arXiv:2602.02093v1 [cs.CE] 2 Feb 2026

Cell-JEPA：面向单细胞转录组学的潜在表示学习

Ali ElSheikh[*, 1], **Rui-Xi Wang**[*, 3, 6], **Weimin Wu**[*, 1], **Yibo Wen**[1], **Payam Dibaeinia**[2], **JenniferYuntong Zhang**[5], **Jerry Yao-Chieh Hu**[1], **Mei Knudson**[2,4], **Sudarshan Babu**[2], **Shao-HuaSun**[6], **Aly A. Khan**[2,4], **Han Liu**[†, 1]

[1]西北大学， [2]Biohub， [3]麻省理工学院， [4]芝加哥大学， [5]多伦多大学， [6]台湾大学

摘要：单细胞基础模型通过重建掩蔽基因表达进行学习，从而隐式地将技术噪声视为信号。当丢失率超过90%时，重建目标会促使模型编码测量伪影，而非稳定的细胞程序。我们提出Cell-JEPA，一种联合嵌入预测架构，将学习从重建稀疏计数转向在潜在空间中进行预测。其关键洞见是，细胞身份在各基因中被冗余编码。我们表明，从部分观测预测细胞级嵌入会迫使模型学习抗丢失特征。在细胞类型聚类上，Cell-JEPA在零样本迁移中取得0.72的AvgBIO，而scGPT为0.53，相对提升36%。在单细胞系内的扰动预测上，Cell-JEPA改善了绝对状态重建，但并未改善效应量估计，这表明表征学习与扰动建模针对的是细胞预测中互补的方面。

关键词： 单细胞；基础模型

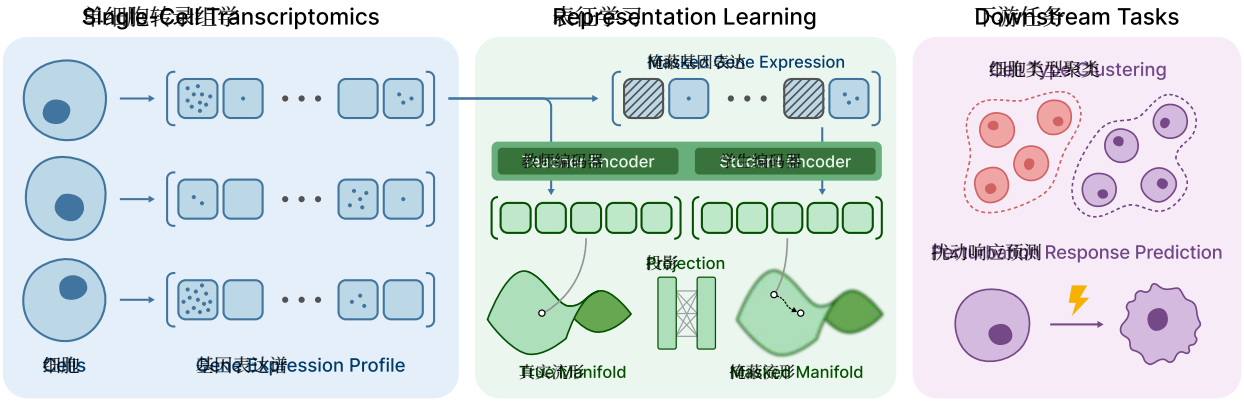


图1：**Cell-JEPA流程概览。** 原始生物细胞经过单细胞RNA测序，生成高维、稀疏的基因表达谱。这些表达谱输入联合嵌入架构，其中学生编码器接收掩码输入，并根据未掩码数据预测教师编码器生成的稳定潜在表征。这种预测对齐学习到稳健的生物流形，从而支持包括细胞类型聚类和扰动响应预测在内的下游任务。

[*]贡献相同。
[†]通讯作者：Han Liu（hanliu@northwestern.edu）。

1 引言

我们提出Cell-JEPA，一种面向单细胞转录组学的自监督基础模型。单细胞转录组学能够用于研究细胞异质性、发育和疾病 [Ding等人，2022，Svensson等人，2018，Tang等人， 2011]。每个细胞由跨越数万个基因的表达谱表示，从而提供细胞状态的高维快照。然而，单细胞测量因分子捕获有限、随机抽样和实验变异而稀疏且含噪 [Luecken和Theis， 2019]。因此，核心挑战在于从高维且含噪的观测中推断稳定且具有生物学意义的细胞程序。

近年来，针对单细胞数据的大规模自监督基础模型取得了进展。一种常见策略是通过预测转录组中被掩码或带噪声的部分来学习表征。例如，scGPT [Cui 等人， 2024]采用生成式预训练：给定观测到的“已知”基因，预测一部分“未知”基因的表达值；为此，它使用专门的注意力掩码和迭代生成过程，逐步纳入高置信度预测结果。Geneformer[Theodoris 等人， 2023] 采用 BERT 式掩码学习目标，对按排名排列的基因令牌进行训练，使模型能够根据每个细胞中的其余上下文恢复被掩码基因。UCE [Rosen 等人， 2023]通过掩码表达基因并训练模型预测基因层面的表达存在性来实现自监督，同时致力于在多样的组织和物种之间提供可迁移的嵌入空间。与此同时，scVI[Lopez 等人， 2018]通过对计数进行变分推断训练概率生成模型，从而学习低维细胞状态。尽管这些方法的建模假设各不相同，但它们有一个共同的关键特征：预训练与重建紧密对应观测测量空间的信息（例如掩码计数、掩码基因标识，或基因层面的存在/缺失）。

这一范式的一个局限在于，单细胞表达值会受到测序深度、捕获效率和技术噪声的显著影响。因此，强调准确进行基因层面重建的目标，可能会促使模型表征测量特异性变异，而非优先学习定义细胞身份和调控程序的稳定结构。这促使我们采用能够约束表征空间的学习目标，而不是追求对含噪观测值的忠实还原。

计算机视觉领域的最新进展表明，联合嵌入预测架构（JEPA）通过在学习得到的潜在空间中预测被掩蔽的内容，而非重建原始像素输入，从而提供强有力的自监督信号 [Assran 等人，2025年，2023，Lecun， 2022]。JEPA 将预测转移到嵌入空间，从而促使模型在抽象特征层面保持一致，并减少对精确输入值的依赖；当观测数据含噪且不完整时，这一特性尤为理想。这一特性与单细胞建模的目标高度契合：从稀疏且含噪的基因表达测量中恢复稳健的细胞状态。

受这些进展的启发，我们提出**Cell-JEPA**，这是一种基于 scGPT 构建的 JEPA 风格单细胞表征学习框架。Cell-JEPA 在重建式预训练的基础上，加入了潜在空间预测目标，使部分观测条件下预测的细胞表征与目标细胞表征保持一致。我们假设，将 scGPT 可扩展的转录组建模与 JEPA 风格的潜在预测相结合，能够获得更好地捕捉稳定细胞程序和高阶生物结构的嵌入。在各项下游任务中，我们表明 Cell-JEPA 学到的表征更加稳健；与仅重建基线相比，它在细胞状态预测问题上的表现更佳。

总而言之，我们有以下三项主要贡献。

- 我们提出 **Cell-JEPA**，一种用于单细胞转录组学的预测式自监督框架， t

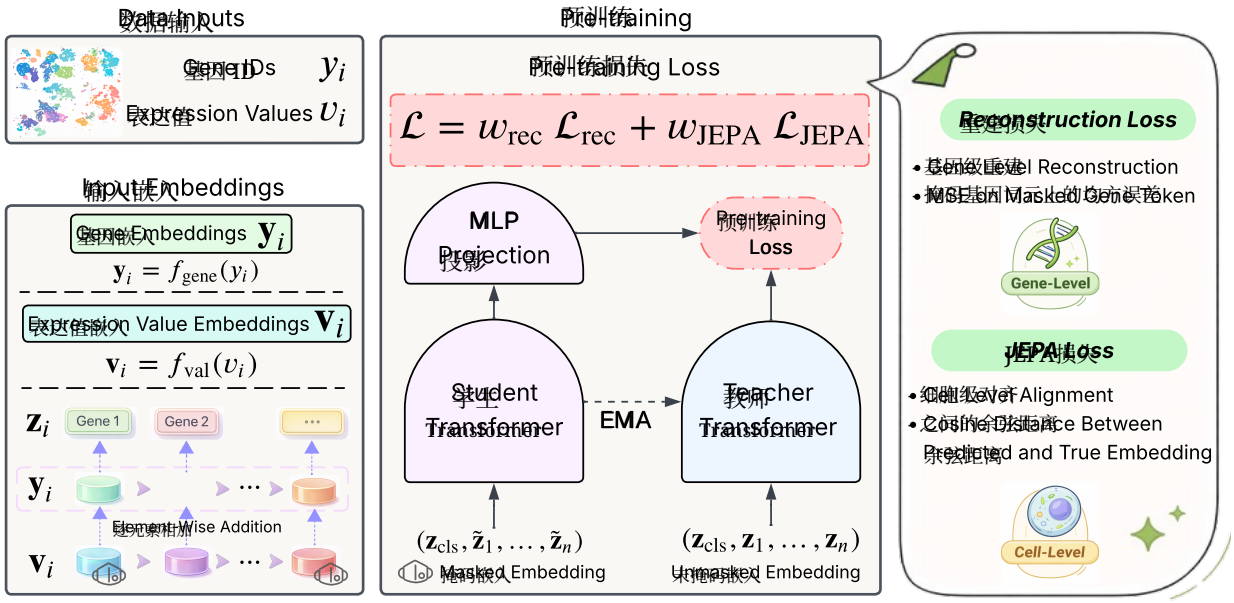


图2：**Cell-JEPA 模型架构与训练流程。** Cell-JEPA 基于 scGPT，采用学生–教师 Transformer 架构。基因标识和离散化表达值经过嵌入并求和，形成输入词元嵌入。学生编码器处理掩码输入，而教师编码器处理相应的未掩码输入，并通过指数移动平均（EMA）进行更新。训练过程优化表达空间中的基因层面重建损失，以及根据学生掩码视图预测教师细胞嵌入的 JEPA 损失。

通过潜在空间预测补充基于重建的预训练。它能够从高维、稀疏且含噪的基因表达数据中，稳定地学习有意义的细胞表征。

- 我们整理了一个大规模人类肾脏单细胞 RNA 测序语料库，包含 800,000 个细胞，涵盖多项研究、多个供体和多种实验条件。基于此，我们采用这一全新的预训练范式进行预训练，并发布我们的单细胞基础模型。
- 我们在三种下游任务设置中评估 Cell-JEPA：（i）用于细胞类型聚类的监督微调；（ii）用于细胞类型聚类的零样本迁移；以及（iii）扰动响应预测。在所有实验中，Cell-JEPA 的表现均优于基线模型 scGPT。尤其是在零样本设置下，Cell-JEPA 的主要评估指标相对提升了 35%。

组织结构。 我们将在第2节介绍预训练和微调方法，在第3节展示实验结果，并在第B节回顾相关工作。

2 方法

本节首先介绍第2.1节中用于 Cell-JEPA 预训练的数据预处理。随后，我们将在第2.2节介绍词元化流程和模型架构。接下来，我们将在第2.3节详细说明预训练目标。最后，我们将在第2.4节介绍下游微调损失。

数据集格式化。 每个单细胞RNA测序数据集都表示为一个稀疏的细胞-基因计数矩阵。遵循先前的研究 [Cui 等人， 2024], 我们排除每个细胞中计数为零的基因，并存储

其余非零条目，并将其保存为相互对应的基因标识符和表达值列表。我们将每个细胞编码为一个稀疏序列 $\{(y_k,\ v_k)\}_{k=1}^{L}$，其中 y_k 表示基因身份，$v_k \in \mathbb{R}_+$ 表示其对应的表达值。

2.1 预训练数据集处理

我们遵循 scGPT 的预处理策略 [Cui 等人，2024] 以（i）通过逐细胞分位数分箱对表达值进行离散化，以及（ii）对基因进行子采样，从而控制不同细胞间的序列长度。

基于分位数的分箱。 遵循 scGPT [Cui 等人，2024], 我们使用逐细胞分位数分箱对非零基因表达值进行离散化，以提高对不同细胞间尺度差异的稳健性。对于每个细胞，我们根据其非零表达值计算分位数分箱边界，并将每个值映射到$\{1,\ldots,B\}$中的箱索引，默认 $B = 50$。这些箱索引作为数值输入 v_k，在保留细胞内表达水平相对顺序的同时，降低对原始计数幅度的依赖。

基因子采样。 细胞表达的基因数量各不相同，从而导致序列长度不同。遵循 scGPT [Cui 等人，2024], 我们将每个细胞的表达基因数量上限设为 $L_{\max}$（设为$L_{\max}$=600）。对于超出该上限的细胞，我们避免选择固定的基因子集（例如前$L_{\max}$个），因为这可能引入偏差。相反，我们从表达基因集合中均匀随机子采样 $L_{\max}$ 个基因令牌，确保每个表达基因被纳入的概率相等。这种随机子采样也可视为一种生物学随机失活：在不同训练轮次中，模型观察到同一细胞的不同部分视图，并且必须从不完整信息中学习。因此，学习到的表示对稀疏性和表达变异性更加稳健。

2.2 词元化与模型架构

概述。 Cell-JEPA 直接基于 scGPT [Cui 等人，2024]。scGPT 是一种双向 Transformer 编码器，在掩码输入下通过基因表达预测目标进行训练（参见第 C 节）。我们采用与 scGPT 相同的词元化、嵌入设计和编码器骨干网络，并通过实例化两个 scGPT 编码器副本来引入联合嵌入预测架构（JEPA）：一个由梯度下降优化的学生网络，以及一个缓慢演化、通过指数移动平均（EMA）更新的教师网络。这种学生–教师框架使我们能够在 scGPT 原有的重建式损失之外，加入表征级 JEPA 目标。下文将详细介绍词元化流程、教师–学生模型设计和编码器架构。

2.2.1 词元化与输入嵌入

词汇表与特殊词元。 遵循 scGPT [Cui 等人，2024], 我们构建了统一的基因词汇表，并添加特殊词元，包括 <cls> 和 <pad>。对于每个细胞，我们在序列前添加一个<cls> 词元，并将一个小批次中的序列填充至长度 $L_{\max} + 1$，使用 <pad>。

基因和数值嵌入。 给定基因身份 y_i 及其（预处理后的）表达值 v_i，我们采用与 scGPT 相同的设计来嵌入基因标记和值：

$$\mathbf{y}_i = f_{\text{gene}}(y_i), \quad \mathbf{v}_i = f_{\text{val}}(v_i),$$

其中 f_{gene} 是一个可学习嵌入查找表，f_{val} 是一个小型多层感知机，用于将标量输入映射到模型的隐藏维度。基因 i 的输入表示为以下两者之和：

$$\mathbf{z}_i = \mathbf{y}_i + \mathbf{v}_i. \tag{2.1}$$

<cls> 词元嵌入从同一基因嵌入表获得：

$$\mathbf{z}_{\text{cls}} = f_{\text{gene}}(\texttt{<cls>}).$$

掩码。 在预训练期间，我们对仅表达值进行掩码，遵循 scGPT [Cui 等人，2024]。对于被掩码的基因 i，我们将 v_i 替换为哨兵值 v_{mask}（设为 -1），从而生成掩码值嵌入：

$$\mathbf{v}_{\text{mask}} = f_{\text{val}}(v_{\text{mask}}),$$

以及掩码标记表示：

$$\tilde{\mathbf{z}}_i = \mathbf{y}_i + \mathbf{v}_{\text{mask}}.$$

令 $\mathbf{Z} = (\mathbf{z}_{\text{cls}}, \mathbf{z}_1, \ldots\ldots, \mathbf{z}_n)$ 表示未掩码的标记序列，$\tilde{\mathbf{Z}} = (\mathbf{z}_{\text{cls}}, \tilde{\mathbf{z}}_1, \ldots\ldots, \tilde{\mathbf{z}}_n)$ 表示相应的掩码序列。

2.2.2 学生-教师编码器

编码器。 学生编码器 g_{S} 与教师编码器 g_{T} 均共享与 *scGPT* 相同的架构 [Cui 等人，2024]（即双向 Transformer 编码器堆栈）。学生处理 $\mathbf{Z}$，而教师处理 $\tilde{\mathbf{Z}}$ 以生成稳定的目标表示：

$$\hat{\mathbf{H}} = g_{\text{S}}(\tilde{\mathbf{Z}}), \quad \mathbf{H} = g_{\text{T}}(\mathbf{Z}),$$

我们使用 <cls> 输出作为细胞级嵌入：

$$\hat{\mathbf{e}} = \hat{\mathbf{H}}_0, \quad \mathbf{e} = \mathbf{H}_0. \tag{2.2}$$

指数移动平均教师更新。 教师参数通过对学生参数进行指数移动平均来更新（JEPA 风格方法中通常采用这种方式）：

$$\theta_{\text{T}} \leftarrow m\,\theta_{\text{T}} + (1-m)\,\theta_{\text{S}},$$

其中 $m \in [0, 1)$ 是动量系数，而 θ_{S}，θ_{T} 分别是学生编码器和教师编码器的参数。这样便得到了一个缓慢演化的目标教师编码器，从而稳定表示预测目标。

2.2.3 Transformer 块与注意力掩码

两个编码器均采用与 scGPT [Cui 等人提出的相同 2024]Transformer 块结构。每个块由自注意力层、多层感知机、非线性变换和层归一化组成。为便于说明，下面将详细介绍自注意力层。

自注意力。对于隐藏序列 $\mathbf{H} \in \mathbb{R}^{(L+1)\times D}$，其中 D 表示嵌入维度。我们使用$\mathbf{A}_{\text{mask}}$ 表示结构化注意力掩码，以防止被掩码目标之间发生信息泄漏。设$\mathcal{U}_{\text{mask}}$ 表示当前步骤中被掩码基因的索引，则$\mathbf{A}_{\text{mask}}$中的相应条目为：

$$a_{i,j} = \begin{cases} 0, & \text{if } j \notin \mathcal{U}_{\text{mask}}, \\ 0, & \text{if } i = j \text{ and } j \in \mathcal{U}_{\text{mask}}, \\ -\infty, & \text{otherwise.} \end{cases}$$

这种设计确保每个被掩码的词元都可以关注所有未被掩码的词元（以及自身），但不会关注其他被掩码的词元。然后，我们将注意力模块计算为：

$$\text{Attn}(\mathbf{H}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{A}_{\text{mask}}\right)\mathbf{V},$$

其中 $\mathbf{Q} = \mathbf{H}\mathbf{W}_Q$、$\mathbf{K} = \mathbf{H}\mathbf{W}_K$、$\mathbf{V} = \mathbf{H}\mathbf{W}_V$以及 $\mathbf{W}_Q$、$\mathbf{W}_K$、$\mathbf{W}_V \in \mathbb{R}^{D\times d}$为投影矩阵，$d$表示注意力头维度。

2.3 预训练损失计算

Cell-JEPA 在原始 scGPT 预训练目标的基础上 [Cui 等人， 2024] 加入了 JEPA 风格的表征预测损失。具体而言，我们通过优化学生编码器来实现以下目标：（i）基因层面重建损失，其理念与 scGPT 相同；（ii）细胞级 JEPA 目标，即从学生掩码视图预测教师的潜在嵌入。

JEPA 目标。令$\hat{\mathbf{e}}$和$\mathbf{e}$分别表示来自（<cls>）的学生和教师嵌入2.2。我们引入一个预测器头 $p(\cdot)$，并将其应用于学生嵌入：

$$\tilde{\mathbf{e}} = p(\hat{\mathbf{e}}).$$

然后，我们最小化预测的学生嵌入与教师目标嵌入之间的余弦距离损失：

$$\mathcal{L}_{\text{JEPA}} = 1 - \frac{\tilde{\mathbf{e}}^\top \text{sg}(\mathbf{e})}{\|\tilde{\mathbf{e}}\|_2 \, \|\text{sg}(\mathbf{e})\|_2}, \tag{2.3}$$

其中，sg(·) 表示停止梯度（教师嵌入被视为固定目标）。不同于在表达空间中定义的重构损失，$\mathcal{L}_{\text{JEPA}}$ 仅在表征空间中运作，并促使学生在掩码条件下匹配教师的细胞层面语义。

基因层面重建损失。除了 $\mathcal{L}_{\text{JEPA}}$之外，我们还保留了 scGPT 式的基因表达预测目标 [Cui 等人， 2024]。不同于 scGPT 的迭代生成过程，我们在单次前向传播中预测所有 被掩码基因，采用 BERT式掩码预测方案。

$\hat{}$~给定学生词元输出 $\mathbf{H} = g_S(\mathbf{Z})$，我们使用值预测头$r(\cdot)$（逐行应用的多层感知机）来预测每个词元的表达值：

$$\hat{\mathbf{v}} = r(\hat{\mathbf{H}}),$$

其中，$\hat{\mathbf{v}}_i$ 表示基因标记 i的重构表达值。我们在掩码基因集 $\mathcal{U}_{\text{mask}}$上计算均方误差：

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{U}_{\text{mask}}|} \sum_{i \in \mathcal{U}_{\text{mask}}} (\hat{\mathbf{v}}_i - v_i)^2 . \tag{2.4}$$

这一项通过明确鼓励根据已观察到的上下文准确预测掩码表达值，保留了 scGPT 的训练信号。

联合预训练目标。最终的预训练目标是重建损失与 JEPA 损失的加权组合：

$$\mathcal{L}_{\text{pre-train}} = w_{\text{rec}} \mathcal{L}_{\text{rec}} + w_{\text{JEPA}} \mathcal{L}_{\text{JEPA}},$$

其中，w_{rec} 和 w_{JEPA} 控制各项的相对权重。直观而言，$\mathcal{L}_{\text{rec}}$ 将学习锚定在基因水平表达信号上，与 scGPT 中的做法一致；而 $\mathcal{L}_{\text{JEPA}}$ 则确保细胞层面的表征在潜在空间中保持一致。联合优化两者，可获得既保留用于表达恢复的细粒度信息，又通过表征层面的预测提升鲁棒性和语义结构的嵌入。

2.4 下游微调

在下游微调期间，我们使用四个目标优化 Cell-JEPA：（i）基因表达预测（GEP）；（ii）细胞建模基因表达预测（GEPC）；（iii）弹性细胞相似性（ECS）；以及（iv）JEPA预测损失。我们采用 scGPT 中的 GEP、GEPC 和 ECS [Cui 等人， 2024], 并通过额外保留 $\mathcal{L}_{\text{JEPA}}$来扩展微调目标。

2.4.1 GEP 和 GEPC 损失

基因表达预测（GEP）。GEP 等价于预训练期间使用的重建损失 $\mathcal{L}_{rec}$ （2.4）。使用学生模型预测$\hat{\mathbf{v}}_i$，我们在掩码基因集上最小化均方误差 $\mathcal{U}_{\text{mask}}$：

$$\mathcal{L}_{\text{GEP}} = \frac{1}{|\mathcal{U}_{\text{mask}}|} \sum_{i \in \mathcal{U}_{\text{mask}}} (\hat{\mathbf{v}}_i - v_i)^2 .$$

细胞建模基因表达预测（GEPC）。GEPC 是来自 scGPT 的互补目标，用于将基因级预测明确关联到全局细胞表征 [Cui 等人， 2024]。对于每个基因标记 i，我们从其基因嵌入中计算一个查询向量$\mathbf{y}_i$ ，并通过与细胞嵌入$\hat{}\mathbf{e}$进行内积来预测表达：

$$\tilde{\mathbf{v}}_i = f(\mathbf{y}_i)^T \mathbf{W} \hat{\mathbf{e}},$$

其中，$f(\cdot)$ 是一个多层感知机，$\mathbf{W}$ 是一个可学习权重矩阵。GEPC 损失是被掩码基因上的均方误差：

$$\mathcal{L}_{\text{GEPC}} = \frac{1}{|\mathcal{U}_{\text{mask}}|} \sum_{i \in \mathcal{U}_{\text{mask}}} (\tilde{\mathbf{v}}_i - v_i)^2.$$

GEP 和 GEPC 共同提供互补的监督：GEP 确保准确恢复被掩码值，而 GEPC 则鼓励全局细胞嵌入能够预测基因水平表达。

2.4.2 ECS 损失

弹性细胞相似性（ECS）。ECS 通过鼓励生物学上相似的细胞具有更接近的表征，对细胞嵌入空间进行正则化。虽然 scGPT 提出了基于成对相似度阈值的 ECS 正则项 [Cui 等人， 2024], 但根据第 C.5 节所述的直觉，我们将该组件修改为以下形式。考虑一个包含 N 个细胞及其嵌入 $\{(\mathbf{e}_i,\ c_i)\}_{i=1}^{N}$的小批次，其中$\mathbf{e}_i$ 表示细胞 i 的嵌入（如（2.2）中所定义），而 c_i 是其对应的细胞类型标签。我们将细胞 i 与 j 之间的余弦相似度定义为：

$$s_{ij} = \frac{\cos\left(\mathbf{e}^i, \mathbf{e}^j\right)}{\tau},$$

其中 τ 是温度超参数。

我们依次将小批次中的每个细胞视为锚点。对于给定的锚点细胞 i，我们将小批次中与其具有相同细胞类型的细胞确定为：

$$p_{ij} = \begin{cases} \frac{1}{|\mathcal{P}(i)|}, & \text{if } c_j = c_i, \\ 0, & \text{otherwise.} \end{cases}$$

锚点细胞 i的对比损失为：

$$\mathcal{L}_{\mathrm{ECS},i} = -\sum_{j=1}^{N} p_{ij} \log \frac{\exp(s_{ij})}{\sum_{k=1}^{N} \exp(s_{ik})}.$$

我们通过对小批次中的所有锚点取平均值，计算弹性细胞相似性损失：

$$\mathcal{L}_{\mathrm{ECS}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{ECS},i}.$$

这一目标充当连续的正则项，以确保学习到的细胞嵌入空间具有结构化的相似性。

2.4.3 JEPA与总体损失

我们在微调期间保留 JEPA 损失项 $\mathcal{L}_{\mathrm{JEPA}}$ （2.3） 。这使模型能够调整嵌入空间映射，以适应数据集内的分布差异。

总体微调目标 $\mathcal{L}_{\mathrm{finetune}}$ 定义为四项损失的加权和，如下所示：

$$\mathcal{L}_{\mathrm{finetune}} = w_{\mathrm{GEP}}\mathcal{L}_{\mathrm{GEP}} + w_{\mathrm{GEPC}}\mathcal{L}_{\mathrm{GEPC}} + w_{\mathrm{ECS}}\mathcal{L}_{\mathrm{ECS}} + w_{\mathrm{JEPA}}\mathcal{L}_{\mathrm{JEPA}}.$$

对于 scGPT 基线，我们使用相同的微调目标，但不包括 $\mathcal{L}_{\mathrm{JEPA}}$：

$$\mathcal{L}_{\mathrm{finetune}}^{\mathrm{scGPT}} = w_{\mathrm{GEP}}\mathcal{L}_{\mathrm{GEP}} + w_{\mathrm{GEPC}}\mathcal{L}_{\mathrm{GEPC}} + w_{\mathrm{ECS}}\mathcal{L}_{\mathrm{ECS}}.$$

2.5 扰动预测

扰动目标。 我们进一步微调预训练的 scGPT 和 Cell-JEPA 模型，以预测基因扰动效应。为融入扰动信息，我们在预测模型中使用标准 ID嵌入层 $f_{\text{perturb}}(\cdot)$ 来嵌入基因 i的扰动标签 p_i：

$$\mathbf{p}_i = f_{\text{perturb}}(p_i).$$

我们修改（2.1）中的基因词元表示，将扰动嵌入整合其中：

$$\tilde{\mathbf{z}}_i^{\text{pert}} = \mathbf{y}_i + \mathbf{p}_i + \mathbf{v}_i.$$

我们将扰动后的基因嵌入传入学生编码器，以生成预测的扰动基因值。令 $\tilde{\mathbf{Z}}^{\text{pert}} = (\mathbf{z}_{\text{cls}}, \tilde{\mathbf{z}}_1^{\text{pert}}, \ldots, \tilde{\mathbf{z}}_n^{\text{pert}})$，预测过程如下：

$$\hat{\mathbf{H}}_{\text{pert}} = g_{\text{S}}(\tilde{\mathbf{Z}}_{\text{pert}}), \quad \hat{\mathbf{v}}_i^{\text{pert}} = r_{\text{pert}}(\hat{\mathbf{H}}_{\text{pert}}),$$

其中， r_{pert} 是一个 MLP 值头。我们将扰动重建损失定义为 a s:

$$\mathcal{L}_{\text{pert-pred}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{v}}_i^{\text{pert}} - v_i^{\text{pert}})^2,$$

其中， v_i^{pert} 是基因 i的真实扰动后表达值。

JEPA 和总体损失。与预训练和微调设置不同，我们不对基因值输入应用掩码。我们将学生扰动细胞嵌入定义为：

$$\hat{\mathbf{e}}_{\text{pert}} = \hat{\mathbf{H}}_0^{\text{pert}}.$$

我们将目标扰动基因嵌入构造如下：

$$\mathbf{z}_i^{\text{pert}} = \mathbf{y}_i + \mathbf{p}_i + \mathbf{v}_i^{\text{pert}},$$

其中，$\mathbf{v}_i^{\text{pert}}$ 表示基因 i 的扰动基因表达嵌入。令 $\mathbf{Z}^{\text{pert}} = (\mathbf{z}_{\text{cls}}, \mathbf{z}_1^{\text{pert}}, \ldots\ldots, \mathbf{z}_n^{\text{pert}})$。目标编码器生成：

$$\mathbf{H}_{\text{pert}} = g_{\text{T}}(\mathbf{Z}_{\text{pert}}), \quad \mathbf{e}_{\text{pert}} = \mathbf{H}_0^{\text{pert}}.$$

我们引 入一个预测器头 $p_{\text{pert}}(\cdot)$（多层感知机），并将其应用于学生的扰动嵌入 ：

$$\tilde{\mathbf{e}}_{\text{pert}} = p_{\text{pert}}(\hat{\mathbf{e}}_{\text{pert}}).$$

我们以类似于（2.3）的方式计算扰动 JEPA损失：

$$\mathcal{L}_{\text{JEPA}}^{\text{pert}} = 1 - \frac{\tilde{\mathbf{e}}_{\text{pert}}^{\top}\,\text{sg}(\mathbf{e}_{\text{pert}})}{\|\tilde{\mathbf{e}}_{\text{pert}}\|_2\,\|\text{sg}(\mathbf{e}_{\text{pert}})\|_2}.$$

我们还采用第2.4.2 节中的 ECS 公式来定义扰动 ECS 损失，其中两个细胞 i 与 j 之间的相似度计算为其扰动细胞嵌入之间的余弦相似度

$\hat{\mathbf{e}}_i^{\text{pert}}$ 和 $\hat{\mathbf{e}}_j^{\text{pert}}$ 。

整体扰动目标是所有扰动损失项的加权和：

$$\mathcal{L}_{\text{pert}} = w_{\text{pert-rec}}\mathcal{L}_{\text{pert-rec}} + w_{\text{JEPA}}^{\text{pert}}\mathcal{L}_{\text{JEPA}}^{\text{pert}} + w_{\text{ECS}}\mathcal{L}_{\text{ECS}}.$$

对于 scGPT 基线，我们使用相同的目标，但不包含 JEPA 项：

$$\mathcal{L}_{\text{pert}}^{\text{scGPT}} = w_{\text{pert-rec}}\mathcal{L}_{\text{pert-rec}} + w_{\text{ECS}}\mathcal{L}_{\text{ECS}}.$$

3 实验研究

本节首先在第3.1节介绍基础模型预训练方案。随后，我们在三种下游任务设置中评估预训练模型：(i) 在第3.2节进行监督微调以实现细胞类型聚类，(ii) 在第3.3节进行细胞类型聚类的零样本迁移，以及 (iii) 在第3.4节进行扰动响应预测。

3.1 预训练方案

为验证 Cell-JEPA 框架的优势，我们使用相同的预训练数据集和训练方案预训练两个基础模型：Cell-JEPA 和 scGPT [Cui 等人， 2024]。这两个模型的唯一差异在于预训练目标，如第2.3节所示。

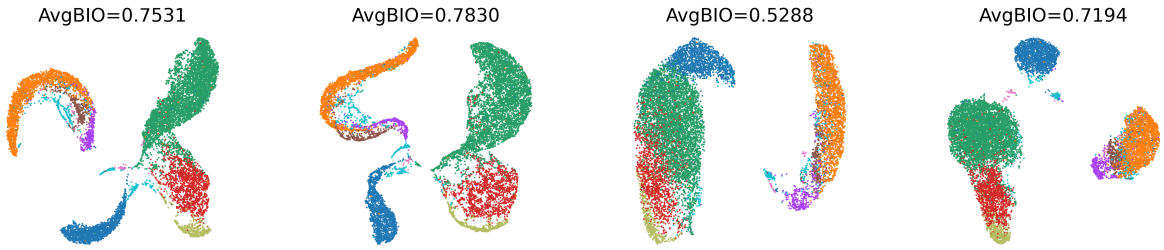
预训练数据集。 为了预训练基础模型，我们从 CELLxGENE Census 收集了一个大规模的人类肾脏 scRNA-seq 语料库 [Program 等人， 2024] (版本 `2023-05-15`）。我们按生物体（智人）和组织（肾脏）筛选数据集，并限定为 RNA 模态。由此得到一个异质性肾脏细胞集合，涵盖多项研究、多个供体和不同实验条件，所有组成数据集均列于第E.1节。预训练语料库共包含**80 万**个细胞。

预训练方案。 我们使用第2.3节所述的预训练目标训练 Cell-JEPA。对于 scGPT 基线，我们遵循第C节以及Cui 等人 [2024]所述的训练目标。Cell-JEPA 模型的训练实现细节和超参数设置见第E.1节。

3.2 任务 1：在外周血单个核细胞上微调

任务描述。 我们从 Cell-JEPA 和 scGPT 中提取细胞嵌入，并主要通过定量聚类指标评估表征质量，同时辅以定性的 UMAP 可视化。我们参照 Cui 等人 [2024], Luecken 等人 [2022]报告具有生物学意义的评估指标。

数据集。 我们在一个外周血单个核细胞（PBMC）数据集 PBMC-10k 上进行评估。该数据集由来自一位健康供体的两个单细胞 RNA 测序批次组成（名义上分别包含 4k 和 8k 个细胞） [10xGenomics， 2017a， b]。数据按照`scvi-tools` PBMC 流程 [Gayoso 等人， 2022]进行预处理。处理后的数据集包含 3,346 个高变基因，第一批次有 7,982 个细胞，第二批次有 4,008 个细胞 [Gayoso 等人， 2022]。细胞类型注释涵盖主要免疫细胞群，包括 B 细胞、CD4+ 和 CD8+ T 细胞、NK 细胞、单核细胞、树突状细胞等。



(a) 微调后的 scGPT (b) 微调后的 Cell-JEPA (c) scGPT 的零样本评估 (d) Cell-JEPA 的零样本评估

图 3：**scGPT 与 Cell-JEPA 细胞嵌入的比较**。PBMC-10K 上细胞嵌入的 UMAP 可视化。(a) 微调后的 scGPT（任务 1），(b) 微调后的 Cell-JEPA（任务 1），(c) scGPT 的零样本评估（任务 2），以及 (d) Cell-JEPA 的零样本评估（任务 2）。

训练/验证集划分。 我们采用 [Cui 等人的划分策略， 2024]。我们以 9:1 的比例将 PBMC 10K 数据集随机划分为训练集和验证集，并使用固定随机种子 42 以确保结果可复现。UMAP 可视化使用完整的 PBMC 10K 数据集生成。

微调方案。我们使用第 2.4 节中所述的微调目标，在每个数据集上对预训练的 Cell-JEPA 和 scGPT 检查点进行微调。Cell-JEPA 模型微调所用的实现细节和超参数见第 E.2 节。

评估指标。 我们考虑四项嵌入质量指标：平均生物学保守性（AvgBIO）、归一化互信息（$\mathrm{NMI}_{\mathrm{cell}}$）、平均轮廓宽度（$\mathrm{ASW}_{\mathrm{cell}}$）和调整兰德指数（$\mathrm{ARI}_{\mathrm{cell}}$）。详细定义见第D.1节。

结果。 我们为PBMC-10K生成单细胞嵌入，并使用UMAP进行可视化，结果表明两个模型所学得的潜在空间均对注释细胞类型呈现出定性分离。见图3a和3b。与基线scGPT模型相比，Cell-JEPA在全部四项聚类指标上均取得了更好的性能（表1）。这些结果表明，用JEPA目标增强scGPT有助于从单细胞基因表达数据中学习信息更丰富的表征。

表1：**任务1的性能：在外周血单个核细胞上微调**。Cell-JEPA与scGPT在PBMC-10K数据集上微调后的 AvgBIO 分数、$\mathrm{NMI}_{\mathrm{cell}}$、$\mathrm{ASW}_{\mathrm{cell}}$和$\mathrm{ARI}_{\mathrm{cell}}$指标。经过特定任务微调后，Cell-JEPA在全部四项指标上均优于scGPT。

模型	AvgBIO	$\mathrm{NMI}_{\mathrm{cell}}$	$\mathrm{ASW}_{\mathrm{cell}}$	$\mathrm{ARI}_{\mathrm{cell}}$
scGPT	0.7531	0.7652	0.7100	0.7842
Cell-JEPA	**0.7830**	**0.7761**	**0.7256**	**0.8472**

表2：**任务2：在外周血单个核细胞上进行零样本评估的表现**。预训练的 Cell-JEPA 和 scGPT 模型在 PBMC-10K数据集上进行零样本评估时的 AvgBIO 分数、$\mathrm{NMI}_{\mathrm{cell}}$、$\mathrm{ASW}_{\mathrm{cell}}$ 和 $\mathrm{ARI}_{\mathrm{cell}}$ 指标。Cell-JEPA 的表现优于 scGPT。

模型	AvgBIO	$\mathrm{NMI}_{\mathrm{cell}}$	$\mathrm{ASW}_{\mathrm{cell}}$	$\mathrm{ARI}_{\mathrm{cell}}$
scGPT	0.5288	0.5585	0.5329	0.4951
Cell-JEPA	**0.7194**	**0.7690**	**0.5811**	**0.8081**

3.3 任务2：外周血单个核细胞零样本学习

接下来，我们在不进行任何特定任务微调的情况下，评估零样本表征质量。我们使用预训练检查点进行推理，并使用与第3.2节相同的指标评估嵌入。

结果。 我们通过零样本推理为 PBMC-10K 生成单细胞嵌入，并使用 UMAP 对其进行可视化（图3c 和3d），结果表明，与 scGPT 相比，Cell-JEPA 对注释细胞类型具有更强的零样本分离能力。此外，在各项定量指标上，Cell-JEPA 在零样本设置下始终优于 scGPT；例如，相较于 scGPT，其 AvgBIO 提升了 36% （表2）。总体而言，这些结果表明，JEPA 目标提升了学习到的表示的迁移能力，使模型在无需下游微调的情况下也能实现强劲的性能。

3.4 任务3：扰动预测

任务描述。 我们探究在固定的生物学背景下，改进表征是否有助于扰动响应预测。给定细胞的基线表达和一种遗传扰动，目标是预测扰动后表达。两个评估数据集（Norman、Adamson）均使用 K562 细胞系；因此，我们评估的是对留出扰动的背景内泛化，而非跨背景迁移。

数据集。 我们在 K562 白血病细胞系中测定的两个 Perturb-seq CRISPR筛选数据集上评估扰动响应预测：Adamson [Adamson 等人、 2016] 以及 Norman [Norman 等人，2019]。Adamson 是一个 CRISPR干扰 Perturb-seq 数据集，包含 86 种单基因扰动（经 GEARS 处理）以及匹配的对照细胞。Norman 是一个 CRISPR激活 Perturb-seq 数据集，包含 105 种单基因扰动和 131 种双基因扰动（经 GEARS 处理），每种扰动平均有数百个细胞。所有数据集均使用 GEARS流程进行处理 [Roohani 等人、 2023]。

训练/验证/测试集划分。我们遵循 GEARS 模拟划分协议，该协议在扰动层级而非细胞层级划分数据。在此设置下，测试集中的扰动在训练期间被留出。在 Norman 数据集中，留出的双基因扰动还会根据训练期间两个组成基因、其中一个组成基因或两个都未作为单基因扰动出现过，进一步分为`combo_seen2`、`combo_seen1`和`combo_seen0`。 [Roohani 等人， 2023]。

微调方案。我们使用官方的 scGPT 扰动微调流程对*scGPT* 和我们的*Cell-JEPA* 模型进行微调。该流程利用 GEARS 进行数据集预处理、划分构建和评估。在此设置中，基于 Transformer 的预测器（scGPT，也经过调整以用于 Cell-JEPA）经过训练，用于建模扰动后基因表达。每个微调数据点包括：(i) 未扰动（对照组）的基因表达谱，(ii) 指明受扰动基因的二元扰动指示变量，以及 (iii) 真实的扰动后表达谱。两个模型均训练 15 个训练轮次，学习率为 10^{-4}，批次大小为 64。训练在单张2 A100 GPU 上耗时。我们在第E.3节提供完整的训练配置详情。

评估指标。 遵循 GEARS [Roohani 等人， 2023]，我们使用预测值与观测到的扰动后基因表达向量之间的皮尔逊相关系数来评估扰动预测。我们报告 皮尔逊相关系数 （基于所有基因），差异表达皮尔逊相关系数 （仅限于差异表达基因），以及前20个非缺失差异表达基因 （在过滤脱落基因后，仅限于排名最高的差异表达基因）。完整细节见第D.2节。

结果。 表 3 总结了 GEARS `unseen_single` 扰动在 Adamson 和 Norman 数据集上的表现。与 scGPT 基线相比，Cell-JEPA 在绝对状态预测和以差异表达为重点的恢复方面均持续取得提升，其中 Norman 数据集上的提升最大。这些结果表明，增强

表 3：**任务3：未见过的单基因扰动预测**。与 scGPT 相比，Cell-JEPA 提升了 Norman 和 [Norman 等人、 2019]以及 Adamson [Adamson 等人， 2016] 数据集上的扰动后绝对状态预测和以差异表达为重点的恢复表现。 `pear` 表示皮尔逊相关系数，`pear_de`表示差异表达皮尔逊相关系数，而`top20`表示前20个非缺失差异表达基因。

数据集	模型	pear	pear_de	top20
Norman	scGPT	0.631	0.276	0.278
Norman	Cell-JEPA	**0.787**	**0.592**	**0.565**
Adamson	scGPT	0.905	0.699	0.677
Adamson	Cell-JEPA	**0.937**	**0.741**	**0.720**

表4：**未见过的单基因扰动的基于差值的指标**。两种模型的差值相关性仍然较低，且 Cell-JEPA 在两个数据集上的表现均有所下降，其中 Adamson 数据集上的降幅最大 [Adamson 等人， 2016]。Δ_pear表示皮尔逊差值， Δ_pear_de 表示皮尔逊差异表达差值，以及 Δ_top20表示皮尔逊差值前20差异表达非脱落。

数据集	模型	Δ_pear	Δ_pear_de	Δ_top20
Norman	scGPT	**0.103**	**0.105**	**0.169**
Norman	JEPA	0.059	0.041	0.115
Adamson	scGPT	**0.173**	**0.101**	**0.120**
Adamson	JEPA	0.120	−0.176	−0.152

采用潜在空间 JEPA 目标对 scGPT 进行微调，能够获得对未见过的遗传扰动泛化能力更强的表征。

我们还报告了基于差值的指标（皮尔逊差值、`pearson_de_delta`、`delta_top20_de_non_dropout`），这些指标评估相对于对照状态的表达变化（扰动后减去对照）的相关性，因此强调效应量准确性，而非扰动后状态预测的绝对值。如表4所示，scGPT 和 Cell-JEPA 在未见过的单基因扰动上均表现出较低的差值相关性，表明在此情形下，准确估计相对于对照的变化仍然颇具挑战。虽然 Cell-JEPA 改善了绝对状态指标（表3），但它并未持续改善基于差值的指标，且在 Adamson 数据集上的降幅最大。

一种合理的解释是，JEPA 目标鼓励学习在不同掩码视图间保持不变的特征，这有助于识别细胞状态，但可能会降低对扰动所致偏差的敏感性。这表明，表征学习（这个细胞处于什么状态？）和扰动建模（细胞会如何变化？）分别针对细胞预测的互补方面，可能需要采用不同的方法。总体而言，我们认为这些结果凸显了绝对状态保真度与效应量准确性之间的权衡，并指出，要充分发挥潜在空间预测学习在扰动泛化方面的优势，可能需要扩展到更大的扰动数据集，和/或使用显式考虑变化量的目标进行训练。

4 结论

我们提出了 Cell-JEPA，这是一种联合嵌入预测框架，以潜在空间预测补充基于重建的预训练。Cell-JEPA 通过预测细胞级嵌入而非重建含噪计数，学习能够抵御主导单细胞测量的丢弃伪影的表征。

我们的主要发现是零样本细胞类型聚类相对提升了 36%，这表明潜在预测能够生成无需特定任务微调即可迁移的表征。在单一细胞系内进行扰动预测时，我们观察到绝对状态重建有所改善，但效应量估计并未改善。这表明，学习稳健的细胞表征与建模扰动效应是互补的挑战，可能需要采用不同的方法。

这些结果表明，潜在空间预测是单细胞表征学习的一个有前景的方向，同时也凸显了扰动建模仍是一个尚未解决的问题，可能需要显式的上下文建模，或在训练数据中提高上下文多样性。局限性见第 A 节，相关工作见第 B节。

致谢

我们感谢西北大学的 Quest 高性能计算设施提供计算资源和人员支持；Quest 由教务长办公室、研究办公室和西北大学信息技术部门共同支持。

H.L. 获得 NIH R01LM1372201、NSF AST-2421845、Simons Foundation MPS-AI-00010513、艾伯维、杜比以及陈-扎克伯格生物中心芝加哥分中心奖的支持。A.A.K. 获得 NIH DP2AI177884、陈-扎克伯格研究员奖以及自身免疫共同机制洞察奖的支持。M.K. 获得 NIH T32GM139782 的支持。

本文内容完全由作者负责，不一定代表资助机构的官方观点。

参考文献

10x Genomics。来自一位健康供体的 4k 个外周血单个核细胞（使用 Cell Ranger v2.1.0 生成的单细胞基因表达数据集）。10x Genomics 公共数据集，2017a。

10x Genomics。来自一位健康供体的 8k 个外周血单个核细胞（由 Cell Ranger v2.1.0 生成的单细胞基因表达数据集）。10x Genomics 公共数据集，2017b。

Britt Adamson、Thomas M. Norman、Marco Jost、Min Y.Cho、James K. Nuñez、Yuwen Chen、Jacqueline E. Villalta、Luke A. Gilbert、Max A. Horlbeck、Marco Y. Hein、Ryan A. Pak、Andrew N. Gray、Carol A. Gross、Atray Dixit、Oren Parnas、Aviv Regev 和 Jonathan S. Weissman。多重单细胞 CRISPR 筛选平台可对未折叠蛋白反应进行系统性解析。细胞，167(7)：1867–1882.e21，2016 年 12 月。

Constantin Ahlmann-Eltze、Wolfgang Huber 和 Simon Anders。基于深度学习的基因扰动效应预测尚未超越简单的线性基线。《自然·方法》，22(8)：1657–1661，2025年。

Mahmoud Assran、QuentinDuval、Ishan Misra、Piotr Bojanowski、Pascal Vincent、Michael Rabbat、Yann LeCun 和 Nicolas Ballas。利用联合嵌入预测架构从图像进行自监督学习，2023。

Mido Assran、Adrien Bardes、David Fan、Quentin Garrido、Russell Howes、Mojtaba、Komeili、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zholus、Sergio Arnaud、Abha Gejji、Ada Martin、Francois Robert Hogan、Daniel Dugas、Piotr Bojanowski、Vasil Khalidov、Patrick Labatut、Francisco Massa、Marc Szafraniec、Kapil Krishnakumar、Yong Li、Xiaodong Ma、Sarath Chandar、Franziska Meier、Yann LeCun、Michael Rabbat 和 Nicolas Ballas。《V-JEPA 2：自监督视频模型助力理解、预测与规划》，2025年。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、AmandaAskell 等。《语言模型是少样本学习者》。神经信息处理系统进展，33：1877–1901，2020年。

Junyue Cao、Diana R. O'Day、Hannah A. Pliner、Paul D. Kingsley、Mei Deng、Riza M. Daza、Michael A. Zager、Kimberly A. Aldinger、Ronnie Blecher-Gonen、Fan Zhang、Malte Spielmann、James Palis、Dan Doherty、Frank J. Steemers、Ian A. Glass、Cole Trapnell 和 Jay Shendure。胎儿基因表达的人类细胞图谱。《科学》，370(6518)，2020年11月。

The Tabula Sapiens联盟*、Robert C. Jones、Jim Karkanias、Mark A. Krasnow、Angela Oliveira Pisco、Stephen R. Quake、Julia Salzman、Nir Yosef、Bryan Bulthaup、Phillip Brown、William Harper、Marisa Hemenez、Ravikumar Ponnusamy、Ahmad Salehi、Bhavani A. Sanagavarapu、Eileen Spallino、Ksenia A. Aaron、Waldo Concepcion、James M. Gardner、Burnett Kelly、Nikole Neidlinger、Zifa Wang、Sheela Crasta、Saroja Kolluru、Maurizio Morri、Serena Y. Tan、Kyle J. Travaglini、Chenling Xu、Marcela Alcántara-Hernández、Nicole Almanzar、Jane Antony、Benjamin Beyersdorf、Deviana Burhan、Kruti Calcuttawala、Matthew M. Carter、Charles K. F. Chan、Charles A. Chang、Stephen Chang、Alex Colville、Rebecca N. Culver、Ivana Cvijović、Gaetano D'Amato、Camille Ezran、Francisco X. Galdos、Astrid Gillich、William R. Goodyer、Yan Hang、Alyssa Hayashi、Sahar Houshdaran、Xianxi Huang、Juan C. Irwin、SoRi Jang、Julia Vallve Juanico、Aaron M. Kershner、Soochi Kim、Bernhard Kiss、William Kong、Maya E. Kumar、Angera H. Kuo、Rebecca Leylek、Baoxiang Li、Gabriel B. Loeb、Wan-Jin Lu、Sruthi Mantri、Maxim Markovic、Patrick L. McAlpine、Antoine de Morree、Karim Mrouj、Shravani Mukherjee、Tyler Muser、Patrick Neuhöfer、Thi D. Nguyen、Kimberly Perez、Ragini Phansalkar、Nazan Puluca、Zhen Qi、Poorvi Rao、Hayley Raquer-McKay、Nicholas Schaum、Bronwyn Scott、Bobak Seddighzadeh、Joe Segal、Sushmita Sen、Shaheen Sikandar、Sean P. Spencer、Lea C. Steffes、Varun R. Subramaniam、Aditi Swarup、Michael Swift、Will Van Treuren、Emily Trimm、Stefan Veizades、Sivakamasundari Vijayakumar、Kim Chi Vo、Sevahn K. Vorperian、Wanxin Wang、Hannah N. W. Weinstein、Juliane Winkler、Timothy T. H. Wu、Jamie Xie、Andrea R. Yung、Yue Zhang、Angela M. Detweiler、Honey Mekonen、Norma F. Neff、Rene V. Sit、Michelle Tan、Jia Yan、Gregory R. Bean、Vivek Charu、Erna Forgó、Brock A. Martin、Michael G. Ozawa、Oscar Silva、Angus Toland、Venkata N. P. Vemuri、Shaked Afik、Kyle Awayan、Olga Borisovna Botvinnik、Ashley Byrne、Michelle Chen、Roozbeh Dehghannasiri、Adam Gayoso、Alejandro A. Granados、Qiqing Li、Gita Mahmoudabadi、Aaron McGeever、Julia Eve Olivieri、Madeline Park、Neha Ravikumar、Geoff Stanley、Weilun Tan、Alexander J. Tarashansky、Rohan Vanheusden、Peter Wang、Sheng Wang、Galen Xing、Les Dethlefsen、Po-Yi Ho、Shixuan Liu、Jonathan S. Maltzman、Ross J. Metzger、Koki Sasagawa、Rahul Sinha、Hanbing Song、Bruce Wang、Steven E. Artandi、Philip A. Beachy、Michael F. Clarke、Linda C. Giudice、Franklin W. Huang、Kerwyn Casey Huang、Juliana Idoyaga、Seung K. Kim、Christin S. Kuo、Patricia Nguyen、Thomas A. Rando、Kristy Red-Horse、Jeremy Reiter、David A. Relman、Justin L. Sonnenburg、Albert Wu、Sean M. Wu 和 Tony Wyss-Coray。Tabula Sapiens：人类多器官单细胞转录组图谱。《科学》，376(6594)，2022年5月。

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。《自然·方法》，21 (8):1470–1480，2024年8月。

Jun Ding、Nadav Sharon 和 Ziv Bar-Joseph。利用单细胞转录组学进行时间建模。Nature Reviews Genetics，23(6):355–368，2022年6月。

Adam Gayoso、Rom ain Lopez、Galen Xing、Pierre Boyeau、Valeh Valiollah Pour Amiri、Justin Hong、Kather-

ine Wu、Michael Jayasuriya、Edouard Mehlman、Maxime Langevin、Yining Liu、Jules Samaran、Gabriel Misrachi、Achille Nazaret、Oscar Clivio、Chenling Xu、Tal Ashuach、Mariano Gabitto、Mohammad Lot- follahi、Valentine Svensson、Eduardo da Veiga Beltrame、Vitalii Kleshchevnikov、Carlos Talavera-López、Lior Pachter、Fabian J. Theis、Aaron Streets、Michael I. Jordan、Jeffrey Regier 和 Nir Yosef。用于单细胞组学数据概率分析的 Python 库。《自然·生物技术》，40(2):163–166，2022年2月。

Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, Yincong Zhou, Fang Ye, Mengmeng Jiang, Junqing Wu, Yanyu Xiao, Xiaoning Jia, Tingyue Zhang, Xiaojie Ma, Qi Zhang, XueliBai, Shujing Lai, ChengxuanYu, Lijun Zhu, Rui Lin, Yuchi Gao, Min Wang, Yiqing Wu, Jianming Zhang, Renya Zhan, Saiyong Zhu, Hailan Hu, Changchun Wang, Ming Chen, He Huang, Tingbo Liang, Jianghua Chen, Weilin Wang, Dan Zhang 和 Guoji Guo。单细胞水平人类细胞图谱的构建。《自然》，581(7808):303–309，2020年3月。

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang 和 Le Song。大规模单细胞转录组学基础模型。《自然·方法》，21(8):1481–1491，2024年6月。

Hai Huang、Yann LeCun 和 Randall Balestriero。LLM-JEPA：大语言模型与联合嵌入预测架构相遇，2025年。

Matthias Kretzler、Edgar Otto、Christopher O'Connor、Markus Bitzer 和 Rajasree Menon。HCA肾脏种子网络：密歇根大学，2025年。

David Lähnemann、Johannes Köster、Ewa Szczurek、Davis J McCarthy、Stephanie C Hicks、Mark D Robinson、Catalina A Vallejos、Kieran R Campbell、Niko Beerenwinkel、Ahmed Mahfouz 等。单细胞数据科学面临的十一项重大挑战。《基因组生物学》，21(1):31，2020年。

Blue B. Lake、Rajasree Menon、Seth Winfree、Qiwen Hu、Ricardo Melo Ferreira、Kian Kalhor、Daria Barwinska、Edgar A. Otto、Michael Ferkowicz、Dinh Diep、Nongluk Plongthongkum、Amanda Knoten、Sarah Urata、Abhijit S. Naik、Sean Eddy、Bo Zhang、Yan Wu、Diane Salamon、James C. Williams、Xin Wang、Karol S. Balderrama、Paul Hoover、Evan Murray、Anitha Vijayan、Fei Chen、Sushrut S. Waikar、Sylvia Rosas、Francis P. Wilson、Paul M. Palevsky、Krzysztof Kiryluk、John R. Sedor、Robert D. Toto、Chirag Parikh、Eric H. Kim、Evan Z. Macosko、Peter V. Kharchenko、Joseph P. Gaut、Jeffrey B. Hodgin、Michael T. Eadon、Pierre C. Dagher、Tarek M. El-Achkar、Kun Zhang、Matthias Kretzler 和 Sanjay Jain。人类肾脏中健康与受损细胞状态及生态位图谱。2021年7月。

Yann Lecun. 迈向自主机器智能之路. 2022。

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。*bioRxiv*，2025年。

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。《自然·方法》，15(12):1053–1058，2018年11月。

Malte D Luecken 和 Fabian J Theis。单细胞RNA测序分析的当前最佳实践：教程。《分子系统生物学》，15(6):e8746，2019年6月。

Malte D. Luecken、M. Büttner、K. Chaichoompu、A. Danese、M. Interlandi、M. F. Mueller、D. C. Strobl、L. Zappia、M. Dugas、M. 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估。《自然·方法》，19(1):41–50，2022年1月。

Yoshiharu Muto、Parker C. Wilson、Nicolas Ledru、Haojia Wu、Henrik Dimke、Sushrut S. Waikar和BenjaminD. Humphreys。单细胞转录组与染色质可及性分析重新定义成年人体肾脏中的细胞异质性。《自然·通讯》，12(1)，2021年4月。

Thomas M. Norman、Max A. Horlbeck、Joseph M. Replogle、Alex Y. Ge、Albert Xu、Marco Jost、Luke A. Gilbert和Jonathan S. Weissman。探索由丰富单细胞表型构建的遗传相互作用流形。《科学》，365(6455):786–793，2019年。

James D Pearce、Sara E Simmonds、Gita Mahmoudabadi、Lakshmi Krishnan、Giovanni Palla、Ana-Maria Istrate、Alexander Tarashansky、Benjamin Nelson、Omar Valenzuela、Donghui Li、Stephen R Quake和Theofanis Karaletsos。跨越15亿年进化的跨物种生成式细胞图谱：TranscriptFormer单细胞模型。*bioRxiv*，2025年。

CZI细胞科学计划、Shibla Abdulla、Brian Aevermann、Pedro Assis、Seve Badajoz、Sidney M Bell、Emanuele Bezzi、Batuhan Cakir、Jim Chaffer、Signe Chambers、J Michael Cherry、Tiffany Chi、Jennifer Chien、Leah Dorman、Pablo Garcia-Nieto、Nayib Gloria、Mim Hastie、Daniel Hegeman、Jason Hilton、Timmy Huang、Amanda Infeld、Ana-Maria Istrate、Ivana Jelic、Kuni Katsuya、Yang Joon Kim、Karen Liang、Mike Lin、Maximilian Lombardo、Bailey Marshall、Bruce Martin、Fran McDade、Colin Megill、Nikhil Patel、Alexander Predeus、Brian Raymor、Behnam Robatmili、Dave Rogers、Erica Rutherford、Dana Sadgat、Andrew Shin、Corinn Small、Trent Smith、Prathap Sridharan、Alexander Tarashansky、Norbert Tavares、Harley Thomas、Andrew Tolopko、Meghan Urisko、Joyce Yan、Garabet Yeretssian、Jennifer Zamanian、Arathi Mani、Jonah Cool和Ambrose Carr。CZ Cellxgene Discover：用于聚合数据可扩展探索、分析与建模的单细胞数据平台。《核酸研究》，53(D1):D886–D900，2024年11月。

Yusuf Roohani、Kexin Huang 和 Jure Leskovec。《利用 GEARS 预测新型多基因扰动的转录结果》。《自然·生物技术》，2023。

Yanay Rosen、Yusuf Roohani、Ayush Agrawal、Leon Samotorcan、Tabula Sapiens Consortium、Stephen R. Quake 和 Jure Leskovec。《通用细胞嵌入：细胞生物学基础模型》。2023年11月。

Benjamin J. Stewart、John R. Ferdinand、Matthew D. Young、Thomas J. Mitchell、Kevin W. Loudon、Alexandra M. Riding、Nathan Richoz、Gordon L. Frazer、Joy U. L. Staniforth、Felipe A. Vieira Braga、Rachel A. Botting、Dorin-Mirel Popescu、Roser Vento-Tormo、Emily Stephenson、Alex Cagan、Sarah J. Farndon、Krzysztof Polanski、Mirjana Efremova、Kile Green、Martin Del Castillo Velasco-Herrera、Charlotte Guzzo、Grace Collord、Lira Mamanova、Tevita Aho、James N. Armitage、Antony C. P. Riddick、Imran Mushtaq、Stephen Farrell、Dyanne Rampling、James Nicholson、Andrew Filby、Johanna Burge、Steven Lisgo、Susan Lindsay、Marc Bajenoff、Anne Y. Warren、Grant D. Stewart、Neil Sebire、Nicholas Coleman、MuzlifahHaniffa、Sarah A. Teichmann、Sam Behjati 和 Menna R.Clatworthy。《人类肾脏的时空免疫分区》。《科学》，365(6460):1461–1466，2019年9月。

Chenqu Suo、Emma Dann、Issac Goh、Laura Jardine、Vitalii Kleshchevnikov、Jong-Eun Park、Rachel A. Botting、Emily Stephenson、Justin Engelbert、Zewen KelvinTuong、KrzysztofPolanski、NadavYayon,

Chuan Xu、Ondrej Suchanek、Rasa Elmentaite、Cecilia Domínguez Conde、Peng He、Sophie Pritchard、Mohi Miah、Corina Moldovan、Alexander S. Steemers、Pavel Mazin、Martin Prete、Dave Horsfall、JohnC. Marioni、Menna R. Clatworthy、Muzlifah Haniffa 和 Sarah A. Teichmann。《跨器官绘制发育中的人类免疫系统图谱》。《科学》，376(6597)，2022年6月。

Valentine Svensson、Roser Vento-Tormo 和 Sarah A. Teichmann。过去十年单细胞RNA测序的指数级增长。《自然·实验方法》，13(4):599–604，2018年4月。

Fuchou Tang、Kaiqin Lao 和 M. Azim Surani。单细胞转录组分析的发展与应用。《自然·方法》，8(4):S6–S11，2011年4月。

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。迁移学习助力网络生物学预测。《自然》，618(7965):616–624，2023年5月。

Ramon Viñas Torné、Maciej Wiatrak、Zoe Piran、Shuyang Fan、Liangze Jiang、Sarah A Teichmann、Mor Nitzan 和 Maria Brbić。Systema：一个用于评估超越系统性变异的遗传扰动响应预测的框架。《自然·生物技术》，第1–10页，2025年。

Parker C. Wilson、Yoshiharu Muto、Haojia Wu、Anil Karihaloo、Sushrut S. Waikar 和 Benja min D. Humphreys。多模态单细胞测序表明，染色质可及性和遗传背景与糖尿病肾病进展有关。《自然·通讯》，13(1)，2022年9月。

Weimin Wu、Xuefeng Song、Yibo Wen、Qinjie Lin、Zhihan Zhou、Jerry Yao-Chieh Hu、Zhong Wang 和 Han Liu。Genome-factory：用于调优、部署和解读基因组模型的集成式库。*arXiv* 预印本 *arXiv:2509.12266*，2025年。

Fan Yang、Wenchuan Wang、Fang Wang、Yuan Fang、Duyu Tang、Junzhou Huang、Hui Lu 和 Jianhua Yao。scBERT：一种用于单细胞RNA测序数据细胞类型注释的大规模预训练深度语言模型。《自然·机器智能》，4(10):852–866，2022年9月。

Xiaodong Yang、Guole Liu、Guihai Feng、Dechao Bu、Pengfei Wang、Jie Jiang、Shubai Chen、Qinmeng Yang、Hefan Miao、Yiyang Zhang、Zhenpeng Man、Zhongming Liang、Zichen Wang、Yaning Li、Zheng Li、Yana Liu、Yao Tian、Wenhao Liu、Cong Li、Ao Li、Jingxi Dong、Zhilong Hu、Chen Fang、Lina Cui、Zixu Deng、Haiping Jiang、Wentao Cui、Jiahao Zhang、Zhaohui Yang、Handong Li、Xingjian He、Liqun Zhong、Jiaheng Zhou、Zijian Wang、Qingqing Long、Ping Xu、Xin Li、Hongmei Wang、Baoyang Hu、Wei Li、Fei Gao、JingtaoGuo、Leqian Yu、Qi Gu、Weiwei Zhai、Zhengting Zou、Guihai Feng、Wenhao Liu、Yao Tian、Chen Fang、Jingxi Dong、Yana Liu、Jingqi Yu、Wenhui Wu、Xinxin Lin、Cong Li、Yu Zou、Yongshun Ren、Fan Li、Yixiao Zhao、Yike Xin、Longfei Han、Shuyang Jiang、Kai Ma、Qicheng Chen、Haoyuan Wang、Huanhuan Wu、Chaofan He、Yilong Hu、Shuyu Guo、Yiyun Li、Yuanchun Zhou、Yangang Wang、Xuezhi Wang、Pengfei Wang、Fei Li、Zhen Meng、Zaitian Wang、Ping Xu、Wentao Cui、Zhilong Hu、Huimin He、Shan Zong、Jiajia Wang、Yan Chen、Chunyang Zhang、Chengrui Wang、Ran Zhang、Meng Xiao、Yining Wang、Yiqiang Chen、Yi Zhao、Xiaodong Yang、Dechao Bu、Xin Qin、Jiaxin Qin、Zhaohui Yang、Chenhao Li、Zhufeng Xu、Zeyuan Zhang、Xiaoning Qi、Shubai Chen、Wuliang Huang、Yaning Li、Ge Yang、Jing Liu、Guole Liu、Liqun Zhong、Yaoru Luo、Jiaheng Zhou、Zichen Wang、Qinxuan Luo、Ziwen Liu、Ao Li、Teng Wang、Yiming Huang、Handong Li、Yong Wang、Shihua Zhang、Jiahao Zhang、Yiyang Zhang、Shirui Li、Zhongming Liang、Zhenpeng Man、Kangning Dong、Qunlun Shen、Hongmei Wang、Zhen Meng、Xuezhi Wang、Yangang Wang、Yong Wang、Shihua Zhang、Jingtao Guo、Yi Zhao、

Yuanchun Zhou、Fei Li、Jing Liu、Yiqiang Chen、Ge Yang和Xin Li。GeneCompass：利用知识引导的跨物种基础模型解析通用基因调控机制。《细胞研究》，34(12):830–845，2024年10月。

Matthew D. Young、Thomas J. Mitchell、Felipe A. Vieira Braga、Maxine G. B. Tran、Benjamin J. Stewart、John R. Ferdinand、Grace Collord、Rachel A. Botting、Dorin-Mirel Popescu、Kevin W. Loudon、Roser Vento-Tormo、Emily Stephenson、Alex Cagan、Sarah J. Farndon、Martin Del Castillo Velasco-Herrera、Charlotte Guzzo、NathanRichoz、Lira Mamanova、TevitaAho、JamesN.Armitage、AntonyC.P.Riddick、Imran Mushtaq、Stephen Farrell、Dyanne Rampling、James Nicholson、Andrew Filby、Johanna Burge、Steven Lisgo、Patrick H. Maxwell、Susan Lindsay、Anne Y. Warren、Grant D. Stewart、Neil Sebire、Nicholas Coleman、Muzlifah Haniffa、Sarah A. Teichmann、Menna Clatworthy 和 Sam Behjati。来自人类肾脏的单细胞转录组揭示了肾肿瘤的细胞身份。《科学》，361(6402)：594–599，2018年8月。

Matthew D. Young、Thomas J. Mitchell、Lars Custers、Thanasis Margaritis、Francisco Morales-Rodriguez、Kwasi Kwakwa、Eleonora Khabirova、Gerda Kildisiute、Thomas R. W. Oliver、Ronald R. de Krijger、Marry M. van den Heuvel-Eibrink、Federico Comitani、Alice Piapi、Eva Bugallo-Blanco、Christine Thevanesan、Christina Burke、Elena Prigmore、Kirsty Ambridge、Kenny Roberts、Felipe A. Vieira Braga、Tim H. H. Coorens、Ignacio Del Valle、Anna Wilbrey-Clark、Lira Mamanova、Grant D. Stewart、Vincent J. Gnanapragasam、Dyanne Rampling、Neil Sebire、Nicholas Coleman、Liz Hook、Anne Warren、Muzlifah Haniffa、Marcel Kool、Stefan M. Pfister、John C. Achermann、何晓玲、Roger A. Barker、Adam Shlien、Omer A. Bayraktar、Sarah A. Teichmann、Frank C. Holstege、Kerstin B. Meyer、Jarno Drost、Karin Straathof 和 Sam Behjati。人类肾肿瘤中的单细胞来源的mRNA信号。《自然·通讯》，12(1)，2021年6月。

张玉萍、Sathiya P. Narayanan、Rahul Mannan、Gregory Raskind、王晓明、Pankaj Vats、苏凤云、Noshad Hosseini、曹旭红、Chandan Kumar-Sinha、Stephanie J. Ellison、Thomas J. Giordano、Todd M. Morgan、Sethuramasundaram Pitchiaya、Ajjai Alva、Rohit Mehra、Marcin Cieslik、Saravana M. Dhanasekaran 和 Arul M. Chinnaiyan。单细胞分析肾细胞癌，揭示了关于肿瘤微环境、起源细胞和治疗反应的见解。《美国国家科学院院刊》，118(24)，2021年6月。

Zhihan Zhou、Weimin Wu、Harrison Ho、Jiayi Wang、Lizhen Shi、Ramana V Davuluri、Zhong Wang 和 Han Liu。Dnabert-s：利用基因组基础模型学习物种感知的 DNA 嵌入。*arXiv*预印本 *arXiv:2402.08777*，10，2024。

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber- Gordon, Mutian Yu, Harrison Ho, Fengchen Liu 等。Genomeocean：一种基于大规模宏基因组组装序列训练的高效基因组基础模型。*bioRxiv*，第 2025–01 页，2025年。

附录

A 局限性 20B 相关工作 21CscGPT
22C.1 词元化与输入表示 .22C.2 通过掩码注意力实现自回归基因生成 .22C.3 预训练目标 .23C.4 迭代自回归推理.23C.5 scGPT 中 ECS 损失的问题.23

D 评估指标 **24**
D.1 平均生物学保守性（AvgBIO）.24D.2 扰动预测评估指标.24

E 训练方案 **26**
E.1 预训练方案.26E.2 PBMC-10K 微调方案.26E.3 扰动微调方案.26

A 局限性

尽管 Cell-JEPA 展现出诸多改进，但仍存在五项局限性。

- **预训练数据集有限**。 受计算资源限制，我们在肾脏特异性数据集上对模型进行预训练，而未使用Cui 等人 [2024], 所采用的完整人类单细胞图谱；该图谱包含约 3300 万个细胞。从生物学角度来看，Cell-JEPA 接触到的细胞类型多样性有限，因此其泛化到其他组织、稀有细胞群体和疾病状态的能力尚未得到系统评估。为确保在这一限制下与 scGPT 基线进行公平比较，我们采用相同的超参数和小批量大小，因此无法针对 Cell-JEPA 进行特定任务的超参数优化。

- **模型规模**。 我们将模型规模限制为 2800 万个参数（单个学生模型或教师模型），这可能不足以充分利用规模显著更大、异质性更高的预训练数据集。将 Cell-JEPA 扩展至更大的模型容量并开展大规模基础模型预训练，将留待未来工作完成。

- **跨模型家族的评估范围有限**。 我们的实验评估仅将 JEPA 训练框架应用于 scGPT，并在这一固定骨干网络下比较使用与不使用 JEPA 时的性能。我们并未将 JEPA 目标扩展到其他单细胞基础模型架构，评估其在其他训练流程中的通用性仍是未来探索的重要方向。

- **基于变化量的评估指标**。 尽管我们的方法改善了扰动后状态的绝对预测，但在基于变化量的评估指标上的表现仍然有限。这表明，相对于对照条件准确建模效应量变化仍存在持续挑战。

- **残余技术与生物学混杂因素**。 尽管潜在空间 JEPA 目标具有稳健性，学得的嵌入仍可能反映残余批次效应、供体特异性变异以及测序深度差异，这些因素可能影响下游生物学解释。

• **扰动评估范围**。 我们的扰动实验使用来自单一细胞系（K562）的两个数据集，评估模型在固定生物学背景下对留出扰动的泛化能力。我们未评估跨背景迁移（例如跨细胞类型、供体或激活状态的迁移）；近期研究表明，这才是扰动预测面临的主要挑战 [Ahlmann-Eltze 等人，2025]。我们也未与均值模型或加性模型等简单基线进行比较，而这些基线在系统评估中已被证明具有竞争力 [Viñas Torné 等人， 2025]。因此，我们的扰动结果应被视为初步探索，而非对扰动预测问题的解决方案。

B 相关工作

本节回顾单细胞基础模型的发展，以及联合嵌入预测架构在其他领域的应用。

单细胞基础模型。大型基础模型在语言和生物学等多个领域的复杂数据建模方面展现出巨大潜力 [Zhou 等人，2025年，Wu 等人，2025年，Zhou 等人，2024，Brown 等人， 2020]。将大型基础模型应用于转录组数据，重塑了单细胞生物学中的表征学习。早期的深度概率方法（如 scVI [Lopez et al.，2018] ）引入了采用编码器–解码器架构的变分推断，用于建模单细胞 RNA 测序计数，并学习有助于处理批次效应和文库大小变异的潜在表征。在此基础上，基于 Transformer 的方法将每个细胞视为由基因令牌组成的集合（通常排列成伪序列），并采用自监督目标来学习上下文感知表征。在这一范式下，scBERT [Yang et al.,2022] 和 Geneformer [Theodoris et al.，2023] 等掩码建模方法通过预训练 Transformer 编码器，根据细胞上下文恢复被掩码基因；而 scGPT [Cui 等人， 2024] 和 TranscriptFormer [Pearce et al.，2025] 等生成式方法则通过结构化条件方案预测缺失值，从而学习对基因表达进行建模。近期研究着重于提升模型容量，并融入跨物种知识。例如，scFoundation [Hao et al.，2024] 利用 xTrimoGene 将预训练规模扩展至 5000 多万人类细胞 [Hao et al.，2024], ；而 GeneCompass [Yang et al.，2024] 则提出了一个利用知识、在超过 1.2 亿个转录组上训练的跨物种框架 [Yang et al.，2024]。类似地，通用细胞嵌入（UCE） [Rosen et al., 2023] 利用蛋白质语言模型为细胞构建统一潜在空间，使其能够泛化至不同组织和物种。

一个尚待解决的挑战是，当前许多预训练目标都与观测测量空间中的基因级重建密切相关——例如预测掩码计数、掩码基因标识或相关的逐项目标。由于 scRNA-seq 测量数据稀疏，且受到技术变异（例如捕获效率和抽样噪声）的强烈影响，优化此类目标可能会促使表征除了保留生物学信息外，也保留干扰变异 [Lähnemann 等人， 2020]。这促使我们探索互补目标，直接在表征空间中施加不变性。相比之下，我们的方法通过潜在空间预测学习单细胞表征，旨在降低对表层测量噪声的敏感性，同时更好地捕获稳定的高层级细胞程序。

联合嵌入预测架构。 联合嵌入架构（JEA）为自监督表征学习方法提供了一个统一框架。在 JEPA 中，模型会将同一数据点的多个视图编码并投影到共享的潜在空间中，其训练目标则鼓励不同视图之间保持一致。此类方法面临的核心挑战之一是表征坍塌，即编码器

无论输入如何，最终都生成完全相同的表征。 **联**合嵌入**预**测**架**构（联合嵌入预测架构（**JEPA**） [Lecun， 2022]通过将学习信号从表示匹配转向潜在空间预测，克服了直接对齐目标的局限。JEPA 不会强制要求不同视图的嵌入彼此相似，而是训练一个预测器，将上下文视图的表征映射到目标视图的表征。在常见的实现中，独立的目标编码器会生成目标表示，并通过对上下文编码器权重进行指数移动平均来更新；预测损失则完全在表征空间而非输入空间中计算。这种预测式方法无需负样本，也避免了对观测令牌或数值进行显式重建，否则这可能会使学习偏向低层细节。JEPA 已在多个领域取得成功，包括用于图像的 I-JEPA [Assran 等人， 2023] 用于视频的 V-JEPA [Assran 等人， 2025] 以及用于语言建模的 LLM-JEPA [Huang 等人， 2025] 。它展现出优异的性能和更强的可扩展性。JEPA 专注于预测嵌入，而非重建原始测量值，从而促进有意义且具有组合性的特征学习。

在我们的设定中，基因表达值反映了复杂的调控程序和细胞状态特异性信号。有效的建模需要表征同时捕捉基因层面变异和细胞层面语义。GeneJEPA [Litman 等人， 2025] 是近期一种用于转录组学的 JEPA 式基础模型，它将潜在空间预测应用于掩码基因子集，从而在各类下游任务中实现了良好的迁移能力。尽管在理念上与我们超越纯粹基于重建的预训练这一目标一致，我们的方法在一个关键方面有所不同：Cell-JEPA 将 JEPA 表示层面目标与显式的基因层面重建损失相结合。我们发现，保留重建锚点有助于保留细粒度的基因表达信息并稳定优化，而 JEPA 式预测则促使学习到的表征具备不变性和更高层次的结构。这种混合目标使 Cell-JEPA 能够学习到既能抵御测量噪声、又能为下游细胞预测任务提供信息的细胞嵌入。

C scGPT

与语言不同，基因表达向量本质上是无序的（基因并不构成自然序列），因此 scGPT [Cui 等人， 2024] 引入了一种专门的注意力掩码和迭代生成方案，使模型按基因集合进行自回归生成，而非依照固定的基因顺序。

C.1 词元化与输入表示

每个细胞表示为一组基因令牌及其对应的表达值。Transformer 的输入通常包括：(i) 一个专用的 <cls> 词元，用于生成全局细胞嵌入；(ii) 一部分已知基因及其观测到的表达值；以及 (iii) 一部分未知 基因，其表达值经过掩码处理，必须进行预测。基因标识通过可学习的基因嵌入表进行嵌入，而表达值则通过数值嵌入模块进行嵌入（还可选择添加批次或扰动标签等条件令牌）。

C.2 通过掩码注意力进行自回归基因生成

为实现不依赖任意基因排序的生成，scGPT 使用结构化注意力掩码来强制执行已知 $\rightarrow$ 未知依赖关系。预测未知基因时，其词元可以关注所有已知基因（以及自身），但不能关注其他未知基因。令 $\mathcal{U}_{\text{unk}}$ 表示

当前步骤中未知基因的索引。注意力掩码定义如下：

$$a_{i,j} = \begin{cases} 0, & \text{if } j \notin \mathcal{U}_{\text{unk}}, \\ 0, & \text{if } i = j \text{ and } j \in \mathcal{U}_{\text{unk}}, \\ -\infty, & \text{if } i \neq j \text{ and } j \in \mathcal{U}_{\text{unk}}. \end{cases} \tag{C.1}$$

这可以防止同时被掩码的目标之间发生信息泄漏，同时允许每个未知基因以完整的已知上下文为条件。

C.3 预训练目标

给定 Transformer 的最终隐藏状态 $\{h_j\}$，scGPT 通过多层感知机头预测未知基因的表达值，并针对掩码位置优化均方误差：

$$\mathcal{L}_{\text{GEP}} = \frac{1}{|\mathcal{U}_{\text{unk}}|} \sum_{j \in \mathcal{U}_{\text{unk}}} \left(\text{MLP}(h_j) - v_j\right)^2,$$

其中，v_j 是基因 j 的真实表达值。

除了 基因提示 预测（根据观测到的基因预测被掩码基因）之外，scGPT 还描述了一种细胞提示生成模式，其中全局细胞嵌入（来自 `<cls>`）被用作提示，以生成其余基因表达值 [Cui 等人，2024]。在实践中，这些模式可以按顺序应用，并在预训练期间合并其损失。

C.4 迭代自回归推理

在推理时，scGPT 会以 K 迭代轮次的方式进行生成，从而在基因子集上执行自回归过程，而不是遵循固定的基因顺序。在每次迭代 t中，模型会使用（C.1）中的掩码，为剩余的未知基因计算预测结果。随后，模型选取置信度最高的部分预测结果，并将其纳入已知集合，供下一轮迭代使用。具体而言，scGPT 采用基于置信度的扩展策略，每轮添加置信度最高的约 $1/K$ 比例的基因，直到生成所有基因 [Cui 等人，2024]。这形成了一个自回归细化过程：高置信度的基因会先生成，随后作为上下文，帮助模型在后续轮次中预测较难预测的基因。

C.5 scGPT 中 ECS 损失的问题

原始的 scGPT 框架 [Cui 等人，2024] 采用了 ECS 损失的另一种表述形式：

$$\text{ECS}_{\text{scGPT}} = \left(\cos\left(\mathbf{e}^i, \mathbf{e}^j\right) - \beta\right)^2,$$

其中，$\mathbf{e}^i$ 和$\mathbf{e}^j$ 分别表示细胞 i 和 j的细胞嵌入，β 是一个固定的相似度阈值，设为 0.3。尽管这种表述明确地将成对余弦相似度约束到预设目标，但它可能会仅凭手动设定的超参数，人为地将不同的细胞状态推得更近或更远，从而意外扭曲具有生物学意义的关系。相比之下，我们的方法无需依赖此类固定阈值，因此能够在学习得到的表征空间中更稳健地保留内在的生物结构。

D 评估指标

我们在此提供详细的评估指标计算方法。

D.1 平均生物学保守性（AvgBIO）

我们采用 [Cui 等人， 2024] 提出的 AvgBIO 分数来评估单细胞嵌入的质量。AvgBIO 分数是 [Luecken 等人 提出的三种细胞类型聚类指标的平均值， 2022],包括归一化互信息（$\mathrm{NMI}_{\mathrm{cell}}$）、调整兰德指数（$\mathrm{ARI}_{\mathrm{cell}}$）和平均轮廓宽度（$\mathrm{ASW}_{\mathrm{cell}}$）。具体而言，AvgBIO 分数定义如下：

$$\mathrm{AvgBIO} = \frac{\mathrm{NMI}_{\mathrm{cell}} + \mathrm{ARI}_{\mathrm{cell}} + \mathrm{ASW}_{\mathrm{cell}}}{3}$$

归一化互信息 $\mathrm{NMI}_{\mathrm{cell}}$。 为了量化真实细胞类型注释与根据整合后的细胞嵌入得出的聚类结果之间的一致性，我们计算归一化互信息（NMI）分数。我们在 0.1 到 2.0 的分辨率范围内，以 0.1 为步长进行 Louvain 聚类，并报告所有分辨率下的最高 NMI 分数。该指标称为 $\mathrm{NMI}_{\mathrm{cell}}$，用于衡量预测的簇分配与真实细胞类型标签之间的共享信息。$\mathrm{NMI}_{\mathrm{cell}}$ 的取值范围为 0 到 1，数值越高，表示簇与注释细胞类型之间的对应关系越强。

调整兰德指数$\mathrm{ARI}_{\mathrm{cell}}$。除 NMI 外，我们还使用调整兰德指数（ARI）来评估经过优化以实现 NMI 的 Louvain 簇与标注细胞类型标签之间的一致性。与标准兰德指数不同，ARI 会校正偶然产生的一致性，因此能更稳健地衡量聚类准确率。细胞类型的 ARI 得分记为 $\mathrm{ARI}_{\mathrm{cell}}$，范围为 0 到 1，其中 0 表示标签一致性相当于随机结果，1 则表示预测簇与真实标注完全匹配。

平均轮廓宽度$\mathrm{ASW}_{\mathrm{cell}}$。为评估整合嵌入空间中细胞类型的分离质量，我们计算平均轮廓宽度（ASW）。每个细胞的轮廓宽度衡量其簇内距离与到最近相邻簇距离之间的相对差异，从而同时反映簇紧密度和分离程度。对所有细胞的轮廓宽度取平均后，即可得到细胞类型的 ASW 得分，记为 $\mathrm{ASW}_{\mathrm{cell}}$。该得分范围为 -1 到 1，数值越高表示簇分离越好且内部越紧密；接近或低于零的数值则表明簇之间存在重叠或可能发生了错误分类。

D.2 扰动预测的评估指标

遵循 GEARS [Roohani 等人， 2023], 我们通过使用皮尔逊相关系数比较预测与观测到的扰动后基因表达谱，来评估扰动响应预测。对于评估集中的每个扰动 p，设 $x^{(p)} \in \mathbb{R}^G$ 表示观测到的扰动后基因表达向量，涵盖 G个基因。设$\hat{}x^{(p)} \in \mathbb{R}^G$ 表示相应的模型预测。GEARS 评估套件会报告基于不同基因子集计算的多种皮尔逊相关系数指标。皮尔逊相关系数的计算范围包括：所有基因（全局状态匹配）、差异表达（DE）基因（扰动特异性效应），以及去除易发生脱落的基因后的排名前 20 的 DE 基因。

对于任意两个向量 $a,\ b \in \mathbb{R}^d$，皮尔逊相关系数为：

$$\rho(a,b) = \frac{\sum_{i=1}^{d}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{d}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{d}(b_i - \bar{b})^2}},$$

其中 $\bar{a}$ 和 b 为逐分量均值。除非另有说明，所报告的指标是通过对相应 GEARS 测试子集（即 unseen_single）中的各扰动取平均进行汇总的，即对 p求$\rho(\hat{x}^{(p)}, x^{(p)})$的平均值。

全基因相关 皮尔逊相关系数。 皮尔逊相关系数使用全部基因计算 $\rho(\hat{x}^{(p)}, x^{(p)})$：

$$\texttt{pearson} = \mathbb{E}_p\big[\rho(\hat{x}^{(p)}, x^{(p)})\big].$$

Th该指标强调与扰动后绝对转录st的一致性 ate.

限于差异表达基因的相关性 差异表达皮尔逊相关系数。 对于每个扰动 p，一组差异表达（DE）基因 $\mathcal{D}_p$ 通过比较扰动条件与对照条件而被确定 [Roohani 等人， 2023]。差异表达皮尔逊相关系数计算限于该差异表达基因集合的皮尔逊相关系数：

$$\texttt{pearson_de} = \mathbb{E}_p\big[\rho(\hat{x}^{(p)}_{\mathcal{D}_p}, x^{(p)}_{\mathcal{D}_p})\big],$$

其中$x_{\mathcal{D}_p}$ 表示 x 中由 $\mathcal{D}_p$索引的子向量。该指标将评估重点放在对扰动表现出转录响应的基因上。

经过非脱落过滤的排名前列差异表达基因 前20个非缺失差异表达基因。 GEARS 提供了一个经过过滤的排名最高的差异表达基因子集，并移除了易发生脱落的基因。令 $\mathcal{T}_p \subseteq \mathcal{D}_p$ 表示所得集合（通常为$|\mathcal{T}_p| = 20$），该集合由 GEARS 预处理提供 [Roohani 等人， 2023]。 前20个非缺失差异表达基因在该集合上计算相关性：

$$\texttt{top20_de_non_dropout} = \mathbb{E}_p\big[\rho(\hat{x}^{(p)}_{\mathcal{T}_p}, x^{(p)}_{\mathcal{T}_p})\big].$$

变化量变体（相对于对照的变化相关性）。GEARS 还会报告基于差值的指标，用于评估预测的相对于对照的变化。令 $x^{(\mathrm{ctrl})} \in \mathbb{R}^G$ 表示对照表达参照（由 GEARS 流程提供 [Roohani 等人， 2023]）。我们将扰动 p的观测变化量和预测变化量定义为：

$$\Delta x^{(p)} = x^{(p)} - x^{(\mathrm{ctrl})}, \qquad \Delta\hat{x}^{(p)} = \hat{x}^{(p)} - x^{(\mathrm{ctrl})}.$$

变化量指标计算与上文相同的相关性，但比较对象为 $\Delta\hat{x}^{(p)}$ 与 $\Delta x^{(p)}$ ：

$$\begin{aligned}
\texttt{pearson_delta} &= \mathbb{E}_p\big[\rho(\Delta\hat{x}^{(p)}, \Delta x^{(p)})\big],\\
\texttt{pearson_de_delta} &= \mathbb{E}_p\big[\rho(\Delta\hat{x}^{(p)}_{\mathcal{D}_p}, \Delta x^{(p)}_{\mathcal{D}_p})\big],\\
\texttt{delta_top20_de_non_dropout} &= \mathbb{E}_p\big[\rho(\Delta\hat{x}^{(p)}_{\mathcal{T}_p}, \Delta x^{(p)}_{\mathcal{T}_p})\big].
\end{aligned}$$

这些指标强调 delta（相对于对照组的变化）估计，并且对相对于对照组基线的效应量准确性十分敏感。

E 训练方案

我们在此提供预训练和微调的详细超参数设置。

E.1 预训练方案

预训练模型的嵌入维度为 512，由 12 个堆叠的 Transformer 块组成，每个块包含 8 个注意力头，隐藏维度为 512。嵌入对齐多层感知机由两个全连接层组成，中间采用 GELU 激活函数。我们使用 AdamW 优化模型，小批量大小为 32，词元掩码比例为 0.15，初始学习率为 1×10^{-4}。我们采用 2×10^{-4} 的权重衰减，并在每个训练轮次后将学习率乘以 0.9。我们在 Transformer 头和嵌入对齐多层感知机的每一层之间应用丢弃概率为 0.2 的丢弃。模型在肾脏数据集上训练，共进行 4 个训练轮次。为补偿数值尺度差异，我们将 JEPA 损失项的权重设为 1000，将重建损失项的权重设为 1。整个训练过程在单个 A5000 GPU 上耗时约一天。

我们谨向以下作者致以谢意，感谢他们的数据集用于训练我们的基础模型及撰写本文： [Kretzler 等人，2025年；Consortium* 等人，2022；Wilson 等人，2022；Suo 等人，2022；Young 等人，2021；Lake 等人，2021；Zhang 等人，2021；Muto 等人，2021；Han 等人，2020年；Cao 等人，2020年；Stewart 等人，2019；Young 等人， 2018]。

E.2 PBMC-10K微调方案

我们从一个在肾脏数据集上训练得到的预训练检查点开始微调模型，具体描述见第E.1节。我们使用40%的掩码比例，在PBMC-10K数据集上训练模型。我们沿用预训练时的JEPA损失权重，并将所有其他损失项的权重设为1。我们使用Adam 优化器优化模型，小批量大小为64 ，初始学习率为 1×10^{-4}，并在每个训练轮次后将学习率乘以0.9 。模型总共训练30 个训练轮次，微调过程在单张NVIDIA RTX A5000 GPU上约需2 小时。

E.3 扰动微调方案

我们使用 Adamson 和 Norman 数据集对模型进行扰动预测微调。模型初始化自一个在肾脏数据集上预训练的检查点。对于 JEPA 增强模型，我们使用随机权重初始化投影器和预测器头（如果有可用的检查点，则从检查点推断），并与学生编码器一同训练。我们使用 Adam 进行优化，学习率为 1×10^{-4}，批次大小为 64。学习率每个训练轮次衰减为原来的 0。9 倍。与预训练阶段不同，在进行扰动预测时，我们不对输入基因使用掩码（掩码比例设为 0。0）。对于 JEPA 配置，学生模型经过训练以预测冻结的教师模型（学生模型的副本）的表征。JEPA 损失权重设为 1。0（以防止分布外细胞系导致表征漂移），重建损失权重设为1。0。对比式 ECS 损失已启用，权重设为 0。8。训练进行15个训练轮次，在单个 NVIDIA RTX A100 GPU 上约需2小时。模型性能使用 GEARS 评估套件（第 D.2节）在留出测试集上进行评估。