一个划分，由真实的生物学簇$\mathcal{C}_1$、……、$\mathcal{C}_K$组成，其中每个簇 $\mathcal{C}_k$的大小为 n_k。低秩亲和矩阵 $A_0 \in \mathbb{R}^{n\times n}$ 可表示为块形式$A_0 := \text{diag}(A^{(1)}, A^{(2)}, \ldots\ldots, A^{(K)})$，其中每个对角块 $A^{(k)} \in \mathbb{R}^{n_k\times n_k}$ 表示生物学簇 k 内部的亲和度，其定义为 $A^{(k)} = p_k \cdot \mathbf{1}_{n_k\times n_k}$，其中$p_k > 0$，且 $\mathbf{1}_{n_k\times n_k} \in \mathbb{R}^{n_k\times n_k}$ 为全 1 矩阵。A_0 的所有非对角块均为零。整体矩阵 A_0 是对称半正定的，且秩为K。

请注意，尽管 A_0 的非对角块为零，但这并不意味着不同簇中的数据点之间不存在亲和度。相反，块对角结构是一种建模选择，旨在选取与给定数据点距离较近且相似度较高的数据点。

我们仍需对噪声矩阵进行建模 E。一种标准的噪声建模方法是使用高斯正交系综（GOE）（Mehta，2004年）。高斯正交系综是对称随机矩阵，其高斯条目相互独立（除对称性约束外），并被广泛用于建模噪声。重要的是，它们的奇异值以高概率集中在 $O(\sqrt{n})$ 附近，因此是一种实用的模型。然而，在我们的设定中，矩阵的条目必须严格为正，˜因为 A 仍然是一个亲和矩阵。因此，我们采用一种具有正条目的随机化类比模型，同时保留高斯正交系综的谱尺度。

假设2.3（指数噪声矩阵）。令 $E \in \mathbb{R}^{n\times n}$ 为一个对称随机矩阵，其中对所有 $i < j$，$E_{ij} = E_{ji} \sim$相互独立且服从 Exp(λ)，并且 $E_{ii} \sim$ 服从 Exp($\lambda/2$)，其中 $\lambda > 0$为某个值。

这确保了 E 是对称的，具有与高斯正交系综相似的奇异值，同时满足逐项非负性。我们的目标是利用来自 A的一小部分行˜恢复 A_0。为了近似 A_0，我们采样一个大小为 s 的子集 $S \subseteq [n]$，并构造 Nyström 估计量（Kumar 等人，2012年）$\widehat{A} := \widetilde{A}_S^{\top}\widetilde{A}_S^{+}{}_S\widetilde{A}_S$，其中 $\widetilde{A}_S \in \mathbb{R}^{s\times n}$ 包含

~采样行以及$A_{S,\ S} \in \mathbb{R}^{s\times s}$是它们构成的子矩阵。这里，$A^{+}$表示 Moore–Penrose 伪逆（Cour-rieu，2008），它是满足 $AA^{+}A = A$ 和$A^{+}AA^{+} = A^{+}$的唯一矩阵。这就得到了A_0 的低秩近似，该近似使用由 S定义的子空间。

我们如何评估这个估计量的质量？重构质量取决于采样行对数据连通性的捕捉效果。由于 $\sim A$ 的每一行都表示一个点与数据集中其余点之间的亲和度，因此，优秀的采样算法会从数据的不同区域采样信息丰富的行。为此，我们定义覆盖度的概念。令 cov(i) 表示点 i 从先前采样的行中获得的累积亲和度，即 cov(i) : $= \sum_{j\in S_{\text{past}}} A_{ji}$，其中 S_{past} 表示迄今为止采样的索引集合。直观而言，累积覆盖度较低的点尚未在近似中得到充分体现。采样以块为单位进行，每个块的大小为 K。每个块开始时，临时覆盖向量初始化，采样分布设为均匀分布。在块内的每个步骤 t，从分布 $q^{(t)}$ 中抽取一个索引 i_t；该分布优先选择覆盖不足的点，其中 $q_i^{(t)} \propto 1/\text{cov}^{(t)}(i)$，而 $\text{cov}^{(t)}(i)$ 表示当前块内累积的覆盖度。每个块结束后，覆盖向量重置为 0，采样分布也重置为均匀分布。虽然假设2.3保证 ~ 亲和矩阵 A 的元素严格为正（因此对所有 i，cov(i) > 0~），但实际应用中会对 A 进行稀疏化（见第3节），这可能导致零覆盖。在这种情况下，我们会按照算法 5 所述修改采样规则。这种交替方案确保采样能够探索数据集中的新区域，并将重点放在覆盖不足的区域上。我们分析了这一新颖的采样方案，并证明 ^，只用少量行得到的估计量 A 就能以谱保证近似 A_0。

2.3. 理论结果

现在，我们将形式化地阐述基于覆盖度的采样在近似底层亲和矩阵 A_0方面所提供的保证。回顾一下，数据集被划分为 K个生物学簇 $\mathcal{C}_1$，……，$\mathcal{C}_K$其大小分别为 n_1，……，n_K，总大小为 $n = \sum_{k=1}^{K} n_k$。令 m表示采样行数（我们将在第 节讨论如何通过停止准则确定 3）。

定理 2.4 (簇覆盖保证)。令 T_k 表示从 $\mathcal{C}_k$中采样的行数。那么，对于常数$\delta > 0$，存在另一个常数 $C := C(\delta)$ 使得若 $m \geq CtK\log K$，则 $\mathbb{P}(T_k \geq t,\ \forall\ k) \geq 1 - \delta$.

证明（附录A）基于经过改进的优惠券收集论证（Boneh 和 Hofri，1997年）。与随机采样不同，自适应策略无论簇大小如何，都能实现 $\mathcal{O}(K\log K)$ 覆盖度，因此达到阶最优。我们还在附录A中证明，这一性质保证

^我们的估计量 A 在算子范数下与 A_0 高度近似。令 $\|\cdot\|_{\text{op}}$ 表示算子范数。

^**定理 2.5** A_0 $A =$。$\widetilde{A}_S^{\top}\widetilde{A}_S^{+}{}_S\widetilde{A}_S$ (的谱近似)。^令，表示通过对 $m \geq C(\delta)tK\log K$ 个自适应采样的~行 A进行低秩重构所形成的估计量。于是，

$$\|\widehat{A} - A_0\|_{\text{op}} \leq \frac{Dn}{\sqrt{t}\cdot\min_k p_k}, \tag{1}$$

以高概率成立，其中 p_k 是簇 $\mathcal{C}_k$的簇亲和度，而 D : $= D(\lambda, K, \vec{p}) > 0$ 取决于噪声矩阵的参数 λ、簇的数量 K以及亲和度向量 $\vec{p}$。

定理2.5 保证了在抽取足够多的行时，谱重构误差能够得到控制。具体而言，当样本数量 $m = O(tK\log K)$ 满足定理2.4中的条件时，使用这些样本足以实现较小的谱误差，其衰减速率为 $O(1/\sqrt{t})$，同时确保所有簇都有样本代表。

3. BALANS 算法

算法 1展示了 BALANS 的完整实现。给定数据矩阵 $X \in \mathbb{R}^{n\times d}$ 和批次标签$\vec{b} \in \{1,\ldots,B\}^n$，该算法还将最近邻参数 k、停止阈值 τ和自适应采样参数 J作为输入。也可以选择将数据投影到低维主成分分析子空间中。算法维护三个变量：(i) 一个累积覆盖向量 $c \in \mathbb{R}^n$，用于跟踪每个点的累积亲和度；(ii) 一个临时覆盖向量 $c_K \in \mathbb{R}^n$，用于计算跨 J 个步骤的自适应抽样概率；以及 (iii) 一个抽样索引集合 S，初始化为空。若连续τ 次迭代中都没有新点被覆盖，主循环便会终止。

在每次迭代中，使用基于覆盖度的自适应策略（第 i_t 2.2节，算法5）对一个索引进行采样。对于选中的点 x_{i_t}，使用批次相关局部尺度核（第A_{i_t}2.1节，算法3）计算稀疏亲和度行：。为减少噪声和内存占用，通过肘部检测启发式方法（Truong 等，2020）对亲和度进行稀疏化。首先按降序对条目排序，仅保留高于陡降阈值的值。稀疏化后，Δ统计首次获得非零亲和度的新索引数量。如果 $\Delta = 0$，则递增停滞计数器；否则，将其重置。连续 τ 轮未出现新的覆盖度后，循环停止，这作为一种收敛启发式方法。收集到的亲和度行构成稀疏矩阵$A_S \in \mathbb{R}^{m\times n}$，该矩阵紧凑且易于存储。

理论上，校正后的数据通过低秩估计器 $\widehat{X} := (A_S^{\top}A_S^{+}$, $_S A_S)X$ 计算，其中 $A_S \in \mathbb{R}^{m\times n}$ 包

含抽样后经行归一化的亲和度行，而$A_{S,\ S} \in \mathbb{R}^{m\times m}$ 是基于抽样索引得到的子矩阵。这种表述方式支持高效的谱重构，因为 A_S 是稀疏的，而 A_S^+ ，S 是一个小型稠密矩阵。实际应用中，为避免其 $O(m^3)$ 开销，我们省略 A_S^+ ，S，将估计量近似为 $\widehat{X} \approx (A_S^\top A_S)X$。其中，$A_S$ 对 X 进行平滑，而 $A_S^\top$将结果传播至所有点。为保证稳定性，这两个矩阵均按行归一化；我们通过先计算 $A_S X$，再计算 $A_S^\top(A_S X)$，避免构造稠密矩阵 $A_S^\top A_S$。由于 A_S 是稀疏的，所有运算都能保持高效。

算法1 BALANS 批次校正

输入： 数据 $X \in \mathbb{R}^{n\times d}$，批次标签 $\vec{b} \in \{1,\dots,B\}^n$，参数 k, τ, J
输出： 校正后的数据 $\widehat{X} \in \mathbb{R}^{n\times d}$
对 X 执行主成分分析（可选）；设置 c，$c_K \leftarrow \mathbf{0}$；$S \leftarrow \emptyset$；不变 $\leftarrow 0$
当不变 $\le \tau$ **时**
  对于 $j = 1$至 J **执行**
    通过自适应采样（算法 i_t 进行采样（算法5）；计算计算$A_{i_t,:}$：通过局部尺度核（算法 3）
    将 $A_{i_t,:}$ 追加到 A_S；更新 $S \leftarrow S \cup \{i_t\}$
    $\Delta \leftarrow \#\{j:\ A_{i_t,j} > 0$ 和 $c_J[j] = 0\}$
    如果 $\Delta = 0$**则**
      未变化计数 $\leftarrow$ 未变化计数 $+ 1$
    else
      未变化计数 $\leftarrow 0$
    结束条件判断
    $c \leftarrow c + A_{i_t,:}$；$c_J \leftarrow c_J + A_{i_t,:}$
  结束循环
  重置 $c_J \leftarrow \mathbf{0}$
结束循环
通过低秩补全计算 X（算法 4）；**返回** $\widehat{X}$

重要的是，BALANS 对超参数具有鲁棒性。消融研究见附录E。现在，我们说明 BALANS 的总体计算复杂度，详见附录 D。

定理 3.1（BALANS 的计算复杂度）。设 n 为数据点数量，d 为特征维度，$m = |S|$ 为 *BALANS* 收敛后自适应采样的行数。则 BALANS 的总计算复杂度为 $\mathcal{O}$（nm（$d + \log n$））， 其中第一项表示稀疏亲和度构建和矩阵乘法的开销，第二项表示最近邻搜索的开销。当 $m \ll n$ 时，该复杂度关于 n 近似呈线性。

根据定理 2.5，采样 $m = \mathcal{O}(K$ 对数 $K)$ 行即可在高概率下重建 A_0，同时覆盖大量点，从而使算法能够终止。因此，总体计算复杂度为 $\mathcal{O}(nK$ 对数 $K(d+$对数 $n))$，这相对于数据集大小 n 近似为线性。

4. 实验

我们在来自 7 个数据集的 8 个真实场景中评估 BALANS，涵盖 CellProfiler 和 DeepProfiler 特征。其中，5 个场景来自 JUMP Cell Painting 联盟（约有 140,000 个扰动、115 TB 图像），另有 3 个场景使用来自 BBBC 数据集的 DeepProfiler 嵌入（约有 1000 个扰动，包含 50000 个孔均值化特征谱）。我们还纳入了 6 个基于高斯混合的合成场景，数据点数量从 500 到 500 万不等。所有方法均使用综合评估批次校正效果和生物信号保留情况的指标进行评估。完整详情见附录 F。我们将 BALANS 与多种批次校正方法进行比较，包括统计方法（Combat（Zhang 等人，2020）、球化（Kessy 等人，2018年））、线性投影方法（典型相关分析、稳健主成分分析（Satija 等人，2015）），以及基于图或 K近邻的方法（FastMNN、Scanorama（Hie 等人，2019年）、Harmony（Korsunsky 等人，2019年））。这些方法的底层假设各不相同。附录3 中的表格总结了这些特性；完整的实现细节和参考文献见附录 G。

我们使用一系列成熟的指标评估性能（Luecken 等人，2022）。这些指标包括图连通性、kBET、LISI（批次和标签）、轮廓系数（批次和标签），以及基于聚类的度量，例如 Leiden 调整兰德指数和 Leiden NMI。每项指标均被归一化到范围[0, 1], 内，数值越高表示性能越好。需要指出的是，由于资源限制，在大规模场景（例如 JUMP 2）中无法计算 kBET。指标的完整定义和实现细节见附录 H。附录 4 中的表格 H总结了评估指标。表 1中展示了三个代表性场景（完整详情见附录 I）。**JUMP 1**：中等规模的 CellProfiler（25K 个点、3 个来源、302 种化合物）。**JUMP 2**：大规模、高多样性（350K 个点、5 台显微镜、80K+种化合物，重叠度低）。**DEEP 1**：基于微孔板行分组产生批次效应的 `BBBC036` DeepProfiler 版本。我们还比较了表现最佳方法的实际运行时间（表2）。在各数据集中，BALANS 的表现均优于其他竞争方法。

指标说明：我们区分批次指标与生物学指标，以便对整合质量形成平衡的认识。批次指标（例如，图连通性、批次 LISI）用于评估虚假技术变异的去除程度，但若孤立看待则可能产生误导。抹除全部结构的方法（例如通过添加噪声）尽管丢弃了生物学信号，仍可能获得高分。相比之下，标签指标（例如，调整兰德指数、归一化互信息、标签轮廓系数）则直接评

表 *1.* 三个场景下的评估分数：**JUMP 2**（批次规模大且多样）、**DEEP 1**（批次多样性低，使用 DeepProfiler）以及 **JUMP 1**。由于大多数方法具有确定性，标准差（SD）为 0；受篇幅所限，此处省略。包含标准差的表格见附录。

方法	连通性	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
BALANS	0.33	**0.48**	**0.91**	**1.00**	0.01	**0.46**	0.32	**0.57**	**0.45**	**0.50**
Scanorama	**0.34**	0.41	0.83	**1.00**	0.01	0.30	0.27	0.53	0.40	0.45
SCVI	0.28	0.44	0.82	0.99	0.01	0.28	0.29	0.53	0.39	0.44
fastMNN	**0.34**	0.45	0.79	**1.00**	**0.02**	0.26	0.22	0.53	0.38	0.44
Harmony	**0.34**	0.25	0.84	**1.00**	0.01	0.27	0.32	0.48	0.40	0.43
球化	**0.34**	0.00	0.84	**1.00**	0.00	0.22	**0.38**	0.39	0.40	0.40
Combat	**0.34**	0.02	0.81	**1.00**	0.00	0.24	0.31	0.39	0.39	0.39
基线	**0.34**	0.01	0.80	**1.00**	0.00	0.23	0.31	0.38	0.39	0.38
DESC	0.29	0.40	0.80	0.99	0.00	0.25	0.28	0.50	0.39	0.42

方法	连通性	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
BALANS	0.20	0.61	0.89	**1.00**	**0.00**	0.32	0.39	**0.57**	0.43	**0.49**
球化	**0.30**	0.43	**0.90**	0.99	**0.00**	0.30	**0.46**	0.54	**0.44**	0.48
Seurat RPCA	0.26	**0.62**	0.79	0.99	**0.00**	0.25	0.36	0.56	0.40	0.47
Harmony	0.26	**0.62**	0.80	0.99	**0.00**	0.23	0.37	0.56	0.40	0.47
Seurat CCA	0.27	0.60	0.79	0.99	**0.00**	0.27	0.36	0.55	0.40	0.47
基线	0.27	0.57	0.80	0.99	**0.00**	0.23	0.37	0.55	0.40	0.46
Combat	0.27	0.58	0.80	0.99	**0.00**	0.24	0.37	0.55	0.40	0.46
fastMNN	0.26	0.60	0.78	0.99	**0.00**	0.24	0.35	0.55	0.40	0.46
Scanorama	0.19	0.56	0.80	**1.00**	**0.00**	**0.34**	0.33	0.52	0.42	0.46
SCVI	0.24	0.58	0.82	0.99	**0.00**	0.29	0.34	0.55	0.41	0.47
DESC	0.23	0.55	0.79	0.99	**0.00**	0.26	0.33	0.53	0.40	0.46

方法	连通性	KBET	批次 LISI	批次轮廓系数	LISI_t	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
Seurat CCA	0.59	**0.61**	0.50	0.88	**0.98**	**0.05**	0.40	0.47	**0.65**	0.48	**0.56**
BALANS	0.54	0.45	0.44	**0.89**	**0.98**	**0.05**	**0.46**	**0.53**	0.58	**0.51**	0.54
Seurat RPCA	0.59	0.46	0.41	0.88	**0.98**	0.03	0.39	0.47	0.59	0.47	0.53
fastMNN	0.53	**0.61**	0.46	0.83	0.97	0.03	0.35	0.43	0.61	0.45	0.53
Harmony	**0.60**	0.41	0.43	0.88	**0.98**	0.03	0.39	0.47	0.58	0.47	0.52
Scanorama	0.35	0.56	**0.52**	0.80	**0.98**	0.03	0.34	0.43	0.56	0.45	0.50
SCVI	0.50	0.40	0.45	0.85	0.97	0.03	0.38	0.46	0.58	0.47	0.51
基线	0.51	0.17	0.13	0.78	**0.98**	0.02	0.32	0.47	0.40	0.45	0.42
Combat	0.54	0.08	0.11	0.80	**0.98**	0.02	0.33	0.47	0.38	0.45	0.42
球化	0.48	0.10	0.06	0.79	**0.98**	0.01	0.28	0.47	0.36	0.44	0.40
DESC	0.49	0.35	0.42	0.84	0.97	0.02	0.36	0.45	0.57	0.46	0.50

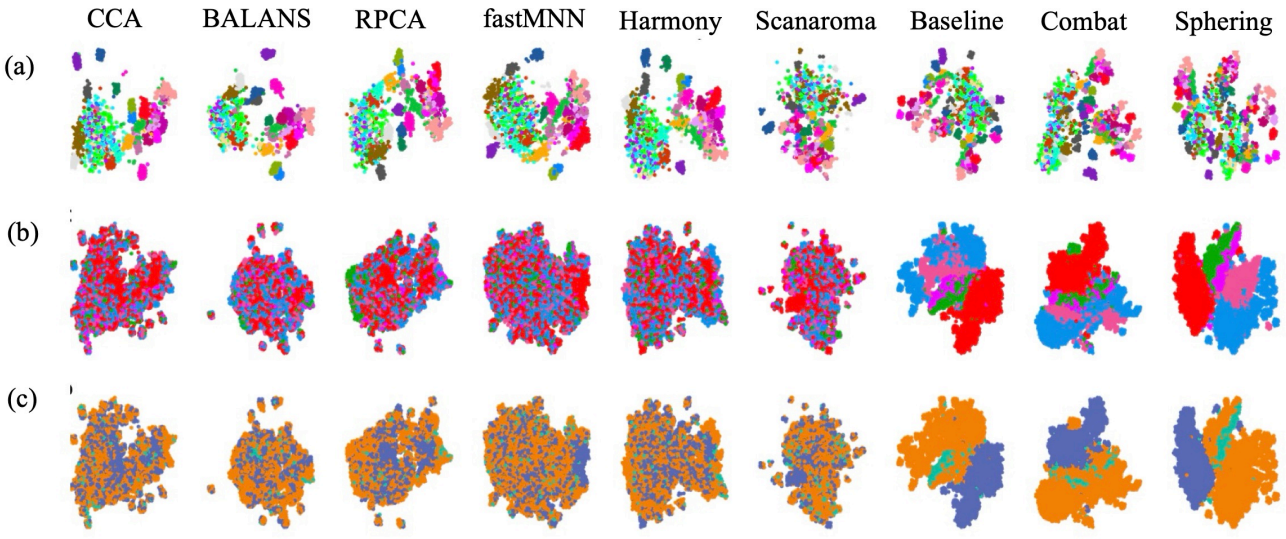


图*2*。图3：JUMP 1 的 UMAP 可视化。(a) 按生物学簇着色，展示了若干代表性化合物之间的分离情况。(b) 按来源实验室着色。(c) 按显微镜着色。这些图展示了批次效应如何在不同技术来源中显现。值得注意的是，BALANS 形成了紧密且连贯的生物学簇，同时显著减少了实验室之间的分离。

估有意义的生物学结构是否得以保留，因而具有更大的解释权重。因为没有单

单一指标无法涵盖所有方面，因此我们着重关注比较模式，并将总体平均值作为更稳健的总结

表 2. 真实场景中表现最佳方法的墙钟运行时间（分钟）。

方法	场景 JUMP1	场景 JUMP2	场景 DEEP1
BALANS	**00:00:59**	**00:55:30**	**00:00:44**
Harmony	00:12:35	01:44:53	00:22:43
Scanorama	00:01:33	01:13:02	00:01:57
Seurat CCA **	00:34:37	12:00:00+	01:03:24
Seurat RPCA **	00:14:03	12:00:00+	00:21:29

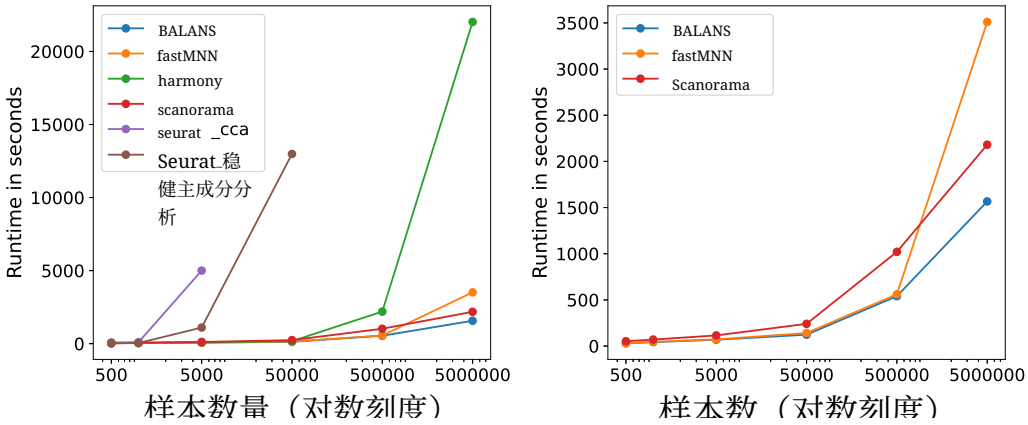


图3. 合成数据上的运行时间比较（10种化合物、5个批次、10个特征）。左：所有方法。右：仅最快的方法。X轴：数据点数量（对数刻度）；Y轴：运行时间（秒）。

在局部与全局结构之间的权衡。最终，我们优先考虑标签指标的改进，因为恢复生物学上连贯的变异是批次校正的首要目标。值得注意的是，在所有数据集上，BALANS 的标签类生物学指标平均值（标签平均值）始终表现最佳，这突显了其保留生物学上连贯变异的能力。关于指标的详细讨论见附录F。

指标解读：作为生物学可解释性的一个例证，BALANS能够更清晰地找回表型已知的阳性对照化合物。例如，Aurora激酶抑制剂AMG900 ()形成了一个独特的簇，其轮廓系数较高（0.82，而Harmony为0.45）。这凸显了BALANS保留有意义的生物结构的能力。更多化合物层面的比较和热图见附录**??**。

与基线方法相比，BALANS 在异质性较高的数据集上表现最佳。尤其是，JUMP-1 和 JUMP-2 上的结果清楚地体现了这一点。JUMP-2 的批次效应更强，生物和技术异质性也更高，结果显示 BALANS 在所有指标上都具有明显优势。在异质性较为适中的 JUMP-1 上，BALANS 能够稳定地进行校正，但并未大幅超越所有基线方法。总的来说，这些结果表明，随着数据集异质性的增加，BALANS 的表现尤为出色，凸显了它在规模庞大且多样化的真实场景中的适用性。

为评估可扩展性，我们基于结构化高斯混合模型构建了一个合成基准。该基准支持最多包含 500 万个数据点、批次复杂度各异的数据集（详情见附录J）。主实验在 10 个维度上使用了 5 个批次中的 10 个复合簇。如图 3所示，即使处理 500 万个数据点，BALANS 的运行时间仍然最短，且校正质量不受影响。本实验及其他场景的聚合得分、运行时间详情（包括图形处理器硬件）及其聚合得分见附录 J。

5. 总结

我们提出了BALANS，这是一种新方法，将局部亲和结构与自适应采样相结合，为细胞绘画数据的批次校正提供了一种可扩展且具有理论依据的方法。

6. 影响声明

本文旨在推动机器学习领域的发展。我们的工作可能会带来许多社会影响，但我们认为没有哪项影响需要在此特别指出。

参考文献

Aviv Regev、Sarah A Teichmann、Eric S Lander、Ido Amit、Christophe Benoist、Ewan Birney、Bernd Bodenmiller、Peter Campbell、Piero Carninci、Menna Clatworthy 等。《人类细胞图谱》。《*elife*》，6:e27041，2017年。

MohammadHosseinRohban、ShantanuSingh、XiaoyunWu、Julia B Berthet、Mark-Anthony Bray、Yashaswi Shrestha、Xaralabos Varelas、Jesse S Boehm 和 Anne E Carpenter。通过细胞绘画系统分析人类基因与等位基因功能的形态学特征。*Elife*，6:e24060，2017年。

Mark-Anthony Bray、Shantanu Singh、Han Han、Chadwick T Davis、Blake Borgeson、Cathy Hartland、Maria Kost-Alimova、Sigrun M Gustafsdottir、Christopher C Gibson 和 Anne E Carpenter。细胞绘画：一种利用多重荧光染料进行形态学分析的高内涵图像检测方法。*Nature protocols*，11 (9):1757–1774，2016。

Juan C Caicedo、Sam Cooper、Florian Heigwer、Scott War- chal、Peng Qiu、Csaba Molnar、Aliaksei S Vasilevich、Joseph D Barry、Harmanjit Singh Bansal、Oren Kraus 等。基于图像的细胞分析的数据分析策略。*Naturemethods*，14(9):849–863，2017年。

Beth A Cimini、Srinivas Niranj Chandrasekaran、Maria Kost-Alimova、Lisa Miller、Amy Goodale、Briana Fritch-man、Patrick Byrne、Sakshi Garg、Nasim Jamali、David J Logan 等。优化用于基于图像的表型分析的细胞绘画检测。《*Nature protocols*》，18(7):1981– 2013，2023年。

Marie FA Cutiongco、Bjørn Sand Jensen、Paul M Reynolds 和 Nikolaj Gadegaard。利用纳米形貌诱导的细胞形态反应预测基因表达。《*Nature communications*》，11(1):1384，2020。

Mathias J Wawer、Kejie Li、Sigrun M Gustafsdottir、Vebjorn Ljosa、Nicole E Bodycombe、Melissa A Marton、Katherine L Sokolnicki、Mark-Anthony Bray、Melissa M Kemp、Ellen Winchester 等。利用多重高维剖析，构建适用于细胞表型筛选的性能多样化小分子文库。《美国国家科学院院刊》，111(30): 10911–10916，2014。

Milad Dagher、Grant Ongo、Nathaniel Robichaud、Jinglin Kong、Woojong Rho、Ivan Teahulos、Arya Tavakoli、Samantha Bovaird、Shahem Merjaneh、Andrew Tan 等。nELISA：一个支持分泌组定量分析的高通量、高多重化平台。*bioRxiv*，2023年。

Anne E Carpenter、Thouis R Jones、Michael R Lamprecht、ColinClarke、InHan Kang、OlaFriman、David A Guertin、Joo Han Chang、Robert A Lindquist、Jason Moffat 等。Cellprofiler：用于识别和量化细胞表型的图像分析软件。*Genome Biology*，7：1–11，2006年。

Nikita Moshkov、Michael Bornholdt、Santiago Benoit、Matthew Smith、Claire McQuin、Allen Goodman、Re- becca A Senft、Yu Han、Mehrtash Babadi、Peter Horvath 等。学习用于扰动图像分析的表征。《*Nature Communications*》，15(1):1594，2024年。

Mingxing Tan 和 Quoc Le。EfficientNet：重新思考卷积神经网络的模型缩放。发表于国际机器学习会议，第 6105–6114 页。PMLR，2019年。

John Arevalo、Ellen Su、Jessica D Ewald、Robert van Dijk、Anne E Carpenter 和 Shantanu Singh。评估基于图像的细胞表型分析的批次校正方法。《Nature Communications》，15(1):6516，2024年。

Srinivas Niranj Chandrasekaran、Jeanelle Ackerman、Eric Alix、D Michael Ando、John Arevalo、Melissa Bennion、Nicolas Boisseau、Adriana Borowa、Justin D Boyd、Laurent Brino 等。JUMP 细胞绘画数据集：136,000 种化学和遗传扰动的形态学影响。《BioRxiv》，第 2023–03 页，2023年。

David Van Dijk、Roshan Sharma、Juozas Nainys、Kristina Yim、PoojaKathail、AmbroseJ Carr、Cassandra Burdziak、Kevin R Moon、Christine L Chaffer、Diwakar Pattabira- man 等。利用数据扩散从单细胞数据中恢复基因相互作用。《Cell》，174(3):716–729，2018年。

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, 和 Nancy R Zhang。SAVER：单细胞RNA测序的基因表达恢复。*Nature Methods*，15(7):539–542，2018年。

Wei VivianLi 和 Jingyi Jessica Li。用于单细胞 RNA 测序数据的准确且稳健的插补方法 scImpute。《*Nature Communications*》，9(1):997，2018年。

Hoa Thi Nhu Tran、Kok Siong Ang、Marion Chevrier、Xi- aomeng Zhang、Nicole Yee Shin Lee、Michelle Goh 和 Jinmiao Chen。单细胞RNA测序数据批次效应校正方法的基准评估。*Genome Biology*，21:1–32，2020。

Christopher Williams 和 Matthias Seeger。使用 nystr¨om 方法加速核机器。神经信息处理系统进展，13，2000年。

Vebjorn Ljosa、Katherine L Sokolnicki 和 Anne E Carpen- ter。用于验证的带注释高通量显微成像图像集。《Nature Methods》，9(7):637，2012年。

Michael Doron、Th´eo Moutakanni、Zitong S Chen、Nikita Moshkov、Mathilde Caron、Hugo Touvron、Piotr Bo- janowski、Wolfgang M Pernice 和 Juan C Caicedo。利用自监督视觉变换器实现无偏的单细胞形态分析。《bioRxiv》，2023年。Daniel R Wong、David J Logan、Santosh Hariharan、Robert Stanton、Djork-Arn´e Clevert 和Andrew Kiruluta。深度表征学习通过细胞绘制图像确定药物作用机制。《Digital Discovery》，2(5): 1354–1367，2023年。

Oren Kraus、Kian Kenyon-Dean、Saber Saberian、Maryam Fallah、Peter McLean、Jess Leung、Vasudev Sharma、Ayla Khan、Jia Balakrishnan、Safiye Celik 等。用于显微成像的掩码自编码器是可扩展的细胞生物学学习器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第11757–11768页，2024年。

D Michael Ando、Cory Y McLean 和 Marc Berndl。改善高内涵成像筛选中的表型测量。*BioRxiv*，第161422页，2017年。

Jelena Cuklina、Chloe H Lee、Evan G Williams、Tatjana Sajic、Ben C Collins、Mar´ıa Rodr´ıguez Mart´ınez、Varun S Sharma、Fabian Wendt、Sandra Goetze、Gregory R Keele 等。大规模蛋白质组学研究中批次效应的诊断与校正：教程。*Molecular Systems Biology*，17(8):e10240，2021年。

Ying Yu、Naixin Zhang、Yuanbang Mai、Luyao Ren、Qiaochu Chen、Zehui Cao、Qingwang Chen、Yaqing Liu、Wanwan Hou、Jingcheng Yang 等。使用基于参考材料的比值方法校正大规模多组学研究中的批次效应。*Genome Biology*，24(1):201，2023年。

Lihi Zelnik-Manor 和 Pietro Perona。自调节谱聚类。神经信息处理系统进展，17，2004年。

Eva Herrmann。核回归估计中的局部带宽选择。计算与图形统计学杂志，6(1):35–54，1997年。

Hans Knutsson、C-F Westin 和 G¨osta Granlund。局部多尺度频率与带宽估计。载于第一届国际图像处理会议论文集，第1卷，第36–40页。IEEE，1994年。Petros Drineas、Michael W Mahoney 和 Nello Cristianini。关于用于改进基于核的学习、近似Gram矩阵的Nyström方法。《机器学习研究杂志》，6(12)，2005年。Sanjiv Kumar、Mehryar Mohri 和 Ameet Talwalkar。Nyström方法的采样方法。《机器学习研究杂志》，13(1)：981–1006，2012年。

Cameron Musco 和 Christopher Musco。Nyström方法的递归采样。神经信息处理系统进展，30，2017年。Krzysztof Pola´nski、Matthew D. Young、Zhichao Miao、Kerstin B. Meyer、Sarah A. Teichmann 和 Jong-Eun Park。BBKNN：单细胞转录组的快速批次对齐。生物信息学，36(3)：964–965，2020年2月。doi：10.1093/bioinformatics/btz625。网址：https://doi.org/10.1093/bioinformatics/btz625。Thomas Hofmann、Bernhard Sch¨olkopf 和 Alexander J Smola。机器学习中的核方法。2008。

Mohammad Reza Abbasifard、Bijan Ghahremani 和 Has- san Naderi。最近邻搜索方法综述。《国际计算机应用杂志》，95(25)，2014年。

Brian Kulis 等。度量学习：综述。*Foundationsand Trends® in Machine Learning*，5(4):287–364，2013年。

Madan Lal Mehta。《随机矩阵》，第 142 卷。爱思唯尔，2004年。

Pierre Courrieu。Moore-Penrose 逆矩阵的快速计算。*arXiv* 预印本 *arXiv:0804.4809*，2008。

Arnon Boneh 和 Micha Hofri。重新审视优惠券收集问题——工程问题与计算方法综述。《随机模型》，13(1):39–66，1997年。

Charles Truong、Laurent Oudre 和 Nicolas Vayatis。离线变点检测方法的选择性综述。《信号处理》，167:107299，2020。

Yuqing Zhang、Giovanni Parmigiani 和 W Evan Johnson。Combat-seq：RNA-seq 计数数据的批次效应校正。《*NAR* 基因组学与生物信息学》，2(3):lqaa078，2020。

Agnan Kessy、Alex Lewin 和 Korbinian Strimmer。最佳白化与去相关。《美国统计学家》，72(4):309–314，2018年。

Rahul Satija、Jeffrey A Farrell、David Gennert、Alexander F Schier 和 Aviv Regev。单细胞基因表达数据的空间重建。《自然-生物技术》，33(5)：495–502，2015年。Brian Hie、Bryan Bryson 和 Bonnie Berger。利用 Scanorama 高效整合异质性单细胞转录组。《自然-生物技术》，37(6)：685–691，2019年。Ilya Korsunsky、Nghia Millard、Jean Fan、Kamil Slowikowski、Fan Zhang、Kevin Wei、Yuriy Baglaenko、Michael Brenner、Po-ru Loh 和 Soumya Raychaudhuri。

利用 Harmony 快速、灵敏且准确地整合单细胞数据。*Nature Methods*，16(12)：1289–1296，2019年。

Malte D Luecken、Maren B¨uttner、Kridsadakorn Chaichoompu、Anna Danese、Marta Interlandi、Michaela F M¨uller、Daniel C Strobl、Luke Zappia、Martin Dugas、Maria Colom´e-Tatch´e等。单细胞基因组学中图谱级数据整合的基准评测。*Nature Methods*，19(1):41–50，2022。

Art B Owen与Patrick O Perry。奇异值分解与非负矩阵分解的双向交叉验证。2009。

Dennis S Bernstein。矩阵数学：理论、事实与公式。普林斯顿大学出版社，2009。

Laleh Haghverdi、Aaron TL Lun、Michael D Morgan与John C Marioni。单细胞RNA测序数据中的批次效应通过匹配互为最近邻加以校正。《自然-生物技术》，36(5):421–427，2018年。

Maren B¨uttner、Zhichao Miao、F Alexander Wolf、Sarah A Teichmann 和 Fabian J Theis。用于评估单细胞 RNA 测序批次校正的测试指标。*Nature Methods*，16(1):43–49，2019年。

Edward H Simpson。多样性的测量。《自然》，163(4148):688–688，1949年。

Peter J Rousseeuw。轮廓系数：用于解释和验证聚类分析的图形辅助工具。计算与应用数学杂志，20:53–65，1987年。

Vincent A Traag、Ludo Waltman 和 Nees Jan Van Eck。从 Louvain算法到 Leiden算法：保证社区强连接。《科学报告》，9(1):1–12，2019年。

Erik Serrano、Srinivas Niranj Chandrasekaran、Dave Bun-ten、Kenneth I Brewer、Jenna Tomkinson、Roshan Kern、Michael Bornholdt、Stephen J Fleming、Ruifan Pei、John Arevalo 等。使用 Pycytominer 实现可复现的基于图像的表型分析。*Nature Methods*，第 1–4 页，2025年。

Alex Gittens 与 Michael W Mahoney。重探 nystr¨om 方法以改进大规模机器学习。《机器学习研究杂志》，17(1)：3977–4041，2016。

Kevin R. Moon、David Van Dijk、Zheng Wang、Scott Gigante、Daniel B. Burkhardt、Weiwei S. Chen、Kim La Yim、Aron van den Elzen、Matthew J. Hirn、Ronald R. Coifman、Aniruddh P. Patel、Guy Wolf和Smita Krishnaswamy。高维生物数据中结构与转变的可视化。《自然-生物技术》，37：1482–1492，2019年。doi: 10.1038/s41587-019-0336-3。

A. 定理的证明

A.1. 定理2.4的证明

定理1（簇覆盖保证，重述）。 设 T_k 表示来自簇 $\mathcal{C}_k$的采样行数。则对任意常数 $\delta > 0$，存在常数 $C := C(\delta)$，使得若 $m \geq CtK \log K$，则（在自适应采样 5下）

$$\mathbb{P}\left(T_k \geq t \text{ for all } k \in [K]\right) \geq 1 - \delta.$$

在证明该定理之前，我们先陈述几个较弱的假设。

假设1.1（增长簇情形）。 固定一个与 n 无关的整数 $t \geq 1$。令 $\{n_k\}_{k=1}^K$ 表示簇大小，并假设

$$\min_{1 \leq k \leq K} n_k \to \infty \quad \text{as the total sample size } n = \sum_{k=1}^K n_k \to \infty.$$

这是一个非常宽松的假设：它允许簇大小高度不均衡，包括一个或多个簇相对于 n 极小的情况，只要每个簇仍无限增长。

假设 1.2（低噪声情形）。假设噪声水平满足

$$p_{\min} \cdot \lambda > \delta n$$

其中某个常数 $\delta > 0$与 n无关。

直观地说，这保证真实的亲和度不会被噪声淹没。

证明。回顾

$$\tilde{A} = A_0 + E,$$

是我们从中抽取行的含噪块亲和度矩阵。令 m 表示采样步的总数，每个块包含 K个采样步。考虑其中一个块，并令 $s \in [K]$ 表示该块内的第 s^{th} 个采样步。定义该块中步骤 s 时点 x_i 的覆盖度为

$$\text{cov}^{(s)}(i) := \sum_{j \in S_{\text{past}}} \widetilde{A}_{ji},$$

其中 $S_{\text{past}} := \{i_j \mid 1 \leq j < s\}$ 表示当前块中在步骤 s之前抽取的索引集合。令

$$\text{cov}^{(s)} := \left(\text{cov}^{(s)}(1), \text{cov}^{(s)}(2), \ldots, \text{cov}^{(s)}(n)\right)$$

表示步骤 s 时的覆盖向量。逆覆盖采样分布定义为

$$q^s := (q^{(s)}(1), q^{(s)}(2), \ldots q^{(s)}(n))$$

在步骤 s 时定义为

$$q^{(s)}(i) := \frac{\text{cov}^{(s)}(i)^{-1}}{\sum_{j=1}^n \text{cov}^{(s)}(j)^{-1}},$$

当所有元素均满足 $\text{cov}^{(s)}(j) > 0$ 时。虽然这一表达式在实践中定义良好（因为实践中任意两点之间的亲和度始终为正），但理论上可能存在某些情况，例如 $E = 0$，其中一个或多个元素满足 $\text{cov}^{(s)}(j) = 0$。在这种情况下，$\text{cov}^{(s)}(j)^{-1} = \infty$，使得该表达式在标准意义下未定义。假设对于一部分索引，$\text{cov}^{(s)}(i) = 0$，而对于其余索引，$\text{cov}^{(s)}(i) > 0$。我们将采样分布 q^s 定义为

$$q_i^{(s)} := \begin{cases} \frac{1}{|\{j : \text{cov}^{(s)}(j) = 0\}|} & \text{if } \text{cov}^{(s)}(i) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

也就是说，我们在未覆盖点中均匀采样。这对应于未覆盖条目占主导时逆覆盖度分布的极限行为。为了建立直觉，下面考虑无噪声情形，此时 $E = 0$。

~**无噪声情形**。 假设 $E = 0$。因此，观测到的亲和矩阵 A 为

$$\widetilde{A} = A_0. \tag{2}$$

回顾一下，矩阵 $A_0 \in \mathbb{R}^{n\times n}$ 是块对角的，包含 K 个不相交的簇 $\mathcal{C}_1, \dots, \mathcal{C}_K$，即

$$A_0 := \operatorname{diag}(A^{(1)}, A^{(2)}, \dots, A^{(K)}),$$

其中每个对角块 $A^{(k)} \in \mathbb{R}^{n_k\times n_k}$ 表示簇 k 内的亲和度，并由 $A^{(k)} = p_k \cdot \mathbf{1}_{n_k\times n_k}$,给出，其中 $p_k > 0$ ，且 $\mathbf{1}_{n_k\times n_k} \in \mathbb{R}^{n_k\times n_k}$ 是全**1**矩阵。因此当 $E = 0$时，对于 $i \in [n]$ 和 $j \in [n]$

$$\widetilde{A}_{i,j} > 0 \quad \Longleftrightarrow \quad x_i, x_j \text{ belong to the same cluster.} \tag{3}$$

考虑一个采样块以及该块中的当前步骤 $s \in [K]$。将该块中抽取的索引序列记为$i_1, \dots, i_K$.

在无噪声情形下，由于 A_0的块对角矩阵结构，只有与先前抽样索引属于同一簇的点才会获得非零覆盖度。因此，在每一步骤 $s > 1$，未覆盖集合 $Z^{(s)}$ 由尚未从中抽取任何索引的簇中的点组成。

因此，采样分布就是在 $\{j$ 上均匀分布：$\mathrm{cov}^{(s)}(j) = 0\}$ 由下式给出

$$q_i^{(s)} := \begin{cases} \frac{1}{|\{j:\mathrm{cov}^{(s)}(j)=0\}|} & \text{if } \mathrm{cov}^{(s)}(i) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

令 $\mathcal{U}^{(s)}$ 表示截至步骤 s 尚未被采样的簇的集合，即:

$$\mathcal{U}^{(s)} := \{\, k \in [K] : \mathcal{C}_k \cap \{i_1, \dots, i_{s-1}\} = \varnothing \,\}.$$

根据构造， $\mathcal{U}^{(1)} = [K]$ ，并且每一步都会从 $\mathcal{U}^{(s)}$ 中移除一个新簇。因此，

$$|\mathcal{U}^{(s)}| = K - (s-1),$$

而步骤 s 的采样分布在这些剩余的 $|\mathcal{U}^{(s)}|$ 个簇的并集上均匀分布。经过 K 步后，$\mathcal{U}^{(K+1)} = \varnothing$，且每个簇恰好贡献一个样本。

每个块都会从每个 K 簇中采样一行。注意，共有 m/K 个块。因此，从每个簇中至少采样 t 个不同样本的问题，等价于在我们以有放回的方式从 t 次抽取中均匀随机抽样时，需要抽取多少次才能从每个簇中至少采样 m/K 个不同块的问题。这让人联想到经典的优惠券收集问题（Boneh 和 Hofri，1997年）。更正式地说，我们将陈述假设并证明以下结果。

引理 1.3 (部分目标的优惠券收集问题，(Boneh 和 Hofri，1997年))。设 $N \geq t$ 并从集合 $[N] = \{1, \dots, N\}$ 中有放回地抽取独立同分布的优惠券。设 S_t 为观察到 t 种不同优惠券所需的抽取次数。定义 N^{th} 调和数为

$$H_N = \sum_{i=1}^{N} 1/i$$

那么，

$$\mathbb{E}[S_t] = N(H_N - H_{N-t}) = N \sum_{i=N-t+1}^{N} \frac{1}{i} = t + \mathcal{O}\left(\frac{t^2}{N}\right).$$

此外，对于每个 $\gamma > 0$,

$$\Pr(S_t > (1+\gamma)N(H_N - H_{N-t})) \leq \exp\left(-\frac{\gamma^2 t}{2(1+\gamma)}\right). \tag{4}$$

证明：我们在此概述证明过程。完整证明请参见Boneh 和 Hofri （1997年）。
我们可以写成 $S_t = \sum_{j=0}^{t-1} X_j$，其中 X_j 是在收集到 j 个不同优惠券后观察到新优惠券所需的等待时间。每个 X_j 都服从成功概率为 $p_j = 1 - \frac{j}{N}$ 的几何分布，因此

$$\mathbb{E}[X_j] = \frac{1}{p_j}, \quad \mathrm{Var}(X_j) = \frac{1-p_j}{p_j^2} \le \frac{1}{p_j^2}.$$

然后根据线性性，

$$\mathbb{E}[S_t] = \sum_{j=0}^{t-1} \frac{1}{1-\frac{j}{N}} = N \sum_{i=N-t+1}^{N} \frac{1}{i}.$$

对几何随机变量应用 Bernstein 界或切尔诺夫界可得到尾部不等式。 □

在每个块中，我们从每个簇抽取一个样本，因此经过 m 次总抽样后，每个簇 k 会被抽样 $m_k = m/K$ 次。固定 $\gamma > 0$ 并设

$$m_k = \lceil (1+\gamma) n_k \left(H_{n_k} - H_{n_k - t}\right) \rceil, \tag{5}$$
$$m = K \cdot m_k. \tag{6}$$

那么，在假设 1.1下，

$$\Pr\left(\text{any cluster has fewer than } t \text{ distinct samples}\right) \le K \cdot \exp\left(-\frac{\gamma^2 t}{2(1+\gamma)}\right). \tag{7}$$

失败 概率变为 $O\left(\exp\left(-\gamma t\right)\right)$，其中 γ 为常数。由于 t 固定且 $n_k \to \infty$，我们有

$$n_k \left(H_{n_k} - H_{n_k - t}\right) = t + \mathcal{O}\left(\frac{t^2}{n_k}\right) = t\left(1 + o(1)\right). \tag{8}$$

因此，

$$m = K(1+\gamma)t(1+o(1)) = O(Kt), \tag{9}$$

个样本就足够了，也就是说，块的总数随 K （以及 t）线性增长即可，且与簇大小无关，从而以 $1 - \exp(-\gamma t)$ 的概率保证从每个簇中至少获得 t 个样本。
备注 1.4. 这一集中界也作为上界是紧的。特别是，我们可以预期，以高概率，从每个簇中抽取的样本数不会超过 $O(t)$。这一上界将在谱近似保证的证明中使用（定理2.5）。

含噪情形。 现在我们考虑含噪情形，其中 $E \neq 0$。这一设定更具挑战性，因为采样过程可能不再保证每个块中恰好包含来自每个簇的一个样本。为解决这一问题，我们证明，尽管存在噪声，每个块仍以一个常数概率（与 n 无关）至少包含来自每个簇的一个样本。这一关键性质使我们的分析能够以类似于无噪情形的方式继续进行。

现在，我们在假设 2.3 下分析含噪情形，其中噪声矩阵的各项满足 $E_{ij} = E_{ji} \sim \mathrm{Exp}(\lambda)$，且相互独立同分布，并且与采样过程相互独立。考虑 v 个连续采样块，每块大小为 K。令 T_k^v表示从簇 $\mathcal{C}_k$ 中跨这 v 个块抽取的不同样本数。

定义事件 $\mathcal{E}_k := \{T_k^v = 0\}$，即簇 $\mathcal{C}_k$ 在这 v 个块中从未被选中。令 $\mathcal{E} := \bigcup_{k=1}^{K} \mathcal{E}_k$ 表示某个簇完全未被选中的失败事件。
根据并集界，

$$\Pr(\mathcal{E}) \le K \cdot \max_{1 \le k \le K} \Pr(\mathcal{E}_k). \tag{10}$$

为界定 $\Pr(\mathcal{E}_k)$，注意到

$$T_k^v \geq \sum_{j=1}^{v} Z_j, \tag{11}$$

其中， $Z_j := \mathbf{1}\{$是在块 $j\}$ 中被选中的簇 $\mathcal{C}_k$ 的指示变量。这意味着

$$\Pr(\mathcal{E}_k) = \Pr\left(\sum_{j=1}^{v} Z_j = 0\right) = \Pr(Z_1 = 0)^v,$$

由于 Z_j 是独立同分布的。现在我们来计算 $\Pr(Z_1 = 0)$。这里

$$\begin{aligned}
&\Pr(Z_1 = 0) = \Pr(\mathcal{C}_k \text{ is not selected in block } 1) \\
&= \Pr\left(\bigcap_{s=1}^{K} \left\{ \mathcal{C}_k \text{ is not selected in step } s \,\middle|\, \mathcal{C}_k \text{ is not selected in steps 1 to } s-1 \right\}\right) \\
&\leq \Pr\left(\{\mathcal{C}_k \text{ is not selected in step } K | \mathcal{C}_k \text{ is not selected in steps 1 to } K-1\}\right).
\end{aligned} \tag{12}$$

定义

$$\Pr\left(\{\mathcal{C}_k \text{ is not selected in step } K | \mathcal{C}_k \text{ is not selected in steps 1 to } K-1\}\right) := r_k$$

固定簇 $\mathcal{C}_k$，并以该簇在一个块的前 $K-1$ 个步骤中未被选中为条件。令 $\mathcal{S}_{\text{past}} = \{i_1, \ldots, i_{K-1}\}$表示迄今为止的抽样索引，其中 $\mathcal{S}_{\text{past}} \cap \mathcal{C}_k = \emptyset$。回顾步骤 K 时的覆盖度：

$$\text{cov}^{(K)}(i) := \sum_{j \in \mathcal{S}_{\text{past}}} \widetilde{A}_{ji} = \sum_{s=1}^{K-1} (A_{0,\, i_s i} + E_{i_s i}).$$

由于 A_0 是块对角的，且 $\mathcal{C}_k$ 此前未被选中，因此我们有：

$$\text{cov}^{(K)}(i) = \begin{cases} \sum_{s=1}^{K-1} E_{i_s i} := S_i, & i \in \mathcal{C}_k, \\ \sum_{s=1}^{K-1} (p_{c(i)} + E_{i_s i}) = S_i + p_{c(i)}, & i \notin \mathcal{C}_k, \end{cases}$$

其中， $p_{c(i)} \in [p_{\min}, p_{\max}]$ 和 $S_i \sim$ 伽马分布$(K-1,\ \lambda)$在 i上相互独立。

备注 1.5。 实践中，在计算 p,f 时，也应考虑当前块之外各行的噪声贡献。不过，这些项可以直接纳入计算，且不会影响核心论证。为简洁起见，主文的分析中省略了这些项。

步骤 K 的逆覆盖采样概率为

$$q^{(K)}(i) = \frac{\text{cov}^{(K)}(i)^{-1}}{\sum_{j=1}^{n} \text{cov}^{(K)}(j)^{-1}}.$$

因此， $\mathcal{C}_k$ 在步骤 K 未被选中的概率（即条件漏选概率）为

$$r_k \leq \mathbb{E}\left[\frac{\sum_{i \notin \mathcal{C}_k} (S_i + p_{c(i)})^{-1}}{\sum_{i \in \mathcal{C}_k} S_i^{-1} + \sum_{i \notin \mathcal{C}_k} (S_i + p_{c(i)})^{-1}}\right]. \tag{13}$$

利用 $p_{c(i)} \in [p_{\min}, p_{\max}]$，我们得到以下界

$$r_k \leq \mathbb{E}\left[\frac{\sum_{i \notin \mathcal{C}_k} (S_i + p_{\min})^{-1}}{\sum_{i \in \mathcal{C}_k} S_i^{-1} + \sum_{i \notin \mathcal{C}_k} (S_i + p_{\min})^{-1}}\right]. \tag{14}$$

对于所有 $i \notin \mathcal{C}_k$，我们有 $(S_i + p_{\min})^{-1} \le p^{-1}$ 最小值。因此，

$$\sum_{i \notin \mathcal{C}_k} (S_i + p_{\min})^{-1} \le \frac{n - n_k}{p_{\min}} \le \frac{n}{p_{\min}}.$$

现在我们对 $\sum_{i \in \mathcal{C}_k} S_i^{-1}$ 给出下界。由于 $S_i \sim$ 伽马分布$(K-1,\ \lambda)$，因此对于 $K \ge 3$，有 $\mathbb{E}[S_i^{-1}] = \lambda/(K-2)$。于是，

$$\mathbb{E}\left[\sum_{i \in \mathcal{C}_k} S_i^{-1}\right] = \frac{n_k \lambda}{K-2}.$$

引理 1.6（逆伽马和的集中性）。设 S_1、…、S_{n_k} 是独立同分布随机变量，且 $S_i \sim$ 服从伽马分布(α, λ)，其中 $\alpha > 2$ 且 $\lambda > 0$。那么，以至少 $1 - \exp(-cn_k)$ 的概率，对于某个仅依赖于 α 的常数 $c > 0$，有

$$\sum_{i=1}^{n_k} \frac{1}{S_i} \ge \frac{1}{2} \cdot \frac{n_k \lambda}{\alpha - 1}.$$

这里用到了以下事实：$\mathbb{E}[S_i^{-1}] = \frac{\lambda}{\alpha-1}$，且 $\mathrm{Var}(S_i^{-1}) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}$；并对独立同分布的次指数变量之和应用标准集中性。由上述集中界可得，

$$\sum_{i \in \mathcal{C}_k} S_i^{-1} \ge \frac{1}{2} \cdot \frac{n_k \lambda}{K-2}.$$

在分母至少为 $2\,\frac{n_k \lambda}{K-2}$（）的高概率事件上，我们得到：

$$r_k \le \frac{\frac{n}{p_{\min}}}{\frac{n_k \lambda}{2(K-2)}} + \frac{\delta}{2} = \frac{2n(K-2)}{p_{\min} \lambda n_k} + \frac{\delta}{2}.$$

为确保 $r_k \le \delta$，只需满足

$$\frac{2n(K-2)}{p_{\min} \lambda n_k} \le \frac{\delta}{2} \quad \Rightarrow \quad p_{\min} \lambda \ge \frac{4n(K-2)}{\delta n_k},$$

这由假设1.2可知成立

现在固定 $v = C \log K$。于是，

$$\Pr(\mathcal{E}_k) \le \delta^{\log K} = K^{\log \delta} = K^{-\gamma}, \quad \text{where } \gamma := -C \log \delta > 0.$$

选择 C 使得 $-C \log \delta > 1$。根据并集界，

$$\Pr(\mathcal{E}) \le K \cdot \max_k \Pr(\mathcal{E}_k) \le K \cdot K^{-\gamma} = K^{1-\gamma}.$$

由于 $\gamma > 1$，我们有 $\Pr(\mathcal{E}) \to O(K^{1-\gamma})$。

在存在噪声的情况下，大小为 K 的每个块不再保证恰好从每个簇中抽取一个样本。然而，我们证明了以高概率而言，$C \log K$ 个这样的块足以覆盖所有 K 个簇。

一旦某个簇在一个块中被选中，从该簇抽取的样本便是从其 n_k 个成员中均匀抽取的。因此，给定某个簇被选中这一条件后，从该簇抽取行的过程相当于从 $[n_k]$ 中进行有放回的均匀抽样。

现在，我们在每个簇内重复优惠券收集论证。令 $m_k = C \log K$ 表示簇 $\mathcal{C}_k$ 被选中的块数（因此 m_k 是随机的，但以高概率，对所有 k，都有 $m_k \ge C \log K$）。

为了获得 t 来自每个簇的不同行，只需确保至少有 t 个不同元素被均匀地抽取自$[n_k]$ 经过 m_k 次这样的试验。这与引理1.3中的情形相同。根据该结果，只需满足

$$m_k \ge (1+\gamma) n_k (H_{n_k} - H_{n_k - t}) \approx (1+\gamma) t$$

失败概率至多为 $\exp(-\gamma t/2)$。由于 $m_k = C\log K$ 以高概率成立，且 t 是固定的，因此对于充分大的 K，该条件成立。

综合以上结果，我们得出

$$m = K \cdot C(1+\gamma)t\log K$$

总计只需 t 个样本，就能以高概率从每个簇中获得互不相同的行，即使存在噪声也是如此。这与无噪声情形相似，但为确保覆盖所有簇，我们现在需要 $\log K$ 个块（而非直接需要 K 个样本）。

备注 1.7。 相比之下，在没有任何自适应结构的均匀独立同分布采样下，单次抽样从特定簇 $\mathcal{C}_k$ 中选取一个样本的概率与该簇的大小成正比，即 n_k/n。

对于较小的簇，这一概率可能小到趋近于零，因此在固定大小的块中，甚至很难从此类簇中获得一个样本。因此，要确保覆盖所有簇，就需要显著增加样本数量，从而导致样本复杂度保证大幅减弱，也更难扩展。

备注 1.8。 不难证明，以高概率，每个块都会产生一个独一无二的采样索引。这证明了累积覆盖所含噪声相互独立这一假设的合理性。

□

B. 定理 2.5 的证明

定理 2 （对 A_0 的谱近似）。 设 $\widehat{A} = \widetilde{A}_S^{\top}\widetilde{A}_S^{+}{}_{,S}\widetilde{A}_S$ 表示由对 A 的 m（$\geq C$（δ）$tK\log K$）自适应采样的行进行低秩重构$\widetilde{}$所得到的估计量。则，

$$\|\widehat{A} - A_0\|_{\mathrm{op}} \leq \frac{Dn}{\sqrt{t}\cdot \min_k p_k}, \tag{15}$$

以高概率成立，其中 p_k 是簇 $\mathcal{C}_k$ 的簇亲和度，且 $D := D$（λ、K、$\vec{p}$）> 0 取决于噪声矩阵的参数 λ、簇的数量 K 以及亲和度向量 $\vec{p}$。

证明。 为建立直观理解，我们首先考虑 $E = 0$ 的情形。在此情形下

$$\widehat{A} = A_0$$

当每行有 t 个可用样本时，在我们的块结构假设下，可以说 $\mathrm{rank}(A_S) = K$，即不同簇的数量。下述引理使我们能够据此完美重构 A_0。

引理 2.1 (Owen 和 Perry (2009))。 设$\mathbf{M} \in \mathbb{R}^{n\times n}$ 是一个矩阵，并假设$\mathbf{M}_1 \in \mathbb{R}^{s\times n}$ 由 s 行构成，这些行是$\mathbf{M}$ 的一个子集，且$\mathrm{rank}(\mathbf{M}_1) = \mathrm{rank}(\mathbf{M})$。令$\mathbf{M}_{11} \in \mathbb{R}^{s\times s}$ 表示由$\mathbf{M}_1$形成的对应子矩阵。于是，

$$\mathbf{M} = \mathbf{M}_1^{\top}\mathbf{M}_{11}^{+}\mathbf{M}_1.$$

取 $M_1 = A_S$ 和 $M_{11} = A_{S,\,S}$。以高概率，定理 2.4 推出 $\mathrm{rank}(A_{S,\,S}) = \mathrm{rank}(A) = K$. 由此可得结果：$A_0 = \widetilde{A}_S^{\top}\widetilde{A}_S^{+}{}_{,S}\widetilde{A}_S$，且该结论以高概率成立。

$\widetilde{}$ $\widetilde{}$现在我们转向含噪情形，此时 $E \neq 0$。在这种情况下，我们将 A_S 和 $A_{S,\,S}$ 的无噪版本定义为 $A_{S,\,S}$和 A_S

定义：

$$A_S = A_0[S,:], \qquad A_{S,S} = A_0[S,S],$$
$$E_S = \widetilde{A}_S - A_S, \qquad E_{S,S} = \widetilde{A}_{S,S} - A_{S,S}.$$

然后 $\widehat{A} = \widetilde{A}_S^{\top}\widetilde{A}_S^{+}{}_{,S}\widetilde{A}_S = (A_S + E_S)^{\top}(A_{S,S} + E_{S,S})^{+}(A_S + E_S)$。

令 $M := A_{S,S} + E_{S,S}$。展开并重新分组后，

$$\begin{aligned}\widehat{A} - A_0 &= A_S^\top (M^+ - A_{S,S}^+) A_S + A_S^\top M^+ E_S + E_S^\top M^+ A_S + E_S^\top M^+ E_S \\ &= T_1 + T_2 + T_3 + T_4.\end{aligned}$$

取算子范数，

$$\|\widehat{A} - A_0\| \le \|A_S\|^2 \cdot \|M^+ - A_{S,S}^+\| + 2\|A_S\| \cdot \|M^+\| \cdot \|E_S\| + \|E_S\|^2 \cdot \|M^+\|.$$

现在我们对每一项进行界定。首先分析分块常数模型下矩阵 A_S 和 $A_{S,S}$ 的结构。由于 A_0 是一个具有 K 个簇的块对角矩阵，且每个块 $A^{(k)} = p_k \cdot \mathbf{1}_{n_k \times n_k}$，因此 A_S 中对应于簇 $\mathcal{C}_k$ 的任意一行 i 都具有如下形式

$$(A_S)_{i,j} = \begin{cases} p_k & \text{if } j \in \mathcal{C}_k, \\ 0 & \text{otherwise.} \end{cases}$$

因此，$\|(A_S)_{i,:}\|_2 = p_k \cdot \sqrt{n_k} \le p_{\max}\sqrt{n}$，因为 $n_k \le n$ 且 $p_k \le p_{\max}$，其中 $\|\|_2$ 为欧几里得范数。

每个簇有 t 行，因此对于 K 个簇，$A_S \in \mathbb{R}^{tK \times n}$。请记住，该上界是紧的（参见定理 2.5证明后的备注）。使用该界

$$\|A_S\| \le \max_i \|(A_S)_{i,:}\|_2 \cdot \sqrt{tK} \le p_{\max}\sqrt{n} \cdot \sqrt{tK} = p_{\max}\sqrt{tn}.$$

接下来，考虑 $A_{S,S}$，即 A_0 在行 S上的 $tK \times tK$ 子矩阵。$A_{S,S}$ 由 K 个互不相交的块组成，每个块的大小为 $t \times t$ 且等于 $p_k \cdot \mathbf{1}_{t\times t}$。每个这样的块的秩为 $\mathbf{1}$，最大特征值为 tp_k。

因此，

$$\sigma_{\min}(A_{S,S}) = tp_{\min}, \quad \sigma_{\max}(A_{S,S}) = tp_{\max}, \quad \|A_{S,S}^+\| = \frac{1}{tp_{\min}}.$$

噪声界。 我们对具有独立、均值为零的次指数条目的矩形矩阵使用矩阵Bernstein不等式（Bernstein，2009）。

引理2.2（矩阵Bernstein不等式，矩形情形（Bernstein，2009））。 设 $Z \in \mathbb{R}^{m\times n}$ 为一个条目相互独立且均值为零的随机矩阵： $\mathbb{E}[Z_{ij}] = 0$ 对所有 i， j。假设每个条目都是次指数的，且

$$\|Z_{ij}\|_{\psi_1} \le v \quad \textit{for all } i, j,$$

其中，$\|\cdot\|_{\psi_1}$ 表示与次指数函数 $\psi_1(x) = \exp(x) - 1$相关的*Orlicz*范数：

$$\|X\|_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}[\exp(|X|/t) - 1] \le 1\right\}.$$

定义行/列方差代理

$$\sigma^2 := \max\left\{\max_i \sum_{j=1}^n \mathbb{E}[Z_{ij}^2], \ \max_j \sum_{i=1}^m \mathbb{E}[Z_{ij}^2]\right\}.$$

Then 存在一个绝对常数 $c > 0$，使得对所有 $t > 0$， Z 的算子范数满足 s

$$\mathbb{P}\left(\|Z\| > t\right) \le (m+n)\exp\left(-c\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{v}\right\}\right).$$

E_S 和 $E_{S,\ S}$ 的各项均为中心化指数随机变量，其 ψ_1 范数为 $O(1/\lambda)$，方差为 $O(1/\lambda^2)$。因此，以高概率，

$$\|E_S\| \le \frac{C\sqrt{n}}{\lambda}, \quad \|E_{S,S}\| \le \frac{C\sqrt{s}}{\lambda} = \frac{C\sqrt{tK}}{\lambda}.$$

Neumann 级数展开 展开。 若 $\|E_{S,\ S}\| < \sigma_{\min}(A_{S,\ S})$，则 $M = A_{S,\ S} + E_{S,\ S}$ 可逆，我们可以展开：

$$M^{-1} = A_{S,S}^{-1} \sum_{k=0}^{\infty} \left(-E_{S,S} A_{S,S}^{-1}\right)^k,$$

其在$\|E_{S,S} A_{S}^{-1}{}_{,S}\| < 1$时收敛。

由于$\|A_{S}^{-1}{}_{,S}\| = 1/(tp_{\min})$，只需满足

$$\|E_{S,S}\| < \frac{1}{2} \cdot \frac{1}{\|A_{S,S}^{-1}\|} = \frac{tp_{\min}}{2},$$

当 $t \geq CK/(\lambda^2 p_{\min}^2)$时，该条件成立。

因此，

$$\|M^+\| \leq \frac{2}{tp_{\min}}, \quad \|M^+ - A_{S,S}^+\| \leq \|A_{S,S}^+\|^2 \cdot \|E_{S,S}\| \cdot \sum_{k=0}^{\infty} \|E_{S,S} A_{S,S}^+\|^k \leq \frac{2\|E_{S,S}\|}{(tp_{\min})^2}.$$

代入噪声界可得：

$$\|M^+ - A_{S,S}^+\| \leq \frac{2C\sqrt{K}}{t^{3/2} p_{\min}^2 \lambda}.$$

综合以上各项。

$$\|T_1\| \leq \|A_S\|^2 \cdot \|M^+ - A_{S,S}^+\| \leq p_{\max}^2 tn \cdot \frac{2C\sqrt{K}}{t^{3/2} p_{\min}^2 \lambda} = \frac{2Cp_{\max}^2 n\sqrt{K}}{p_{\min}^2 \lambda \sqrt{t}},$$

$$\|T_2\| = \|T_3\| \leq \|A_S\| \cdot \|M^+\| \cdot \|E_S\| \leq p_{\max}\sqrt{tn} \cdot \frac{2}{tp_{\min}} \cdot \frac{C\sqrt{n}}{\lambda} = \frac{2Cp_{\max} n}{p_{\min}\lambda\sqrt{t}},$$

$$\|T_4\| \leq \|E_S\|^2 \cdot \|M^+\| \leq \left(\frac{C\sqrt{n}}{\lambda}\right)^2 \cdot \frac{2}{tp_{\min}} = \frac{2C^2 n}{\lambda^2 p_{\min} t}.$$

最后一项为低阶项，因此我们将其吸收进常数中。

定义：

$$D(\lambda, K, \vec{p}) := \frac{2Cp_{\max}^2\sqrt{K}}{\lambda} + \frac{4Cp_{\max}}{\lambda} + \frac{2C^2\sqrt{K}}{\lambda^2}.$$

那么，以很高的概率，

$$\|\widehat{A} - A_0\| \leq \frac{D(\lambda, K, \vec{p}) \cdot n}{\sqrt{t} \cdot p_{\min}}.$$

□

C. 算法

算法2 概述了用于批次校正的扩展 BALANS 流程。该算法的输入包括数据矩阵$X \in \mathbb{R}^{n\times d}$、相应的批次标签 $\vec{b}$以及三个超参数：局部邻域大小 k、用于控制收敛的停止阈值 τ和用于自适应采样的块长度 J。

总体而言，BALANS 通过选择性采样信息丰富的行，迭代构建稀疏亲和矩阵。采样以大小为 J的块为单位进行，并使用临时覆盖向量 c_J，优先采样数据中代表不足的区域。对于每个采样的索引，BALANS 使用批次条件局部尺度核计算一行亲和度（见算法3）。当连续 τ 个块均未引入任何新的亲和度时，采样便会停止。

达到足够的覆盖度后，低秩亲和矩阵补全过程（算法4）会重建完整的亲和$\widehat{}$矩阵，再据此获得校正后的表示 X。

算法 2 BALANS 批次校正

输入： 数据 $X \in \mathbb{R}^{n\times d}$，批次标签 $\vec{b} \in \{1,\dots,B\}^n$，最近邻参数 k，停止准则 τ，重置间隔 K
输出： 校正数据矩阵 $\widehat{X} \in \mathbb{R}^{n\times d}$
对 X 执行主成分分析（可选）
初始化累积覆盖度 $c \leftarrow \mathbf{0}$，采样索引集 $S \leftarrow \emptyset$
初始化临时覆盖率 $c_J \leftarrow \mathbf{0}$
初始化未变化计数 $\leftarrow 0$
当未变化计数 $< \tau$**时**
    对 $j = 1$到 J**执行**
        使用自适应采样（算法 i_t （算法5）
        计算稀疏亲和度行$A_{i_t,:}$ 使用局部尺度核（算法 3）
        $A_S \leftarrow$ 追加$A_{i_t,:}$ 作为新行
        令 $\Delta \leftarrow$ 满足以下条件的索引数量：$A_{i_t,:} > 0$ 和 $c_J = 0$
        如果 $\Delta = 0$**则**
            未变化计数 $\leftarrow$ 未变化计数 $+ 1$
        else
            未变化计数 $\leftarrow 0$
        结束如果
        更新累积覆盖度：$c \leftarrow c + A_{i_t,:}$
        更新临时覆盖率：$c_J \leftarrow c_J + A_{i_t,:}$
        $S \leftarrow S \cup \{i_t\}$
    结束 for
    重置临时覆盖率： $c_J \leftarrow \mathbf{0}$
结束 while
使用低秩亲和度补全（算法4）以获得校正后的 $\widehat{X} \in \mathbb{R}^{n\times d}$
返回 $\widehat{X}$

C.1. 批次依赖的局部尺度计算

算法3 详细说明了如何计算以数据点 x_i 为中心的单个稀疏亲和度行，并纳入批次依赖的局部尺度。对于每个批次，使用该批次中第 k 个最近邻的距离来估计单独的核带宽 $\sigma_{i,b'}$。随后，根据每个数据点的批次标签为其分配相应的批次依赖尺度，并据此计算亲和度。然后使用肘部阈值法对所得亲和度行进行稀疏化，即检测亲和度急剧下降的截断点。具体而言，我们使用 ruptures 库中的滑动窗口变点检测算法：按降序排列亲和度值，并在该序列上滑动窗口。在每个步骤中，我们计算候选变点前后双段分段常数拟合的均方误差。在每个步骤中，我们计算候选变点前后的段内总方差，并寻找超过基于五分位数的阈值的下降幅度（扫描过程中所有此类下降幅度的第80百分位数）。这种批次感知设计在降低亲和度计算中的批次间偏差的同时，保留了局部几何结构。

算法 3 单行的批次条件局部带宽亲和度

输入： 采样行 $x_i \in \mathbb{R}^d$、数据矩阵 $X \in \mathbb{R}^{n\times d}$、批次标签 b、带宽参数 k
输出： 稀疏亲和度行$A_{i,:} \in \mathbb{R}^n$
计算距离向量：$d_j \leftarrow \|x_i - x_j\|^2$ 对于所有 $j = 1$ 至 n
对于每个批次 b'，计算 $\sigma_{i,b'} \leftarrow \|x_i - x_{b'}^{(k)}\|$ ▷ 批次 b' 中第 k 个最近邻
将 $\sigma_{i,j} \leftarrow \sigma_{i,b_j}$ 广播到所有 j
为所有 j 计算 $A_{i,j} \leftarrow \underline{\exp}\left(-\frac{d_j}{\sigma_{i,j}^2}\right)$
使用肘部阈值法稀疏化$A_{i,:}$。
返回 $A_{i,:}$

C.2. 低秩平滑变换

算法4介绍了用于从稀疏采样亲和矩阵 X 中恢复完整校正数据矩阵 A_S的低秩补全步骤。该方法无需构建完整的 $n \times n$ 亲和矩阵，而是利用矩阵 A_r进行高效投影。该过程首先通过 $A_r \cdot X$将数据映射到低维子空间，然后使用 $A_r^\top$进行反投影，以获得平滑重构。最后进行重归一化，确保有效传播矩阵 $P = A_r^\top A_r$ 为行随机矩阵。

算法4 低秩平滑

输入： 稀疏亲和矩阵 $A_S \in \mathbb{R}^{s\times n}$，数据矩阵 $X \in \mathbb{R}^{n\times d}$
输出： 补全后的数据矩阵 $\widehat{X} \in \mathbb{R}^{n\times d}$
$r \leftarrow$ A_S的行和向量 ▷ $r_i = \sum_j A_S[i,j]$
$D_r \leftarrow$diag$(1./r)$ ▷ 行归一化缩放
$A_r \leftarrow D_r \cdot A_S$ ▷ 每一行的 A_r 之和为 $\mathbf{1}$
$Y \leftarrow A_r \cdot X$ ▷ 将数据投影到低秩子空间
$c \leftarrow A_r^\top \cdot \mathbf{1}_s$ ▷用于重归一化的列权重
$Z \leftarrow A_r^\top \cdot Y$ ▷ 反投影到全维度
$\widehat{X} \leftarrow$ diag$(1./c) \cdot Z$ ▷对行进行归一化，使 $P = A_r^\top A_r$ 成为行随机
返回 $\widehat{X}$

C.2.1. 基于覆盖的自适应采样

算法 5 根据覆盖向量在每个采样步中选择信息丰富的行 c。当存在未覆盖点（即具有 $c_j = 0$ 的条目）时，算法会在这些点中进行均匀采样，以确保覆盖扩展。否则，它会分配与覆盖度成反比的采样权重，优先选择采样频率较低的点。

算法5 基于覆盖的自适应采样

输入： 覆盖向量 $c \in \mathbb{R}^n$
输出： 在步骤 t采样的索引 i_t
$Z \leftarrow \{j \in [n] : \ c_j = 0\}$ ▷零覆盖的索引
如果 $|Z| > 0$，**则**
    若 $j \in Z$则设置 $p_j = \frac{1}{|Z|}$，否则 $p_j = 0$
否则
    对每个 $j \in \{1, \ldots, n\}$ **执行**
        计算权重 $w_j = \frac{1}{c_j}$
    结束循环
    归一化权重：$p_j = \frac{w_j}{\sum_k w_k}$
结束 if
从 $i_t \sim$ 分类分布$(p_1, \ldots, p_n)$中采样
返回 i_t

D. 运行时间

定理 3 (BALANS 的计算复杂度)。 设 n 为数据点数量， d 为特征维度，$m = |S|$ 为 *BALANS* 收敛后自适应采样的行数。那么BALANS 的总计算复杂度为 $\mathcal{O}(nm(d + \log n))$，其中第一项表示稀疏亲和度构建和矩阵乘法的开销，第二项表示最近邻搜索的开销。当 $m \ll n$ 时，该复杂度近似与 n 呈线性关系。

证明。 设 $m = |S|$ 为满足停止准则前选取的索引数量。令 k 表示亲和度计算中使用的局部邻域大小（复杂度分析中假设其为常数）。我们考察算法 2–4的各个主要步骤：

在每个 m 采样步中，算法执行以下操作：

- 采样索引： 从包含 n 个元素的自适应分布中抽取样本，耗时为 $O(n)$。
- 距离计算： 计算所有 $j \in [n]$ 的 $\|x_i - x_j\|^2$ 需要 $O(nd)$ 时间。
- 局部带宽估计： 对每个批次内的距离进行排序以找到第 k 个近邻，总共需要 $\mathcal{O}(n\log\ n)$ 时间（各批次共享）。
- 亲和度计算： 应用核并进行归一化需要 $\mathcal{O}(n)$ 时间，这部分开销包含在排序中。
- 肘部稀疏化： 通过在排序后的亲和度上使用滑动窗口进行变点检测，采用基于方差的分段和提前停止策略，实际每行耗时 $\mathcal{O}(K)$，最坏情况下为 $\mathcal{O}(n)$。

因此，每次迭代耗时 $\mathcal{O}(n(d + \log n))$，而 m 次这样的迭代耗时

$$\mathcal{O}(mn(d + \log n)).$$

给定稀疏矩阵 $A_S \in \mathbb{R}^{m\times n}$，将执行以下矩阵运算：

- 行归一化：计算行和并对 A_S 进行归一化，在最坏情况下的代价为 $\mathcal{O}(mn)$。然而，由于稀疏化后 A_S 的每一行最多包含 K 个非零元素，实际代价为 $\mathcal{O}(mK)$。
- 投影：计算 $Y = A_r X$ 的最坏情况代价为 $\mathcal{O}(mnd)$，但由于稀疏性，实际代价仅为 $\mathcal{O}(mKd)$。
- 反向投影：计算 $Z = A_r^\top Y$ 的实际代价同样为 $\mathcal{O}(mKd)$。
- 最终归一化：用 $\mathrm{diag}(1./c)$ 对 Z 进行缩放需要访问 Z 中的所有元素，代价为 $\mathcal{O}(nd)$。

因此，尽管最坏情况复杂度为 $\mathcal{O}(mnd)$，但由于亲和度行的稀疏性受到控制，实际运行时间通常要低得多——约为 $\mathcal{O}(mKd)$。

3. 主成分分析（可选）。若在开始时对 X 进行一次主成分分析以将维度降至 d_{PCA}，则通过随机奇异值分解的代价为 $\mathcal{O}(ndd_{\text{PCA}})$。

4. 总复杂度。 合并所有主导项：

$$\mathcal{O}(mn(d + \log n)) + \mathcal{O}(mnd) = \mathcal{O}(mn(d + \log n)).$$

由于 $m \ll n$ 基于覆盖度的收敛准则，总体成本与 n近似呈线性关系。 □

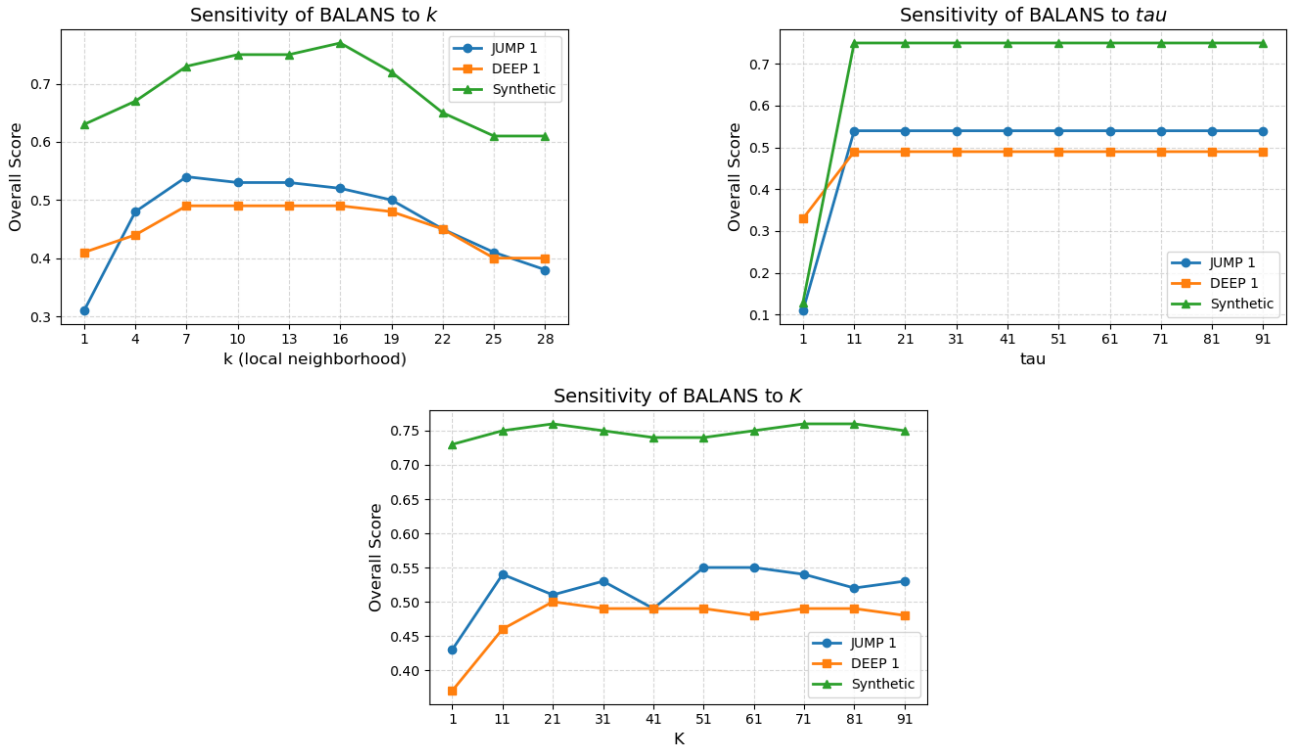


图 4. BAL 的敏感性　　ANS 对超参数 k 的敏感性，涵盖三个数据集：JUMP 1、DEEP 1 和合成数据集。每条曲线展示了　　e 不同取值下的总体得分　　k，表明该方法在较大范围内具有稳健性。

E. 消融实验

我们评估了 BALANS 对三个主要超参数的敏感性：（a）用于计算局部带宽的最近邻参数 k，（b）停止准则 τ，定义为采样集没有变化的连续自适应采样轮次，以及（c）块长度 K，它决定了每个自适应采样块中抽取的样本数量。

为评估稳健性，我们在真实世界数据集 JUMP 1 和 DEEP 1 以及合成数据集上开展实验。真实世界基准的详细信息见附录 I。对于合成数据设置，我们使用附录J中所述的结构化高斯混合模型，其中包含 50,000 个数据点、100 种化合物和 5 个批次。

对于每种设置，我们系统地改变一个超参数，同时固定其他超参数，并根据总体得分报告性能得分。

我们发现，BALANS 在各种超参数设置下始终表现出色。在真实场景和合成场景中均观察到这种稳健性，表明该方法在所有超参数的宽泛取值范围内对参数调整具有稳健性。结果汇总见图 4。

根据这些结果，我们建议在实践中采用以下默认范围：

$$k \in [5, 15], \qquad \tau \in [20, 100], \qquad K \in [20, 50].$$

注： 除非另有说明，我们在所有实验中均使用 $k = 5$、$\tau = 50$和 $K = 50$　　s.

F. 实验详情

所有实验均使用 4 个英伟达图形处理器（A100 或 V100）运行，每种方法的最长实际运行时间限制为 12 小时。对于大规模数据集，部分基线方法（典型相关分析、稳健主成分分析）无法在此时间限制内完成，因此在适用情况下将其报告为超时。

我们在各种真实世界数据集和合成数据集上评估性能。有关完整的数据集描述，包括 CellProfiler和 DeepProfiler特征、化合物和批次参数以及数据维度的详细信息，请参阅附录I 和附录J 。所有必要的预处理和后处理步骤均在相应附录中说明。

BALANS 使用 Python 实现，并采用高效的 NumPy 和 SciPy 程序。所有亲和度计算和矩阵运算均使用由 ‘ scipy.sparse‘ 提供的压缩稀疏行（CSR）格式，通过稀疏矩阵代数完成。自适应采样使用 NumPy 向量化运算实现，亲和度稀疏化则利用通过 ‘pd‘ 计算的最近邻距离。除非另有说明，我们使用 $k = 5$ （用于局部带宽估计的邻居数）、$\tau = 50$（基于采样行收敛情况的提前停止阈值）和 $K = 50$ （自适应采样的块大小）。所有实验均通过 Snake make 工作流运行，以实现任务编排和可复现性，并包括自动日志记录、检查点保存和内存管理。我们的实验未使用主成分分析（PCA），不过代码提供了使用该方法的功能。我们发现，使用 PCA（保留所有维度时）不会改变结果。

我们将 BALANS 与多种具有竞争力的批次校正方法进行比较。附录 G 提供了完整的实现细节，包括超参数配置、提前停止准则，以及各基线方法所采用的评估方案。对于这些方法，我们采用了作者推荐的默认设置，这些设置主要取自（ Arevalo等人，2024年）。
所有评估指标均在附录 H 中定义。这些指标包括生物信号保留和批次校正的相关指标。

G. 评估方法

表 3. 基线与竞争批次校正方法的比较。

方法	类别	需要批次标签	需要重新计算
BALANS	核 + 稀疏采样	Yes	No
Combat（Zhang 等人，2020）	统计	Yes	No
球化（Kessy 等人，2018年）	统计	No	No
典型相关分析（Satija 等人，2015）	线性投影	Yes	No
稳健主成分分析（Satija 等人，2015）	线性投影	No	No
fastMNN（Haghverdi 等人，2018年）	基于图 / K近邻	Yes	Yes
Scanorama（Hie 等人，2019年）	基于图 / K近邻	Yes	Yes
Harmony（Korsunsky 等人，2019年）	基于图 / K近邻	Yes	Yes

我们全面回顾了批次校正方法，重点关注为单细胞RNA测序（scRNA-seq）开发的方法，因为该领域的文献更为成熟且内容广泛。基于一项近期的基准分析（Luecken 等人，2022），我们选取了符合以下标准的代表性高性能方法子集：(i) 可通过 Python 或 R 使用；(ii) 根据引用次数具有较高知名度；(iii) 涵盖多种方法学途径。这些方法包括线性模型、混合模型和基于最近邻的技术。表3 总结了这些方法。

线性方法

- **Combat** (Zhang 等人，2020年）：将批次效应建模为加性和乘性噪声，并使用贝叶斯框架通过线性建模拟合并校正这些效应。没有需要调优的超参数。
- **球化**(Kessy 等人，2018年）：将根据阴性对照样本推导出的白化变换应用于整个数据集。该方法要求每个批次都包含阴性对照。我们对正则化参数进行了网格搜索，并根据总体得分选取了表现最佳的取值。

基于混合模型的方法

- **Harmony** (Korsunsky 等人，2019年）：交替进行数据聚类，将数据划分为批次多样化的组，并通过期望最大化（EM）框架在簇内进行校正。关键超参数包括簇的数量和多样性惩罚参数 λ。我们采用了（Arevalo等人，2024年）使用的超参数设置，该研究建议采用默认值。

基于最近邻的方法

- **fastMNN** (Haghverdi 等人，2018年）：fastMNN 是 MNN 的一个更快变体，它在最近邻计算之前进行主成分分析，从而更好地适用于大型数据集。关键超参数包括互为最近邻的数量（k）以及降维过程中保留的主成分数量（d）。我们采用了（Arevalo等人，2024年）中使用的设置。

- **Scanorama** （Hie 等人，2019年）：在所有数据集的低维空间中进行近似最近邻搜索，放宽了对批次对之间存在共享亚群体的假设。关键超参数包括邻居数量（k）以及整合的目标维度。默认值由 Arevalo等人 （2024年）提供。

- **Seurat-CCA** （Satija 等人，2015）：采用典型相关分析来寻找联合低维空间，并假设不同批次间的细胞类型有较大重叠。超参数包括典型相关向量的数量（`num.cc`）和整合锚点的数量（`k.anchor`）。我们采用了（Arevalo等人，2024年）建议的取值。

- **Seurat-RPCA** （Satija 等人，2015）：采用互惠主成分分析，允许批次间存在更大的异质性，并在大型数据集上提供更短的运行时间。主要超参数是主成分数量（`n.pcs`）和整合锚点数量（`k.anchor`），两者均设为作者推荐的默认值。

这些方法在假设、计算需求和输入依赖方面差异很大。大多数方法需要明确的批次标签，但**球化**除外，它依赖于阴性对照。四种方法（**fastMNN**、**MNN**、**Harmony**、**Scanorama**）在添加新样本时需要在整个数据集上重新计算批次校正。表3 总结了这些方法。

H. 指标

表 4. 本研究中使用的评估指标摘要。

指标	描述	衡量
图连通性 (Luecken 等人，2022)	校正后同一批次细胞的连通性	批次
kBET(B¨uttner 等人，2019年)	局部邻域中的批次混合（统计检验）	批次
LISI（批次）（Simpson，1949年）	基于邻域多样性的局部批次熵	批次
轮廓系数（批次）（Rousseeuw，1987年）	嵌入空间中各批次之间的分离	批次
LISI（标签）（Simpson，1949年）	用于评估生物学聚类的局部标签熵	Bio
轮廓系数（标签）（Rousseeuw，1987年）	生物学类别或标签的分离程度	Bio
Leiden 调整兰德指数（Traag 等，2019年）	与真实生物学标签的一致性	Bio
Leiden NMI（Traag 等人，2019年）	与真实标签的归一化互信息	Bio

为评估批次校正方法在基于图像的表型分析中的性能，我们采用一套共十项指标，分为两类：一类用于衡量批次去除效率，另一类用于评估生物变异的保留。这些指标改编自单细胞 RNA 测序领域的先前工作（Luecken 等人，2022年），并使用`scib` 软件包（Luecken 等人，2022年）。每项指标均归一化至 $[0, 1]$ 范围，其中 1 表示理想性能。表4汇总了这些指标。

批次去除指标

- **批次轮廓系数（ASW-Batch）**（Rousseeuw，1987年）：使用批次标签作为簇分配计算平均轮廓宽度。对于每个点 i，令 a（i）为到同一批次中其他点的平均距离，b（i）为到不同批次中各点的最小平均距离。则轮廓系数 s（i）定义为：

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

  ASW-Batch 分数是所有点的 $s(i)$ 的平均值。数值越低，表示批次混合效果越好。但请注意，在我们的研究中，所有分数都会重新归一化到 $[0, 1]$ 之间，使得数值越高始终表示性能越好

- **批次 LISI (bLISI)**(Simpson，1949年)：局部逆辛普森指数 (LISI) 用于量化局部邻域内的有效批次数。对于每个点 i，令 $\{p_{i,b}\}_{b=1}^{B}$ 表示属于批次 b的邻居所占比例：

$$\text{bLISI}(i) = \left(\sum_{b=1}^{B} p_{i,b}^2 \right)^{-1}.$$

  较高的 bLISI 值表示邻域内的批次混合更加均匀。

- **kBET（k 近邻批次效应检验）**（B¨uttner 等人，2019年）：对于每个点，零假设认为其 k-近邻邻域中的批次标签来自全局批次分布。对每个点进行一次 χ^2 检验，未拒绝零假设（即接受零假设）的检验所占比例即为 kBET 接受率。记作：

$$\text{kBET} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\text{local distribution} \approx \text{global}\}.$$

  接受度越高，表示批次效应越小。

- **图连通性** (Luecken 等人，2022)：构建 k近邻图，并评估每种化合物内的平均连通性。对于具有点 c的化合物 $\mathcal{I}_c$，设 G为邻接矩阵：

$$\text{Conn}(c) = \frac{1}{|\mathcal{I}_c|^2} \sum_{i,j \in \mathcal{I}_c} G_{ij}, \quad \text{GC} = \frac{1}{C} \sum_{c=1}^{C} \text{Conn}(c).$$

  数值越高，表示生物学上相似的细胞在校正后仍保持连通。
  如果它们都彼此接近，则意味着它们来自不同的批次。

生物学保真度指标

• **轮廓标签（ASW-Label）**（Rousseeuw，1987年）：与 ASW-Batch 相同，但使用化合物标签（例如扰动身份）作为簇分配。ASW-Label 得分较高表示生物学簇分离良好：

$$\text{ASW-Label} = \frac{1}{n}\sum_{i=1}^{n} s(i), \quad s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

• **LISI 标签（lLISI）**（Simpson，1949年）：使用化合物标签计算 LISI。对于每个点 i，设 $\{p_{i,c}\}_{c=1}^{C}$ 为其邻域中化合物 c 的比例：

$$\text{lLISI}(i) = \left(\sum_{c=1}^{C} p_{i,c}^2\right)^{-1}.$$

lLISI 值越高，表示生物学邻域越集中。

• **Leiden 调整兰德指数**（Traag 等人，2019年）：对校正后的数据应用 Leiden 聚类算法，以获得簇分配 $\hat{z}_i$，并使用调整兰德指数将其与真实化合物标签 z_i 进行比较：

$$\text{ARI} = \text{AdjustedRandIndex}(\{\hat{z}_i\}, \{z_i\}).$$

ARI 越高，表示与生物结构的一致性越好。我们发现，在所有方法中，ARI 通常都过低。

• **Leiden NMI**（Traag 等，2019年）：同样使用 Leiden算法进行聚类，并计算预测簇与真实化合物标签之间的归一化互信息（NMI）：

$$\text{NMI} = \frac{2I(\hat{Z}; Z)}{H(\hat{Z}) + H(Z)},$$

其中 I 是互信息，H 是熵。较高的归一化互信息表示生物学信息得到了良好保留。

I. 真实世界实验细节与完整结果

我们在从高通量形态学分析中获得的八个主要真实世界数据集上评估Chandrasekaran 等人，2023年）和Broad生物图像基准集合（BBBC）（Ljosa 等人，2012年）。

该**JUMP联盟**是一项大规模产学合作，涵盖多个实验室与实验条件，包括多样的化合物扰动、批次和成像平台。它包含来自13个来源的超过100,000个孔，是稳健批次校正的一项具有挑战性的测试平台。其累计有4 个数据集。CellProfiler（Carpenter等人，2006年）被用于从JUMP联盟提取特征谱

该**BBBC 数据集**在更受控的条件下提供了额外基准，包括带有标注扰动标签的标准Cell Painting实验。其累计有3 个数据集。这些数据集规模较小，但提供更干净的真实标签，且站点间变异更少，因此有助于在噪声较低的条件下评估生物信号保留。DeepProfiler（Moshkov等人，2024年）被用于从BBBC数据集提取特征谱。

我们将全部数据集划分为八个结构化评估场景，合并来自多个批次、来源和成像模态的数据。每个场景旨在测试批次校正方法在不同形式异质性下的表现，范围从合作方特异性混杂到显微镜变异以及特征提取多样性。这些场景包括源自JUMP数据集的五个设置（JUMP 1–5），以及三个使用基于深度学习的特征提取的设置（DEEP 1–3）。BALANS与基线方法在每一场景上均使用全套生物学保真度与批次校正指标进行评估。表5 总结了每个场景的构成，以及其结果是在主文还是补充材料中报告。

I.1. CellProfiler

CellProfiler（Carpenter 等人，2006）是一款开源软件工具，旨在从显微图像中提取可解释的特征。凭借模块化设计以及在单细胞和孔级别生成具有生物学意义的测量值的能力，它广泛应用于高内涵筛选和形态学分析流程。

在我们的实验中，我们使用 CellProfiler 处理多个数据集中的 Cell Painting 检测。提取的特征集包含每孔约 4,762 个特征，这些特征由多个细胞汇总而成。这些特征涵盖多种生物学和形态学类别，包括：

- **形状**：细胞组分的面积、周长、偏心率和紧致度。
- **强度**：各通道和细胞区室（细胞核、细胞质和全细胞）中像素强度的均值、中位数、最大值和标准差。
- **纹理**：基于灰度共生矩阵计算的 Haralick特征，用于描述纹理模式。
- **颗粒度**：量化细胞内颗粒状结构的大小和丰度，这些结构通常与细胞器或亚结构相关。
- **径向分布**：测量强度从细胞中心向外的空间分布。
- **邻居与邻接关系**：描述相邻细胞之间空间关系的统计量。

特征提取按通道进行，涵盖五种荧光染色剂：细胞核（Hoechst）、核仁/细胞质 RNA、内质网、高尔基体/质膜以及线粒体。分别从每种染色剂和细胞区室（细胞核、细胞质、全细胞）中提取特征，然后使用细胞间的中位数，按孔进行汇总。

结果是每个孔的生物学信息丰富的高维向量表示，捕捉对扰动的多样形态学响应。

I.2. DeepProfiler

DeepProfiler（Moshkov等人，2024年）是一种深度学习框架，可从高内涵细胞显微图像中生成紧凑的特征表示。与依赖人工设计特征的CellProfiler不同，DeepProfiler

采用EfficientNet架构直接从原始图像中学习特征嵌入，从而能够捕捉复杂的非线性形态模式。

在这项工作中，我们使用（Moshkov等人，2024年）中介绍的 DeepProfiler。它利用从实验元数据（例如化合物和浓度标签）中获得的弱监督来训练深度网络。所用架构是 ResNet-50 的一个变体，先在 ImageNet 上预训练，再通过分类头使用扰动标签进行微调。

输入处理。每张五通道细胞绘画图像都作为一个堆栈进行处理，其中每个通道对应一个细胞区室或一种染色剂。DeepProfiler 可以：

- 独立处理每个通道（逐通道 ResNet 编码），
- 或将这些通道组合成一个五通道输入张量，供统一的 CNN 主干网络处理。

训练目标。网络经过训练，以 softmax 交叉熵损失区分不同处理，其中每个唯一的化合物-浓度对都被视为一个单独的类别。嵌入取自网络的倒数第二层（分类头之前），并用作特征表示。

特征输出。 输出为一个 384 维特征向量。通过平均池化对多个视野和细胞的嵌入进行聚合，从而为每个孔生成一个紧凑向量。

在我们的分析中，我们直接评估 DeepProfiler 表征。

I.3. DeepProfiler 架构：Cell Painting 卷积神经网络

集成于 DeepProfiler 框架中的细胞绘画卷积神经网络模型，旨在从高内涵细胞图像中提取有意义的形态学特征。该模型采用 EfficientNet-B0 架构，以兼顾性能与计算效率而闻名。EfficientNet-B0 由一个初始卷积层和随后七个倒置残差块组成，这些块采用了带有挤压与激励优化的 MobileNetV2 结构，最后接一个分类层。该模型约有 2.4 百万个参数。

输入处理： 该模型处理五通道细胞涂染图像，每个通道对应特定的细胞组分：DNA（细胞核）、RNA（核仁和细胞质 RNA）、内质网（ER）、高尔基体和质膜（AGP），以及线粒体。每张单细胞图像都会裁剪为 128×128 像素大小，以细胞核为中心，同时保留周围背景信息且不进行缩放。通道顺序如下：DNA、RNA、ER、AGP 和线粒体。

训练策略：该模型采用弱监督学习方法进行训练，其中每种独特的处理条件（化合物-浓度对）都作为一个独立类别。训练目标是最小化分类交叉熵损失，以促使模型学习能够区分不同处理条件的特征。为了解决类别不平衡问题，训练过程中每个周期都会为每种处理条件采样相同数量的细胞，并对代表不足的处理条件进行过采样。训练中采用随机裁剪、水平翻转、旋转和亮度/对比度调整等数据增强技术，以提高泛化能力并防止过拟合。

特征提取：训练完成后，从 EfficientNet-B0 模型的 block6a 激活层提取特征。该层为每张单细胞图像生成一个 384 维特征向量，捕捉由不同处理条件引起的细微形态变化。这些特征可以在多个细胞或视野之间进行聚合（例如通过平均池化），从而生成适用于下游分析的孔级特征谱。

实现细节：该模型使用 TensorFlow 2.5 实现，要求 Python 3.6 或更高版本。建议使用图形处理器加速，以提高处理效率。提供了 Cell Painting 卷积神经网络的预训练权重，可用于新数据集的特征分析，无需重新训练。

I.4. JUMP 联盟数据

JUMP Cell Painting 联盟汇集了大量使用 Cell Painting 检测生成的形态学特征谱。本研究主要关注 JUMP Cell Painting（JUMP-CP）项目的数据集，尤其是 cpg0016-jump 数据集（简称 cpg0016），该数据集是在核心数据生成阶段生成的。该数据集涵盖 13 个不同的数据生成站点（或来源），并包含多种生物学扰动：

- **小分子化合物：**116,753 项扰动。
- **ORF基因过表达：**15,142 种扰动。
- **CRISPR基因敲除：**7,977 种扰动。

为促进批次对齐并确保纵向一致性，每个批次均加入一块名为**JUMP-Target-2-Compound**（称为*Target2* 板）的阳性对照板。该板包含302种不同的化合物，可用于在数据集内进行对齐，也有助于与联盟之外未来生成的数据集进行对齐。其余包含全部116,753种化学扰动的板称为**生产板**。

每块生产板都包含一组阴性和阳性对照，用于识别或校正实验伪影。具体而言，经二**甲基亚砜（DMSO）**处理的孔作为阴性对照。这些对照有助于检测并校正板间变异，并作为形态学基线，用于识别具有可检测效应的扰动。

I.4.1. JUMP 的预处理

预处理流程总体遵循Arevalo等人(2024年）。JUMP 细胞绘画数据集中的每个孔均由 CellProfiler 提取的 4,762 维特征向量表示，涵盖形状、强度、纹理和像素级统计量。

(1) 变异过滤。为去除无信息的特征，我们根据对照孔计算稳健的逐板变异得分。对于每个特征 X，令 $X_e =$ 中位数(X) 和 $\sigma_e =$ 中位数$(|X_i - X_e|)$。变异系数定义为：

$$\mathrm{Cvar}(X) = \frac{\sigma_e}{X_e}.$$

在任一板上 Cvar $< 10^{-3}$ 的特征均予以剔除。

(2) 中位数绝对偏差（MAD）归一化。我们使用对照孔对每块板上的特征值进行标准化。具体而言，对于每个特征 X 和孔 i，我们应用以下变换：

$$\bar{X}_i = \frac{X_i - X_e}{\sigma_e},$$

其中， X_e 和 σ_e 的定义同上，且仅使用对照孔计算。

(3) 基于秩的逆正态变换（INT）。我们使用基于秩的 INT 对每个特征进行非参数归一化：

$$Y_i = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right),$$

其中， r_i 是样本 i的秩， N 是样本数量，且 $c = 3/8$ 。 Φ^{-1} 表示标准正态分布的逆累积分布函数。

(4) 特征选择。 为消除冗余，我们采用相关性阈值函数，该函数来自Pycytominer(Serrano 等人，2025)。我们识别出皮尔逊相关系数高于 0.9 的特征对，并在每组中剔除与所有其他特征的总相关性最高的特征。

该流程可确保归一化的稳健性，剔除信息量低和冗余的特征，并促进异质批次和不同来源数据的下游分析。

I.5. BBBC数据

Broad生物图像基准集合（BBBC）提供公开可用的显微镜数据集，用于评估图像分析算法。在研究“学习用于扰动图像分析的表征”（Moshkov等人，2024年）中，使用了三个BBBC 数据集来评估所提出的DeepProfiler框架的有效性：

• **BBBC037**：该数据集包含经各种基因过表达处理的U2OS细胞图像。具体而言，它包括302种独特的遗传扰动，每种均有多个重复样本。该数据集对于评估模型捕捉由基因修饰引起的表型变异的能力具有重要作用。

• **BBBC022**：该数据集包含经多种小分子处理的U2OS细胞图像，涵盖1,600多种化学扰动。它可作为基准，用于评估模型识别化学处理所致形态变化的能力。

• **BBBC036**：该数据集包含暴露于多种化合物的A549细胞图像，共约1,200种扰动。它提供了一个平台，用于评估所学习表征在不同细胞系和处理类型之间的泛化能力。

每个数据集都包含五通道细胞涂染图像，可捕捉细胞核、内质网和线粒体等多种细胞组分。这些数据集在扰动类型和细胞系方面的多样性，有助于全面评估DeepProfiler的表征学习能力。

I.6. 评估场景

我们定义了八种评估场景，旨在涵盖真实世界数据集和基准数据集中的多种技术与生物学变异。这些场景包括五种源自JUMP Cell Painting 联盟的场景（JUMP 1–5），以及三种源自公开BBBC 数据集的场景（DEEP 1–3）。每种JUMP场景对应特定的实验室、显微镜和扰动规模设置，而DEEP场景则与DeepProfiler研究中使用的BBBC数据集相对应（Moshkov等人，2024年）。

为清晰起见并确保可比性，我们明确将这些场景对应到批次校正方法基准研究中的场景 1–5（Arevalo 等人，2024年），如下所示：

• **JUMP 3 ↔ 场景 1**：来自单个实验室、使用单台显微镜的数据，包含 302 种化合物扰动。作为变异最小的基线。

• **JUMP 4 ↔ 场景 2**：单个实验室使用多台显微镜。包含 302 种化合物。成像硬件带来了中等程度的技术变异。

• **JUMP 5 ↔ 场景 3**：来自多个实验室的数据，每个实验室均使用自己的显微镜。包含超过 80,000 种化学扰动，体现了大规模的跨合作方变异。

• **JUMP 1 ↔ 场景 4**：三个实验室使用共享显微镜型号，包含 302 种化合物。反映了标准化设备条件下的实验室间变异。

• **JUMP 2 ↔ 场景 5**：五个实验室使用相同的显微镜类型，规模扩大至超过 80,000 种化合物。这代表了受控硬件条件下最丰富的批次多样性。

• **DEEP 1 (BBBC036)**：A549细胞中的化学扰动，涵盖约1,200种化合物。可用于在不同细胞系和生物学背景下进行评估。

• **DEEP 2 (BBBC037)**：U2OS细胞中的遗传扰动谱分析。由302种基因过表达处理组成，用于评估弱监督下的生物学信号保留。

• **DEEP 3 (BBBC022)**：U2OS细胞中的小分子扰动。包括1,600多种具有已知注释的化合物。

这组结构化场景涵盖单实验室至多实验室条件、显微镜变异，以及工程化和学习得到的特征空间。它为在真实且具有挑战性的条件下评估批次校正性能提供了一个全面的框架。

表*5.* JUMP 和 DEEP 评估场景概要。每种场景均结合了多种来源、实验室和显微镜，涵盖多种扰动和分析方法。

场景	# 来源	# 实验室	# 显微镜	# 扰动	# 数据点	分析方法	主文/补充材料
JUMP 1	3	3	3	302	30,000	CellProfiler	Main
JUMP 2	5	5	3	80,000+	400,000	CellProfiler	Main
JUMP 3	1	1	1	302	5,000	CellProfiler	补充材料
JUMP 4	3	1	1	302	150,000	CellProfiler	补充材料
JUMP 5	3	3	1	80,000+	300,000	CellProfiler	补充材料
DEEP 1	1	1	1	1,000+	20,000	DeepProfiler	Main
DEEP 2	1	1	1	300+	5,000	DeepProfiler	补充材料
DEEP 3	1	1	1	1,500+	10,000	DeepProfiler	补充材料

I.7. 完整结果

这里提供了汇总所有未纳入主文正文的实验场景性能指标的表格，具体包括 **JUMP 3–5** 和**DEEP2–3**。

BALANS 在所有设置下均持续表现出色。在批次多样性较高的场景中，例如 **JUMP 3-5**，其表现最佳；在这些场景中，批次效应更加显著，也更具挑战性。因此，BALANS 尤其适用于细胞绘画等应用，因为不同扰动类型、板和成像条件之间的批次可能存在显著差异。BALANS 在批次多样性较低的设置中也保持了有竞争力的准确性，例如**DEEP2–3**，展现出其在不同批次结构条件下的稳健性。

BALANS 在真实世界成像数据集上广泛有效。

表 *6.* 在真实世界批次校正基准（**JUMP 3**）上的评估分数。BALANS 在批次指标上表现最佳，并在保留生物结构的同时取得了出色的整体性能。

方法	连通性	KBET	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
BALANS	0.76	**1.00**	**0.48**	0.81	**0.98**	**0.09**	**0.63**	**0.59**	**0.76**	**0.57**	**0.67**
Seurat CCA	0.80	**1.00**	0.46	0.77	**0.98**	0.07	0.60	0.52	0.76	0.54	0.65
Combat	0.78	0.99	0.41	**0.84**	**0.98**	0.06	0.56	0.53	0.76	0.53	0.65
Seurat RPCA	**0.81**	0.99	0.43	0.78	**0.98**	0.06	0.60	0.53	0.75	0.54	0.64
fastMNN	**0.81**	**1.00**	0.44	0.76	**0.98**	0.05	0.59	0.55	0.75	0.54	0.64
基线	0.77	0.99	0.39	0.85	**0.98**	0.06	0.55	0.52	0.75	0.53	0.64
Harmony	0.79	0.99	0.40	0.74	**0.98**	0.06	0.60	0.55	0.73	0.55	0.64
球化	0.74	0.99	0.37	**0.87**	**0.98**	0.06	0.54	0.52	0.74	0.53	0.64
Scanorama	0.45	0.89	0.50	0.62	**0.98**	**0.09**	0.52	0.44	0.61	0.51	0.56

表 *7.* 使用 DeepProfiler 特征时，大规模、多样化批次设置下的评估得分。BALANS 在批次校正和标签结构方面表现出色，同时保持了强劲的整体性能。**JUMP 4**

方法	连通性	KBET	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
Seurat CCA	0.64	0.66	0.55	**0.86**	**0.98**	0.04	0.45	**0.49**	**0.68**	0.49	**0.59**
BALANS	0.59	0.54	0.54	**0.86**	**0.98**	**0.05**	**0.50**	**0.54**	0.63	**0.52**	0.57
Seurat RPCA	**0.65**	0.57	0.46	0.84	**0.98**	**0.05**	0.45	**0.49**	0.63	0.49	0.56
Harmony	0.64	0.62	0.55	**0.86**	**0.98**	0.04	0.42	0.48	0.67	0.48	0.58
Scanorama	0.52	**0.68**	**0.68**	0.80	**0.98**	0.04	0.42	0.47	0.67	0.48	0.58
fastMNN	0.56	**0.70**	0.52	0.80	0.97	0.04	0.39	0.45	0.64	0.46	0.55
Combat	0.62	0.24	0.16	0.76	**0.98**	0.03	0.42	0.48	0.45	0.48	0.46
基线	0.60	0.28	0.17	0.75	**0.98**	0.04	0.42	0.48	0.45	0.48	0.46
球化	0.58	0.22	0.13	0.76	**0.98**	0.03	0.39	0.48	0.42	0.47	0.45

I.8. UMAP 图

为定性评估批次混合和生物结构的保留情况，我们展示了若干选定场景的二维 UMAP 嵌入。如图所示，BALANS 能有效对齐各批次，同时保持生物标签之间清晰的区分。

I.9. 与 BBKNN 的比较

BBKNN 是一种广泛使用的预处理工具，可构建批次感知的 k-最近邻图，因此适用于可视化以及 UMAP 等嵌入方法。然而，其功能与目标从根本上不同于

表 *8.* 大规模、多样化批次设置（**JUMP 5**）的评估分数。在这一具有挑战性的设置中，BALANS 在所有批次指标和生物结构指标上表现最佳。由于我们强制执行的 12 小时截止时间，CCA/RPCA 无法运行

方法	连通性	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
BALANS	0.33	**0.48**	**0.91**	**1.00**	0.01	**0.46**	**0.32**	**0.57**	**0.33**	**0.45**
Scanorama	**0.34**	0.41	0.83	**1.00**	0.01	0.30	0.27	0.53	0.31	0.42
fastMNN	**0.34**	0.45	0.79	**1.00**	**0.02**	0.26	0.22	0.53	0.30	0.41
Harmony	**0.34**	0.25	0.84	**1.00**	0.01	0.27	0.32	0.48	0.32	0.40
球化	**0.34**	0.00	0.84	**1.00**	0.00	0.22	**0.38**	0.39	0.31	0.35
Combat	**0.34**	0.02	0.81	**1.00**	0.00	0.24	0.31	0.39	0.31	0.35
基线	**0.34**	0.01	0.80	**1.00**	0.00	0.23	0.31	0.38	0.30	0.34

表 *9.* 低多样性批次设置（**DEEP 2**）的评估得分。在这一具有挑战性的设置中，BALANS 在所有批次和生物结构指标上均表现最佳。由于我们设定了 12 小时截止时间，无法运行 CCA/RPCA

方法	连通性	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
BALANS	0.24	**0.63**	0.87	**1.00**	**0.03**	0.35	0.41	**0.58**	**0.45**	**0.51**
球化	**0.29**	0.44	**0.89**	0.98	**0.03**	0.31	**0.46**	0.54	0.44	0.49
Seurat RPCA	0.27	0.61	0.79	0.97	0.02	0.28	0.38	0.56	0.41	0.48
Harmony	0.26	0.60	0.80	0.97	0.02	0.26	0.39	0.55	0.41	0.48
Seurat CCA	0.27	0.58	0.78	0.96	0.02	0.29	0.38	0.54	0.41	0.48
基线	0.26	0.56	0.80	0.97	0.02	0.27	0.39	0.54	0.41	0.47
Combat	0.26	0.57	0.81	0.96	0.02	0.28	0.39	0.55	0.41	0.48
fastMNN	0.25	0.59	0.77	0.97	0.02	0.26	0.37	0.54	0.40	0.47
Scanorama	0.22	0.53	0.81	0.99	0.02	**0.36**	0.34	0.52	0.43	0.47

BALANS 的情况也是如此。这里我们总结两者的主要区别，并在两种方法有所重叠的有限范围内进行公平比较。

尽管 BALANS 和 BBKNN 都依赖基于 K近邻的构造，但两者的目标截然不同。BALANS 旨在通过构建并将批次感知亲和算子应用于数据，生成校正后的特征谱。相比之下，BBKNN 只输出图，不会生成校正后的嵌入或特征谱。原则上，可以尝试将 BBKNN 的亲和矩阵与数据相乘以获得校正后的特征谱，但以下几点说明了为何这在实际中不可行。

BBKNN 通过合并各批次的 K近邻图来构建完整的批次平衡亲和矩阵，但由于需要对一个 $N \times N$ 矩阵进行显式运算，因此无法高效地将该矩阵应用于数据。BALANS 避开了这一瓶颈。BALANS 不构建完整矩阵，而是计算以下形式的表达式：

$$\hat{X} \approx (A_S^\top A_S)X,$$

如第 328–332 行所述，使用低秩采样算子 A_S，无需构建完整矩阵或与之相乘，即可将亲和度应用于数据。因此，关键区别不在于能否定义亲和矩阵（BBKNN 显然可以定义），而在于它是否能用于可扩展批次校正。BALANS 专为此目的而设计；BBKNN 则不是。

同样，BBKNN 所报告的近线性运行时间仅适用于通过 Annoy 构建近似 K近邻。由于 BBKNN 从未将其亲和矩阵应用于数据，因此这种复杂度并不适用于批次校正。生成校正后的特征谱无法在线性时间内完成，因为即便使用稀疏亲和矩阵，也需要超过 $O(N)$ 的运算量。BALANS 可在线性时间内构建亲和矩阵（参见补充材料D），随后利用低秩近似，使亲和矩阵与数据的乘法运算具备可扩展性。在我们的实验中，使用 BBKNN 的完整亲和矩阵校正特征谱，在 JUMP 2 等大型数据集上会慢得难以接受。

另一个区别在于，BALANS 使用批次条件局部带宽，从而支持一种基于原理的肘部检测阈值设定机制（补充材料C.1）。BBKNN 不计算批次条件尺度；它会针对每个批次识别近似邻居，并保留固定数量的邻居。对于不平衡群体或异质密度，这种固定且非自适应的规则可能表现不佳。因此，BALANS 的批次条件缩放是 BBKNN 所不具备的独特且有理有据的组成部分。

尽管 BBKNN 无法生成校正后的特征谱，我们仍会评估其在可根据其 K近邻图计算的指标上的表现，例如批次混合评分；在可行的情况下，还会尝试进行替代性亲和矩阵与数据乘法。在这些基于图的指标上，BALANS 的表现始终优于 BBKNN；而尝试扩展 BBKNN 以生成校正后的特征谱

表*10*。低多样性批次设置（**DEEP3**）下的评估得分。在这一具有挑战性的设置中，BALANS 在所有批次和生物结构指标上均表现最佳。由于我们设定了 12 小时截止时间，因此无法运行 CCA/RPCA

方法	连通性	批次 LISI	批次轮廓系数	LISI-标签	ARI	NMI	标签轮廓系数	批次平均值	标签平均值	总体平均值
BALANS	0.28	**0.65**	0.85	**1.00**	**0.04**	0.36	0.42	**0.59**	**0.46**	**0.53**
球化	**0.30**	0.40	**0.88**	0.98	0.03	0.30	**0.46**	0.53	0.44	0.49
Seurat RPCA	0.27	0.63	0.78	0.97	0.03	0.29	0.39	0.56	0.42	0.49
Harmony	0.27	0.62	0.79	0.97	0.03	0.27	0.40	0.56	0.42	0.49
Seurat CCA	0.28	0.60	0.78	0.97	0.03	0.30	0.39	0.55	0.42	0.49
基线	0.27	0.58	0.79	0.97	0.03	0.28	0.40	0.55	0.42	0.48
Combat	0.27	0.59	0.80	0.96	0.03	0.29	0.40	0.55	0.42	0.48
fastMNN	0.26	0.61	0.76	0.97	0.03	0.27	0.38	0.54	0.41	0.48
Scanorama	0.23	0.54	0.79	**1.00**	0.03	**0.37**	0.34	0.52	0.44	0.48

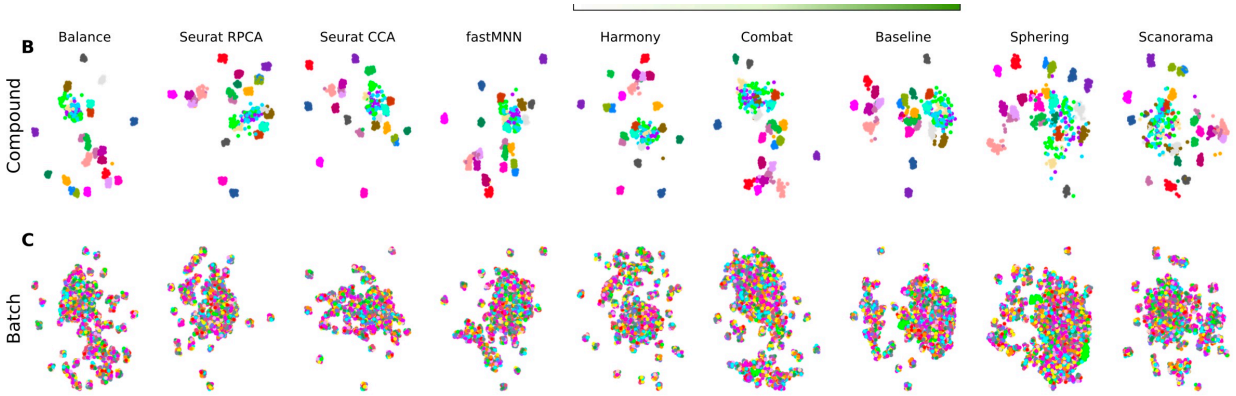


图*5*. JUMP 3 的 UMAP 可视化化合物。(b) 按批次着色。结果。(a) 按生物学簇着色，展示了几个代表性

e

在大型数据集上，计算上不可行。

表 *11*. BBKNN 和 BALANS 在 JUMP-1 和 JUMP-2 上的评估指标。括号中的数字表示排名。

数据集 / 方法	图连通性	批次 LISI	LISI 标签	Leiden 调整兰德指数	Leiden NMI
JUMP-1（场景 4）					
BBKNN	0.25	0.63	0.97	0.03	0.12
BALANS	0.54	0.44	**0.98**	0.05	0.46
JUMP-2（场景 5）					
BBKNN	0.33	0.006	1.000	0.000	0.229
BALANS	0.33	**0.48**	**1.00**	0.01	**0.46**

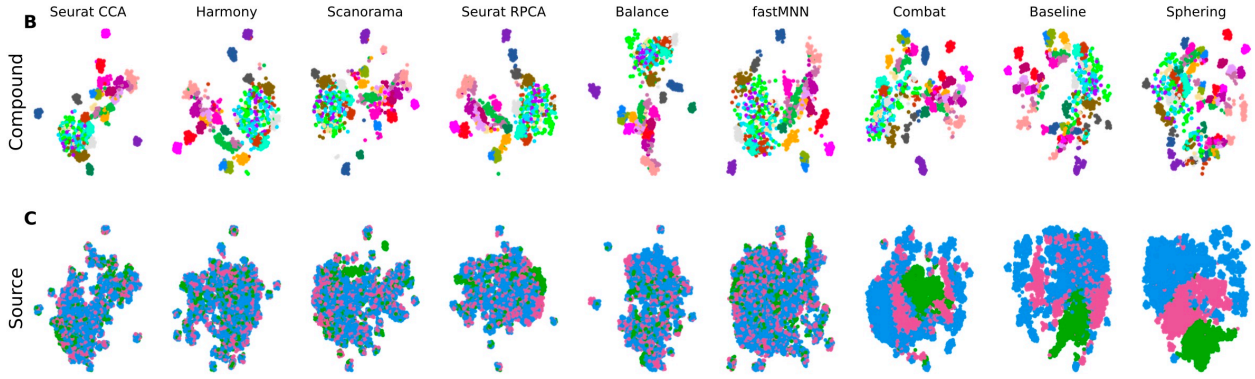


图 *6.* JUMP 4 的 UMAP 可视化。(a) 按生物学簇着色，展示了若干代表性化合物之间的分离情况。(b) 按来源实验室着色。

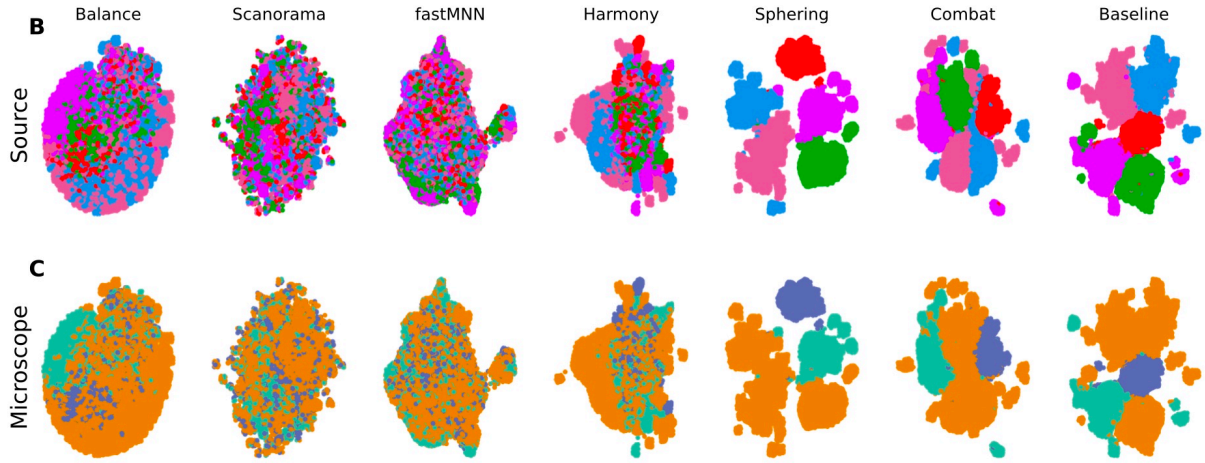


图 *7.* JUMP 5 的 UMAP 可视化。(a) 按来源实验室着色。(b) 按显微镜着色。化合物数量过多，因此无法绘制按化合物着色的 UMAP 图

J. 合成实验与结果

我们使用层次高斯混合模型（GMM）生成合成数据集，以反映受控条件下的标签和批次结构。生成过程如下：

- **标签级生成：** 对于每个语义标签 $\ell \in \{1, \ldots, L\}$，我们在 d 维空间中采样一个标签级均值 $\mu_\ell \sim \mathcal{N}(0,\ \sigma^2_{\text{标签}} \cdot I_d)$。

- **批次级生成：**对于分配给标签 ℓ 的每个批次 $b \in \{1, \ldots, B\}$，我们采样一个批次特定的均值$\mu_{\ell,\ b} \sim \mathcal{N}(\mu_\ell,\ \sigma^2_{\text{批次}} \cdot I_d)$。

- **数据点级生成：** 批次 b 中标签为 ℓ 的每个数据点 x 均按 $x \sim \mathcal{N}(\mu_{\ell,\ b},\ \sigma^2_{\text{噪声}} \cdot I_d)$ 抽取。

这种设置使我们能够通过 σ^2_{label} 控制标签内变异性，通过 σ^2_{batch} 控制批次引起的扰动，并通过σ^2_{noise}控制整体噪声。

我们在图中总结的多种合成数据情境下评估了性能。与其他方法相比，BALANS 始终更快。

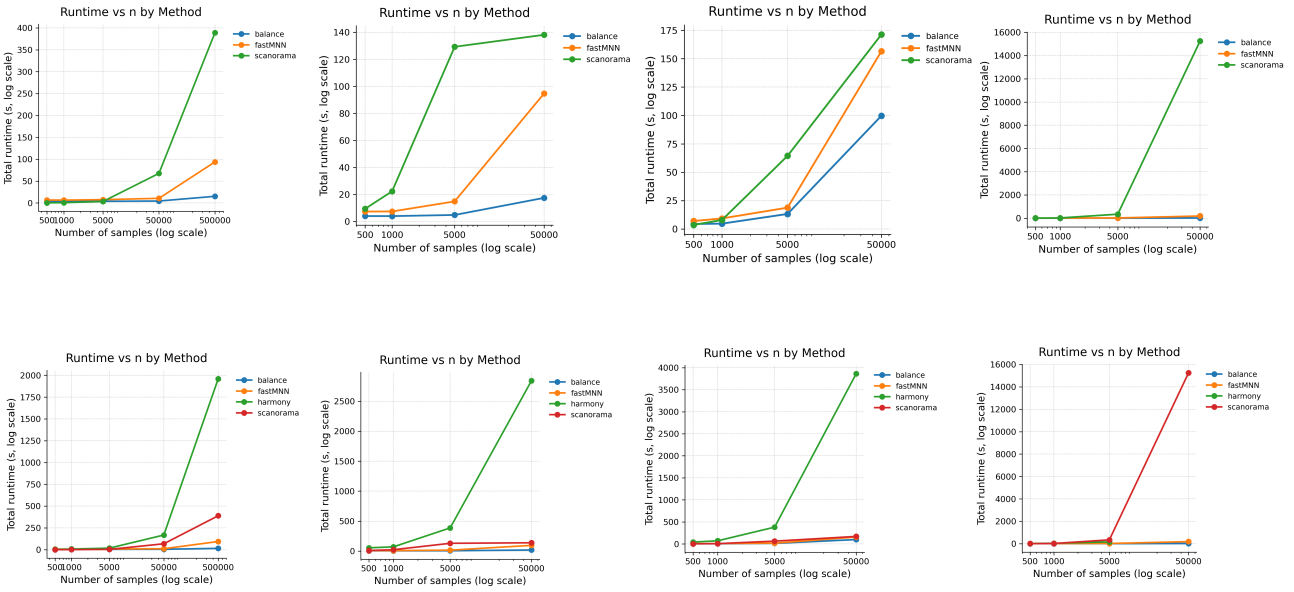


图 *8.* **不同合成场景下的运行时间比较。** 每列展示两个运行时间条：使用 Harmony 和不使用 Harmony，因为 Harmony 通常比其他方法慢得多。从左到右，我们依次考虑：**(1)** 包含 10 个化合物、10 个批次和 10 个特征维度的小规模设置。**(2)** 包含 100 个化合物、5 个批次和 10 个维度的中等规模设置。**(3)** 在固定结构下改变批次特异性噪声水平 σ_{batch} 的稳健性测试。**(4)** 包含 1000 个特征、10 个化合物和 5 个批次的高维设置。需要注意的是，在高维情况下，与其他方法相比，Scanorama 的运行速度明显较慢。

K. 相关工作

K.1. 用于细胞绘画的深度学习表征

近年来，深度学习被广泛用于从细胞绘画图像中提取丰富表征，研究重点也从人工设计特征扩展至此。相关工作包括 DeepProfiler (Moshkov 等人，2024年），该方法提出了基于卷积神经网络的特征提取方案，并使用扰动分类进行训练。Moshkov等人(2024年) 利用 ResNet 等预训练网络进行迁移学习，展现出更优的形态学分析性能。近期，Wong 等人(2023年) 提出了针对生物学重复样本的对比预训练策略，而Doron 等人(2023年) 提出了一种掩码自编码方法，可从带噪声且未经标注的图像中学习。Kraus 等人(2024年) 通过在预训练过程中整合结构保持损失项和批归一化，进一步改进了自监督学习流程。

K.2. 生物数据中的批次校正

批次校正是基因表达、蛋白质组学和基于图像的表型分析等领域中得到广泛研究的问题。早期的线性方法，如 Combat（Zhang 等人，2020年），使用经验贝叶斯模型去除加性和乘性效应。Ando 等人 （2017年）提出了先进的矩阵分解模型，用于对高通量ˇ测定进行去噪。Cuklina 等人 （2021年）引入对抗学习，在保留生物信号的同时去除批次身份；Tran 等人 （2020）对 30 多种单细胞RNA测序整合方法进行了严格比较，强调其对复杂批次结构的稳健性。Yu 等人 （2023年）通过整合空间信息与元数据，解决高维显微成像所特有的挑战。总体而言，这些方法涵盖了用于批次校正的统计、概率和表征学习方法。

K.3. 局部带宽估计

局部尺度选择是流形学习中自适应亲和度计算的基础。Zelnik-Manor 和 Perona 的自调节核方法（Zelnik-Manor 和 Perona，2004年）根据局部 k近邻距离调整高斯带宽，是一种广泛使用的基线。Herrmann (1997年) 提出了用于自适应平滑的尺度自适应核。Knutsson 等人(1994年) 提出了将其用于正交小波。

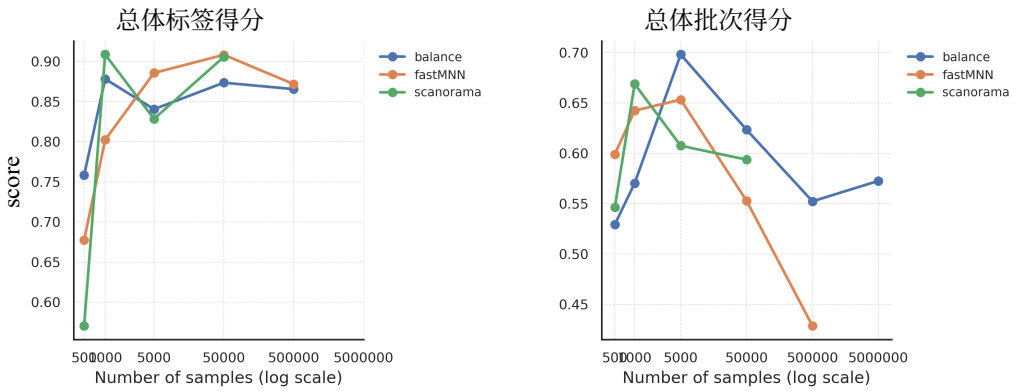


图*9.* 主文运行时间评估中所有方法的准确度得分。表现快速的方法得分相当一致，这表明它们在计算资源受限的情况下仍能保持稳定表现。由于发生错误，HARMONY 得分分析未能完成，但我们正在努力修复。

K.4. 低秩近似

低秩近似长期以来一直是可扩展学习的支柱。Nystr¨om 方法（Williams 和 Seeger，2000年）是一种开创性技术，通过地标点选择近似大型核矩阵。Drineas 等人(2005年) 将其扩展为随机采样策略，以提高精度。Kumar 等人(2012年) 提出了基于杠杆分数的采样，以减少近似误差。Musco和Musco (2017年) 提出了递归 Nystr¨om 近似，并给出了可证明的谱界；而Gittens和Mahoney (2016) 则提出了更完善的观点，以便将这种采样方法应用于大规模数据集。PHATE 是一种用于高维数据非线性可视化的工具，其遵循的方法由Moon等人(2019年) 提出。

K.5. 理论保证与实证表现之间的联系

在本节中，我们阐明主文所建立的理论结果与BALANS实证表现之间的联系。实践中使用的算法组件直接源自理论保证，而我们分析中的假设本身也立足于来自生物数据的实证观察。

自适应采样直接源自该理论。第2.3节（定理1）中的理论结果确立，自适应采样方案能够对底层亲和矩阵实现高效且近乎最优的覆盖度。这一保证推动了BALANS中所用采样过程的设计：算法1（第3节）中的实证算法精确实现了理论所支持的自适应策略。更多实现细节见补充材料C.2.1节。因此，实践中使用的采样机制并非启发式的，而是理论上最优过程的直接实例化。

L. 理论保　　证为无批次亲和矩阵的实证近似提供了依据　　.

理论进一步表明，在假设 2下，BALANS过程恢复的亲和矩阵以高概率逼近真实的无批次亲和矩阵。这一结果促使我们将所恢复的矩阵用于实证批次校正：BALANS产生的校正后的嵌入是对理想的无批次结构具有理论依据的近似。实证评估表明，这种近似在保留生物学信号的同时改善了批次混合。

这些假设由实证观察提供依据并得到验证。 重要的是，假设 2 并非任意提出，而是直接以阳性对照的实证观察为依据。正如主文第 187–191 行所述，在多种扰动和多个板中，我们观察到簇内距离在不同批次间保持稳定，而批次效应主要表现为平滑的外在畸变。这一实证规律为理论中采用的建模假设提供了依据，并支持其对生物单细胞和基于图像的数据集的适用性。

总结。 理论分析为 BALANS 算法的设计提供了指导，该算法实现了理论上最优的组成部分，而理论所依据的假设则由实证观察提供依据并得到验证。这些联系共同确保了 BALANS 的理论与实证部分紧密契合。

M. 关于去除逆的计算

在我们的框架中，去除 Moore–Penrose 伪逆会得到一种在特定理想化条件下具有理论依据的近似——具体而言，当噪声可忽略不计，且各簇的规模与亲和结构相对均匀时。在此类情形下，仿射算子条件良好，伪逆除数值细化外几乎没有额外贡献。我们已在附录中补充了一段简短讨论，明确阐述这些条件，以及近似可能减弱的情形，例如存在强烈不平衡簇或亲和度存在显著异质性时。这一澄清明确了简化算子仍忠实于原始表述的适用情形。