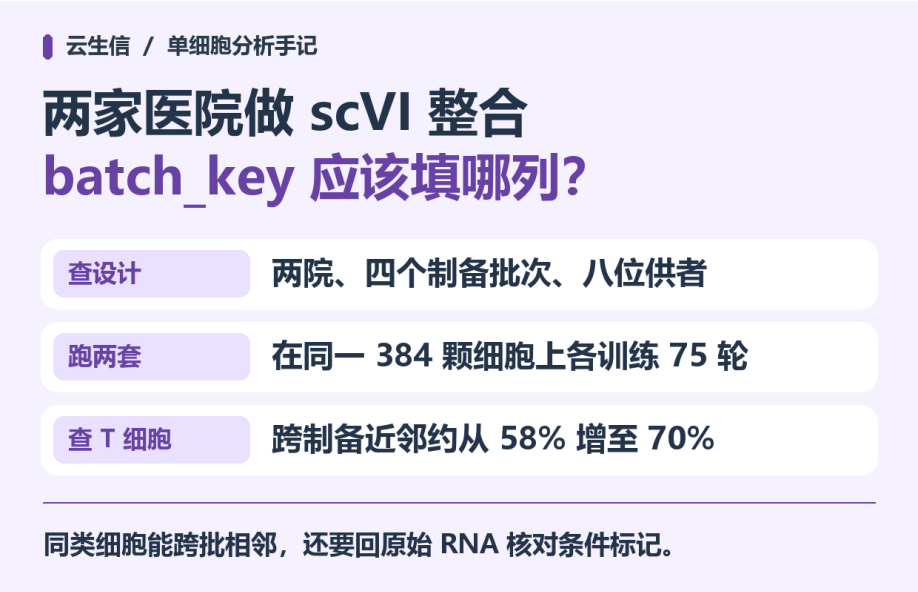
用 scVI 整合两家医院的数据，哪一列该填进 batch_key？

从相同原始 UMI 训练两套 scVI，分别以医院和实际制备批次作为 batch_key，在同一细胞类型里核对混合程度、疾病状态和供者原始标记。



运行教学示例

轻量核对需 numpy、scipy、matplotlib。重训另建 Python 3.12 环境，安装 requirements-model.txt，训练代码和已实跑的两个模型坐标都在包内。

```
python run.py --input example --out results
```

在解压后的资料包根目录运行。example 是教学输入，results 保存本次输出，expected 保存可供核对的参考结果。

输入字段

counts/、cells.tsv、settings.json 原始 UMI、条码、供者、医院、制备批次、病例对照、可信细胞大类。

model/ 与 fit_scvi.py 两份实际训练出的六维表示、训练版本和可重新拟合的完整代码。

核对输出，再换自己的数据

design_table.tsv、raw_markers.tsv 每个候选批次变量与条件的独立供者覆盖，以及原始 RNA 上逐供者的标记均值和检出率。

embedding_audit.tsv、cell_audit.tsv、within_type_edges.tsv 两模型的谱系边界与同类型内近邻组成、逐细胞记录和实际邻居条码名单。

用示例验算

1 两院各有两个病例和两个对照供者；每个技术批次都有两组，部分供者跨轮次重复。

2 T 细胞跨制备批次的近邻比例约 0.580 与 0.702，同条件比例约 0.995 与 0.997。

3 病例 T 细胞在四位供者中 ISG15 的检出率均为 1。此处回到原计数核验，没有用模型表达替代。

换用真实数据

用原始整数计数训练，按供者确认医院、技术批次和条件的对应关系。先按同类型比较两个设置，再将假设相关信号带回供者层核对。

训练随机性、设备和版本会改变潜在坐标。资料包核对脚本验算固定的实跑结果，不能用邻居混合程度代替样本级差异检验。

方法资料

scvi-tools 的建模与输入示例

SCVI 模型和 setup_anndata 接口