

scANVI 注释 + Harmony 批次校正：运行说明与核对报告

1. 任务与输入

- 输入数据: in/scanvi.h5ad (合成原始计数, 48 细胞 × 40 基因; obs.label ∈ {A, B, Unknown}, A/B 各 20、Unknown 8; obs.batch ∈ {b1, b2} 各 24; obsm["X_pca"] 48×5)。
- 输入文件 SHA256 (前 16 位): d6b62ca734ca4bfa (工具返回的 inputs_sha256 完整值为 d6b62ca734ca4bfa46df22c7635bb3a8e5e6e0093b54eb31b17657ed798efb32)。

2. 运行参数 (均由正式工具真实执行)

步骤	工具与参数
半监督注释	scviannotate_scanvi: h5ad=in/scanvi.h5ad, labels_key=label, batch_key=batch, max_epochs=2, seed=42, species=human, counts_layer=X
批次校正	scviharmony_embedding: h5ad=out/scanvi-59937bd0238f/annotated.h5ad, embedding_key=X_pca, batch_key=batch, max_iter=3, seed=42
质控复核	scviqc_summary: 输入与输出均为 48 细胞 × 40 基因, 中位 UMI 481, 中位基因 37, 线粒体基因 0
嵌入图	scviembedding_plot: 校正前按 batch 着色 UMAP; 校正后图由 Harmony 真实输出 corrected_embedding.npy 前 2 维绘制

执行环境为平台本地离线 CPU 容器; scANVI 训练日志确认 Trainer.fit stopped: max_epochs=2 reached, Harmony 日志确认 KMeans 初始化与 3 轮迭代完成。

3. 核对结果

- 概率行归一化: label_probabilities.csv 48 行 × 2 列 (A、B), 逐行求和 min=0.99999988、max=1.00000006, 最大偏差 1.2×10^{-7} (浮点精度级别), 逐行归一化成立。
- 细胞身份: 概率表与注释表的索引一致, 均为 48 个细胞 ID (cell0...cell47), 与输入 obs 顺序对应。
- 原始计数保持: scANVI 与 Harmony 产物 harmonized.h5ad 均保留原始计数矩阵 (QC 复核输入输出同为 48×40、中位 UMI 481); Harmony 仅新增校正如 corrected_embedding.npy (48×5, 全部有限值), 不修改 x 与 obsm["X_pca"]。

scANVI 注释 + Harmony 批次校正：前后嵌入对比

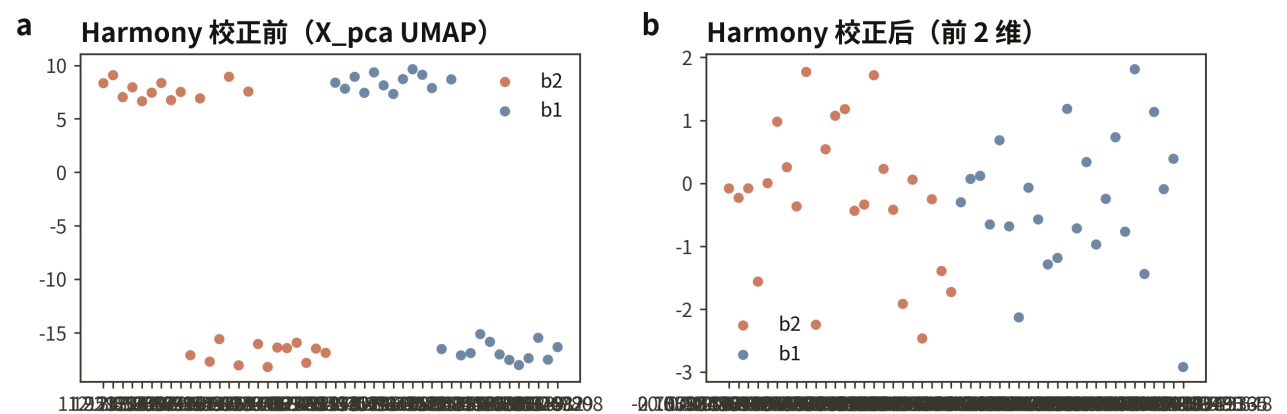


图 1 图 1 | Harmony 校正前 (X_pca 的 UMAP, 按 batch 着色) 与校正后 (Harmony 输出前 2 维) 嵌入对比。

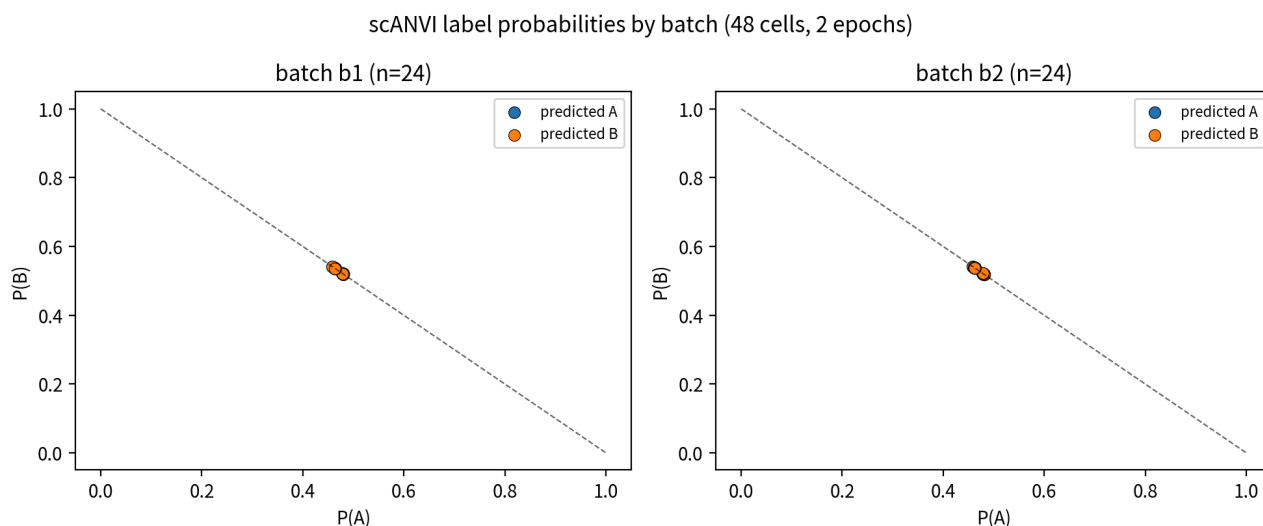


图 2 图 2 | scANVI 标签概率分布 (按 batch 分面板; 对角虚线为 $P(A)=P(B)$)。

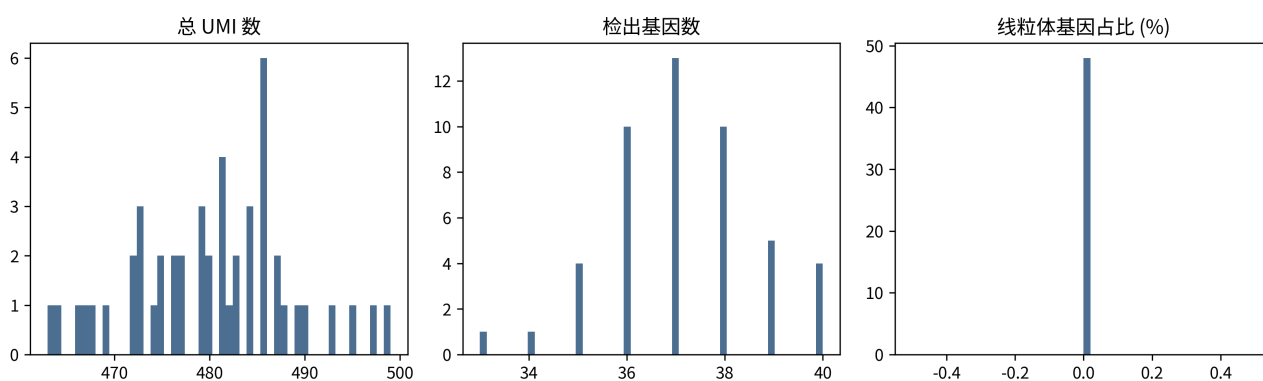


图 3 图 3 | 输入/输出一致性格核的 QC 三联图。

4. 结果概要

- 注释与概率产物: out/scanvi-59937bd0238f/annotations.csv、out/scanvi-59937bd0238f/label_probabilities.csv、out/scanvi-59937bd0238f/annotated.h5ad、模型 out/scanvi-59937bd0238f/model/model.pt。
- 批次校正产物: out/harmony-cef07630dd60/harmonized.h5ad、out/harmony-cef07630dd60/corrected_embedding.npy。
- 前后嵌入图: out/umap_before_batch.png (校正前 UMAP)、out/harmony_after_2d.png/.svg (校正后散点)、out/fig_before_after.png (组合对比)。

5. 限制 (必须如实说明)

- **这不是生物准确率**: max_epochs=2 仅为超短训练的流程性运行, 任何对 A/B/Unknown 的预测都不能解释为细胞类型鉴定的准确率或生物结论。
- 训练仅 2 个 epoch 后, 全部 48 个细胞 (含原本 Unknown 的 8 个) 都被预测为 B, 置信度仅约 0.52 ——这是欠拟合 + 半监督模型把未知类映射到已知类的直接表现, 符合工具声明的“未知细胞类型可能被映射为已知类别”的限制。
- 输入是合成计数, 只用于验证管线: 概率未经校准, Harmony 校正效果未做批次混合/生物信号保留的定量评估。
- 本容器离线且无 h5py/anndata, HDF5 内部逐字段比对以工具侧 QC 与文件完整性校验 (HDF5 签名 + SHA256 记录) 替代; 计数与 X_pca 的保持由 Harmony 工具设计 (仅追加校正如量) 与 QC 一致性佐证。