arXiv:2611.30955v1 [q-bio.GN] 23 Nov 2026

LeCellModel：基因表达流形上的可解释密度估计

Gil Karin* **Artemii Bakulin*** **NirYosef**魏茨曼科学研究所
{gil.karin, artemii.bakulin, nir.yosef}@weizmann.ac.il

摘要

单细胞 RNA 测序的表征学习通常通过其捕获的身份——细胞类型、细胞谱系和扰动——而非其对基因表达流形上细胞状态分布的忠实建模程度来进行评估。因此，密度估计、稀有状态检测、扰动评分和基因级归因都作为独立模型附加在冻结的嵌入之上，与数据的几何结构相互割裂。我们认为，对细胞状态分布进行建模应当是首要目标，而表征质量则由此而来。在 LeJEPA 的基础上——其 SIGReg 目标函数能够强制得到具有可证明最优性的各向同性高斯潜变量——我们提出 LeCellModel。该模型在这一高斯潜变量下读取编码器的雅可比对数体积（JEPA 分数），将其作为相对典型性分数。在多个数据集上，LeCellModel 在 scIB 基准测试的表征质量评估中优于已有方法。此外，我们表明，LeCellModel 能够利用 JEPA 分数作为典型性度量来检测稀有表达状态，并将其应用于受控的细胞因子面板，在不同少数类比例下识别受刺激细胞，同时追踪通路响应幅度。我们提出了典型性分数的基因级可解释性——据我们所知，这是首个在特征空间中分析 JEPA 分数的方法。借助这一框架，LeCellModel 重现了已知的促纤维化巨噬细胞程序与特发性肺纤维化之间的关联，并发现了一个 $CD177^{+}$ 激活的调节性 T 细胞亚群；此前，这一亚群仅在肿瘤背景下被描述过。这两项发现都直接源于所学习嵌入的几何结构，无需依赖监督式注释。代码可在 https://github.com/gilk717/lecellmodel_paper 获取。

1 引言

单细胞 RNA 测序测量数千个单个细胞的基因表达，生成高维计数数据，捕捉每个细胞的分子状态。一项核心计算任务是学习这些状态的表征，以形成具有生物学意义的结构：区分细胞类型、解析发育轨迹，并揭示细胞对扰动的响应。

主流方法一直是训练以重构为目标的编码器-解码器模型，从 scVI [Lopez et al., 2018] 和 scANVI [Xu et al., 2021] 等变分自编码器，到 Geneformer [Theodoris et al., 2023] 和 scGPT [Cui et al.,2024]等 BERT 风格基础模型。然而，重构完全在基因空间中进行，这会促使编码器在保留生物学信号的同时，也保留批次效应和丢失模式等干扰变异。scDREAMER [Shree et al., 2023] 等对抗式批次校正方法在一定程度上缓解了这一问题，但对抗器所对抗的是一个重构目标，而该目标本身会奖励保留与批次相关的变异，以尽量减少误差。最后，也出现了一些新兴方法，例如 GeneJEPA [Litman et al., 2025]

*贡献相同。

第40届神经信息处理系统大会（NeurIPS 2026）。

以及 CellJEPA [ElSheikh et al., 2026],，它们遵循联合嵌入预测架构（JEPA）范式，完全舍弃解码器，通过潜在预测而非重构来塑造嵌入空间。关键在于，所有这些方法的评判标准主要是其表征捕捉到的身份信息，而不是它们描述细胞状态分布的准确程度。

JEPA 有一项显著却尚未得到充分利用的性质：当潜在空间被约束为各向同性高斯分布时，编码器可以被严格证明能够估计训练分布的密度 [Balestriero 等人， 2025]。我们首次在实际的单细胞场景中利用这一性质，推出 LeCellModel：一种基于 JEPA、用于学习细胞状态分布的模型。LeCell Model 基于 LeJEPA [Balestriero 和 LeCun， 2025], ；与以往的单细胞 JEPA 不同，它无需使用停止梯度等启发式方法来防止模型坍塌，而是通过草图化各向同性高斯正则化（SIGReg） [Balestriero 和 LeCun， 2025]，以有理论依据的方式约束各向同性高斯潜变量。此外，我们证明，与扩展至大型 Transformer 架构相比，轻量级多层感知机编码器已足以处理单细胞数据，同时使模型易于计算且雅可比矩阵可计算。在 scIB 基准 [scverse 贡献者， 2024],的多个数据集上，LeCellModel 均取得了最先进的表征质量；同时，在细胞类型内解读其典型性分数，可用于实际检测稀有细胞状态。此外，我们还表明，对输入基因表达求分数的导数，可以将典型性的局部变化归因于具体基因，从而在输入的原生坐标中获得逐基因归因。因此，单个训练完成的编码器可同时充当表征、典型性分数和归因图。

贡献。

- **一种用于估计细胞状态密度的策略**。 我们提出一种策略，利用 JEPA-SCORE [Balestriero 等人， 2025], 即高斯嵌入 JEPAs 的隐式密度估计器，来估计细胞状态的相对密度。结合具有生物学意义的数据增强和批次效应控制，JEPA-SCORE 可在学习到的流形上得出细胞状态密度的相对估计；在单一细胞类型内，我们将其解读为典型性，尽管数据具有随机、稀疏且受批次混杂影响的特性。
- **密度的可微分基因级归因**。 LeCellModel 的一项关键创新是，密度分数本身可对基因表达矩阵求导。因此，我们对其求导，以识别每个细胞中局部表达变化对其 JEPA-SCORE 影响最大的基因，从而将细胞级相对密度估计转化为直接的基因级归因。与差异表达分析不同，这种方法不需要预先设定对比组。
- **LeCellModel：用于单细胞 RNA 测序的高斯潜变量 JEPA**。 我们通过一种无教师的单网络 JEPA 实现这一构想，该模型结合了轻量级多层感知机主干网络、基因空间增强、SIGReg [Balestriero 和 LeCun， 2025], 以及对抗式批次校正。该架构足够轻量，可针对每个数据集单独训练，并保留易于处理的雅可比矩阵，这使密度读数切实可行（第 2.3 节）。
- **在图谱尺度保留细胞身份**。LeCellModel 在图谱尺度整合不同批次的同时保留细胞身份，在 HLCA Core（584K 个细胞）和 Heart v2（704K 个细胞，32 份制备样本）上取得最高的 scIB Total 评分，并在胰腺数据集上与 scVI 并列；它的表现也优于使用相同数据训练的 GeneJEPA，以及基础模型 scGPT 和 UCE（第3.2节；附录 K）。

- **疾病组织中的密度驱动发现**。JEPA-SCORE 能在受控的细胞因子面板中识别受刺激细胞，并追踪通路响应幅度。将其应用于特发性肺纤维化肺组织时，它发现了一个促纤维化单核细胞来源巨噬细胞程序，以及一个 $CD177^{+}$ 活化 Treg 亚群；此前，该亚群仅在肿瘤中有过描述——两者均完全依据嵌入的几何结构识别（第3.3–3.4节）。

2 方法

LeCellModel 构建于联合嵌入预测架构（JEPA）框架之上 [Balestriero 和 LeCun， 2025]。JEPA 通过要求同一输入的增强视图在嵌入空间中能够相互预测来学习表征——无需重构输入。对于单细胞数据而言，这一方法十分自然：由于细胞的身份信息编码在许多相关基因之中，经过随机掩码和泊松重采样的视图仍保留足够的信号，从而能够对同一细胞状态达成一致。在 LeCellModel 中，所有增强视图均经过训练以收敛到共享投影中心；不使用动量教师或非对称预测器头。SIGReg 将投影分布正则化为各向同性高斯分布——防止坍塌，并确保嵌入维度之间以及不同

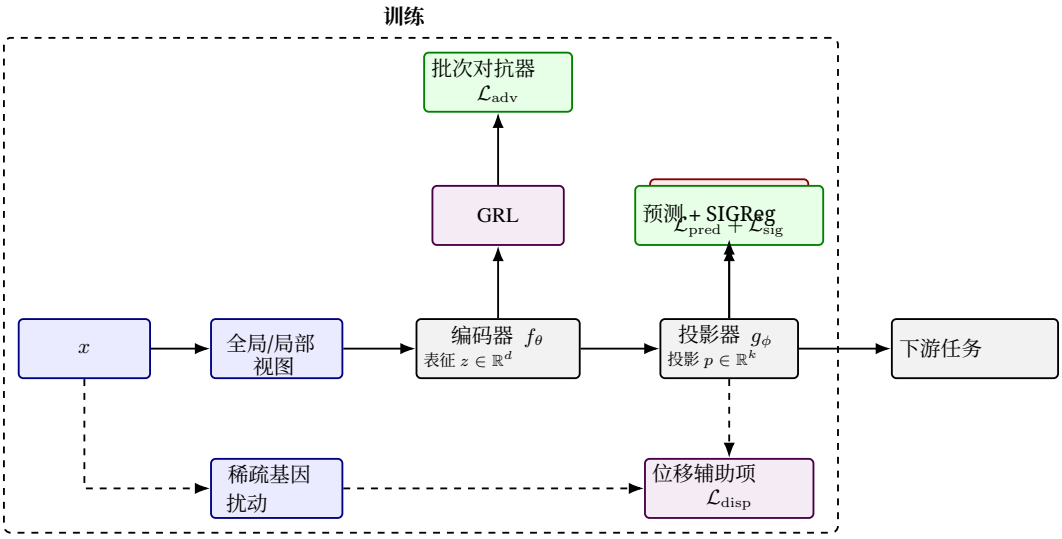

图1：**LeCellModel 概览**。输入细胞被增强为全局视图和局部视图。编码器 f_θ 将每个视图映射为表征 $z \in \mathbb{R}^d$，投影器 g_ϕ 将 z 映射为投影$p \in \mathbb{R}^k$。训练期间，批次监督通过显式的梯度反转层（GRL）作用于 z，JEPA预测和 SIGReg 正则化作用于 p；可选的位移辅助项则预测由施加于输入细胞的稀疏基因空间扰动所引起的投影偏移。投影 p 可直接用作下游任务的嵌入，而 JEPA 分数则根据 $F = g_\phi \circ f_\theta$的雅可比矩阵单独计算。

细胞之间的多样性，同时使 JEPA 分数能够作为密度代理指标（第 2.4 节）。为处理图谱级批次结构，对抗式批次校正会惩罚编码器保留批次预测特征，使其不再编码技术变异。综合而言，这些组件既带来了可靠的细胞状态嵌入，也提供了有原则的密度读数（图 1）。

2.1 问题设定与记号

令 $x \in \mathbb{R}^G$ 表示一个细胞在 G 个高变基因上的原始表达向量，$b \in \{1, \ldots, B\}$ 表示其批次标签，$c \in \{1, \ldots, C\}$ 表示可选的生物学分组（细胞类型，或谱系等更粗粒度的分组）。我们训练一个编码器 f_θ：$\mathbb{R}^G \to \mathbb{R}^d$ 和一个投影器 g_ϕ：$\mathbb{R}^d \to \mathbb{R}^k$，并将二者的复合映射记为 $F = g_\phi \circ f_\theta$，因此 $z = f_\theta(x)$ 和 $p = F(x)$。对抗器头 h_ψ：$\mathbb{R}^d \to \mathbb{R}^B$ 根据表征预测批次。编码器、投影器和对抗器头均为轻量级的两层多层感知机；完整架构见附录C。

2.2 视图生成与数据增强

每个细胞通过对原始计数进行符合生物学原理的数据增强，生成 $n_g = 1$ 个全局视图和 $n_\ell = 6$ 个局部视图：泊松重采样（体现测序深度随机性）、文库大小归一化后取 $\log(1 + x)$，以及以 $m_g = 0$（全局视图）和 $m_\ell = 0.6$（局部视图，扰动更强）为掩蔽率的随机基因掩蔽。在图谱级基准测试（HLCA Core、Heart v2）中不进行泊松重采样。全局视图不进行掩蔽，使其成为局部视图的富含信息的目标（式 1）；两个掩蔽程度低于局部视图的全局视图可获得相近的 scIB 评分（附录 I）。

2.3 训练目标

2.3.1 JEPA 预测损失

令 p^g 表示全局视图的投影，$\{p_j^\ell\}_{j=1}^{n_\ell}$ 表示局部视图的投影。我们将预测目标定义为 $\hat{c} = p^g$（更一般地，当 $n_g > 1$ 时为 $\hat{c} = \frac{1}{n_g}\sum_{i=1}^{n_g} p_i^g$），并最小化

$$\mathcal{L}_{\text{pred}} = \frac{1}{n_g + n_\ell} \sum_{v} \|p_v - \hat{c}\|^2, \tag{1}$$

在批次上取平均，其中求和遍历所有 $n_g + n_\ell$ 视图投影。与 I-JEPA 和以往的单细胞 JEPA 不同，LeCellModel 不使用动量/EMA 教师，也没有非对称预测器头：同一个编码器–投影器生成 $\hat{c}$ 和用于预测的视图，梯度对称地流经两者，且 $\hat{c}$ 上没有停止梯度。坍塌预防完全由 SIGReg（式 2）提供：若没有它，常数投影会使 $\mathcal{L}_{\text{pred}}$ 最小化。

2.3.2 SIGReg 正则化

遵循 LeJEPA 的做法，我们使用沿 $M = 1024$ 随机一维投影进行的 Epps–Pulley 特征函数检验，将投影分布正则化至各向同性高斯分布 $\mathcal{N}(0, I_k)$；检验按视图槽位分别计算后取平均：

$$\mathcal{L}_{\text{sig}} = \frac{1}{(n_g + n_\ell)\, M} \sum_{v} \sum_{m=1}^{M} T_{\text{EP}}^{(v)}(\omega_m). \tag{2}$$

经验特征函数、加权、截断和数值积分的具体内容见附录 E。各向同性高斯目标还确定了读取雅可比对数体积作为相对密度时所依据的几何结构 [Balestriero 等人， 2025]；用 VICReg [Bardes 等人， 2022], 替代 SIGReg 后，由于它仅约束前两阶矩，训练稳定性较差，并且 scIB 和评分（第 4 节）均有所下降。

2.3.3 对抗式批次校正

我们训练 h_ψ 通过梯度反转层 GRL_λ [GaninandLempitsky, 2015]，根据表征预测批次标签 b：

$$\mathcal{L}_{\text{adv}} = \text{CE}\big(h_\psi(\text{GRL}_\lambda(z)), b\big), \tag{3}$$

其中，λ 在预热期间从 0 逐渐增大。附录 F 描述了一种保留粗粒度生物学结构的条件变体，以及两种稳定方法（对 h_ψ 进行谱归一化、在交叉熵中使用标签平滑）。SIGReg 不会整合批次；移除对抗器会使 scIB 总分降低 0.09–0.20，其中 Batch 的降幅最大（附录 I）。

2.3.4 位移辅助项

视图预测主要约束投影器沿数据增强所诱导的方向，而对其他局部基因空间扰动的敏感性则未得到充分约束。因此，我们可选择添加位移辅助项：一个小型预测器估计由稀疏加性基因空间扰动引起的投影偏移 $\Delta p = F(x') - p$，并在预热后使用 25% 的批次子样本，通过结合余弦方向与幅度损失进行训练。这使 scIB（图谱规模数据集上的 $+0.02$ Total）略有提升，但对雅可比矩阵进行正则化可能会降低 JEPA 分数作为典型性信号的精确度；损失形式、扰动流程和消融实验见附录 H。

2.3.5 组合目标

最终训练损失为

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{sig}}\, \mathcal{L}_{\text{sig}} + \lambda_{\text{adv}}\, \mathcal{L}_{\text{adv}} + \lambda_{\text{disp}}\, \mathcal{L}_{\text{disp}}. \tag{4}$$

编码器参数通过梯度反转层最小化 $\mathcal{L}$；对抗器参数直接最小化 $\mathcal{L}_{\text{adv}}$。系数 λ_{adv} 和 λ_{disp} 在预热期间从 0 开始线性递增； λ_{sig} 保持固定。默认值见附录。

2.4 JEPA 分数

定义。 对于训练好的模型 $F = g_\phi \circ f_\theta$，我们定义单细胞评分

$$S(x) \;=\; \sum_{i=1}^{k} \log \max\big(\sigma_i(J(x)),\, \varepsilon\big), \qquad J(x) = \frac{\partial F(x)}{\partial x} \in \mathbb{R}^{k \times G}, \tag{5}$$

其中， $\sigma_i(J(x))$ 是 $J(x)$ 的奇异值， $\varepsilon = 10^{-7}$ 会对接近零的奇异值设置下限，以保证对数计算的数值稳定性。雅可比矩阵是针对复合映射 $F = g_\phi \circ f_\theta$ 计算的，即相对于投影器的输出。

解释。 Balestriero 等人 [2025] 表明，在各向同性高斯分布正则化下，训练完成的 JEPA 的雅可比对数体积在最优状态下近似于 $\log p(x)$，并在合成数据和图像数据上进行了全局验证。单细胞数据不符合这一结果的假设（稀疏计数、掩码和重采样视图、对抗式批次校正），因此我们将 $S(x)$ 视为相对评分，而非经过校准的对数密度。细胞还构成了一个由相关类型组成的结构化流形，各类型的丰度差异极大：在不同类型之间，该评分反映了各类型在数据集中的丰度，这既体现了采样情况，也体现了生物学特征（第 4 节）；而具有生物学意义的信号很少是全局稀有的细胞，更多是在已知类型中连贯的稀有程序。因此，我们在细胞类型内部将该评分解读为典型性。

有两种机制会使稀有且具有预测性的状态获得较低评分。SIGReg 的经验统计量是在批次上取平均，因此某个亚群对它的影响仅与其混合权重成正比；所以，稀有成分受到全局高斯匹配目标的约束较弱。此外，如果稀有状态的跨视图可预测性由紧凑基因程序承载，那么只需少数特征即可满足预测损失，作用方向也相同。因此，稀有且具有预测性的状态会集中到少数占主导地位的雅可比矩阵方向上，从而得到较低的对数体积评分；我们在第 4 节通过实证予以验证，而第 3.4 节则利用该评分定位疾病富集的程序，并通过 $\nabla_x S$ 逐个基因地解读这些程序。

计算。 稳定化得分通过可组合的函数变换（vmap、jacrev、svdvals）进行一次向量化遍历即可计算；详见附录 G。对于 GeneJEPA 这样规模的 Transformer（超过 100M 个参数），每个细胞的雅可比矩阵计算开销约为我们的多层感知机的 $2,000\times$ 倍，因此不适合在图谱规模上进行评分。

2.5 基因层面的梯度归因

对于低评分的细胞 x，我们计算评分相对于基因表达的梯度，

$$\nabla_x S(x) \;=\; \nabla_x \left[\sum_{i=1}^{k} \log \max\left(\sigma_i\left(\frac{\partial F(x)}{\partial x} \right), \varepsilon \right) \right] \;\in\; \mathbb{R}^G, \tag{6}$$

通过二阶自动微分实现，将雅可比矩阵（通过 jacrev 计算）与作用于稳定化对数体积的外层梯度组合起来。

$\nabla_x S(x)$ 中的每个条目都衡量了 $S(x)$ 对相应基因表达的局部敏感度；负值表示这些基因的表达会在局部降低评分。

3 实验

3.1 数据集与评估

数据集。 我们在三个整合基准测试上评估 LeCellModel，这些基准测试的规模相差数个数量级，且主要技术变异来源各不相同：人类肺细胞图谱（HLCA）核心数据集 [Sikkema 等人， 2023]（584，944 个细胞，14 项研究）；心脏细胞图谱 v2 [Litviˇnuková 等人，2020]（704，296 个细胞，来自 3 种测序模式的 22 位供体，共 32 个制备样本）；以及一个多研究人类胰腺基准数据集 [Luecken 等人， 2022]（16，382 个细胞，9 个批次，涵盖六种测序方案）。为开展稀有程序发现，我们进一步组建了两个疾病靶向子集，均取自扩展版 HLCA 发布数据集 [Sikkema 等人， 2023], ，其中结合了健康供体和特发性肺纤维化供体，疾病细胞约占各子集的 8%：*HLCA* 单核细胞–巨噬细胞 子集（257，893 个细胞，125 位供体，13 项研究，20，313 个 IPF细胞）以及*HLCA* 淋巴系 子集（159，652 个细胞，171 位供体，

表 1：三个基准数据集上的批次整合表现，均值 ± 标准差（基于 4 个随机种子：42、1、2、3）。scIB聚合指标基于各方法的输出嵌入计算（定义见第3.2节前言）；粗体表示每个数据集上的最佳结果；各指标的表格见附录L。

	HLCA Core			Heart v2			胰腺		
方法	Bio	批次	总计	Bio	批次	总计	Bio	批次	总计
PCA	0.682±.008	0.332±.003	0.542±.006	0.763±.006	0.314±.002	0.583±.003	0.638±.000	0.254±.000	0.485±.000
Scanorama	0.707±.000	0.388±.017	0.580±.006	0.762±.000	0.342±.009	0.594±.004	0.770±.004	0.556±.006	0.684±.004
Harmony	0.685±.007	0.516±.013	0.618±.009	0.843±.005	0.554±.007	0.727±.004	0.727±.001	0.587±.000	0.671±.001
scVI	**0.729**±.001	0.551±.006	0.658±.002	0.818±.001	0.543±.005	0.708±.002	**0.831**±.003	0.594±.007	0.736±.003
LeCellModel	0.727±.005	**0.632**±.003	**0.689**±.003	**0.849**±.002	**0.627**±.001	**0.760**±.002	0.807±.010	**0.640**±.007	**0.741**±.008

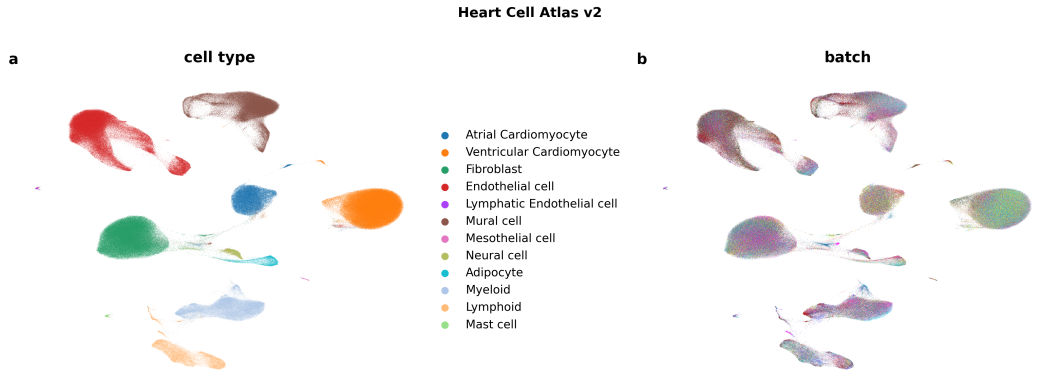

图 2：**LeCellModel 对 HeartCell Atlas v2 的嵌入**。 LeCellModel 编码的 704,296 个细胞的 UMAP。**(a)** 按细胞类型着色。十二种主要心脏细胞群形成了清晰分离的簇，同时保留了生物学结构。**(b)** 按批次着色（涵盖单细胞RNA测序、单核RNA测序和 10x Multiome 的 32 个制备样本）。来自不同批次的细胞在各簇内混合，表明多模态批次效应得到了有效校正。

19 项研究，10，913 个 IPF细胞）；子集筛选标准详见附录 B。Parse Biosciences细胞因子面板将在第 3.3 节中介绍。

基线方法。 我们将 LeCellModel 与四种标准基线方法进行比较：scVI [Lopez 等人， 2018], 一种采用负二项似然并配有显式批次条件解码器的变分自编码器；Harmony [Korsunsky 等人， 2019], 一种在主成分分析空间中运行的迭代批次校正方法；Scanorama [Hie 等人， 2019], 一种在降维空间中进行的互为最近邻对齐方法；以及作为未整合数据下限的原始主成分分析。我们还评估了 GeneJEPA [Litman 等人， 2025]（从头开始训练和微调）、scGPT [Cui 等人， 2024]（零样本和微调）以及 UCE [Rosen 等人， 2026]（零样本）。这些方法均未达到有竞争力的表现：GeneJEPA 的评分低于主成分分析，scGPT 和 UCE 的评分低于 Harmony；单次运行结果见附录 K。

批次整合指标。 我们遵循 scib_metrics [scverse 贡献者的 scIB [Luecken 等人 2022] 重新实现版本，报告十项指标：生物学保留 — 孤立标签、Leiden [Traag 等人 2019] 、归一化互信息、Leiden调整兰德指数、轮廓系数标签、cLISI；批次校正 — BRAS [Rautenstrauchand Ohler 2025]、整合局部逆辛普森指数、kBET、图连通性、PCR比较。生物学和批次分别为各自组内指标的算术平均值；总计为 scIB约定加权平均值（0.6 生物学，0.4 批次）。所有方法均使用四个随机种子进行评估；超参数列于附录中。

3.2 数据集整合

表 1 报告了 LeCellModel 和所有基线方法在三个基准数据集上的生物学保留、批次校正和总计汇总结果；各项指标的详细结果见附录L。

总结。LeCellModel 在所有三个基准测试中均取得最高总分，在图谱尺度上的领先幅度是跨随机种子标准差的许多倍：在 HLCA Core 上领先 scVI $\Delta = +0.031$，在 Heart 上领先 Harmony 2 $\Delta = +0.033$。在小规模胰腺数据集上，LeCellModel 仍具竞争力，在总分上与 scVI 并列（0.741 ± 0.008 对 0.736 ± 0.003）；LeCellModel 在批次校正方面领先，而 scVI 在生物学保留方面领先——在这一情形下，大多数图谱规模方法都处于劣势。scGPT、UCE 和 GeneJEPA 在生物学和批次校正两方面均落后于 LeCellModel（附录 K）。

3.3 使用真实标签验证稀有状态检测

JEPA 分数旨在实现无监督发现，因此需要在存在稀有性标签、但训练时不使用该标签的设定下进行验证。Parse Biosciences PBMC 细胞因子刺激面板 [Parse Biosciences, 2026] 正好提供了这样的条件：受刺激细胞和静息细胞均按暴露条件独立标注——不过，在受刺激细胞群体中，响应的强度和特征各不相同，并非每个标注为受刺激的细胞都表现出可检测的细胞因子特征。我们对受刺激细胞进行子采样，使其占训练混合数据的比例为 $f \in \{0.01, 0.03, 0.05, 0.10\}$，并在训练时不使用标签。我们报告相对于留出标签的 AUROC 和 AUPRC，以及仅在受刺激细胞中，评分与经典通路特征（IL-10 对应 STAT3 靶基因；IL-4 对应 STAT6）之间的斯皮尔曼相关性 ρ ——以考察该评分是否还能按响应的程度对细胞进行排序，而不仅仅是区分受刺激细胞与静息细胞。我们评估了四组（细胞因子、细胞类型）配对，涵盖不同效应量：IL-10 × CD16单核细胞、IL-10 × CD14单核细胞、IL-4 × CD14单核细胞和IL-4 × CD16单核细胞（后者的IL-4响应较弱；配对说明和细胞数量见附录 N）。基线方法包括七种基于密度和重构的异常评分——PCA-kNN、PCA-Mahal、PCA-IForest、scVI-IForest、scVI-kNN、scVI-Mahal和scVI-recon——均在不使用标签的情况下拟合。

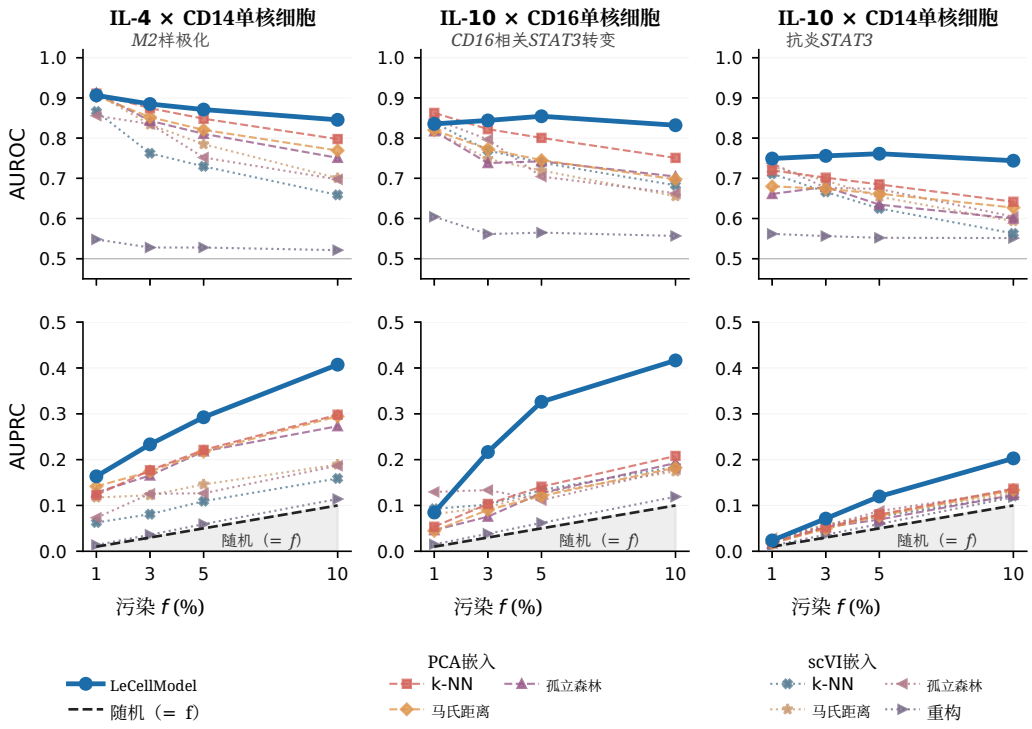

图 3：**稀有状态检测基准。 </style}少数类比例为 %，**3%，5%，10%} 的受刺激细胞；训练期间未见过其标签。AUROC（上）和 AUPRC（下）。其中三组配对具有已确认的生物学响应；CD16 单核细胞中较弱的 IL-4 响应见附录 M。

每种基于密度的基线方法都依赖单一的几何假设；随着 f 增加，受刺激的少数群体会使这些密度估计膨胀，导致性能显著下降（例如，在IL-10 $\times$ CD16上，scVI-kNN的AUROC下降 $0.839 \rightarrow 0.683$ ）。JEPA 分数在不同 f下保持稳定，在三个已确立的响应配对上，AUROC和AUPRC均领先，且超过 $f = 1\%$（图 3；表 12、13）。在LeCellModel自身的嵌入中计算同类评分，无法复现这一结果：在三个已确立的响应配对上，kNN密度和马氏距离的表现低于随机水平，而孤立森林逊于JEPA 分数，仅在低 f时的IL-4 $\times$ CD14上接近后者（表15）。

在受刺激细胞中，JEPA 分数与反应幅度 相关（斯皮尔曼 ρ 与通路特征的相关性；该特征在静息细胞上计算 z 分数，且无需获取基因列表）；LeCellModel 在生物学意义更强的配对上表现最佳（表 14；配对的生物学信息见附录 N）。

3.4 稀有基因程序发现

在同一种细胞类型内，低 JEPA 分数细胞可标示稀有且具有预测性的状态（第 2.4 节）。我们可以从两个互补方向恢复此类区域的基因程序：**基因 $\rightarrow$ 细胞**，即在嵌入 kNN 图中寻找表达空间上与低分共定位的基因（双变量莫兰指数；附录 O，表 16）；或**细胞 $\rightarrow$ 基因**，即提取低分邻域，并依据基因对评分的局部效应对其进行排名（$\nabla_x S$；第 2.5 节）。这两种视角在重叠部分汇聚为同一程序，但 $\nabla_x S$ 以单个细胞为单位运行，能够揭示评分驱动因素，而非共现标志物——这是任何群体层面的相关性都无法做到的。我们将其应用于 HLCA 单核细胞–巨噬细胞队列（13 项研究）以及 HLCA 淋巴系正常+特发性肺纤维化队列（19 项研究）；正是对这些贡献者进行批次校正，才使共享程序得以恢复。

单核细胞来源巨噬细胞。 在单核细胞来源巨噬细胞中，LeCellModel 分离出一个低 JEPA 分数亚群，该亚群对应于促纤维化巨噬细胞程序（SPP1、CHI3L1、CHIT1、ADAMDEC1） [Morse 等人， 2019], 与 HLCA 疾病分析中强调的纤维化相关 SPP1$^+$ 状态一致 [Sikkema 等人， 2023] 并在此基础上进一步纳入基质重塑和免疫抑制效应分子（CTSK、SIGLEC15）。该细胞群中 IPF 细胞富集（在单核细胞来源巨噬细胞中占 31.7%，基线为 12.5%；富集 2.53 倍），涵盖 53 位供体（其中 16 位患有 IPF）。

CD4 T细胞。 在 CD4 T细胞中，LeCellModel 分离出一个由 FOXP3、CD177、CTLA4、CCR8 和 LAYN 定义的低 JEPA 分数细胞群——即此前在肿瘤微环境中表征的*CD177*$^+$ 活化调节性 *T* 细胞 表型 [Kim 等人， 2021]。据我们所知，该表型此前未曾在非癌的纤维化 IPF 肺中报道。LeCellModel 仅凭嵌入几何便识别出该细胞群：逐研究标签中只有一个研究组标注了调节性 T 细胞（Meyer 2019/2021），而 HLCA 的统一词汇表中完全没有此标签，因此，对该队列进行基于标签的分析很可能会完全遗漏这一细胞群。恢复的程序与 Kim 等人描述的 CD177$^+$ 肿瘤浸润调节性 T 细胞状态高度重叠 [2021], , 包括与近期组织浸润相关的高度活化和抑制性表型。该细胞群中的 IPF 细胞也显著富集（在 CD4 T细胞中占 34.7%，基线为 5.3%；富集 6.57 倍），涵盖 95 位供体（其中 17 位患有 IPF）。将该簇限制为 CD177$^+$ 细胞后，疾病关联进一步增强，IPF 细胞比例升至 56.9%（为 CD4 基线的 10.76 倍），涉及 51 位供体（其中 13 位患有 IPF）（图 4）。

基因级归因。 给定一个低 JEPA 分数亚群 C，我们根据簇均值输入×梯度的绝对值对基因进行排名，$|\mathbb{E}_{x \in C}[x \odot \nabla_x S]|$，其中单细胞梯度 $\nabla_x S(x) \in \mathbb{R}^G$ 的定义见第 2.5 节。与针对重构或分类损失计算的积分梯度不同，后者往往突出细胞类型标志物；这种方法揭示了影响细胞局部典型性和稀有状态几何的基因——无需预先定义基因列表，即可将 JEPA 分数从单细胞异常检测器扩展为基因水平解释工具。在 CD4 T细胞中，CD177 和 CCR8 的原始梯度幅值最大，分别为 $|\mathbb{E}_{x \in C}[\partial S / \partial x]| \approx 11.0,\ 9.4$——该评分对活化 Treg 亚群的定义性标志物在局部最为敏感。在单核细胞来源巨噬细胞中，梯度将基质重塑和检查点效应分子（MMP9、CTSK、SIGLEC15）排在前列，而恢复的程序中丰度最高的标志物 SPP1 却未进入梯度排名前列：梯度往往揭示局部驱动评分的因素，而非一般性标志物。该排名找回了已知的通路靶标，并对差异表达分析形成补充：在 CD14 细胞因子对中，它将 STAT3 和 STAT6 靶基因 F13A1 和 ALOX15 排在首位；在特发性肺纤维化中，它将程序基因 CD177

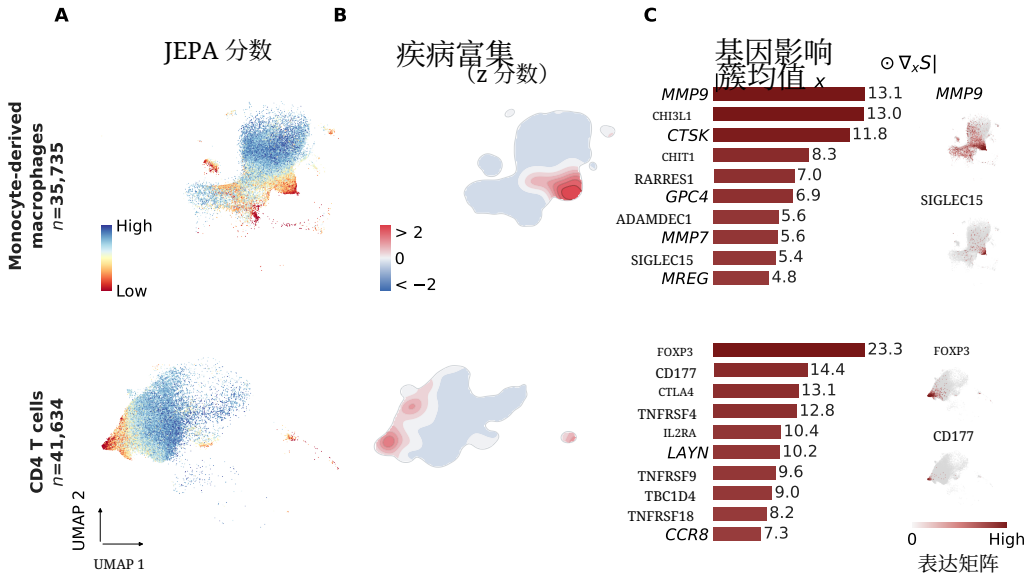

图 4：**特发性肺纤维化肺部的稀有程序发现与基于评分的基因归因**。 对于每个队列，均基于所有细胞计算 UMAP；每一行显示指定细胞类型在完整嵌入中的坐标。 **（A）JEPA 分数：** 低 JEPA 分数的细胞（红色）在该细胞类型中形成一个连贯的邻域。 **（B）疾病富集：** k经 NN 平滑的疾病富集 z分数；与细胞类型基线相比，低 JEPA 分数区域中的 IPF细胞富集（单核细胞来源巨噬细胞：31.7% 对 12.5%， $2.53\times$；CD4 T细胞：34.7%对 5.3%， $6.57\times$）。 **（C）基因影响：** 低 JEPA 分数细胞中，按簇平均输入×梯度绝对值排序的前 10 个基因；每一行右侧显示具有代表性的基因表达UMAP图。在单核细胞来源巨噬细胞中，低 JEPA 分数区域捕获了促纤维化巨噬细胞程序，而梯度优先识别重塑相关基因和检查点基因，如*MMP9*、*CTSK*和*SIGLEC15*，而非丰度最高的标记基因*SPP1*。在 CD4 T细胞中，同样的分析捕获了以*FOXP3*、*CD177*、*CTLA4*、*CCR8*和*LAYN*为核心的 CD177⁺ 活化调节性 T 细胞程序。簇选择仅使用 JEPA 分数和嵌入几何；基因列表不作为输入。

和 SIGLEC15 分别排在第 2 和第 9 位，而差异表达排名分别为第 25 和第 43 位（附录 R）。底层评分在不同随机种子和增强设置下保持稳定（巨噬细胞和淋巴系队列中，四个随机种子间单细胞斯皮尔曼相关系数的平均两两值为 $\rho = 0.965$，且为 0.843；附录 Q）。

4 分析

预测选择性。 在细胞类型尺度上，评分的表现符合 Balestriero 等人提出的全局密度读数预测 [2025]： 在 Heart v2 上，平均 JEPA 分数随细胞类型丰度增加（斯皮尔曼相关性 $\rho = 0.86$ 介于对数$_{10}$ 频率与平均评分之间）。第 2.4 节预测了细胞类型内部 更精细的机制：在 SIGReg 下，跨视图可预测性由紧凑基因程序承载的稀有状态，其表征应比常见状态集中于更少的主导方向。我们直接对此进行检验：对于每个群体，我们测量其嵌入在多大程度上可由最佳匹配的一个或两个基因近似（最佳1基因和最佳2基因余弦相似度，$\cos_1$,$\cos_2$）。这一预测成立。对于 CD177⁺ 调节性 T 细胞（FOXP3 >1 和 CD177 >1, $n = 182$），LeCellModel 在更广泛的 CD4 T细胞群中达到 $\cos_1/\cos_2 = 0.812/0.867$，而 $0.367/0.477$ 为（$n = 500$）；促纤维化单核细胞来源巨噬细胞也呈现相同差距（$0.717/0.821$ 与 $0.385/0.534$）。这种效应只出现在 SIGReg 诱导的几何结构中：在 scVI 下，稀有状态与常见状态对比几乎消失（CD177⁺ 调节性 T 细胞 $0.468/0.593$ 与 CD4 $0.359/0.473$；促纤维化单核细胞来源巨噬细胞 $0.326/0.465$ 与普通单核细胞来源巨噬细胞 $0.335/0.450$）。该评分的

几何机制，而不仅仅是嵌入质量，才是实现同类型稀有状态恢复的关键。

正则化器消融。 将 SIGReg 替换为 VICReg [Bardes et al., 2022]，（它仅约束前两阶矩）会同时削弱评分和表征：最稀有的 1% 中 $CD177^+$ Treg 的占比从 69% 降至 37%，IL-10AUROC 平均下降 0。03，而 HLCA Core scIB Total 从 0。689 降至 0。659（附录 I）。

5 结论

LeCellModel 的关键设计选择在于，SIGReg 既能防止模型坍塌，又赋予雅可比对数体积明确的含义。因此，JEPA 分数及其梯度直接来自训练好的模型，而非事后添加；它们之所以具有可解释性，正是因为模型的训练方式。

迈向图谱规模参考嵌入。 一个自然的下一步是将 LeCellModel 用作图谱参考：在大型多研究队列上训练一次（可采用更庞大的编码器骨干网络，例如 Transformer），冻结模型，再将查询细胞映射到由此得到的坐标系中。新样本将获得整合嵌入、相对于固定参考几何的逐细胞 JEPA 分数，以及基因层面的 $\nabla_x S$ 归因，从而使稀有状态发现成为查询时操作，而不再需要针对每个数据集反复训练。一个尚待解答的问题是，当查询细胞落在训练流形之外时，评分的典型性解释会如何迁移。

局限性。 LeCellModel 在低数据 regime（胰腺 $\sigma = 0.008$）下的 scIB 总分跨随机种子方差大于图谱尺度（Heart v2、HLCA Core $\sigma \leq 0.003$），因此解读小型数据集上的单个随机种子排名时应谨慎。位移辅助消融（附录 H）所用的对抗器和 SIGReg 系数与启用位移的运行略有不同；这些系数是为确保不使用辅助项时的稳定性而选定的，因此比较并非完全匹配超参数。JEPA 分数是训练后模型的几何量，并非有保证的生物学读出：它取决于训练动态以及编码器接触到的数据增强，因此不同运行在稀有尾部边界处可能会有分歧，不过发生变化的细胞仍接近截止线（附录 Q）。我们目前尚无法区分模型不确定性与真实稀有性，也无法检验梯度对有限计数扰动的反应；我们的阴性对照（附录 S）仅确认了已知表型；新的状态仍需复现和实验验证。最后，$\nabla_x S$ 确定了局部表达对评分影响最大的基因；这些是稀有状态几何的解释性标志物，而非经验证的生物学驱动因素。

致谢与经费披露

作者未获得本研究的专项经费，并声明不存在利益冲突。

参考文献

Randall Balestriero 和 Yann LeCun。《LeJEPA：无需启发式方法的可证明且可扩展的自监督学习》。arXiv 预印本 arXiv:2511.08544，2025年。Randall Balestriero、Nicolas Ballas、Mike Rabbat 和 Yann LeCun。《高斯嵌入：JEPA 如何悄然学会数据密度》。arXiv 预印本 arXiv:2510.05949，2025年。Adrien Bardes、Jean Ponce 和 Yann LeCun。《VICReg：用于自监督学习的方差-不变性-协方差正则化》。发表于国际学习表征会议（*ICLR*），2022年。Haotian Cui、Chloe Wang、Hassaan Maan、Kuan Pang、Fengning Luo、Nan Duan 和 Bo Wang。《scGPT：利用生成式人工智能构建单细胞多组学基础模型》。发表于*NatureMethods*，21(8):1470–1480，2024年。doi：10.1038/s41592-024-02201-0。David DeTomaso 和 Nir Yosef。《Hotspot：识别单细胞基因组学跨模态的资讯性基因模块》。发表于*Cell Systems*，12(5):446–456.e9，2021年。doi：10.1016/j.cels.2021.04.005。

Ali ElSheikh、Rui-Xi Wang、Weimin Wu、Yibo Wen、Payam Dibaeinia、Jennifer Yuntong Zhang、Jerry Yao-Chieh Hu、Mei Knudson、Sudarshan Babu、Shao-Hua Sun、Aly A. Khan 和 Han Liu。《Cell-JEPA：单细胞转录组学的潜在表征学习》。arXiv 预印本 arXiv:2602.02093，2026年。Yaroslav Ganin 和 Victor Lempitsky。《通过反向传播实现无监督域适应》。发表于第*32*届国际机器学习会议论文集，ICML，2015年。doi：10.48550/arXiv.1409.7495。Brian Hie、Bryan Bryson 和 Bonnie Berger。《利用 Scanorama 高效整合异质性单细胞转录组》。发表于《自然-生物技术》，37(6):685–691，2019年。doi：10.1038/ s41587-019-0113-3。Myung-Chul Kim、Nicholas Borcherding、Kawther K. Ahmed、Andrew P. Voigt、Ajaykumar Vish- wakarma、Ryan Kolb、Paige N. Kluz、Gaurav Pandey、Umasankar De、Theodore Drashansky 等。《Cd177 调节肿瘤浸润性调节性T细胞的功能与稳态》。发表于《自然-通讯》，12(1):5764，2021年。doi：10.1038/s41467-021-26091-4。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》。发表于*NatureMethods*，16(12):1289–1296，2019年。doi：10.1038/s41592-019-0619-0。Elon Litman、Tyler Myers、Vinayak Agarwal、Ekansh Mittal、Orion Li、Ashwin Gopinath 和 Timothy Kassis。《GeneJEPA：转录组的预测性世界模型》。发表于*bioRxiv*，2025年。doi：10.1101/2025.10.14.682378。Monika Litviˇnuková、Carlos Talavera-López、Henrike Maatz、Daniel Reichart、Catherine L. Worth、Eric L. Lindberg、Masatoshi Kanda、Krzysztof Polanski、Matthias Heinig、Michael Lee、Emily R. Nadelmann、Kenny Roberts、Liz Tuck、Eirini S. Fasouli、Daniel M. DeLaughter、Barbara Mc- Donough、Hiroko Wakimoto、Joshua M. Gorham、Sara Samari、Krishnaa T. Mahbubani、Kourosh Saeb-Parsy、Giannino Patone、Joseph J. Boyle、Hongbo Zhang、Hao Zhang、Anissa Viveiros、Gavin Y. Oudit、Omer Ali Bayraktar、J. G. Seidman、Christine E. Seidman、Michela Noseda、Norbert Hübner 和 Sarah A. Teichmann。《成年人人类心脏的细胞》。发表于*Nature*，588(7838): 466–472，2020年。doi：10.1038/s41586-020-2797-4。Romain Lopez、Jeffrey Regier、Michael B. Cole、Michael I. Jordan 和 Nir Yosef。《用于单细胞转录组学的深度生成建模》。发表于*Nature Methods*，15(12):1053–1058，2018年。doi：10.1038/ s41592-018-0229-2。Malte D. Luecken、Maren Büttner、Ketan Chaichoompu、Alessandro Danese、Marta Interlandi、Mirko F. Mueller、Daniel C. Strobl、Luke Zappia、Martin Dugas、Maria Colomé-Tatché 和 Fabian J. Theis。《单细胞基因组学中的图谱级数据整合基准评测》。发表于*NatureMethods*，19(1):41–50，2022年。doi：10.1038/s41592-021-01336-8。Cameron Morse、Tracy Tabib、Jen Sembrat、Kristy L. Buschur、Hiam T. Bittar、Erika Valenzi、Yong Jiang、Dan J. Kass、Kevin Gibson、Wei Chen 和 Robert Lafyatis。《特发性肺纤维化中增殖的 SPP1/MERTK 表达型巨噬细胞》。发表于《欧洲呼吸杂志》，54 (2):1802441，2019年。doi：10.1183/13993003.02441-2018。Parse Biosciences。《单次实验中的1000万个人外周血单个核细胞》。数据集及分析资源，2026年。URL https://www.parsebiosciences.com/datasets/10-million-human-pbmcs-in-a-single-experiment/。访问日期：2026年4月24日。Pia Rautenstrauch 和 Uwe Ohler。《轮廓系数在单细胞整合基准评测中的局限》。《自然-生物技术》，2025年。doi：10.1038/s41587-025-02743-4。Yanay Rosen、Yusuf Roohani、Ayush Agrawal、Leon Samotorˇcan、Tabula Sapiens联盟、Stephen R. Quake 和 Jure Leskovec。《通用细胞嵌入为细胞生物学基础模型提供支撑》。发表于*Nature*，656:183–191，2026年。doi：10.1038/s41586-026-10689-z。scverse贡献者。《scib-metrics：scIB 单细胞整合基准测试指标的GPU加速重实现》。GitHub代码库，2024年。URL https://github.com/YosefLab/ scib-metrics。访问日期：2026年4月24日。

Ajita Shree、Musale Krushna Pavan 和 Hamim Zafar。利用与对抗分类器配对的深度生成模型，对单细胞数据集进行图谱级整合的 scDREAMER。《自然·通讯》，14(1):7781，2023。doi: 10.1038/s41467-023-43590-8。

Lisa Sikkema、Ciro Ramírez-Suástegui、Daniel C. Strobl、Tessa E. Gillett、Luke Zappia、Elo Madissoon、Nikolay S. Markov、Laure-Emmanuelle Zaragosi、Yuge Ji、Meshal Ansari、Marie-Jeanne Arguel、Leonie Apperloo、Martin Banchero、Christophe Bécavin、Marijn Berg、Evgeny Chichel-nitskiy、Mei-I Chung、Antoine Collin、Aurore C. A. Gay、Janine Gote-Schniering、Baharak Hooshiar Kashani、Kemal Inecik、Manu Jain、Theodore S. Kapellos、Tessa M. Kole、Sylvie Leroy、Christoph H. Mayr、Amanda J. Oliver、Michael von Papen、Lance Peter、Chase J. Taylor、Thomas Walzthoeni、Chuan Xu、Linh T. Bui、Carlo De Donno、Leander Dony、Alen Faiz、Minzhe Guo、Austin J. Gutierrez、Lukas Heumos、Ni Huang、Ignacio L. Ibarra、Nathan D. Jackson、Preetish Kadur Lakshminarasimha Murthy、Mohammad Lotfollahi、Tracy Tabib、Carlos Talavera-López、Kyle J. Travaglini、Anna Wilbrey-Clark、Kaylee B. Worlock、Masahiro Yoshida、Maarten van den Berge、Yohan Bossé、Tushar J. Desai、Oliver Eickelberg、Naftali Kaminski、Mark A. Krasnow、Robert Lafyatis、Marko Z. Nikolic、Joseph E. Powell、Jayaraj Rajagopal、Mauricio Rojas、Orit Rozenblatt-Rosen、Max A. Seibold、Dean Sheppard、Douglas P. Shepherd、Don D. Sin、Wim Timens、Alexander M. Tsankov、Jeffrey Whitsett、Yan Xu、Nicholas E. Banovich、Pascal Barbry、Thu Elizabeth Duong、Christine S. Falk、Kerstin B. Meyer、Jonathan A. Kropski、Dana Pe'er、Herbert B. Schiller、Purushothama Rao Tata、Joachim L. Schultze、Sara A. Teichmann、Alexander V. Misharin、Martijn C. Nawijn、Malte D. Luecken 和 Fabian J. Theis。健康与疾病状态下的肺部整合细胞图谱。《自然·医学》，29:1563–1577，2023。doi: 10.1038/s41591-023-02327-2。

Christina V. Theodoris、Ling Xiao、Anant Chopra、Mark D. Chaffin、Zeina R. Al Sayed、Matthew C. Hill、Helene Mantineo、Elizabeth M. Brydon、Zexian Zeng、X. Shirley Liu 和 Patrick T. Ellinor。迁移学习使网络生物学中的预测成为可能。*Nature*，618(7965):616–624，2023年。doi: 10.1038/s41586-023-06139-9。

Vincent A. Traag、Ludo Waltman 和 Nees Jan van Eck。从 Louvain 到 Leiden：确保形成强连通社区。*ScientificReports*，9:5233，2019年。doi: 10.1038/s41598-019-41695-z。

Chenling Xu、Romain Lopez、Edouard Mehlman、Jeffrey Regier、Michael I. Jordan 和 Nir Yosef。使用深度生成模型对单细胞转录组学数据进行概率协调与注释。*Molecular Systems Biology*，17(1):e9620，2021年。doi: 10.15252/msb.20209620。

A 数据可用性

HLCA Core 可从 CZ CELLxGENE 获取（https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293）；此处使用的确切数据集版本为 b351804c-293e-4aeb-9c4c-043db67f4540.h5ad。HLCA 扩展图谱可从同一门户获取；第 3.4 节中使用的单核细胞–巨噬细胞和淋巴系子集，是通过手动从扩展版中筛选出对应的细胞谱系，并仅保留正常+特发性肺纤维化供体而获得的。心脏细胞图谱 v2 可从 https://www.heartcellatlas.org/ 获取。多研究人类胰腺基准数据集（human_pancreas_norm_complexBatch.h5ad）由 Luecken等人发布于 https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_-_integration_task_

数据集_免疫_和_胰腺_/12420968。Parse Biosciences PBMC 细胞因子刺激面板可从 https://www.parsebiosciences.com/datasets/ 10-million-human-pbmcs-in-a-single-experiment/ 获取；对于稀有状态检测实验（第 3.3 节），我们使用 CellFlow 在 https://cellflow.readthedocs.io/en/stable/user/genapi/cellflow.datasets.pbmc_cytokines.html 发布的 2,000 高变基因版本。

B 数据集预处理详情

使用原始计数矩阵作为输入；在视图生成器中进行文库大小归一化和 log1p 处理（第 2.2 节）。

高变基因选择。 三个整合基准测试使用 2,000 个高变基因，HLCA 单核细胞–巨噬细胞和淋巴系子集使用 3,000 个高变基因；这些基因采用基于批次的 seurat_v3 方法从原始计数中选出。

HLCA 队列子集筛选。 两个子集均从 HLCAextendedrelease 中提取，筛选出相关谱系的细胞类型和正常+特发性肺纤维化供体；从两个队列中排除 Kaminski 等人，2020年的研究，并剔除贡献少于 5,000 个（髓系）或 1,000 个（淋巴系）细胞的数据集。最终，单核细胞–巨噬细胞队列包含 13 项研究，淋巴系队列包含 19 项研究。

图谱规模视图生成。 对于图谱规模数据集（HLCA Core、Heartv2），省略泊松重采样，因为数据集层面的深度差异已能提供足够的视图多样性。

C 架构

该**编码器** f_θ 是一个带末端层归一化的两层多层感知机：

线性层($G \to h$) $\to$ 层归一化(h) $\to$ GELU $\to$ 丢弃 $\to$ 线性层($h \to d$) $\to$ 层归一化(d),

默认值为 $d = 256$, $h = 512$。

该**投影器** g_ϕ 是一个两层多层感知机，不含 末尾的归一化层：

$$\text{Linear}(d \to 512) \to \text{LayerNorm}(512) \to \text{GELU} \to \text{Dropout} \to \text{Linear}(512 \to k),$$

采用默认设置 $k = 128$。

该一 **对抗器头部** h_ψ 是一个两层多层感知机，前置层归一化和带泄漏的整流线性单元激活，针对 B 批次的逻辑值：

$$\text{LayerNorm}(d) \to \text{Linear}(d \to d) \to \text{层归一化}(d) \to \text{带泄漏的整流线性单元} \to \text{丢弃}(0.1) \to \text{Linear}(d \to B).$$

条件对抗器会在输入前添加一个经过学习的细胞类型嵌入 $e_c \in \mathbb{R}^{32}$，并将第一个 LayerNorm 和 Linear 替换为 LayerNorm($d + 32$) $\to$ Linear($d + 32 \to d$)。本文报告的所有实验均未使用条件变体；所有结果均使用无条件对抗器。

D 视图生成详情

对原始计数应用全局和局部视图增强 $x \in \mathbb{R}^G$：(i) 泊松重采样，$\tilde{x}_j \sim$ 每个基因服从 Poisson(x_j)；(ii) 文库大小归一化至目标总和 10^4，然后进行 log($1 + \cdot$)；

(iii) 以 m_g （全局）或 m_ℓ （局部）的比率进行随机基因掩蔽。(m_g， m_ℓ) 在所有三个整合基准测试中均为 (0，0.6)；IPF 和稀有状态检测实验使用 (0.5，0.7)（见表 5 和表 6）。

E SIGReg 详情

令 $\{p_n\}_{n=1}^N \subset \mathbb{R}^k$ 表示大小为 N 的一个批次中单个视图槽位的投影，其中 $p \in \mathbb{R}^k$ 的定义见第 2.1 节。

随机方向。 我们从单位球面上均匀抽取 $M = 1024$ 个单位方向 $\{\omega_m\}_{m=1}^M \subset S^{k-1}$ （每一步重新采样），并将 $\mathbb{R}^k$ 中的各向同性约束简化为对标量投影 $u_n^{(m)} = \omega_m^\top p_n$进行 M 项单变量正态性检验。

经验特征函数。 对于每个方向 ω_m，标量投影 $\{u_n^{(m)}\}_{n=1}^N$ 的经验特征函数为

$$\hat{\varphi}_{\omega_m}(t) \;=\; \frac{1}{N}\sum_{n=1}^{N} \exp\big(i\,t\,u_n^{(m)}\big). \tag{7}$$

Epps–Pulley 统计量。 我们通过加权平方误差积分，将 $\hat{\varphi}_{\omega_m}$ 与标准正态特征函数 $\varphi_{\mathcal{N}}(t) = \exp(-t^2/2)$ 进行比较，

$$T_{\mathrm{EP}}(\omega_m) \;=\; N\int_{-T}^{T} \big|\hat{\varphi}_{\omega_m}(t) - \varphi_{\mathcal{N}}(t)\big|^2\, w(t)\, dt, \tag{8}$$

其中权重为 $w(t) = \exp(-t^2/2)$，截断值为 $T = 5$；该积分通过在 $[-T, T]$ 上的均匀网格上应用 17 点梯形法则进行近似。

跨视图槽位聚合。 在每个训练步骤中，我们针对每个视图槽位计算一次式 8（$n_g + n_\ell$次计算），并对各视图槽位和方向取平均，以得到式 2 中的 $\mathcal{L}_{\mathrm{sig}}$。按视图槽位分别计算，可保留 $\{p_n\}_{n=1}^N$ 之间的独立同分布假设；若将不同视图槽位的投影合并，该假设就会被破坏。

F 对抗式批次校正详情

条件变体。 我们以生物学分组 c（细胞类型，或更粗略的分组，例如谱系——根据数据集选择）作为对抗器的条件：

$$\mathcal{L}_{\mathrm{adv}}^{\mathrm{cond}} = \mathrm{CE}\big(h_\psi([\mathrm{GRL}_\lambda(z)\,;\, e_c]), b\big), \tag{9}$$

其中， e_c 是附录C中定义的已学习组嵌入。这会消除每个分组内部 的批次效应，同时保留不同 分组之间的生物学结构；当存在高置信度的生物学分组时，这一做法很有用。本文报告的实验均未使用条件变体；所有结果均使用无条件对抗器。

稳定化。 采用逐步增大的 λ 计划进行对抗训练容易导致不稳定。我们采用两种标准的稳定化方法：对 h_ψ 的线性层进行谱归一化（限制其 Lipschitz 常数），并在交叉熵中使用标签平滑（降低过度自信的批次预测）。所有实验均采用这两种方法；标签平滑的取值为 0.1。

G JEPA 分数计算

我们在一次向量化遍历中计算稳定化得分（公式 5）。对于包含 N 个细胞的小批量，对各个细胞应用反向模式微分（vmap、jacrev）可得到堆叠雅可比矩阵，而批量奇异值分解（svdvals）则可得到其奇异值：

$$\mathbf{J} = \big(J(x_1), \ldots, J(x_N)\big) \in \mathbb{R}^{N\times k\times G}, \qquad S(x_n) = \sum_{i=1}^{k} \log\max\big(\sigma_i(\mathbf{J}_n),\, \varepsilon\big), \quad n = 1, \ldots, N.$$

这个下限是针对近退化奇异方向进行的数值稳定化，不会改变远离该情况时预期的几何解释。

H 位移辅助项详述

令 $F = g \circ f$ 表示复合编码器–投影器，$R \in \mathbb{R}^{G \times r}$ 表示一个固定的随机矩阵（设定随机种子；在整个训练过程中保持不变），其列经过 ℓ_2 归一化，其中 $r = 64$。预热后，对每个批次进行 25% 的子采样，并对其应用辅助项。对于每个细胞 x（对数归一化计数），步骤如下：

算法 1 位移辅助项（每个细胞 $x \in \mathbb{R}^G$；对数归一化计数的每基因标准差 s；缩放系数 σ；子集大小 n_s）。$\mathrm{sg}(\cdot)$ 表示停止梯度。

均匀随机采样基因子集 $S \subset \{1, \dots, G\}$，$|S| = n_s$；构造指示向量 $v \in \{0,1\}^G$ ▷ 标量，每个细胞

采样 $\alpha \sim \mathcal{N}(0,\ \sigma^2)$

$x' \leftarrow \max(x + \alpha \cdot (v \odot s),\ 0)$ ▷ 按每基因批次标准差缩放

$p \leftarrow F(x)$ ▷ 带梯度

$p' \leftarrow F(x')$ ▷ 无梯度

$\Delta p \leftarrow \mathrm{sg}(p') - \mathrm{sg}(p)$；$\bar{v} \leftarrow v/\|v\|$；$u \leftarrow R^\top \bar{v} \,/\, \|R^\top \bar{v}\|$

$\frown$ $\Delta p \leftarrow h_\theta(p,\ u)$ ▷ 预测器输入：p（带梯度）

h_θ 接收拼接后的 $[p;\ u] \in \mathbb{R}^{k+r}$ 并通过以下方式输出 $\widehat{\Delta p} \in \mathbb{R}^k$

$$\mathsf{Linear}(k + r \to 256) \to \mathsf{LayerNorm}(256) \to \mathsf{GELU} \to \mathsf{Linear}(256 \to k).$$

我们使用 $n_s = 25$ 和 $\sigma = 1.0$。

损失函数结合了余弦方向项和标量幅度项，仅在具有 $\|\Delta p\| > \tau$（$\tau = 10^{-4}$）的细胞上计算：

$$\mathcal{L}_{\mathrm{disp}} = \mathbb{E}_{x,v}\Big[\, 1 - \cos\big(\widehat{\Delta p},\, \Delta p\big) \;+\; \gamma \cdot \big(\|\widehat{\Delta p}\| - \|\Delta p\|\big)^2 \,\Big], \qquad \gamma = 0.1. \tag{10}$$

整个目标 Δp 都已停止梯度，因此预测器会拟合它，但编码器不会通过缩小位移 $\widehat{\ }$ 来最小化它。梯度仅通过 p 流入 F，位于 $\Delta p = h_\theta(p,\ u)$ 中；p' 的计算不涉及梯度，因此该目标促使编码器在潜在空间中将较小的基因空间扰动变得可预测。

消融实验：不使用位移辅助项的 scIB。 表 2 报告了在三个整合基准测试中使用和不使用位移辅助项的 LeCellModel。移除辅助项会降低 HLCA Core 和 Heart v2 上的总计；在胰腺数据集上，总计基本不变（生物学略有提升，抵消了批次表现的下降）。正文中报告的是启用了辅助项的配置。不使用位移辅助项的运行采用了略有不同的系数，这些系数是在不使用辅助项的情况下为确保稳定性而选定的：$\lambda_{\mathrm{sig}} = 0.4$（与 0.5 相比）；对抗器 λ 进度因子为 1.6（HLCA Core）/ 1.6（Heart v2）/ 1.3（胰腺），与 $1.8/1.8/1.5$ 相比；以及胰腺的对抗器权重为 0.8（与 1.0 相比）。

表 2：在三个整合基准测试中，LeCellModel 使用和不使用位移辅助项的结果。基线行数据来自表 1（4 个随机种子的均值）；消融实验行数据来自单个随机种子（种子42）。

	HLCA Core			Heart v2			胰腺		
配置	Bio	批次	总计	Bio	批次	总计	Bio	批次	总计
基线（含离散度）	0.727	0.632	0.689	0.849	0.627	0.760	0.807	0.640	0.741
无位移辅助项	0.716	0.603	0.671	0.845	0.577	0.738	0.820	0.629	0.744

I 其他消融实验

我们在全部三个整合基准测试上，以主论文配置为对照，对两个设计选择进行消融：移除对抗性批次校正头，以及将视图增强方案改为

表3：LeCellModel 在三个整合基准测试上的消融实验。基线行数据来自表 1（4 个随机种子的均值）；消融实验行数据为单个随机种子（种子42）。

	HLCA Core			Heart v2			胰腺		
配置	Bio	批次	总计	Bio	批次	总计	Bio	批次	总计
基线	0.727	0.632	0.689	0.849	0.627	0.760	0.807	0.640	0.741
无对抗器	0.685	0.472	0.600	0.727	0.446	0.614	0.656	0.371	0.542
$n_g = 2$，掩码 0.5/0.7	0.714	0.619	0.676	0.849	0.596	0.748	0.794	0.637	0.731

$n_g = 2$ 全局项，掩码率为 $m_g = 0.5$，$m_\ell = 0.7$（而非 $n_g = 1$、$m_g = 0$、$m_\ell = 0.6$）。The base该行复现了表 1 中多随机种子的总计；消融实验行使用单个随机种子（种子42）。在全部三个数据集中，对抗器都是单个组件中贡献最大的部分：移除对抗器后，HLCA Core、Heart v2 和胰腺的总计分别降低 0.089、0.146 和 0.199，且在批次和生物学两项上均有影响，其中批次上的降幅更大。替代视图机制在 Heart v2 上基本不改变总计（-0.012），并在 HLCA Core（-0.013）和胰腺（-0.010）上造成小幅退化；主论文保留较轻的 $n_g = 1$ 机制。

正则化器消融。对于 VICReg，我们将对抗器进度因子设为 1.0（SIGReg 为 1.6），因为在 1.6 下训练不稳定；其他所有设置均遵循图4配置（种子 1；VICReg$\lambda = 1.0$）。这里的 CD177$^+$ 调节性 T 细胞程序定义为带有 FOXP3 > 2 和 CD177 > 2($n = 162$) 的 CD4 T细胞；其在 SIGReg 下的中位数评分百分位为 0.55%（种子 1；最稀有的 1% 中为 69.1%），在 VICReg 下则变为 1.49%（37.0%）。在细胞因子面板中，在匹配的随机种子（1–2）下进行比较，VICReg 在两个 IL-10配对上使 AUROC 平均降低 0.03（最高降低 0.06），在 IL-4配对上的变化最多为 0.022。在 HLCA Core 上，VICReg 达到 scIB Total 0.659（随机种子 42、1、2 的均值，$\lambda = 2.0$）。在 scVI编码器上计算的相同雅可比评分仍接近随机水平（四个配对、种子42 下，AUROC 为 0.50–0.61，位于 $f = 5\%$；而 LeCellModel 为 0.71–0.87），因此该读出依赖于 SIGReg塑造的表征。

J 训练超参数

LeCellModel。在所有数据集上，我们均使用两层多层感知机编码器和投影器，设置为 $d = 256$,$k = 128$，采用 AdamW（学习率 $= 5 \times 10^{-4}$，权重衰减 10^{-2}）、OneCycle 学习率并进行 10 个 epoch 的预热，将梯度裁剪范数设为 10，并使用 bf16 混合精度。所有设置中均启用对抗器的谱归一化和标签平滑（0.1）。各数据集的具体设置见表 4。

随机种子。整合基准测试结果（表 1）为随机种子 $\{42, 1, 2, 3\}$ 下的平均值——包括框架默认值和最小的三个正整数。

表 4：各数据集的 LeCellModel 训练配置。“Disp.”表示位移辅助项。所有数据集的 Epochs$= 150$。

	HLCA Core	Heart v2	胰腺
批次大小	1024	1024	512
λ_{sig}	0.5	0.5	0.5
对抗器权重	0.5	0.5	1.0
对抗器 λ 进度因子	1.8	1.8	1.5
对抗器预热轮数	0	0	0
n_g/n_ℓ	1 / 6	1 / 6	1 / 6
掩码率 m_g/m_ℓ	0 / 0.6	0 / 0.6	0 / 0.6
泊松重采样	off	off	on
投影器维度 k	128	128	128
位移（权重）	2.0	2.0	2.0

稀有程序发现队列。第 3.4 节中使用的 HLCA 单核细胞–巨噬细胞和 HLCA 淋巴系（正常+特发性肺纤维化）子集，采用表 5 中的配置进行训练。它们共享相同的编码器架构（附录C），但使用更大的投影维度（$k = 256$）和不同的

视图增强方案（$n_g = 2$ 个全局视图，掩码率 $0.5/0.7$），而不是整合基准测试；两者还都使用 3，000 个高变基因（附录 B）。

表 5：IPF 稀有程序发现队列的 LeCellModel 训练配置（第 3.4 节）。约定与表 4 相同。

	HLCA 单核细胞–巨噬细胞	HLCA 淋巴系
批次大小	1024	1024
λ_{sig}	0.5	0.5
对抗器权重	0.7	0.5
对抗器 λ 进度因子	1.6	1.6
对抗器预热轮数	0	0
n_g/n_ℓ	2 / 6	2 / 6
掩码率 m_g/m_ℓ	0.5 / 0.7	0.5 / 0.7
泊松重采样	on	on
投影器维度 k	256	256
位移（权重）	off	off

稀有状态检测实验（第 3.3 节）。四组细胞因子 $\times$ 细胞类型实验均采用表 6 中的配置。每组实验均在某一$f \in \{0.01, 0.03, 0.05, 0.10\}$ 刺激比例下的污染混合物上从头训练 LeCellModel，不使用位移辅助项；对抗器使用供体作为批次标签。

表 6：细胞因子稀有状态检测基准（第 3.3 节）中 LeCellModel 的训练配置。所有少数类比例下的配置均相同；CD14 和 CD16 配对的对抗器 λ 进度因子不同。

	稀有状态检测
批次大小	1024
λ_{sig}	0.35
对抗器权重	0.4
对抗器 λ 进度因子	1.0（CD16 配对）；1.4（CD14 配对）
对抗器预热轮数	5
n_g/n_ℓ	2 / 6
掩码率 m_g/m_ℓ	0.5 / 0.7
泊松重采样	on
投影器维度 k	256
位移（权重）	off

基线方法。 scVI 使用 n_层 $= 2$、n_潜变量 $= 40$、n_隐藏单元 $= 256$ 进行训练，最多训练 200 个 epoch，并采用早停。Harmony 和 Scanorama 使用标准的 scanpy 对数归一化流程（10^4 目标总和、log1p；Harmony 还会将数据缩放至最大值为 10 的单位方差，并以排名前 50 的主成分作为输入）；两者均使用默认设置。主成分分析在经过对数归一化和缩放的表达矩阵上使用 50 个成分。

计算。 所有 LeCellModel 训练任务每个任务使用一块 GPU，GPU 来自共享的机构集群，该集群混合配备了 NVIDIA A40、NVIDIA A100（40 GB 和 80 GB PCIe）以及 NVIDIA H100 NVL。训练和推理均使用 bf16 混合精度。对于图谱级基准测试（HLCA Core、Heart v2），一次完整的 150 epoch 训练大约需要 1 小时；对于胰腺，则约需 3 分钟。实际耗时会因 GPU 型号而异。报告的运行结果对应最终配置；超参数开发还需要进行额外的同等规模运行。

K GeneJEPA 和基础模型基线

我们评估了 GeneJEPA [Litman et al., 2025]零样本（已发布的 Tahoe-100M 检查点），并使用自有训练脚本在我们的基准测试上进行微调和从头开始训练。三种设置的表现均落后于 LeCellModel，且无法区分细胞类型（Leiden NMI 0。06–0。24）；图5展示了 HLCA Core 上的零样本 UMAP。我们还报告了 scGPT [Cui et al., 2024]（零样本和微调）

以及 UCE [Rosen et al., 2026]（零样本，四层模型）。UCE 的 1280 维嵌入使得 scIB 的轮廓系数指标无法在完整的 HLCA Core 上计算，因此我们在分层抽取的十万细胞子样本上对其进行评分，同时让 LeCellModel 在完全相同的细胞上接受评分。表7给出了零样本 GeneJEPA 的各指标细分，表8则汇总了所有设置的结果。

表7：GeneJEPA 在 HLCA Core 上的零样本 scIB 结果（584，944 个细胞），单次运行。为供参考，LeCellModel 行摘自表9。

	生物学保留						批次校正					
方法	孤立标签	NMI	ARI	Sil.	cLISI	**Bio**	BRAS 整合局部逆辛普森指数	kBET	GC	PCR	**批次**	**总计**
GeneJEPA（零样本）	0.450	0.090	0.031	0.440	0.935	0.389	0.793	0.231	0.205	0.089	0.685	0.401	0.394
LeCellModel	0.617	0.804	0.682	0.533	1.000	0.727	0.815	0.119	0.399	0.905	0.921	0.632	0.689

表8：GeneJEPA 和基础模型基线方法，scIB聚合指标，每种方法各单次运行。LeCellModel 行数据来自表 1 （四个随机种子的均值）。 *UCE 在 HLCA Core 上的评分基于分层抽取的十万细胞子样本，并与相同细胞上的 LeCellModel 进行比较。

	HLCA Core			胰腺		
方法	Bio	批次	总计	Bio	批次	总计
GeneJEPA（零样本）	0.389	0.401	0.394	0.389	0.381	0.386
GeneJEPA（从头开始）	0.466	0.507	0.482	0.403	0.158	0.305
GeneJEPA（微调）	0.386	0.396	0.390	0.386	0.377	0.382
scGPT（零样本）	0.598	0.487	0.554	0.682	0.490	0.605
scGPT（微调）	0.706	0.414	0.589	0.760	0.496	0.655
UCE（零样本）	0.698*	0.349*	0.558*	0.600	0.373	0.510
LeCellModel	0.727	0.632	0.689	0.807	0.640	0.741
LeCellModel（100k子采样）	0.732*	0.664*	0.705*	–	–	–

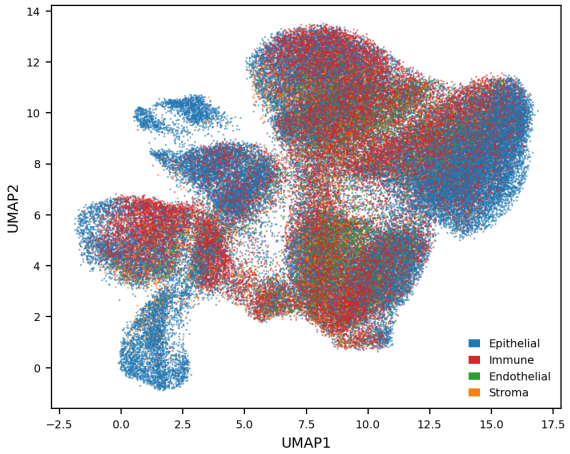

图 5：**GeneJEPA 对 HLCA Core 进行的零样本嵌入，按细胞类型着色**。冻结编码器在原代组织上未生成分离的簇，这与 Tahoe-100M →HLCA 领域差距一致。

L 完整的逐指标 scIB 表格

表 9–11 报告了每种方法在每项基准测试中的十项独立 scIB 指标。生物学保留度是孤立标签、Leiden NMI、Leiden调整兰德指数、轮廓系数标签和 cLISI 的平均值；批次校正是 BRAS（批次标签调整轮廓系数，一种按细胞类型分层的轮廓系数，用于衡量标签内的批次混合情况；scib_metrics.bras）、整合局部逆辛普森指数、kBET、图连通性和 PCR比较的平均值；总计是 scIB 约定加权平均值（ 0.6 生物学 $+0.4$ 批次）。各列汇总值经四舍五入后与表 1 一致。

表9：HLCA Core（584，944 个细胞）：scIB 各指标结果，4 个随机种子的均值 ± 标准差。孤立标签 = 孤立标签；Leiden 上的归一化互信息/调整兰德指数；轮廓系数 =轮廓系数标签；GC =图连通性；PCR = PCR 比较。

	生物学保留						批次校正						
方法	孤立标签	NMI	ARI	Sil.	cLISI	**Bio**	BRAS	整合轮廓密度曲线指数	kBET	GC	PCR	**批次**	**总计**
PCA	0.556±.001	0.763±.015	0.545±.023	0.549±.003	1.000	0.682±.008	0.634±.002	0.001±.001	0.144±.006	0.881±.008	0.000	0.332±.003	0.542±.006
Scanorama	0.562±.001	0.795±.002	0.635	0.545	1.000	0.707	0.778±.002	0.019	0.199±.001	0.791±.003	0.153±.088	0.388±.017	0.580±.006
Harmony	0.538	0.755±.005	0.598±.028	0.537	0.999	0.685±.007	0.812±.001	0.063	0.318±.001	0.830±.001	0.558±.065	0.516±.013	0.618±.009
scVI	0.605±.004	0.810±.001	0.707±.004	0.521±.001	1.000	**0.729**±.001	0.822±.001	0.014	0.184±.001	0.924±.002	0.813±.029	0.551±.006	0.658±.002
LeCellModel	0.617±.017	0.804±.003	0.682±.008	0.533	1.000	0.727±.005	0.815±.003	0.119±.001	0.399±.005	0.905±.006	0.921±.012	**0.632**±.003	**0.689**±.003

Ta表 10：心脏细胞图谱 v2（704，296 个细胞）：各项指标的 scIB 结果，4 个随机种子的均值 ± 标准差 .

	生物学保留						批次校正						
方法	孤立标签	NMI	ARI	Sil.	cLISI	**Bio**	BRAS	整合轮廓密度曲线指数	kBET	GC	PCR	**批次**	**总计**
PCA	0.726±.012	0.820	0.692±.001	0.578±.020	1.000	0.763±.006	0.552±.016	0.021±.001	0.090±.003	0.907±.001	0.000	0.314±.002	0.583±.003
Scanorama	0.642	0.841	0.754	0.572	1.000	0.762	0.672	0.044	0.117	0.796	0.083±.048	0.342±.009	0.594±.004
Harmony	0.691	0.931±.010	0.948±.015	0.643	1.000	0.843±.005	0.851	0.097	0.271±.003	0.854	0.696±.035	0.554±.007	0.727±.004
scVI	0.672±.004	0.938±.001	0.956	0.524±.001	1.000	0.818±.001	0.835±.001	0.066±.001	0.114±.001	0.900±.001	0.801±.025	0.543±.005	0.708±.002
LeCellModel	0.835±.004	0.935±.005	0.952±.007	0.524±.001	1.000	**0.849**±.002	0.777±.003	0.155±.001	0.330±.006	0.919±.007	0.953±.005	**0.627**±.001	**0.760**±.002

表 11：胰腺（16，382 个细胞）：scIB 各项指标结果，4 个随机种子的均值 ± 标准差。

	生物学保留						批次校正						
方法	孤立标签	NMI	ARI	Sil.	cLISI	**Bio**	BRAS	整合轮廓密度曲线指数	kBET	GC	PCR	**批次**	**总计**
PCA	0.656	0.653	0.355	0.528	1.000	0.638	0.461	0.004	0.114	0.689	0.000	0.254	0.485
Scanorama	0.524±.002	0.860±.010	0.904±.010	0.559±.001	1.000	0.770±.004	0.780±.004	0.126	0.367±.005	0.799±.022	0.711±.003	0.556±.006	0.684±.004
Harmony	0.666	0.768±.002	0.617±.003	0.582	1.000	0.727±.001	0.714	0.203	0.457	0.762	0.800	0.587	0.671±.001
scVI	0.719±.010	0.918±.003	0.952±.002	0.568±.002	1.000	**0.831**±.003	0.782±.001	0.131±.001	0.347±.030	0.879±.009	0.829±.005	0.594±.007	0.736±.003
LeCell 模型	0.642±.051	0.908±.004	0.940±.003	0.547±.001	1.000	0.807±.010	0.714±.007	0.232±.006	0.416±.017	0.916±.008	0.923±.005	**0.640**±.007	**0.741**±.00 8

M 完整稀有状态检测结果

在所有四种少数类比例下，四组（细胞因子、细胞类型）配对的完整 AUROC、AUPRC 和斯皮尔曼 ρ（在受刺激细胞中计算，并以静息细胞上的通路特征 z 分数进行比较；seed42，每组配对运行一次）。

表12：按（配对、 f、方法）划分的完整 AUROC。每行最佳结果以粗体标示。

Pair	f	**LeCellModel**	主成分分析-k近邻	主成分分析-Mahal	主成分分析-孤立森林	scVI-孤立森林	scVI-recon	scVI-Mahal	scVI-k近邻
IL-10 × CD16	0.01	0.835	**0.863**	0.820	0.817	0.847	0.604	0.819	0.839
	0.03	**0.844**	0.822	0.773	0.738	0.797	0.561	0.748	0.769
	0.05	**0.854**	0.801	0.745	0.742	0.704	0.565	0.719	0.739
	0.10	**0.832**	0.751	0.698	0.704	0.663	0.557	0.656	0.683
IL-10 × CD14	0.01	**0.749**	0.720	0.680	0.661	0.737	0.562	0.729	0.711
	0.03	**0.756**	0.702	0.674	0.679	0.677	0.556	0.694	0.666
	0.05	**0.761**	0.685	0.662	0.635	0.673	0.552	0.654	0.625
	0.10	**0.744**	0.642	0.627	0.600	0.604	0.551	0.593	0.563
IL-4 × CD14	0.01	0.906	0.911	0.905	**0.912**	0.856	0.548	0.910	0.866
	0.03	**0.885**	0.874	0.852	0.843	0.835	0.528	0.834	0.762
	0.05	**0.871**	0.848	0.821	0.811	0.752	0.528	0.785	0.730
	0.10	**0.845**	0.798	0.769	0.751	0.697	0.521	0.701	0.659
IL-4 × CD16（较弱）	0.01	0.667	0.724	0.721	0.720	0.714	0.467	0.765	**0.775**
	0.03	0.686	**0.729**	0.702	0.694	0.675	0.467	0.713	0.682
	0.05	**0.705**	0.701	0.661	0.658	0.646	0.464	0.680	0.667
	0.10	**0.690**	0.667	0.618	0.606	0.619	0.465	0.608	0.626

为将读出与表征区分开来，我们在 LeCellModel 自身的投影中，对与 JEPA-SCORE 相同的细胞和运行批次，计算 kNN 密度、马氏距离和孤立森林（表15）。在三个已确立的响应对上，两种距离评分在每个比例下都低于随机水平（例如，在 IL-10 × CD16 的 $f = 1\%$ 时 AUROC 为 0.27，而 JEPA-SCORE 为 0.84）：在 SIGReg 下，稀有且具有预测性的状态占据一个紧凑、低容量的区域，基于距离的评分会将其判定为密集区域。孤立森林仅在污染率较低时对 IL-4 × CD14 有竞争力（$f = 1\%$ 时为 0.89），而在 $f = 10\%$ 时降至 0.75；与此同时，JEPA-SCORE 在该配对上保持在 0.84 至 0.91 之间。对于响应较弱的 IL-4 × CD16，四种评分都更接近随机水平。

表 13：按（配对、 f、方法）列出的完整 AUPRC。粗体表示每行中的最佳值。

Pair	f	**LeCellModel**	主成分分析-k近邻	主成分分析-Mahal	主成分分析-孤立森林	scVI-IForest	scVI-recon	scVI-Mahal	scVI-kNN
IL-10 × CD16	0.01	0.085	0.054	0.045	0.046	**0.129**	0.015	0.081	0.093
	0.03	**0.217**	0.103	0.090	0.076	0.134	0.038	0.096	0.102
	0.05	**0.326**	0.141	0.121	0.127	0.112	0.062	0.127	0.134
	0.10	**0.416**	0.208	0.182	0.193	0.180	0.119	0.174	0.185
IL-10 × CD14	0.01	0.024	0.020	0.017	0.017	**0.028**	0.012	0.026	0.024
	0.03	**0.071**	0.052	0.049	0.052	0.055	0.036	0.059	0.055
	0.05	**0.120**	0.080	0.077	0.069	0.088	0.059	0.080	0.076
	0.10	**0.203**	0.137	0.134	0.123	0.132	0.118	0.127	0.121
IL-4 × CD14	0.01	**0.164**	0.123	0.142	0.128	0.073	0.014	0.117	0.063
	0.03	**0.233**	0.177	0.174	0.165	0.125	0.036	0.122	0.081
	0.05	**0.292**	0.221	0.217	0.218	0.127	0.059	0.146	0.109
	0.10	**0.407**	0.298	0.294	0.273	0.187	0.114	0.189	0.159
IL-4 × CD16（弱）	0.01	0.017	0.025	0.029	0.029	0.027	0.009	**0.045**	0.038
	0.03	0.058	0.067	**0.073**	0.071	0.059	0.026	**0.073**	0.058
	0.05	**0.106**	0.092	0.089	0.084	0.087	0.043	0.095	0.087
	0.10	**0.190**	0.151	0.138	0.134	0.144	0.087	0.138	0.144

表 14：仅在受刺激细胞中计算的单细胞评分与通路特征之间的斯皮尔曼 ρ 相关性。每行最佳结果以粗体标示。

Pair	f	**LeCellModel**	主成分分析-k近邻	主成分分析-Mahal	主成分分析-孤立森林	scVI-IForest	scVI-recon	scVI-Mahal	scVI-kNN
IL-10 × CD16	0.01	0.532	0.514	0.565	**0.611**	0.414	0.252	0.384	0.390
	0.03	0.463	0.378	**0.470**	0.449	0.279	0.235	0.275	0.238
	0.05	**0.473**	0.280	0.361	0.339	0.151	0.239	0.183	0.126
	0.10	**0.428**	0.182	0.255	0.232	0.083	0.249	0.071	0.032
IL-10 × CD14	0.01	**0.363**	0.172	0.250	0.265	0.142	0.252	0.133	0.072
	0.03	**0.315**	0.122	0.240	0.238	0.104	0.250	0.029	-0.048
	0.05	**0.312**	0.091	0.210	0.222	0.011	0.255	-0.030	-0.101
	0.10	0.251	0.031	0.138	0.102	-0.030	**0.254**	-0.097	-0.167
IL-4 × CD14	0.01	**0.688**	0.344	0.659	0.605	0.359	0.179	0.410	0.051
	0.03	**0.535**	0.281	0.519	0.484	0.270	0.192	0.218	-0.099
	0.05	**0.473**	0.232	0.430	0.435	0.156	0.180	0.151	-0.104
	0.10	**0.414**	0.186	0.330	0.315	0.056	0.163	0.058	-0.139
IL-4 × CD16（弱）	0.01	0.468	0.522	0.524	**0.525**	0.367	0.145	0.365	0.343
	0.03	0.442	0.515	**0.523**	0.481	0.238	0.206	0.218	0.078
	0.05	0.464	0.498	**0.516**	0.492	0.144	0.193	0.178	0.052
	0.10	0.428	**0.463**	0.436	0.429	0.210	0.185	0.164	0.102

表15：在 LeCellModel 的投影中计算的简单评分用于稀有状态检测：按配对和污染比例 f 列出的 AUROC，取随机种子 1–3 的均值（标准差 ≤ 0.07 ）。每行最佳结果以粗体标示。

Pair	f	JEPA 分数	kNN	马氏距离	孤立森林
IL-10 × CD14	0.01	**0.746**	0.312	0.314	0.525
	0.03	**0.758**	0.277	0.288	0.549
	0.05	**0.761**	0.272	0.289	0.601
	0.10	**0.744**	0.280	0.293	0.569
IL-10 × CD16	0.01	**0.837**	0.273	0.273	0.576
	0.03	**0.865**	0.184	0.223	0.617
	0.05	**0.864**	0.178	0.219	0.639
	0.10	**0.840**	0.203	0.250	0.625
IL-4 × CD14	0.01	**0.907**	0.140	0.290	0.893
	0.03	**0.882**	0.153	0.249	0.863
	0.05	**0.866**	0.164	0.264	0.849
	0.10	**0.844**	0.185	0.288	0.753
IL-4 × CD16（弱）	0.01	**0.630**	0.604	0.577	0.495
	0.03	**0.675**	0.510	0.509	0.497
	0.05	**0.701**	0.412	0.447	0.499
	0.10	**0.688**	0.383	0.421	0.506

N 细胞因子刺激面板：配对详情

所有配对均从固定数量的静息 CD16 单核细胞（16，710 个）和静息 CD14 单核细胞（110，999 个）中抽取。我们考虑四种细胞因子–细胞类型组合：*IL-4* × *CD14 Mono*，这是典型的 M2 样 IL-4/STAT6 应答；*IL-10* × *CD14 Mono*，这是典型的抗炎性 IL-10/STAT3 应答；*IL-10* × *CD16 Mono*，这是在不同 $\mathrm{CD16}^{+}$ 背景下的 IL-10/STAT3 应答；以及*IL-4* × *CD16Mono*，这是附录 M 报告的一种较弱的 IL-4 应答条件。基于密度的基线方法使用 scVI，n_ 个潜在=30，训练 100 个周期。

通路特征。 表 14 中的斯皮尔曼 ρ 采用上调基因集与下调基因集的平均 z 分数之差（以静息细胞为基准进行 z 标准化）：**IL-10**，上调 = {CD163, CD163L1, IRAK3, MARCO, MRC1, F13A1}（STAT3 靶基因），下调 = {IL1B,IL6, CXCL10, CCL3, CCL4, CXCL8}（受 IL-10 抑制的促炎介质）；**IL-4**，上调 ={CCL17, CCL22, MRC1, ALOX15, CCL18, CHI3L1, F13A1}（STAT6 靶基因），无下调基因集。

O 稀有程序发现：统计与簇选择

嵌入图上的双变量莫兰指数。 对于每个基因 g，双变量莫兰指数 I_g 用于衡量基因表达矩阵与 JEPA 分数在嵌入图上是否呈空间负相关： $I_g < 0$表示 g 在邻居 JEPA 分数较低的细胞中表达较高，因此可作为稀有亚群的候选基因。 W 是 LeCellModel 投影上的高斯核加权 kNN 图（$k = 15$，行随机矩阵）；显著性采用 Hotspot 中的分析 z分数 [（DeTomaso 和 Yosef， 2021],），并使用 BH-FDR < 0.05。

在计算 I_g 之前，我们先对原始分数应用逻辑变换，以第 5 百分位数ˆscore $c = F_{5\%}(S)$ 为中心，按分数标准差 σ_S进行缩放，再通过陡度为 $\beta = 2$的 S 形函数进行压缩：

$$\tilde{S}(x) \;=\; \sigma\!\left(\beta\,\frac{S(x)-c}{\sigma_S}\right).$$

这会将稀有异常尾部（$S \ll c$）映射至 0 附近，并将主体部分（$S \gg c$）映射至 1 附近。JEPA 耦合基因（$I_g \;<\; 0$，BH-FDR $\;<\; 0.05$）依据其单细胞耦合特征$\mathbf{q}_g \in \mathbb{R}^N$聚类为多个模块，其中（$\mathbf{q}^g$）$_n = \tilde{x}_{g,\,n} \cdot$（$W\tilde{S}$）$_n$是基因 g 的标准化表达与细胞 n 的邻域平滑分数的乘积。同一模块中的基因在相同的低分细胞中表达升高，而不只是整体共表达。模块 1（按 $|I| \cdot \sqrt{n_{\text{genes}}}$排名）是第 3.4 节（表 16）中报告的恢复程序。

梯度归因的簇选择。 梯度归因需要一个具体的细胞簇。我们通过洼地搜索程序识别 JEPA 分数较低的细胞的连通邻域；簇边界的具体位置不会实质性影响归因基因。该程序在余弦 k 近邻图（$k = 10$）上运行，并将其限制在按 JEPA 分数排序后位于最低 10% 的细胞中（即子图），同时使用数据自适应距离阈值 δ ——子图内 k 近邻边距离的第 10th 百分位数——来控制哪些邻居可被遍历。

1. 平滑 JEPA 分数：将每个细胞的评分替换为其自身与距离 $\leq \delta$范围内邻居评分的均值，执行一遍。2. 从 JEPA 分数最低的 0.5% 的细胞（所有细胞）中选取随机种子，沿距离 $\;\leq \delta$ 范围内评分最低的邻居前进，直到不存在评分更低的邻居；最终到达的细胞即为局部极小值（“洼地底部”）。3. 按洼地底部对随机种子分组；丢弃汇聚随机种子少于 max(3, 5% 的随机种子) 的洼地。4. 迭代合并显著性最低的相邻洼地对：洼地 B 的显著性是 A 与 B 之间最低的跨边界分数减去 B 洼地底部的评分（即离开 B 时评分上升的幅度）；持续合并，直到所有剩余洼地之间的分数差都大于 $0.3\,\sigma_S$，其中 σ_S 是子图内的分数标准差。5. 在距离 δ范围内通过广度优先搜索，将每个保留的洼地扩展为一个核心簇。

表 16：按 $|I_g|$在模块内排序的各队列中恢复的程序（模块 1）前 15 个成员。 I_g 为双变量莫兰指数的原始值（范围为 $[-1,1]$）；负值表示该基因在 JEPA 分数较低（稀有）的细胞中表达升高。模块 1 是 JEPA 耦合基因中的主导共耦合簇（BH-FDR < 0.05, $I_g < 0$）。单核细胞来源巨噬细胞模块 1： $n_{\text{genes}} = 47$，均值 $I = -0.732$。CD4 模块 1： $n_{\text{genes}} = 48$，均值 $I = -0.784$。

单核细胞来源巨噬细胞（模块 1）			**CD4 T 细胞（模块 1）**		
Rank	Gene	I_g	Rank	Gene	I_g
1	CHI3L1	-0.459	1	FOXP3	-0.484
2	ADAMDEC1	-0.429	2	TNFRSF18	-0.417
3	MMP9	-0.415	3	IL2RA	-0.416
4	CHIT1	-0.355	4	TNFRSF9	-0.408
5	CTSK	-0.353	5	TNFRSF4	-0.398
6	MREG	-0.327	6	含有趋化因子结构域的调节T细胞特异性受体	-0.397
7	RARRES1	-0.325	7	CD177	-0.397
8	A2M	-0.325	8	LAYN	-0.385
9	SIGLEC15	-0.319	9	CTLA4	-0.370
10	GPC4	-0.317	10	LINC02099	-0.311
11	LILRB4	-0.315	11	TBC1D4	-0.303
12	MMP7	-0.315	12	IL1R2	-0.297
13	SPP1	-0.297	13	ENTPD1	-0.297
14	PALLD	-0.295	14	CCR8	-0.296
15	FCMR	-0.294	15	IKZF2	-0.281

对于每个核心簇 C，我们通过自动微分计算每个细胞处输入×的梯度归因 $x \odot \nabla_x S(x)$，然后按簇均值的绝对值对基因进行排名， $|\mathbb{E}_{x \in C}[x \odot \nabla_x S]|$。簇选择不使用基因身份；该流程仅依据 JEPA 分数和嵌入几何。第 3.4 节中的所有案例研究结果均采用这些默认设置。

P 预测选择性：计算细节

对于每个细胞 x，其全细胞投影 $p = F(x)$（输入为 log1p-CP10k），$\cos_1$ =max$_g$ cos($F(x^{(g)})$ p)，其中 $x^{(g)}$ 仅保留基因 g；$\cos_2$ 则在 g^*的基础上进一步加入最佳的第二个基因，两者均对所有 G 基因进行穷举搜索，并报告总体中位数。CD177$^+$ Treg 簇使用全部 $n = 182$ 个细胞（FOXP3> 1 & CD177> 1）；CD4 T细胞群和促纤维化单核细胞来源巨噬细胞群均限制为 $n = 500$个（均匀子采样，种子42）。scVI 基线对原始计数输入的后验均值 μ_z 应用相同流程；scVI 使用 n_latent$= 40$、n_layers$= 2$、n_hidden$= 256$进行训练，批次键为供体，最多训练 200 个周期并采用早停。

Q JEPA分数排名的稳定性

我们使用四个随机种子（42、1、2、3）重新训练每个 IPF 队列；在种子42下，还采用另外两种数据增强设置（全局/局部掩码率 $0.2/0.6$ 和 $0.1/0.4$，相较于表 5 中的 $0.5/0.7$），并禁用泊松重采样。各次运行按相同顺序对相同细胞进行评分，因此逐个细胞比较排名（表17）。不同随机种子下，单细胞分数的两两斯皮尔曼相关系数均值为 $\rho = 0.965$（单核细胞–巨噬细胞，257，893 个细胞）和 0.843（淋巴系，159，652 个细胞）。每次运行中最稀有的 1% 细胞，约有 90% 与任何其他运行的结果重合。不同的细胞刚好位于截断点之后：它们在另一轮运行中的中位数评分百分位为 1.13–1.14%，其中 99.9%（单核细胞）和 99.6%（淋巴系）落在该轮运行最稀有的 2% 范围内。在重合的尾部中，排序保持一致（斯皮尔曼相关系数为 0.82–0.89）；三次运行的共识结果可找回留出运行中稀有集合的 91–92%。不同数据增强设置下，稀有尾部也几乎同样稳定：在掩码方案之间稀有 1% 成员身份发生变化的细胞中，90.6%（单核细胞）和 94.2%（淋巴系）仍处于最稀有的 2% 范围内；禁用泊松重采样也得到相同结果。

表17：IPF队列中单细胞 JEPA 分数的稳定性。 ρ：单细胞评分的两两斯皮尔曼相关系数均值；重叠率：某次运行中最稀有的 1% 细胞同时出现在另一次运行中的平均比例； $\leq$2%：稀有 1% 成员身份发生变化的细胞中，落在另一次运行最稀有 2% 范围内的比例。

	单核细胞–巨噬细胞			淋巴系		
比较	ρ	重叠	$\leq$2%	ρ	重叠	$\leq$2%
随机种子42 , 1, 2, 3	0.965	90.2%	99.9%	0.843	89.5%	99.6%
掩码率 $0.1/0.4$, $0.2/0.6$, $0.5/0.7$	0.936	80.6%	90.6%	0.877	88.2%	94.2%
关闭与开启泊松重采样的对比	0.954	87.1%	95.0%	0.894	93.1%	99.5%

R 梯度归因与通路靶标及差异表达

我们按簇平均输入×梯度的绝对值对基因进行排名， $|\mathbb{E}_{x\in C}[x \odot \nabla_x S]|$，并将其与原始梯度 $|\mathbb{E}_{x\in C}[\nabla_x S]|$ 以及威尔科克森差异表达结果进行比较（表18）。在未使用任何通路列表的情况下，对于 CD14 配对，梯度将经典 STAT3 和 STAT6 靶基因排在前列（F13A1 和 ALOX15 位居首位）；对于 CD16 配对，这些基因的排名则较低。当两种排名不一致时，梯度更倾向于在簇的某个子集中表达的基因（CD177、SIGLEC15、MARCO、CCL18），而差异表达则倾向于 SPP1 等广泛表达的标记基因。因此，梯度捕捉的是评分的局部驱动因素，而非组间丰度差异，并且不需要预先定义对照组。

表18：按输入×梯度、原始梯度和差异表达排序的基因排名（威尔科克森秩和检验，按检验统计量排序）。特发性肺纤维化：比较该簇与其细胞类型中的其他细胞，按 3,000 个基因排名。细胞因子面板（$f = 5$%）：比较受刺激细胞与对照细胞，按 2,000 个基因排名。表达矩阵：该簇（特发性肺纤维化）或受刺激细胞（细胞因子）中具有非零表达的细胞占比。

Gene	设置	输入×梯度。	原始梯度。	DE	表达
CD177	IPF, CD4 T细胞	2	1	25	37%
SIGLEC15	特发性肺纤维化，单核细胞来源的巨噬细胞	9	4	43	50%
SPP1	特发性肺纤维化，单核细胞来源的巨噬细胞	414	2,875	8	92%
F13A1	IL-10 × CD14（种子 1）	1	12	1	59%
ALOX15	IL-4 × CD14（种子 1）	1	3	2	68%
CCL18	IL-4 × CD14（随机种子 1）	8	25	93	14%
MARCO	IL-10 × CD16（随机种子 1–3）	3–7	1	26–35	28–31%
CD163	IL-10 × CD16（随机种子 1–3）	4–19	14–65	5–7	69–72%
F13A1	IL-10 × CD16（随机种子 1–3）	74–91	27–43	84–91	16–18%
ALOX15	IL-4 × CD16（随机种子 1–3）	8–19	1–2	43	14–15%

S 健康心脏组织上的阴性对照

作为阴性对照，我们使用健康组织，其中不存在疾病程序：使用随 scvi-tools 一同分发的 Heart Cell Atlas 子采样数据集 [Litviňuková et al., 2020]。我们按照该数据集自身的配置进行训练：批次大小为 512， $\lambda_{\text{sig}} = 0.4$，对抗器权重为 0.8，进度因子为 2.0，使用一个全局视图和四个局部视图，掩码率为 $0.2/0.5$， $d = k = 256$，不使用位移辅助项，共训练 150 个周期。在该模型中，评分最低的区域包括一种活化静脉内皮状态（$n = 46$；输入梯度排名靠前的基因为 ACKR1、SELE 和 VCAM1）。随后，我们移除其 JEPA 分数在各自细胞类型中处于第 20 百分位及以下的所有细胞（18,641 $\rightarrow$ 14,908 个细胞），并在剩余细胞上重新训练。重新训练后的模型只检测到一个低分区域，而非四个；剩余的心室心肌细胞区域位于均值以下 0.67 个标准差处，而剪枝前为 1.71 个标准差，且分数分布更集中（均值 $-818 \rightarrow -747$，标准差 $88 \rightarrow 79$）：均质群体获得的评分更加一致。