# 项目简介

单细胞RNA测序（scRNA-seq）数据分析可以大致分为如下几个步骤：数据预处理->归一化&批次校正->降维聚类->亚群注释&下游分析。本项目旨在研究其中的批次校正算法，解决scRNA-seq中的关键问题。

意义和价值：有效的去批次效应算法能够提高单细胞数据分析的可靠性和准确性，有助于揭示细胞异质性、理解疾病机制，推动精准医学和新型药品研发的进展。

研究方法：本项目将评估对比本文拟提出的算法和部分现有的去批次效应算法，包括基于矩阵分解的方法（如LIGER等）、基于互近邻的算法（如Seurat、Harmony等）和基于深度学习的算法如（如DESC、scDML等）。我们计划结合大、小、单批次、多批次等多种数据集，辅以ARI、NMI等多种评测指标，分析改进算法在去除批次效应方面的效果。

预期结果:提出一种改进的去批次效应算法，并在多种单细胞数据集上验证其有效性，并验证其在较大数据集的可拓展性。

特色和创新：（1）构建独特结构的自编码器，在提取生物学表达特征的同时完成初步批次对齐，从而无需依赖“生物学差异远大于批次效应”这一较强假设。然而很多基于MNN的批次校正方法都依赖于这个假设。（2）在自编码器的训练过程中引入随机的批次噪声，以增强模型的鲁棒性，从而提升其在处理复杂场景数据集时进行批次效应校正的能力。

# 立项背景

scRNA-seq通过高通量检测单细胞基因表达谱，为生命科学和医学研究提供了前所未有的解析力。然而，由于不同实验批次间的技术偏差，直接分析不同样本时可能引入系统性误差，从而导致“批次效应”。随着数据集规模不断扩大，而单个批次的细胞数量有限，实际数据集的数据通常来源于不同实验室或多个批次，批次效应在所难免。这种效应可能混淆真实的生物变异，若未能有效消除，将使下游分析更加复杂，并可能导致对结果的误解。

根据批次校正方法的数学基础和方法视角，可以把现有的批次校正方法分为三类：（1）矩阵分解的方法；（2）深度学习的方法；（3）互近邻的方法。

基于矩阵分解的单细胞转录组批次校正方法，通常通过将原始数据矩阵分解为两个或多个子矩阵，从而获得数据的低维特征空间和共享因子矩阵，并在低维特征空间中进行批次效应的校正。该类方法具有较好的可解释性，但其性能在很大程度上依赖于共享因子数量的选择。若因子数选择不当，可能导致校正效果不佳；此外，因子数量过多还会导致时间复杂度呈指数级增长，从而影响计算效率。代表的算法有2019年Welch J.D等人[1]提出了名为LIGER的批次校正方法。

基于深度学习的批次校正方法常通过对抗生成网络、自编码器和卷积神经网络等模型对数据进行批次校正。深度学习批次校正方法在大多数场景都能很好的对scRNA-seq数据集进行批次校正，但是与其他类型方法相比，基于深度学习的批次校正方法对计算资源的要求较高，缺乏可解释性。代表的算法有2020年Li X等人[2]提出了名为DESC的批次校正方法。

基于互近邻策略的单细胞转录组批次效应校正方法，主要通过在不同批次的数据集中识别相互最近邻（mutual nearest neighbors, MNN）对，以此构建细胞间的社区网络或显式建模批次效应，从而实现跨批次的校正。此类方法在计算效率方面具有显著优势，适用于大规模数据的处理。然而，由于其依赖于局部匹配的互近邻对以实现全局的批次对齐，校正效果在很大程度上取决于互近邻对的数量及其匹配准确性。若匹配不足或误配，可能导致校正结果偏离真实生物差异[3][4][5]。代表算法有Hie B的Scanorama[6]，Korsunsky I的Harmony[7]，Stuart T的Seurat V3[8]，Yu X的scDML[9]。

# 项目方

## 批次校正问题定义

理论上，不同批次中相同类型的细胞应具有相似的基因表达特征，但由于实验条件等非生物因素的差异，实际测量结果往往存在系统性偏差，即批次效应。批次校正的目标是尽可能消除这些非生物差异，以提高多批次单细胞数据整合与分析的准确性。

假设对一个细胞进行两次scRNA测序，测量出的表达值分别为和，由于存在批次效应，导致两次测量出来的表达值有差异。但因为是同一个细胞，其基因理论表达值应该是一样的，设为。批次校正需要去除掉测量值中的批次效应，得到基因理论表达值。实际上我们并不需要求出理论值，只需要让校正后的值相等即可。对于给定的数据集，批次校正算法能够去除数据中的批次效应，得到真实的生物信息。定义公式：

其中表示批次校正模型。

## 批次校正效果可视化方法选择

scRNA-seq数据通常特征维度比较高，常在一万维以上。一般通过降维算法将其高位数据映射至低维空间中，并且尽可能的保留数据的原始内在分布，最后通过可视化的方式将数据进行呈现。

目前业内常用的降维方法有PCA（主成分分析，Principal Component Analysis）、t-SNE（t分布随机邻居嵌入，t-distributed Stochastic Neighbor Embedding）、UMAP（统一流形近似投影，Uniform Manifold Approximation and Projection）。

其中PCA是一种经典的线性降维方法，PCA保留的是数据的线性结构，计算效率高，但精度较低。适用于初步分析或作为其他降维方法的预处理步骤，并不适用于最后的批次校正效果可视化。t-SNE和UMAP属于非线性降维算法，各有优劣。

t-SNE基于概率分布相似性，以Kullback-Leibler（KL）散度为损失函数，通过最小化高维空间与低维空间之间样本对的条件概率分布差异来实现嵌入。该方法能够有效保留局部结构，适合用于发现数据中的细胞簇。然而，KL散度对高维中相距较远而低维中相近的样本对惩罚较轻，容易导致整体差异较小的细胞簇在低维空间中被拉得较远，从而削弱了对全局结构的刻画能力。此外，t-SNE分别采用高斯分布和t分布对高维与低维空间进行建模，这种建模方式虽然增强了局部聚类的效果，但可能会引起高维空间中相近样本在低维空间中的失真。

相较而言，UMAP以拓扑结构保持为核心，构建稀疏图来表示数据的局部邻接关系，并使用二元交叉熵（binary cross-entropy）作为损失函数，优化高低维空间中图结构的相似性。与KL散度相比，二元交叉熵对高低维分布不一致具有更强的惩罚能力，因此UMAP在兼顾局部结构的同时，更能保留数据的全局关系。UMAP还通过设定距离阈值来约束低维嵌入，避免对邻近样本关系的误判，有效降低了低维空间中的结构失真。此外，在计算性能方面，UMAP利用近似最近邻搜索和图简化策略，使得其在处理大规模单细胞数据时具有更低的时间复杂度和更好的可扩展性。

综上所述，我们打算选择UMAP作为校正效果可视化方法。

## 校正算法设计

|  |
| --- |
| 表3-1 算法设计 |
| 输入：scRNA-seq数据集  输出：批次校正后的scRNA-seq数据集  **数据预处理：**  1：对各批次数据集分别进行质量控制，过滤掉低表达基因和低质量细胞  2：特征选择，选择数据集高可变基因的交集作为数据特征  3：将各批次数据进行归一化操作和对数转换  **特征提取及初步批次校正：**  1：构建一个自编码器，把单细胞数据编码成低维嵌入的无批次信息和批次噪声两个部分  2：构建自编码器的损失函数  3：训练自编码器，并在训练中随机添加批次噪声  **选择参考数据集：**  1：选择一个批次数据作为参考数据集，其余批次数据为剩余批次数据集  **搜索互近邻对：**  1：在剩余批次数据集中选择一个批次数据作为待校正数据集  2：在待校正数据集和参考数据集之间搜索互近邻对  **批次效应校正：**  1：使用搜索到的互近邻对计算出一个校正向量对待校正数据集进行校正  2：将待校正数据集与参考数据集合并为新参考数据集，重复上一步骤直到所有批次数据集都合并到参考数据集中，完成批次校正 |

### 数据预处理

**1.质量控制：**数据集中难免会存在较多的技术性噪声。噪声的存在可能会对后续的数据分析产生严重干扰。为了处理噪声和数据缺失等问题，首先，**过滤低质量细胞和低质量基因**，在数据预处理的时候过滤掉基因表达数小于a的细胞和所有细胞上count总数小于b的基因（a,b的值目前还未确定，日后会确定一个值）。研究表明，在scRNA-seq分析中，低表达基因几乎无法提供有价值的信息[10]。

**2.特征选择：**在质量控制之后的基因表达矩阵仍然具有很高的特征维度，且极其稀疏，其中多数基因在细胞间表达变化有限，对后续分析贡献较小。细胞表达量差异较大的基因称为**高可变基因**（Highly Variable Genes, HVGs）。如今几乎所有批次校正方法在实施前都会先筛选一组HVGs。设表达矩阵为,行代表细胞，列代表基因，表示第个细胞上的第个基因表达值。使用数学方差定义第个基因表达差异。

根据下游实验分析不同，可以在下游实验分析中选择1000到6000个高可变基因。有研究结果表明，在高可变基因数量达到一定阈值后，下游实验分析对高可变基因的数量会变得不敏感[11]。本文拟采用默认取所有数据集高可变基因交集的前2000个特征作为特征选择结果。

**3.标准化和对数转换：**在单细胞转录组测序中，由于每个细胞的转录分子量有限，且不同测序平台与文库制备流程存在技术差异，难以保证样本间完全一致性，易引入显著的技术噪声。**数据标准化**的主要目的是减少此类非生物学差异对后续分析的干扰。

单细胞表达数据通常具有高度离散性，若未加以处理，可能导致异常结果和分析偏差。因此，在数据预处理中常采用对数转换以压缩表达值的动态范围，增强数据稳定性与分析可靠性。为避免对数转换过程中对零值或极小值的处理错误，通常在取对数前对原始表达值加一。

其中，表示对数转换后的基因表达值。

通过以上三个步骤，我们将原始数据集转换成处理后的数据集。后续步骤全部在上完成。

### 特征提取及初步批次校正

设有来自个批次的单细胞表达矩阵 ，每个批次包含个细胞与个基因特征。目标是在潜在空间中将生物学“内容”与批次扰动进行解耦，使内容变量不含批次信息，而批次变量收纳与批次相关的变化，其中表示内容嵌入的维度大小。它对应于编码器输出的那部分表示，期望主要携带细胞的生物学差异（真实信号），而不包含批次效应。表示批次嵌入的维度大小，它对应于编码器输出的另一部分表示，专门用来承载批次特有的变化或噪声。随后通过解码器以重构输入。训练时提供每个细胞的批次标签。

模型架构

模型由编码器、解码器与两路批次判别头组成。编码器以两层MLP（线性、LayerNorm、GELU、Dropout）提取表征，隐层维度设为，并在末端通过两条线性分支产生与。训练态对施加Dropout以提升对抗稳定性。解码器将连接后输入两层MLP（）并映射回维表达空间，输出层不加显式激活以配合均方误差重构。两路批次判别头分别作用在内容与批次嵌入上，结构为带隐藏层的线性分类器，并在所有线性层上采用谱归一化以限制Lipschitz常数；进入判别器前，对嵌入进行L2归一化，避免依靠尺度“取巧”。

为消除中的批次信息，在内容支路前引入梯度反转层（GRL），其前向恒等、反向乘以。采用温升调度：令训练进度，则 ，其中。该调度使训练早期以重构与表示稳定为主，随后逐步增强对抗强度以去除批次痕迹。

损失函数

记输入为，解码重构为；内容与批次嵌入分别为、；批次标签为。分类头在与上得到的logits经过softmax分别给出 与的分布。总体损失包含重构损失、两路交叉熵、去相关约束、内容质心对齐以及批次中心约束：

第一项是重构误差 。它把解码器看作条件高斯的均值预测器：假设观测噪声近似各向同性高斯，最大似然会把最后一层变成线性输出并以均方误差为目标。由于输入在预处理后经历了归一化、与z-score标准化，目标是实数域且可正可负，因此线性输出配合MSE不会引入区间饱和或系统性偏差。优化这项迫使与合起来保留足够的信息以还原，从而在其他“去批次”正则施压时提供一个对抗塌缩的锚点。

第二项是作用在内容支路的交叉熵，但它通过梯度反转层（GRL）对编码器产生相反号的梯度。分类头本身在最小化交叉熵，力图从 判出批次；编码器经由GRL在最大化同一个交叉熵，力图让批次“判不出来”。当这场对抗达到均衡，接近均匀分布，损失逼近，意味着与批次标签的可判别性被压到最低。信息论上，这是在把互信息的可实现上界降到接近0。实现采用温升调度控制GRL的强度，使训练先学稳重构与结构，再逐步加大对抗强度避免发散。

第三项是作用在批次支路的交叉熵，这里没有GRL，因而编码器与分类头都在最小化该项。它的作用与上一项互补：明确鼓励主动携带可判别的批次信息。这样一来，批次相关的变化有“容器”可以承载，解码器也能依赖复原被标准化或对抗训练弱化掉的批次因素，从而把“去批次”的压力从 转移出去，真正实现内容/批次的解耦。实践中若过小，批次信号会泄漏回 ；若过大，可能过拟合而影响重构可泛化性，权衡通常与数据规模和批次数相关。

第四项是跨通道去相关 。记小批或全数据上中心化后的矩阵为、，则

最小化这项使内容与批次在线性意义上近似正交，降低简单线性通道之间的“泄漏”。它与对抗项的关系是互补的：对抗项从“可判别性”角度抑制任意可学习模式（含非线性），而xcov从统计二阶矩角度抑制线性相关。两者共同作用既能削弱可分类的批次痕迹，又能在几何上让两个子空间远离。若单独放大这一项而弱化重构与监督，模型可能通过整体缩放或塌缩来“虚假去相关”，因模型在实现中配合了L2归一化与谱归一化来消除靠范数取巧的途径。

第五项是内容质心对齐。这里是每个批次在内容空间的均值向量。实现上通常逐维计算这些均值在个批次上的方差并取平均，记为

最小化它等价于把不同批次的“内容中心”拉拢在一起，显式消解批次间的整体系偏移。与xcov不同，它直接操作在一阶统计量上；与对抗不同，它不依赖判别器能力，对大尺度平移尤其有效。单独使用这一项可能把所有批次的中心压到同一个点并损害可分的生物学结构，但与重构项和对抗项联用时，中心对齐只会消除跨批的系统偏差，而保留类内变异和生物信号的方向性。

第六项是批次中心损失。这里是对应批次的可学习或动量更新的原型中心，代码中采用动量估计（默认 ），并以此惩罚偏离本批中心的程度。它的作用是让在同一批内部形成紧致簇，既稳定了批次分类头的判别边界，又为解码器提供一致的批次通道输入，避免过度依赖去承担批次差异。实现中这项与 协同工作：交叉熵拉大簇间间隔、中心损失收紧簇内半径，两者共同提升“批次通道”的可用性与稳定性。

从整体视角看，这六项构成了一个“守恒—分离—对齐”的闭环。重构项承担信息守恒，保证编码-解码通路不塌缩；内容支路的对抗交叉熵与xcov把批次信息从中驱逐，质心对齐进一步消解跨批的系统偏差；批次支路的交叉熵与中心损失把被驱逐的信息接住并组织成几何上紧致、判别上可用的子空间。为了让这些力同时起效，模型实现还在工程层面加入了三道关键稳定器：其一，对进入分类头的嵌入做L2归一化，使判别更依赖方向而非范数；其二，对分类头权重做谱归一化，从算子范数上限制放大倍数，相当于控制Lipschitz常数以稳住梯度；其三，对GRL系数采温升调度，从弱到强逐步加压，避免训练初期因强对抗而崩溃。默认权重是一组经验上相对稳健的起点，实际任务中可据批次数、样本量与重构难度做比例性微调：重构误差异常升高时降低对抗/对齐强度，批次可判别性迟迟降不下去时适度提高或延长温升时长，并审视是否需要加大与来处理偏移与线性泄漏。

预处理与特征构建

在进入表征学习之前，分别对每个批次执行质量控制以滤除低质量细胞与低表达基因（例如 min\_genes=200、min\_cells=3），随后进行库大小归一化至 与log变换。每个批次独立筛选高变基因（Seurat，默认2000个），并取各批高变基因集合的交集作为特征子集。当交集规模过小时，需降低阈值或检查批次可比性。基于交集子矩阵对各批进行z-score标准化（截断）以获得分布一致的输入，最后沿细胞维度堆叠形成训练矩阵对应批次标签向量。

训练

训练以Adam优化器进行，学习率、权重衰减，批量大小为 256、迭代50个epoch。所有线性层使用Xavier均匀初始化并将偏置置零。为提升鲁棒性与可分性，在训练态向注入方差为的各向同性高斯噪声，并仅对使用Dropout。每步根据训练进度计算，前向得到，同步更新批次中心并最小化总损失。推断阶段关闭噪声与Dropout，并以编码器批处理提取全量嵌入。

搜搜互近邻对

完成训练后，利用编码器分别提取每个批次的内容嵌入。在该内容空间内执行互为近邻（MNN）对齐：选取样本数最大的批次作为参考集，并为参考与目标批各自建立-近邻图；通过互为近邻筛选得到匹配对集合，再对目标批的每个细胞计算其匹配参考邻居的差分平均，作为平移修正项。由此得到校正后的。将校正后的目标批并入参考集并迭代处理剩余批次。若某一目标批与参考集之间不存在MNN，则保留其内容嵌入不变并继续合并流程。

解码回表达矩阵

为了得到批次校正后的表达层表示，将各批校正后的内容嵌入输入解码器，同时将批次通道置零以明确移除批次因素，即计算。该表示可直接用于后续的可视化与差异分析，而未经对齐的 或其对齐版本则适合用于聚类与下游建模。

推断与结果导出

推断阶段以批处理方式提取全数据的与，并根据需要选择是否进行 MNN对齐与解码重建。为便于复现与下游指标计算，将内容嵌入以二进制数组格式（`.npy`）保存，同时持久化模型参数与对齐后的表示，用于后续的聚类一致性评估（如 ARI、NMI）与可视化。

|  |
| --- |
| 图 3‑1 基于MNN的批次校正算法流程图示 |

### 算法评价指标

为了在不同的数据集上对各种竞争方法进行基准测试，采用以下六个评估指标来量化聚类结果的一致性和消除批次效应的能力。

**调整随机数指数（ARI）**用于量化聚类准确率，可以通过Python模块sklearn.metrics.cluster中的adjusted\_rand\_score函数计算。ARI衡量两个聚类结果之间的相似性。ARI的取值范围为[0,1]，值越大表示相似度越高。

**归一化互信息（NMI）**也用于衡量聚类准确率，可以通过Python模块sklearn.metrics.cluster中的normalized\_mutual\_info\_score函数计算。NMI取值范围为[0, 1]，值越高，聚类结果与真实细胞类型的相似度也就越高。类结果与真实细胞类型的相似度也就越高。

**细胞类型平均轮廓宽度（ASW\_celltype）**用于评估细胞类型纯度，可通过Python模块sklearn.metrics中的函数Silhouette\_score计算。ASW\_celltype是所有细胞轮廓宽度的平均值，取值范围为[0, 1]，值越大表示细胞与具有相同标签的细胞越近，与具有不同标签的细胞越远。我们根据低维（默认维度为32）PCA嵌入空间中预定义的细胞类型标签来计算 ASW\_celltype。

**批次平均轮廓宽度（ASW\_batch）**用于评估批次的全局混合程度。ASW\_batch值越高意味着批次间的互斥性越强，反之，ASW\_batch值越低通常表示批次混合效果更好，批次效应校正也更佳。ASW\_batch的值也在[0, 1]范围内。

**逆辛普森积分指数（iLISI）**用于评估积分后批次的局部混合效果，该指数基于预先选定的困惑度选择的局部邻居。然后，使用选定的细胞邻居，根据iLISI指数的批次标签计算LISI，接近预期批次数的分数表示批次混合效果良好。我们使用R包lisi中的compute\_lisi函数为数据集中的所有细胞计算iLISI，输出平均分数，iLISI越高，表示批次混合效果越好。

**KL散度**用于评估各种单细胞聚类算法在消除批次效应方面的性能，即评估不同批次的细胞在每个聚类中混合的随机性。KL散度越小，表示批次混合效果越好，即在消除批次效应方面越有效。

# 实验结果与分析

模型代码和实验代码均已上传到Githubxxx

本章设计实验对方法进行验证与分析。介绍实验环境以及实验数据。在两个真实的数据集上验证了模型的有效性。为了验证模型的校正能力，选择了若干xxxxxx个较为经典的算法，在四个典型场景的单细胞转录组数据集上进行实验分析。选取UMAP作为可视化方法；选择ARI、LISI这xxx种作为批次校正的评价指标。其中xxxx在细胞类型纯度进行了结果评价；xxx在批次间混合成都进行了评价。

### 实验环境

模型基于Python3.10开发，部分实验代码基于R语言实现。模型开发平台与实验环境如表所示。

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

### 实验数据

为了验证模型能够对多样化的单细胞转录组数据集进行批次校正，选取了2个有代表性的数据集来比较模型和其他方法的批次校正能力。这2个数据集分别是鼠源乳腺上皮细胞数据集(bct\_raw)和。实验数据集的信息如表所示，2个实验数据集都包含批次、细胞类型、测序平台以及组织标签等信息。所有参与比较的方法均为无监督方法，数据集的细胞类型仅在分析批次校正结果的时候使用。

鼠源乳腺上皮细胞数据集(bct\_raw)：这份合并数据对应3篇经典的鼠乳腺上皮scRNA-seq论文，每一篇可以看作一个批次，且跨平台，这会带来强烈的效应，很适合做批次校正。数据集有3个批次，3种类型共9288个细胞，表达矩阵有1222 个基因。Pal et al., 2017（Nat Commun）覆盖从青春期前到青春期最后到成体过程；既有C1（Fluidigm）也有10x，成鼠样本中有10x Chromium的大批量细胞。文中明确报告成鼠总上皮细胞的10x数据并解析出Basal、Luminal progenitor（LP）、Mature luminal（ML）等群体；Bach et al., 2017（Nat Commun）聚焦成年过程中的四个生理阶段（未产、妊娠中、泌乳、退化后），使用10x Chromium，共识别到15个簇，强调连续谱的分化过程（尤其是腔面细胞系）；Giraddi et al., 2018（Cell Reports）记录从胚胎期到成体的全时序图谱，同时使用10x Chromium与C1两种平台（论文图示直接写明两平台）。三个batch之间不仅平台不同（10x vs C1），而且生物学阶段不同（胚胎/青春期/成体/妊娠-泌乳-退化），因此批次效应与真实生物差异强耦合——这是做整合评测的理想难例。

### 批次校正实验结果与分析

本节通过UMAP和6个批次校正评价指标，对各批次校正方法结果进行比

较与分析。这里选择了Uncorrect作为批次校正前的对照组。选择了3种传统算法Seurat、scanorama、harmony和1个来自于顶刊Nature communication的最新算法scDML作为本文算法的比较。在前一节的2个多批次不同规格的数据集上作了实验比较。

1.校正多批次小规模数据集

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

首先在小规模数据集验证各个批次校正方法的批次校正能力。数据包含3个批次（spk、vis、wal）与3类细胞（basal、luminal\_mature、luminal\_progenitor）。表xxx中左边的图的不同颜色的点表示来自不同批次的细胞，右边的图的不同颜色的点表示不同类型的细胞。从图中可以直观看出，同一种细胞类型的细胞因为来自不同批次的原因聚成了两个簇。

数据集选用鼠源乳腺上皮细胞数据集(bct\_raw)。使用UMAP对经过各个批次校正方法校正之后的数据进行可视化，实验结果如图xxx所示。UMAP可视化结果表明，除scanorama外，所有的方法均对鼠源乳腺上皮细胞数据集进行批次校正，并很好的区分了不同的细胞类型。

未校正（Uncorrected）的批次效应非常明显——同一种细胞类型被不同批次切成互不重叠的块；但在每个批次内部，细胞类型本身仍有不错的分离度。也就是说，生物学信号存在，但被批次差异“平移、旋转”后彼此错位。

Seurat的批次颜色在各簇内高度混合，说明对齐充分；同时三类细胞依旧形成清晰的大簇，仅在luminal\_mature与luminal\_progenitor的交界处出现一定“桥接”。整体体现出“批次消除充分、类型保持良好”的理想状态。

Scanorama相较未校正有改进，但仍残留明显的批次分块，同一类型在不同批次之间没有很好对齐；右图中细胞类型主要在各自批次岛内分离，提示该方法在此数据上对跨批次的全局对齐力度偏弱。

Harmony的批次混合度与Seurat相近，颜色在簇内充分交织；细胞类型也被清楚地区分为几个主簇，说明在强力对齐的同时没有显著“过校正”。从可视化来看，与Seurat属于同一梯队的稳健表现。

MyModel的整体呈现出较好的折中：批次在簇内基本混合，仍能看到少量批次偏倚的“尾迹”；三类细胞分离清楚，但簇形较为疏松，个别簇之间有细小连接。直观上已接近Seurat/Harmony，但还有进一步压缩批次残差与抑制簇间“拉丝”的空间。

scDML的颜色显示各批次已充分混合，但右图中不同细胞类型彼此大幅混杂，整体被压成两个大的连续团块，类型边界不清。这是典型的“过校正/过对齐”现象：为了对齐批次，牺牲了生物学异质性，导致下游聚类与标注会变得困难。

总的来说，在这套数据上，Seurat与Harmony 表现最平衡，批次混合高、细胞类型边界清晰；MyModel次之，存在轻微批次残留与簇形疏松；Scanorama对齐不足；scDML则出现过校正。若需要一个默认基线用于下游分析与差异表达，建议优先采用Seurat或Harmony；若使用MyModel，建议继续调参以减少批次残留并稳固簇形；如继续尝试 scDML，应加强同类型拉近/异类型拉远的权重、调小对齐强度或改进 triplet 采样来提高“保留生物学结构”的约束，以避免类型被混合。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| method | ARI | NMI | ASW\_celltype | ASW\_batch | iLISI | KL |
| Uncorrect | 0.993 | 0.983 | 0.276 | 0.423 | 1.037 | 1.061 |
| MyModel | 0.935 | 0.884 | 0.325 | 0.504 | 2.598 | 0.203 |
| Harmony | 0.991 | 0.979 | 0.380 | 0.502 | 2.203 | 0.368 |
| Scanorama | 0.992 | 0.981 | 0.334 | 0.427 | 1.029 | 1.066 |
| Seurat | 0.993 | 0.982 | 0.390 | 0.497 | 1.916 | 0.549 |
| scDML(louvain) | 0.987 | 0.972 | 0.920 | 0.854 | 1.752 | 0.690 |

表xxx是各项数值指标。ARI、NMI与ASW\_celltype衡量“保留生物学结构”，ASW\_batch、iLISI、KL衡量“批次混合”，其中ASW、iLISI越大越好，KL越小越好。

未校正（Uncorrect）首先给出几乎满分的ARI/NMI（0.993/0.983），说明同一批次内的三类细胞本身分得很清；但iLISI≈1.04、ASW\_batch≈0.42且KL≈1.06，说明不同批次几乎没有对齐——即“类型清晰但跨批次错位”。这与未校正UMAP的“同类型被分成多个批次”一致，因此单看ARI、NMI会被误导。

Seurat在“保存生物特征”上最好（ASW\_celltype=0.390为各法最高，ARI/NMI≈0.993/0.982几乎与未校正持平），同时也实现了较好的批次对齐（ASW\_batch≈0.497，iLISI≈1.92，KL≈0.55）。整体呈现类型边界清晰、批次已混合，适合作为稳健批次校正方法。

Harmony与Seurat处于同一梯队，ARI/NMI≈0.991/0.979、ASW\_celltype=0.380，且在批次混合上略胜（ASW\_batch≈0.502、iLISI≈2.20、KL≈0.37）。这意味着Harmony在不明显牺牲类型分离度的前提下，进一步压低了批次残差，是本组数据上综合最均衡的候选。

Scanorama的ARI/NMI仍高（≈0.992/0.981），但批次相关指标基本没有提升（iLISI≈1.03≈未校正、ASW\_batch≈0.43、KL≈1.07）。可能的原因是它主要在批次内维持了原有结构，却未实现跨批次的全局对齐，属于对齐不足。

MyModel在“批次混合”上最强（ASW\_batch=0.504全场最高；iLISI=2.60最高；KL=0.203最低），但“生物学结构”显著回落（ARI/NMI≈0.935/0.884、ASW\_celltype=0.325）。显现出典型的“过对齐”取向：批次被充分打散，但同时拉近了不同细胞类型，导致类型边界变松、可分性下降——与可视化中簇形疏松、簇间有细小连桥的现象一致。

scDML（基于 louvain 的评估）介于二者之间：ARI/NMI=0.987/0.972较高、iLISI=1.752 和 ASW\_label/batch=0.854表明批次混合不错，但 BatchKL=0.690不及 Harmon、/MyModel。

时间复杂度上Harmony≈Scanorama<MyModel<scDML≈Seurat。

总的来看，若考虑综合均衡，Harmony≈Seurat＞其余方法。若最看重“批次彻底消除”，MyModel最佳，但需注意对生物学结构的侵蚀。若更看重“下游基因学解释、差异分析的可靠性”，Seurat、Harmony给出更稳的类型保真度。同时，Scanorama在此数据上对齐不足；scDML需要进一步校准以避免类型被拉得过近。

2.校正多批次复杂数据集

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  | |

选用恒河猴视网膜双极细胞数据集（）（4个批次 M1–M4、十余个亚型）进行实验对比和可视化。

未校正的类型簇形本身很紧，但同一亚型被不同批次切成多个岛，跨批次错位明显。右图能看清大多数亚型，但每个亚型常对应多个批次特有的小团块，这是典型的强批次效应。

Seurat的批次颜色在各簇内已充分交织，跨批次对齐明显好转；同时亚型边界大体保留，只是在个别近缘亚型（如DB系之间）出现轻微桥接和簇形被拉长的现象。整体处在“消批次—保类型”的较好平衡点。

Scanorama的改进有限，仍能看到成片的批次专属团块；右图上各亚型主要在各自批次的岛里分离，跨批次融合不足，属于对齐偏弱。

Harmony的批次混合度与Seurat持平甚至略优，颜色在每个簇内均匀交织；同时亚型簇更为紧致，簇间空隙清晰，几乎没有大面积拖尾或粘连。就图像观感看，这是本组里最稳健、最均衡的表现之一。

MyModel的左图显示批次打散得很充分，混合度接近或优于 Seurat、Harmony；但右图可见若干簇被拉成“条状、云状”，近缘亚型边界有被抹平的趋势，个别簇之间出现细小通道。说明去批次力度略大，牺牲了一部分生物学可分性。

scDML的右图亚型形成规则、圆润的小球状簇，边界清晰；左图同一簇内部多批次颜色均匀混合、几乎无批次偏色块。既见到强批次融合，又很好地保住了亚型结构。

时间复杂度上Harmony≈Scanorama<MyModel<scDML≈Seurat。

综合平衡度看，Harmony≈scDML≥Seurat＞MyModel＞Scanorama＞未校正。若以稳健下游分析为目标，可优先选Harmony或scDML。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| method | ARI | NMI | ASW\_celltype | ASW\_batch | iLISI | KL |
| Uncorrect | 0.864 | 0.870 | 0.181 | 0.495 | 1.462 | 1.060 |
| MyModel | 0.110 | 0.112 | -0.042 | 0.504 | 3.169 | 0.175 |
| Harmony | 0.939 | 0.938 | 0.213 | 0.519 | 3.292 | 0.306 |
| Scanorama | 0.425 | 0.572 | 0.113 | 0.474 | 1.342 | 1.114 |
| Seurat | 0.117 | 0.129 | -0.027 | 0.496 | 1.972 | 0.732 |
| scDML(louvain) | 0.934 | 0.922 | 0.783 | 0.933 | 2.364 | 0.410 |

未校正（Uncorrect）的ARI/NMI=0.864/0.870说明在原始空间里亚型本来就有一定可分性，但ASW\_celltype仅0.181且iLISI≈1.46，提示跨批次的“同型对齐”并不充分；ASW\_batch≈0.495又说明该数据集的批次效应本身不算特别重，这也解释了未校正UMAP上已经能看到部分颜色混合。

Harmony是本组综合最优。ARI/NMI=0.939/0.938，在所有方法里同时给出最高的ASW\_batch（0.519）与最高的 iLISI（3.29），KL也较低（0.306）。这意味着它把四个批次很好地揉在了一起，而且没有明显牺牲亚型边界（ASW\_celltype=0.213，为各法中的最高值）。

scDML（基于 louvain 评估）同样很亮眼。ARI/NMI=0.934/0.922、ASW\_label=0.783、ASW\_label/batch=0.933，iLISI=2.364、KL=0.410。类型可分性与Harmony处在同一梯队，批次混合也很强，仅略逊于Harmony的iLISI 峰值。

MyModel体现出过对齐。iLISI=3.17、KL=0.175、ASW\_batch=0.504都很好，说明批次被充分打散；但ARI/NMI仅0.11/0.11，ASW\_celltype为负（-0.042），意味着不同亚型被挤压在一起、边界消失。这与UMAP中“簇被拉成云、近缘亚型相互渗透”的现象相符。

Scanorama与Seurat在该数据集上不理想。Scanorama的iLISI（1.34）和 KL（1.11）几乎等同甚至劣于未校正，说明跨批次融合不足；ARI/NMI（0.425/0.572）也明显下降。Seurat这次的ARI/NMI更是跌到~0.12，且ASW\_celltype为负。

总的来说，在猴视网膜数据上，Harmony ≳ scDML> MyModel> Scanorama/Seurat >Uncorrect。MyModel目前还存在过校正的现象，损耗了生物学信息，稍微回调对齐力度后，有望进入与Harmony、scDML相当的平衡区间。

# 特色与创新

构建独特结构的自编码器，在提取生物学表达特征的同时完成初步批次对齐，从而无需依赖“生物学差异远大于批次效应”这一较强假设。然而很多基于MNN的方法都依赖于这个假设。

在训练过程中添加随机噪声。在自编码器的训练过程中引入随机的批次噪声，以增强模型的鲁棒性，从而提升其在处理复杂场景数据集时进行批次效应校正的能力。

# 代码

# 经费使用情况

无。

给 MyModel 的快速改进思路：适度下调去批次力度、上调“类型保真”约束。例如减少对 z\_c 的对抗权重（w\_adv）或对齐强度、提高区分性损失（如基于已知标签/伪标签的对比/三元组边界或 center/arcface 类损失），并以 ASW\_celltype 与 ARI/NMI 作为早停与调参目标；同时可限制 z\_c 维度或对 z\_b 加强正则，使批次信息更“收敛在 z\_b”而非渗入 z\_c。这样通常能把你现在的“批次极佳、类型偏软”拉回到 Harmony/Seurat 的平衡区间。

若继续打磨 MyModel，建议略降对齐强度、加强“同类内紧、异类间远”的约束，以避免亚型被拉平。

总的来说， Harmony或scDML。前者更稳健、指标更均衡；后者在你最近的实现里也已达到高水平。若继续打磨 MyModel：降低去批次强度（减小对抗项/对齐损失权重、放缓 GRL 日程）、增强“同类内聚/异类分离”的度量学习约束（triplet/contrastive/center/arcface 等）、适度收窄 z\_c 维度并对 z\_b 强正则；以 ASW\_celltype 与 ARI/NMI 作为早停/调参目标。

评价口径：确保各法在同一聚类器与同一超参下计算 ARI/NMI；ASW\_batch 的定义（是否做了 1–silhouette 翻转）与 iLISI 的采样/邻域大小也要统一，否则会造成“数值不一致而趋势一致”的假象。

一句话总结：这套猴视网膜数据上，Harmony ≳ scDML »（Uncorrect）» Scanorama/Seurat；MyModel 当前处于“批次消得太狠”的一侧，稍微回调对齐力度后，有望进入 Harmony/scDML 的平衡区间。

Pal B, Chen Y, Vaillant F, Jamieson P, Gordon L, Rios AC, Wilcox S, Fu N, Liu KH, Jackling FC, Davis MJ, Lindeman GJ, Smyth GK, Visvader JE. Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. Nat Commun. 2017 Nov 20;8(1):1627. doi: 10.1038/s41467-017-01560-x. PMID: 29158510; PMCID: PMC5696379.

Bach K, Pensa S, Grzelak M, Hadfield J, Adams DJ, Marioni JC, Khaled WT. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. Nat Commun. 2017 Dec 11;8(1):2128. doi: 10.1038/s41467-017-02001-5. PMID: 29225342; PMCID: PMC5723634.

Giraddi RR, Chung CY, Heinz RE, Balcioglu O, Novotny M, Trejo CL, Dravis C, Hagos BM, Mehrabad EM, Rodewald LW, Hwang JY, Fan C, Lasken R, Varley KE, Perou CM, Wahl GM, Spike BT. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. Cell Rep. 2018 Aug 7;24(6):1653-1666.e7. doi: 10.1016/j.celrep.2018.07.025. PMID: 30089273; PMCID: PMC6301014.