cell2location多切片分析：先验收参考签名，再评估空间丰度

从原始计数、基因对应和切片标识开始，分阶段检查参考回归、空间映射与后验结果；解释哪些试跑能估资源，哪些不能支持生物学结论。

cell2location通常分成两步：先用单细胞或单核参考数据估计细胞类型表达签名，再用空间转录组计数估计各位置的细胞丰度。资源试跑可以回答“这批数据大致要多少时间和显存”，不能回答“细胞类型注释是否正确”。以下用Visium类spot数据举例；其他平台要重新确认输入和空间单位。

一、先固定输入和版本

阶段	关键输入	主要验收
参考回归	原始计数、细胞类型与批次	重建QC与标签覆盖
空间映射	切片计数、坐标、参考签名	基因交集与位置丰度
后验导出	已验收模型及抽样设置	内存、耗时与不确定性

空间矩阵与参考矩阵都要保留未归一化的非负计数；不要把log-normalized矩阵直接交给模型。记录cell2location、scvi-tools、PyTorch、CUDA版本，并保存AnnData中实际作为输入的X或layer。参考数据至少要有细胞类型标签；若合并了多个样本或技术平台，还应保留相应batch或技术字段。先查看每类细胞数、缺失标签、重复基因名和每个切片的位置数。

多切片不等于把对象简单拼接。给每个位置保留明确的样本/切片字段，并决定是联合拟合还是逐片分析；该选择会影响技术差异如何估计，需写入分析记录。坐标和组织图像也要逐片对应，不能只因obs顺序一致就假定空间坐标没错。

二、参考签名先单独验收

按官方流程用RegressionModel拟合参考数据：batch_key对应真实样本或批次，labels_key对应细胞类型；不要把生物学上想比较的条件误当成应被消除的技术协变量。运行后检查setup摘要、训练记录和plot_QC重建图。明显偏离对角线、某些类型样本太少或参考组织与空间样本不匹配时，先处理参考，不要急着加显卡。

导出的签名表应是“基因×细胞类型”，行名为唯一基因ID，列名为预期细胞类型。先看类型列是否齐全、数值是否有限且非负，再与空间数据取交集。两边用同一基因ID体系，并按同一交集顺序重排；交集很少时，检查Ensembl ID与gene symbol、版本后缀和重复名，而不是默默继续。

三、用小样本估资源，不做结论

先在每张切片抽取一小部分位置、保留较完整的基因和所有参考细胞类型，建立独立测试目录。固定软件、参数和输入后，分别测参考回归与空间模型；改变batch_size时每次只改这一项。短训练只用于排除接口错误和估计大致资源，不应用其后验图解释组织生物学。

```
# ■■■■■■■■■■■■■■■■■API■■■■■■■■■
import inspect
from importlib.metadata import version
from cell2location.models import RegressionModel, Cell2location
from scvi.model.base import PyroSviTrainMixin
print({p: version(p) for p in ("cell2location", "scvi-tools", "torch")})
print(inspect.signature(PyroSviTrainMixin.train))
# ■■■■■■■■■■■■■■■■■ cell_state_df
# model.train(max_epochs=■■■, batch_size=..., accelerator="gpu", device=0)
# ■■■■■■■■■■■■■■■■■■■■■■■■■■
```

当前scvi-tools的Pyro训练接口使用accelerator与device；旧版本文档可能写use_gpu，不能把不同版本参数混用。若本机接口不同，以锁定版本对应文档为准。batch_size=None会将全量数据复制到设备，通常不是降低显存的办法。除训练外，后验采样/导出也会占用资源，应单独计时和记录峰值。

四、先验和结果要结合组织判断

N_cells_per_location表示每个空间位置的预期细胞数先验，应结合spot面积、组织类型及配对组织学估计，不能照抄教程数字。detection_alpha控制位置间RNA检测差异的建模；若不同切片测序深度差异明显，应比较有依据的候选设定，并观察总RNA、组织学细胞密度和模型估计是否相容。不要为了得到“好看”的空间图反复调参。

正式分析至少检查重建QC、每片总丰度分布、空间图是否落在组织区域、低置信度区间以及各细胞类型是否有参考支持。稀有类型铺满全片、背景区域出现高丰度或不同片差异只跟测序深度一致时，应回查参考覆盖、基因匹配、先验和切片批次。丰度是模型估计而非直接计数，需结合组织学或独立标记解释。

五、记录资源与排查顺序

每阶段保存输入文件哈希、版本、切片数、位置数、基因数、细胞类型数、batch_size、训练轮数、耗时、主机RAM峰值、GPU型号与显存峰值。显存不足先区分发生在模型初始化、训练还是后验导出；RAM不足则检查是否复制了稠密矩阵或同时载入多片。接口报错先看当前版本签名；结果不合理先看数据和参考，升级硬件不能修复输入不匹配。

如果代表性试跑确认现有设备无法按项目时限完成，再按实际位置数、基因数、切片数及后验导出需求评估计算资源；咨询时提供这些清单和测试日志，便于核对配置是否适用。

参考资料

cell2location官方多步骤教程 官方文档
cell2location参考签名模型API 官方文档
scvi-tools Pyro训练接口 官方文档