

scGPT：使用生成式 AI 构建单细胞多组学基础模型

收稿日期：2023-07-12

Haotian Cui, Chloe Wang, Hassaan Maan
Fengning Luo, Nan Duan & Bo Wang

 ^{1,3,4}, 宽鹏,





录用日期：2024-01-30

在线发布：2024 年 2 月 26 日

 检查更新

生成式预训练模型在语言和计算机视觉等各个领域都取得了显著的成功。具体来说，大规模多样化数据集和预训练 transformer 的结合已成为开发基础模型的一种很有前途的方法。我们将语言与细胞生物学（其中文本由单词组成；同样，细胞由基因定义）之间进行比较，探讨了基础模型在推进细胞生物学和遗传研究方面的适用性。利用不断增长的单细胞测序数据，我们构建了一个单细胞生物学的基础模型 scGPT，该模型基于超过 3300 万个细胞存储库中的生成式预训练转换器。我们的研究结果表明，scGPT 有效地提炼了有关基因和细胞的关键生物学见解。通过进一步调整迁移学习，可以优化 scGPT 以在各种下游应用程序中实现卓越的性能。这包括细胞类型注释、多批次整合、多组学整合、扰动反应预测和基因网络推断等任务。

单细胞 RNA 测序 (scRNA-seq) 通过对不同细胞类型进行复杂的表征并促进我们对疾病发病机制的理解，为细胞异质性探索、谱系追踪、致病机制阐明以及最终的个性化治疗策略铺平了道路。scRNA-seq 的广泛应用催生了全面的数据图谱，例如人类细胞图谱，现在包含数千万个细胞。测序技术的最新进展促进了数据模式的多样性，并将我们的理解从基因组学扩展到表观遗传学、转录组学和蛋白质组学，从而提供了多模式的见解。这些突破也提出了新的研究问题，如参考映射、扰动预测和多组学整合。同时开发能够有效利用、增强和适应测序数据快速扩展的方法至关重要。

On the Self-attention Transformer Architecture 由于其在学习富有表现力的数据表示方面的有效性，是一类深度学习模型，这些模型在大规模、多样化的数据集上进行了预训练，可以很容易地适应各种下游任务。此类模型最近在各个领域取得了前所未有的成功，例如计算机视觉和自然语言生成 (NLG) 中的 DALL-E 2 和 GPT-4，以及最近用于生物应用的 Enformer。更有趣的是，这些生成式预训练模型的性能始终优于从头开始训练的特定任务模型。这表明对这些领域知识的理解与任务无关，激发我们探索将其用于单细胞组学研究。然而，目前单细胞研究中基于机器学习的方法相当分散，有专门的模型专门用于不同的分析任务。因此，每项研究中使用的数据集在广度和规模上通常受到限制。为了克服这一限制，需要一个基于大规模数据进行预训练的基础模型，并能够理解不同组织中基因之间的复杂相互作用。

解决这一挑战的一种有前途的方法是基础模型的生成式预训练。基础模型，通常构建

1 Peter Munk 心脏中心，大学健康网络，多伦多，安大略省，加拿大。多伦多大学计算机科学系，加拿大安大略省多伦多市。矢量研究所，加拿大安大略省多伦多。多伦多大学医学生物物理学系，加拿大安大略省多伦多市。

5 Microsoft Research，美国华盛顿州雷德蒙德。多伦多大学检验医学和病理生物学系，多伦多，安大略省，加拿大。AI Hub，大学健康网络，加拿大安大略省多伦多。这些作者的贡献相同：Haotian Cui、Chloe Wang。电子邮件：bowang@vectorinstitute.ai

为了增强大规模单细胞测序数据的建模，我们从 NLG 中的自我监督预训练工作流程中汲取灵感，其中自我注意力转换器展示了对单词输入标记进行建模的强大能力。虽然文本由单词组成，但细胞可以通过基因和它们编码的蛋白质产物来表征。通过同时学习基因和细胞嵌入，我们可以更好地理解细胞特征。此外，transformer 输入令牌的灵活特性使附加功能和元信息易于合并。最近在 Geneformer 中也探索了这个方向，其中基于 transformer 的编码器使用按表达水平排序的基因进行训练，并展示了细胞类型和基因功能预测的能力。除此之外，我们设想需要定制一个预训练工作流程，以直接对非序列组学数据的复杂性进行建模，并将其适用性扩展到更广泛的任务。

在这项工作中，我们通过对超过 3300 万个细胞进行预训练，提出了单细胞基础模型 scGPT。我们专门为非序列组学数据建立了统一的生成式预训练工作流程，并调整 transformer 架构以同时学习细胞和基因表示。此外，我们还提供了具有特定任务目标的微调管道，旨在促进预训练模型在一系列不同任务中的应用。

我们的模型 scGPT 通过三个关键方面展示了单细胞基础模型的变革潜力。首先，scGPT 代表了一种大规模的生成基础模型，支持跨各种下游任务的迁移学习。通过在细胞类型注释、遗传扰动预测、批量校正和多组学整合方面实现最先进的性能，我们展示了“普遍预训练，按需微调”方法作为单细胞组学计算应用的通用解决方案的有效性。其次，通过比较微调和原始预训练模型之间的基因嵌入和注意力权重，scGPT 揭示了对特定于各种条件（例如细胞类型和扰动状态）的基因-基因相互作用的有价值的生物学见解。第三，我们的观察揭示了扩展效应：更大的预训练数据大小会产生更好的预训练嵌入，并进一步提高下游任务的性能。这一发现突出了令人兴奋的前景，即基础模型可以随着研究界可用测序数据的扩展而不断改进。基于这些发现，我们设想采用预训练的基础模型将极大地扩展我们对细胞生物学的理解，并为未来的发现奠定坚实的基础。scGPT 模型和工作流程的发布旨在增强和加快这些领域及其他领域的研究。

适用于单细胞研究中各种基本任务的微调管道，包括与批量校正的 scRNA-seq 集成、细胞类型注释、多组学集成、扰动预测和基因调控网络（GRN）推断。为了收集多样化和广泛的测序数据以进行 scGPT 的自我监督预训练，我们从 CELLxGENE 集合 (<https://cellxgene.cziscience.com/>; 图 1d)。这个全面的数据集涵盖了来自 51 个器官或组织和 441 项研究的广泛细胞类型，提供了整个人体细胞异质性的丰富表示。预训练后，我们使用均匀流形近似和投影（UMAP）可视化（图 1e）可视化了 3300 万个细胞中 10% 的人类细胞上的 scGPT 细胞包埋。生成的 UMAP 图表现出有趣的清晰度，细胞类型在局部区域和聚类处由不同的颜色准确表示。考虑到数据集中包含 400 多项研究，这证明了预训练提取生物变异的非凡能力。

scGPT 提高了细胞类型注释的精度

为了微调预训练的 scGPT 以进行细胞类型注释，神经网络分类器将 scGPT 转换器输出细胞嵌入作为输入，并输出细胞类型的分类预测。整个模型在带有专家注释的参考数据集上使用交叉熵进行训练，然后用于预测保留的查询数据分区上的细胞类型。我们对不同的数据集进行了广泛的实验，以评估 scGPT 在细胞类型注释方面的性能。首先，我们调整了 scGPT 来预测人类胰腺数据集中的细胞类型。我们在图 2a 中可视化了预测。值得注意的是，scGPT 对混淆矩阵中显示的大多数细胞类型 (>0.8) 都实现了高精度 (0.8)，但参考分区中细胞数量极低的稀有细胞类型除外。例如，在参考集中的 10,600 个细胞中，只有不到 50 个细胞属于肥大和主要组织相容性 (MHC) II 类细胞类型。图 2c 可视化了微调 scGPT 中的细胞嵌入，它们表现出高度的细胞内类型相似性。

接下来，我们在多发性硬化症 (MS) 的疾病数据集上测试了该模型。该模型在健康人类免疫细胞的参考分区上进行了微调，并评估了对 MS 条件下细胞的预测。微调模型与原始研究提供的细胞类型注释高度一致，并实现了约 0.85 的高精度 (图 2f, g)。此外，我们将该模型应用于更具挑战性的场景，使用肿瘤浸润髓系数据集进行跨疾病类型的泛化。该模型在参考数据分区 (Methods) 中的六种癌症类型上进行了微调，并在三种看不见的癌症类型的查询分区上进行了评估 (图 2d)。结果表明，区分免疫细胞亚型的精度很高 (图 2e, h)，并且细胞包埋在不同细胞类型之间表现出明显的可区分性 (图 2i)。最后，我们在三个数据集 (方法) 中将微调的 scGPT 与另外两种最近的基于 transformer 的方法 TOSICA 和 scBERT 进行了基准测试。scGPT 在所有分类指标上都优于其他方法，包括准确率、精度、召回率和宏 F1 (图 2j)。

除了细胞类型分类之外，我们还进一步探索了 scGPT 通过参考映射将看不见的查询细胞投射到参考数据集的能力 (补充注释 1 和补充图 11)。我们发现，与现有方法相比，仅使用预训练权重的 scGPT 实现了有竞争力的性能。通过对参考数据集进行微调，可以进一步提高性能。

结果

单节电池变压器基础模型概述

单细胞测序能够在单个细胞水平上分析分子特征。例如，scRNA-seq 测量 RNA 转录本的丰度，提供对细胞身份、发育阶段和功能的见解。我们介绍了 scGPT，这是一种采用生成式预训练方法的单细胞领域基础模型。核心模型包含具有多头注意力的堆叠 transformer 层，可同时生成细胞和基因嵌入 (Methods)。

scGPT 包括两个训练阶段：对大型细胞图谱进行初始通用预训练，以及针对特定应用的较小数据集的后续微调 (图 1a-c)。在预训练阶段，我们引入了专门设计的注意力掩码和生成训练管道，以自我监督的方式训练 scGPT，以共同优化细胞和基因表征 (Methods)。该技术解决了基因表达的非顺序性质，以适应顺序预测的 NLG 框架。在训练过程中，模型逐渐学习根据细胞状态或基因表达线索生成细胞的基因表达。在微调阶段，预训练模型可以适应新的数据集和特定任务 (Methods)。我们提供灵活的

scGPT 预测看不见的遗传扰动反应

测序和基因编辑技术的最新进展极大地促进了大规模扰动实验，能够表征细胞对各种遗传扰动的反应。

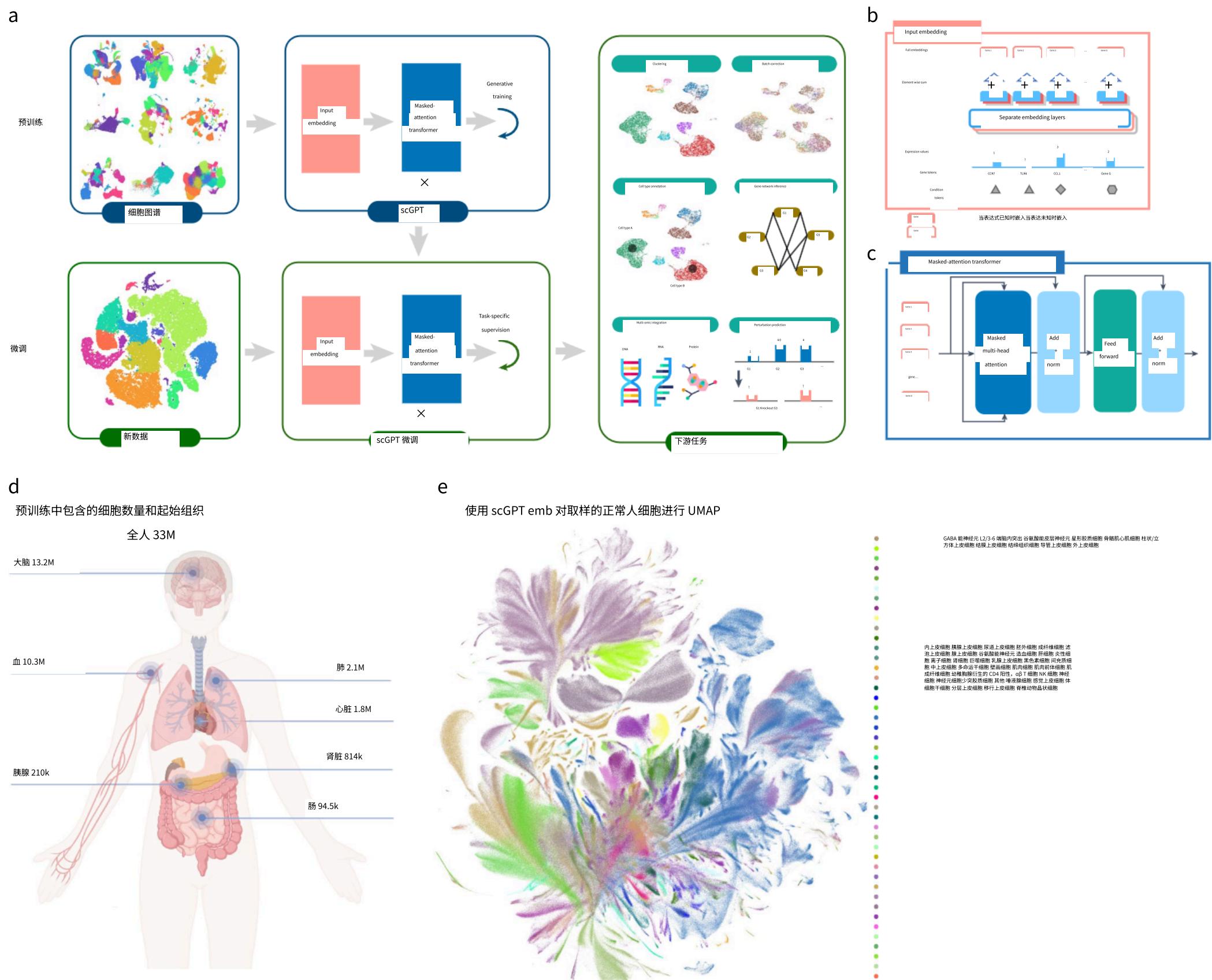


图 1 | 模型原理图。a, scGPT 的工作流程。该模型在来自细胞图谱的大规模 scRNA-seq 数据上进行创式预训练。scGPT 的核心组件包含堆叠的 transformer 块，带有用于生成训练的专用注意力掩码。对于下游应用程序，可以根据新数据微调预训练模型参数。我们将 scGPT 应用于各种任务，包括细胞类型注释、批量校正、多组学整合、遗传扰动预测和基因网络推断。b, 输入数据嵌入的详细视图。输入包含三层信息：基因标记、表达值和条件标记 (modality、batch、perturbation)

条件等)。c, scGPT transformer 层的详细视图。我们在掩码多头注意力块中引入了一个专门设计的注意力掩码，以对单细胞测序数据进行生成式预训练。Norm 表示层归一化。d, 说明训练数据大小和起源器官的图表。scGPT 全人体模型在 3300 万 (M) 正常人类细胞的 scRNA-seq 数据上进行了预训练。k, 千。e, 预训练的 scGPT 细胞嵌入的 UMAP 可视化 (emb; 随机的 10% 子集)，按主要细胞类型着色。GABA, γ -氨基丁酸。

这种方法在发现新的基因相互作用和推进再生医学方面具有巨大的前景。然而，潜在基因扰动的巨大组合空间很快就超过了实验可行性的实际极限。为了克服这一限制，scGPT 可用于利用从已知实验中的细胞反应中获得的知识，并推断它们以预测未知反应。在基因维度上利用自我注意机制可以编码受干扰基因与其他基因反应之间错综复杂的相互作用。通过利用此功能，scGPT 可以有效地从现有实验数据中学习，并准确预测看不见的扰动的基因表达响应。

我们使用白血病细胞系的三个 Perturb-seq 数据集评估了我们的模型：由 87 个单基因扰动组成的 Adamson 数据集，由 1,823 个单基因扰动组成的精选 Replogle 数据集和由 131 个双基因扰动和 105 个单基因扰动组成的 Norman 数据集。为了评估 scGPT 的扰动预测能力，我们在扰动子集上微调了模型，以预测给定输入对照细胞状态和干预基因的扰动表达谱。接下来，在涉及看不见的基因的扰动上测试了该模型（方法）。我们计算了 Pearson 指标，该指标衡量预测和观察到的扰动后表达变化之间的相关性。此外，我们报告了每个扰动变化前 20 个最显着的基因的该指标，在差异表达基因上表示为 Pearson。有关度量计算的详细信息，请参阅补充说明 12。我们进行了一场表演

预测看不见的基因扰动。对于扰动预测任务

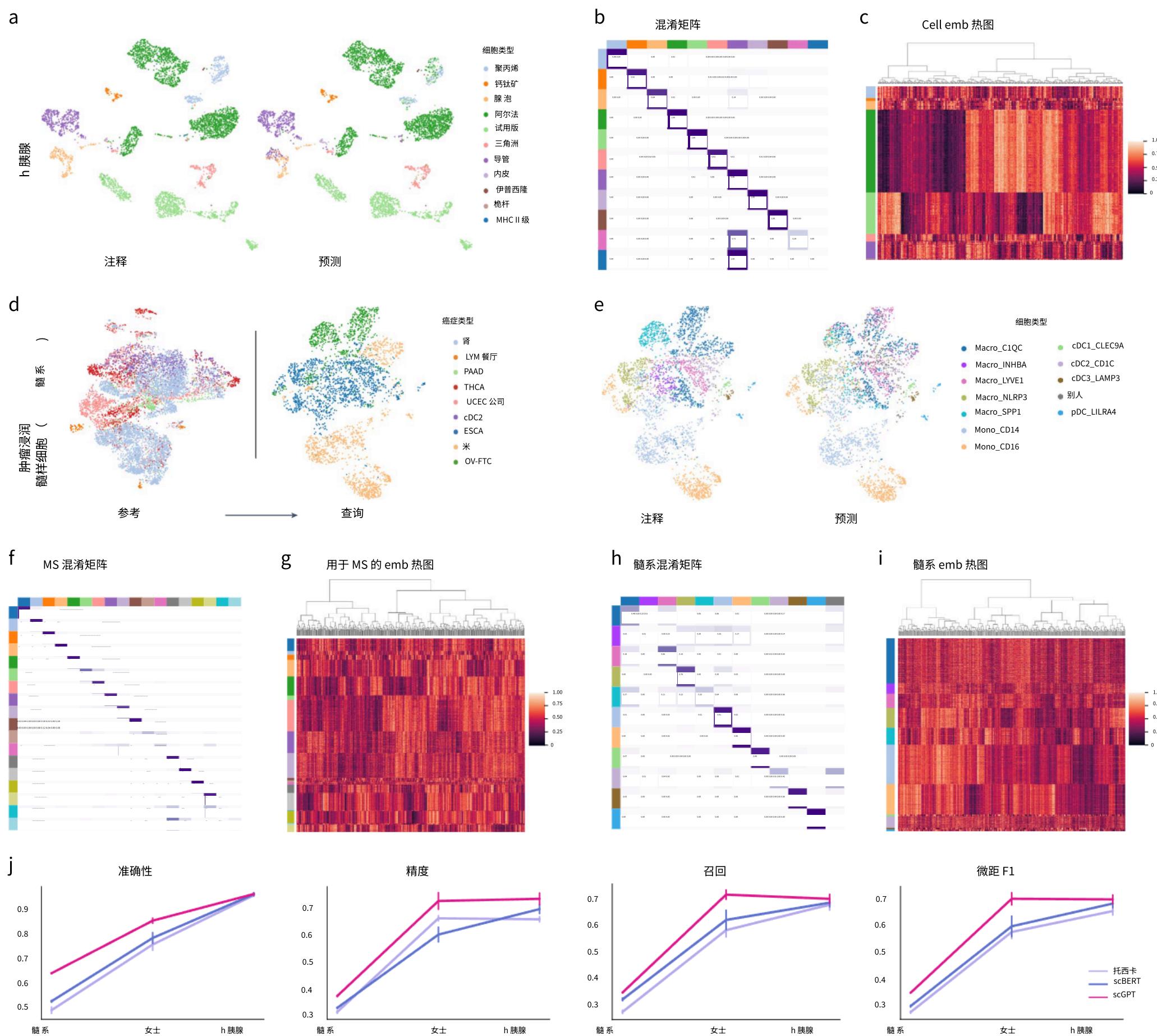


图 2 | 使用 scGPT 的细胞类型注释结果。a, 基因表达的 UMAP

来自人类胰腺数据集的细胞，按原始研究中注释的细胞类型（左）和微调的 scGPT 预测的细胞类型（右）着色。PP, 胰多肽细胞;PSC, 胰星状细胞。b, 人类胰腺数据集中预测和注释细胞类型之间的混淆矩阵。c, 人类胰腺数据集中 scGPT 的 512 维细胞嵌入热图。d, 骨髓数据集的 UMAP 可视化，按癌症类型着色。scGPT 在引用分区（左）上进行了微调，并在查询分区（右）上进行了评估。这两个数据分区包含不同的癌症类型。cDC2,2 型 (CD1ACD172A) 常规树突状细胞;ESCA, 食管癌;LYM, 淋巴瘤;米,

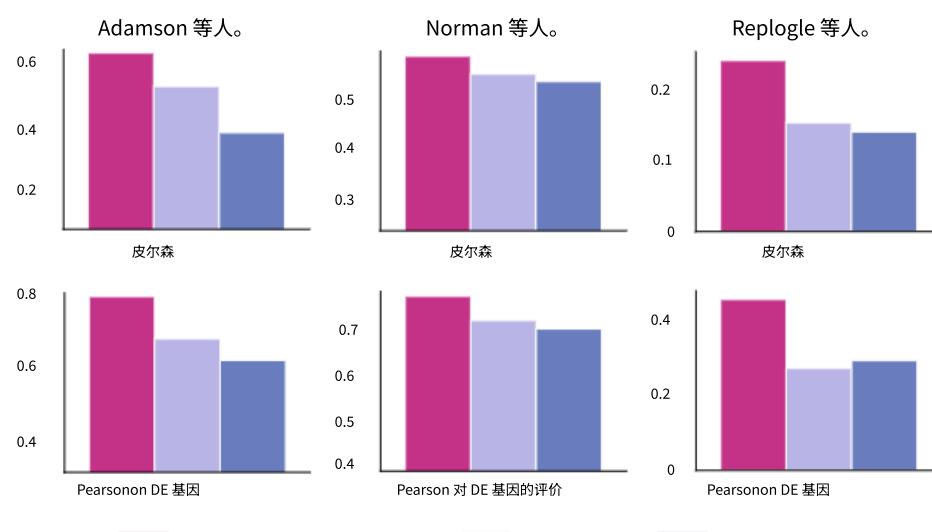
骨髓瘤;OV-FTC, 卵巢或滤泡性甲状腺癌;PAAD, 胰腺癌;THCA, 甲状腺癌;UCEC, 子宫体子宫内膜癌。e, 在查询分区上，UMAP 按原始研究中注释的细胞类型（左）和 scGPT 预测的细胞类型着色。f, h, 分别是 MS 和骨髓数据集的预测细胞类型与实际注释之间的混淆矩阵。g, i, 热图分别显示了 MS 和骨髓数据集中细胞的 scGPT 中的 512 维细胞包埋。

j, 通过在骨髓、MS 和人类胰腺数据集上通过 $n = 5$ 个随机训练验证拆分评估 scGPT 的细胞注释性能。测试集的性能指标以平均值表示 \pm s.e.m.

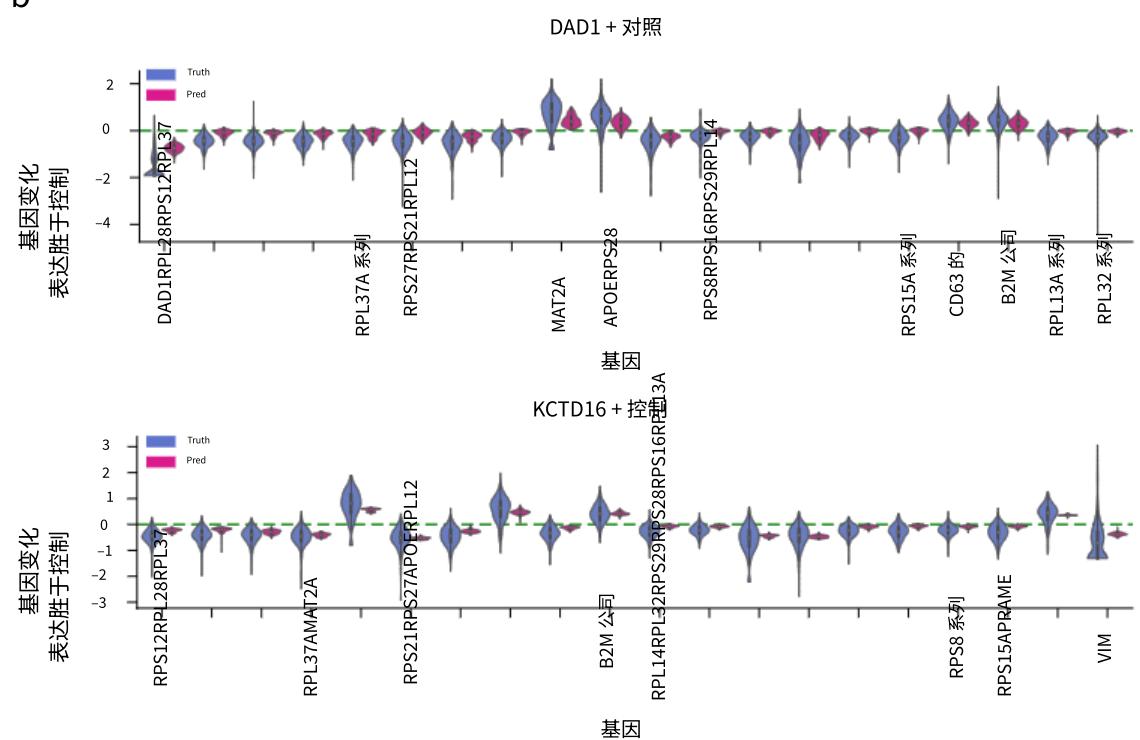
scGPT 与其他两种方法 (GEARS 和线性回归基线) 之间的比较 (方法)。我们的结果表明，scGPT 在所有三个数据集中都获得了最高分 (图 3a 和补充表 6)。特别是，scGPT 在预测扰动后的变化方面表现出色，始终比其他公司高出 5-20%。此外，我们在图 3b 的 Adamson 数据集中可视化了两个示例扰动的预测，其中 scGPT 准确预测了所有前 20 个差异表达基因的表达变化趋势。

预测看不见的扰动响应的能力可以扩大扰动实验的范围，如图 3c 所示。为了探索预测扰动响应的扩展空间，我们使用 Norman 数据集进行了聚类分析，以验证生物学相关的功能信号。最初的 Perturb-seq 研究涵盖了针对 105 个基因的 236 次扰动。然而，考虑到这些靶基因的所有可能组合，总共有 5,565 个潜在的扰动，表明实验 Perturb-seq 数据仅占整个扰动空间的 5%。因此，我们应用了

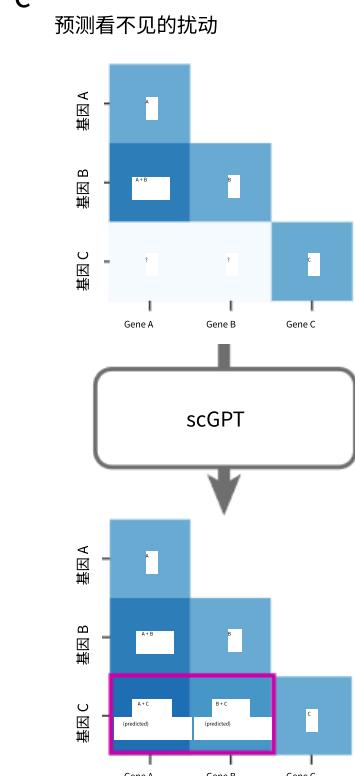
a



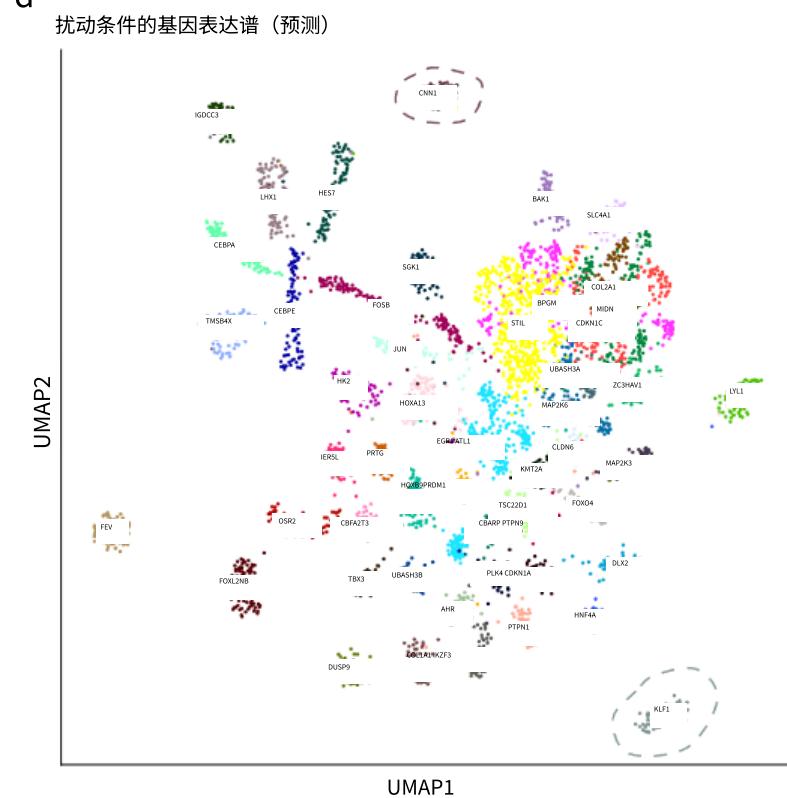
b



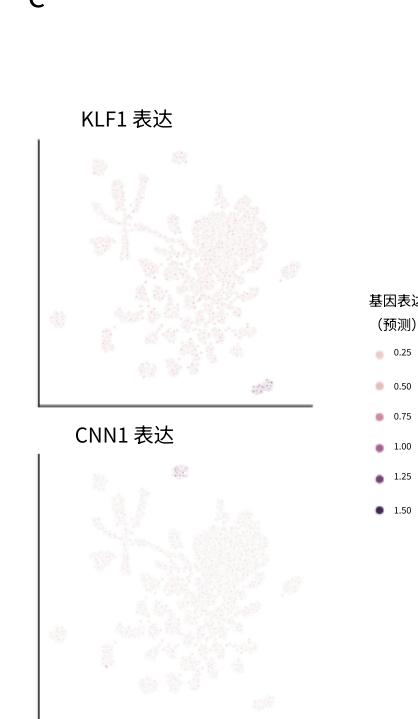
c



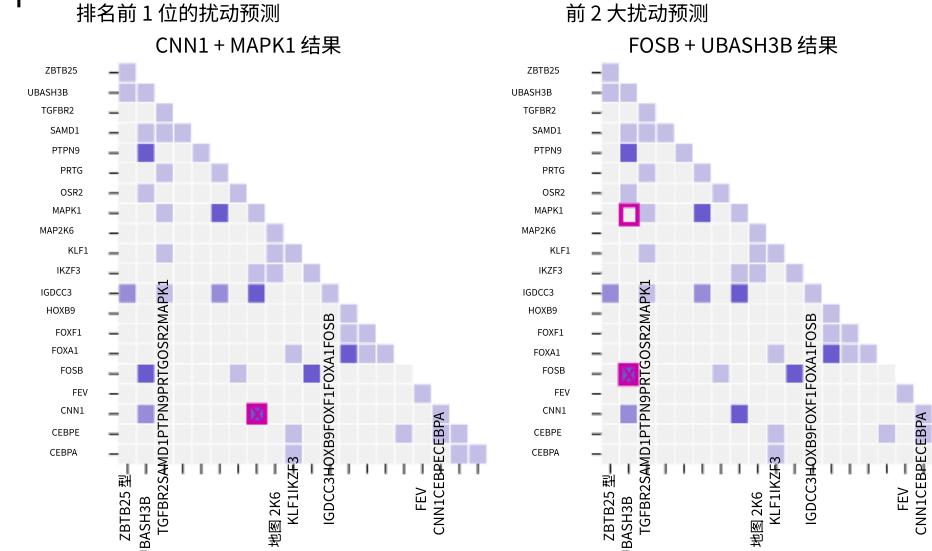
d



e



f



g

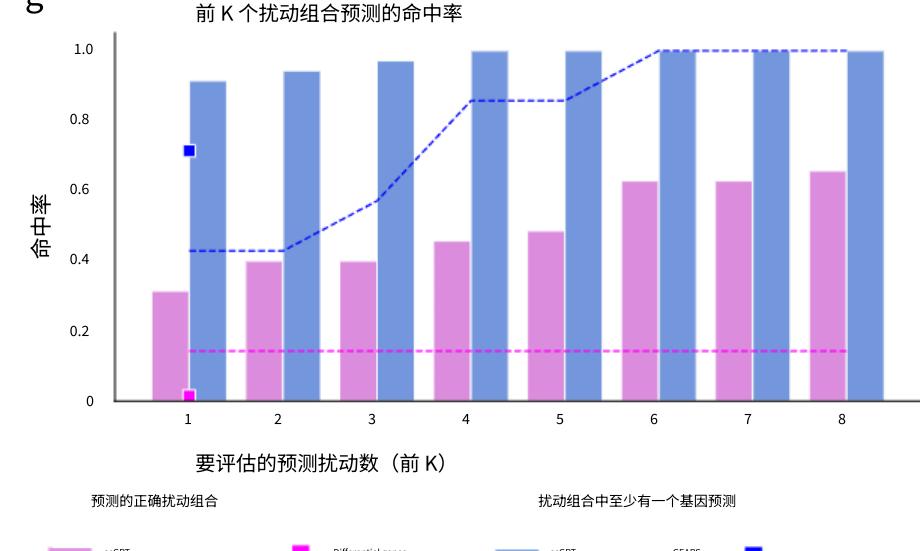


图 3 | 扰动响应和反向扰动的预测结果。

a, scGPT与其他扰动预测方法的比较。报告了预测和实际基因表达变化之间的 Pearson 相关性。该指标分别针对所有基因和排名靠前的差异表达 (DE) 基因进行计算。b, Adamson 检验数据集中的两个示例扰动，预测 (pred; n = 300 个细胞) 和实际基因的分布。

前 20 个差异表达基因的表达变化 (基因 KCTD16 扰动 n = 405 个细胞, 基因 DAD1 扰动 n = 618 个细胞)。该框表示表达式变化的四分位距。中位数由每个框内的中心线标记。须线延伸到四分位间距的 1.5 倍。水平虚线表示基因表达变化的零基线。c, 使用 scGPT 预测看不见的扰动响应的插图。d, 预测基因表达谱的 UMAP 图。

扰动条件。UMAP 图由 Leiden 簇着色，并用每个簇的显性基因标记。e, 两个选定的扰动基因 (KLF1 和 CNN1) 在扰动条件的 UMAP 上的表达模式。f, 在 20 个基因的扰动组合空间上可视化可能的扰动组合。网格按实验类型 (train, valid, test, unseen) 着色。所有预测的扰动都用方框突出显示，实际的源扰动用叉号标记。g, 在 7 个测试用例中，scGPT 对正确和相关预测的前 1-8 准确率，以 GEARS 和前两个差异基因的朴素基线为基准。相关预测 (蓝色) 表明在预测中至少可以找到扰动组合中的一个扰动基因。scGPT 的命中率由条形表示，GEARS 的命中率由线条表示，差异基因 (仅适用于前 1 个预测) 由方形标记表示。

微调 scGPT 以在计算机中扩展扰动，并使用 UMAP 可视化 Fig. 3d 中每个扰动的预测平均响应。使用原始研究的注释，我们发现相同官能团的扰动条件聚集在相邻区域（补充图 4）。接下来，我们使用 Leidenand 对预测的表达进行聚类，观察到这些簇与扰动组合中的“显性基因”表现出高度关联。例如，与 KLF1 基因相关的带圆圈的聚类表明该聚类中的数据点经历了涉及 KLF1 和另一个基因（即 KLF1 + X）的组合扰动。以 KLF1 和 CNN1 簇为两个例子，我们进一步验证了相应的预测表达在这些区域完全很高（图 3e），这与 Norman 数据集中 CRISPRa (CRISPR 介导的转录激活) Perturb-seq 实验的预期结果一致。显性基因簇证明了 scGPT 揭示扰动组合之间关联的能力。

计算机反向扰动预测。scGPT 还能够预测给定结果细胞状态的遗传扰动来源，我们称之为计算机反向扰动预测。进行这种反向预测的理想预测模型可用于推断谱系发育的重要驱动基因或促进潜在治疗基因靶点的发现。这种能力的一个假设示例应用可能是预测影响细胞从疾病状态中恢复的 CRISPR 靶基因。为了展示反向扰动预测的有效性，我们使用了 Norman 数据集的一个子集，重点关注涉及 20 个基因的扰动（图 3f）。这个组合空间由总共 210 个单基因或双基因扰动组合组成。我们使用 39 种（18%）已知的扰动（图 3f 中的训练组）对 scGPT 进行了微调。然后，我们在对看不见的扰动细胞状态的查询上测试了该模型，scGPT 成功预测了将产生观察到的结果的扰动来源（在排名靠前的预测中）。例如，scGPT 将 CNN1 + MAPK1 基因的正确扰动列为一个测试示例的首要预测，而 FOSB + UBASH3B 基因的正确扰动被列为另一个案例的第二个预测（图 3f）。总体而言，scGPT 在前 1 个预测中平均识别了 91.4% 的相关扰动（7 个中的 6.4 个）（图 3g 中的蓝色条）和 65.7% 的正确扰动（7 个测试用例中的 4.6 个）（图 3g 中的粉红色条），性能大大优于 GEARS 和差异基因基线。我们设想这些预测可以通过最大化推导目标细胞状态的可能性来用于规划扰动实验。与随机试验相比，随机试验平均需要该子集中 210 种可能的扰动中的 105.5 次尝试，以更少的尝试找到正确的遗传变化来源为加速发现重要遗传驱动因素和优化扰动实验提供了有价值的工具。

AvgBIO 评分汇总了三个细胞类型聚类指标，标准化互信息（NMI）、调整后的兰德指数（ARI）和平均轮廓宽度（ASW），如补充说明 12 中详述。值得注意的是，即使没有微调，scGPT 在集成 PBMC 10k 数据集方面也表现出了相当大的性能（补充图 5），突出了预训练的通用性。在鼻周皮层数据集的背景下，scGPT 与所有其他方法相比仍然具有竞争力（补充图 6c）。这一发现突出了从全人类数据集中学习的特征在应用于特定器官或组织（如大脑）时的可转移性和稳健性。此外，scGPT 在所有整合指标上始终保持有竞争力的分数，并表现出生物信号的强烈守恒性（补充表 3 和补充图 6 和 7）。此外，我们还开发了加速整合任务微调过程的策略，包括冻结特定模型层和排除不表达的基因，同时保持与我们原始方法相当的结果（补充说明 3）。

单细胞多组学整合。单细胞多组学 (scMultiomic) 数据结合了遗传调控的多种观点，例如表观遗传学、转录组学和翻译活动，在保留生物信号的同时聚合细胞表征提出了独特的挑战。scGPT 通过有效提取不同组学数据集中的集成细胞嵌入来应对这一挑战。在 10x Multiome PBMC 数据集（包括关节基因表达和染色质可及性测量）的情况下，我们将 scGPT 与两种最先进的方法进行了比较，即 scGLUE 和 Seurat (v.4)。如图 4b 所示，scGPT 是唯一成功为 CD8naive 细胞生成独特簇的方法。接下来，我们在来自骨髓单核细胞 (BMMC) 的配对基因表达和蛋白质丰度数据集上测试了 scGPT，如图 4c 所示。该数据集包含大量数据（90,000 个细胞）、多个批次（12 个供体）和精细的子组注释（48 个细胞类型）的额外复杂性。scGPT 呈现出比 Seurat (v.4) 更明确的集群结构，AvgBIO 评分提高了 9%。值得注意的是，scGPT 能够将 CD4 初始 T 细胞和 CD4 活化的 T 细胞分离为两个不同的簇。它还将整合素 β 激活的 CD4T 细胞与其他 CD4T 细胞区分开来，这进一步证实了该模型捕获免疫细胞亚群之间细微差异的能力。在镶嵌数据集成设置中，测序样本共享一些（但不是全部）数据模态，这对集成方法构成挑战。为了展示 scGPT 在这种情况下的能力，我们以 ATAC with select antigen profiling (ASAP) 人类 PBMC 数据集为例。该数据集由四个测序批次组成，具有三种数据模态。在使用 scMoMat 的基准实验中，scGPT 表现出卓越的批量校正性能，如图 4d 所示，尤其是在 B、髓系和自然杀伤 (NK) 细胞组中。总体而言，scGPT 表现出卓越的细胞类型聚类性能，并在各种基准生物保护指标中表现出稳健性（补充表 4）。

scGPT 支持多批次和多组学集成

多批次 scRNA-seq 集成。整合来自不同批次的多个 scRNA-seq 数据集在同时保留整合数据的生物学差异和消除技术批次效应方面提出了独特的挑战。为了整合测序样本，我们通过学习恢复掩蔽基因表达的统一细胞呈递（方法），以自我监督的方式对 scGPT 进行微调。在我们的基准测试实验中，我们将 scGPT 与三种流行的集成方法进行了比较：scVI、Seurat and Harmony。评价是在三个整合数据集上进行的，即 COVID-19 (18 批)、外周血单核细胞 (PBMC) 10k (两批) 和鼻周围皮层 (两批) 数据集。在 PBMC 10k 数据集中，scGPT 成功分离了所有细胞类型（图 4a）。scGPT 卓越的整合性能进一步得到了其高生物保护评分的支持，平均生物评分为 0.821，比比较方法高 5-10%。

scGPT 揭示了特定细胞状态的基因网络

GRN 背后的转录因子、辅因子、增强子和靶基因之间的相互作用介导重要的生物过程。现有的 GRN 推理方法通常依赖于静态基因表达的相关性或伪时间估计作为因果图的代理。scGPT 通过基因表达的生成建模进行优化，在其基因嵌入和注意力图中隐式编码这种关系。因此，我们通过探测来自预训练或微调模型的 scGPT 嵌入和注意力图来提出 GRN 推理工作流程。基因嵌入构建了一个相似性网络，该网络需要在数据集级别进行基因-基因相互作用。注意力图进一步捕获了不同细胞状态中独特的基因网络激活模式。在这项研究中，我们验证了基因网络

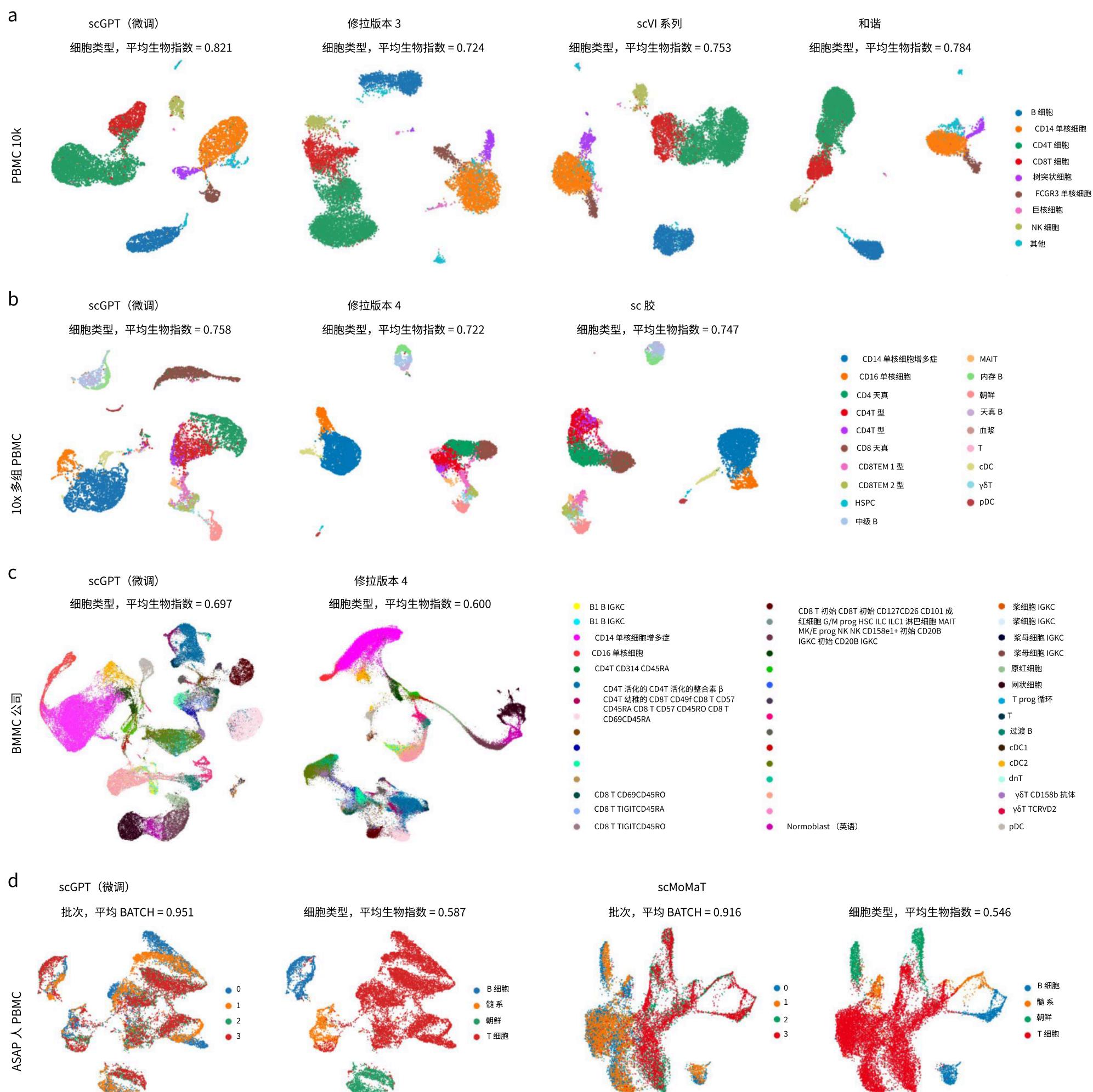


图 4 |多批次和多组学整合的结果。a, 基准

在 PBMC 10k 数据集上微调 scGPT 用于细胞类型聚类任务。学习的单元格嵌入的 UMAP 图按单元格类型着色。b, 在 10x Multiome PBMCdataset (转座酶可及染色质 (ATAC) 数据的配对 RNA 和检测) 上使用 scGLUE 和 Seurat (v.4) 微调 scGPT 模型的基准, 用于细胞类型聚类任务。γδT, γδ T 细胞;HSC, 造血干细胞;HSPC, 造血干细胞和祖细胞;ILC, 先天淋巴细胞;MAIT, 黏膜相关不变 T 细胞;单核细胞、单核细胞;

PDC, 浆细胞样树突状细胞;T, 中枢记忆 T 细胞;T, 效应记忆 T 细胞;T, 调节性 T 细胞。c, 在 BMMC 数据集 (配对 RNA 和蛋白质数据) 上使用 Seurat (v.4) 进行微调的 scGPT 模型的基准测试, 用于细胞类型聚类任务。dnT, 双阴性 T 细胞;G/M, 粒细胞-巨噬细胞;MK/E, 巨核细胞-红细胞;Prog, 祖先。d, 使用 scMoMaton 的 scGPT 对 ASAP PBMC 数据集 (嵌入 RNA、ATAC 和蛋白质数据) 进行基准测试, 用于批量校正和细胞类型聚类任务。学习到的基因嵌入的 UMAP 图按测序批次 (左) 和细胞类型 (右) 着色。

由 scGPT 针对已知生物学提取, 并探索其对基因程序发现的适用性。

scGPT 展示了它通过学习基因标记嵌入对功能相关基因进行分组和功能不同的基因的能力。在图 5a 中, 我们使用来自预训练 scGPT 模型的基因嵌入可视化人类白细胞抗原 (HLA) 蛋白的相似性网络, 从而进行了完整性检查。在此零镜头设置

scGPT 模型成功突出了对应于充分表征的 HLA 类别的两个簇: HLA I 类和 HLA II 类基因。这些类别编码在免疫环境中发挥不同作用的抗原呈递蛋白。例如, HLA I 类蛋白 (由 HLA-A、HLA-C 和 HLA-E 等基因编码) 被 CD8T 细胞识别并介导细胞毒性作用, 而 HLA II 类蛋白 (由 HLA-DRB1、HLA-DRA 和 HLA-DPA1 编码) 被 CD4 T 细胞识别并触发更广泛的辅助功能。此外, 我们在 “免疫人类” 数据集上微调了 scGPT 模型, 并探索了特定于该数据集中存在的免疫细胞类型的 CD 基因网络。我们使用了与此相同的微调策略

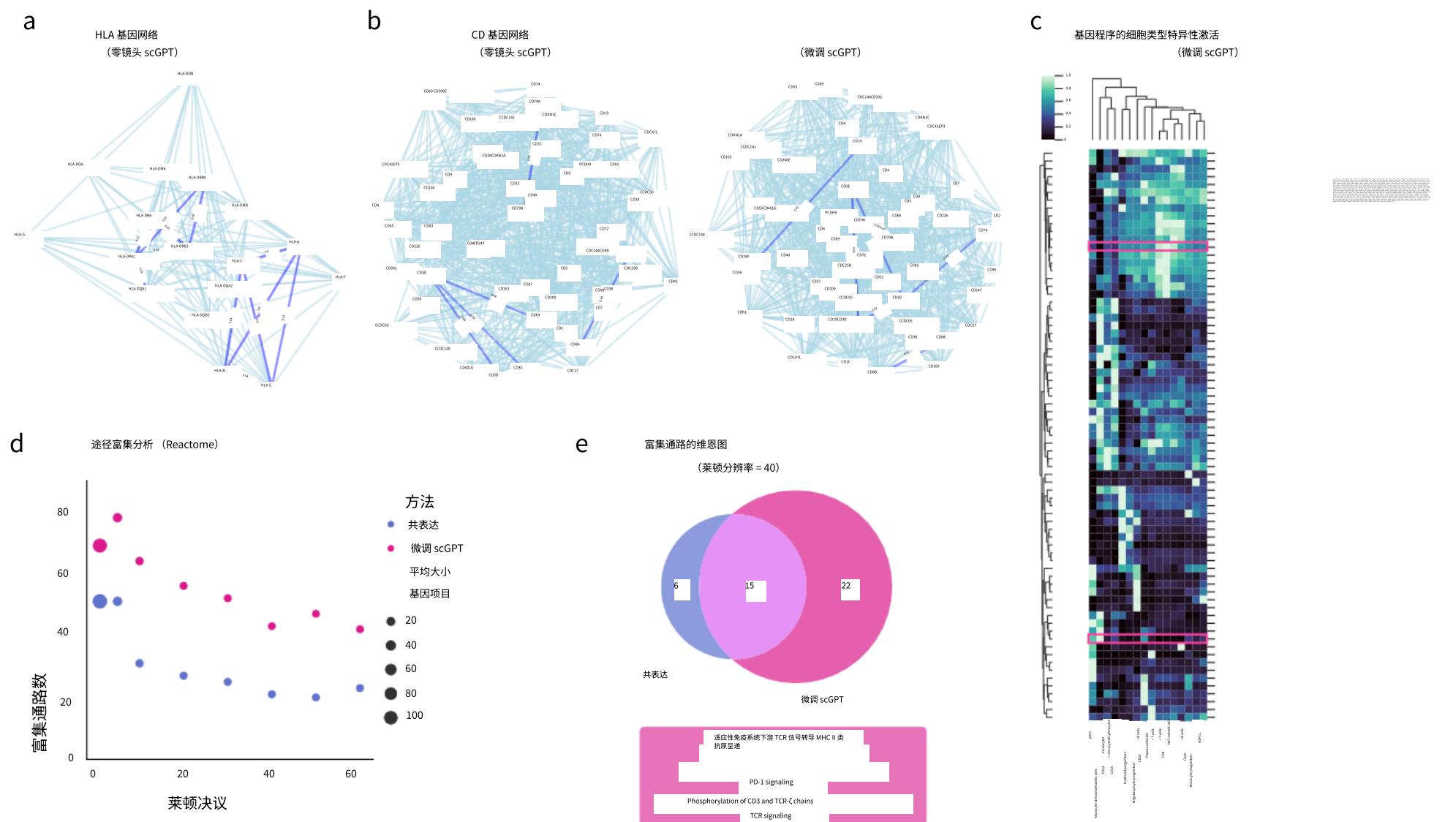


图 5 |基因标记嵌入分析。a, HLA 基因网络来源

零镜头 scGPT。b, 来自免疫人类数据集上的零脉冲（即预训练）和微调 scGPT 的 CD 基因网络。c, 免疫人类数据集中 scGPT 提取的基因程序中的细胞类型特异性激活。着色表示平均基因表达。对于包含 10 个以上基因（用星号表示）的基因程序，为简单起见，将显示前 10 个基因。d, 提取的基因程序的通路富集分析

通过 scGPT 和免疫人类数据集中的共表达网络。将 scGPT 基因程序的富集通路数量与不同 Leiden 分辨率的共表达方法的富集通路数量进行了比较。e, 比较共表达与 scGPT 鉴定的富集通路之间的重叠和差异的维恩图。文本框中突出显示了一些 scGPT 特有和适应性免疫功能特有的示例通路。TCR, T 细胞受体。

用于 GRN 分析的集成任务 (Methods)。预训练的 scGPT 模型成功识别了编码 T 细胞活化的 T3 复合物的基因组 (CD3E, CD3D 和 CD3G)，以及用于 B 细胞信号传导的 CD79A 和 CD79B，以及作为 HLA I 类分子的辅助受体的 CD8A 和 CD8B (图 5b)。此外，微调的 scGPT 模型突出了 CD36 和 CD14 之间的联系 (图 5b)。

scGPT 能够发现表现出细胞类型特异性激活的有意义的基因程序。随后使用来自 scGPT 的基因嵌入选择和聚类基因程序 (方法)。在图 5c 中，我们可视化了微调 scGPT 模型在免疫人类数据集中高度可变基因 (HVG) 上提取的基因程序及其在不同细胞类型中的表达。我们观察到一组 HLA II 类基因被鉴定为第 2 组。同样，参与 T3 复合物的 CD3 基因被鉴定为第 3 组，在 T 细胞中的表达最高。为了系统地验证提取的基因程序，我们对 Reactome 数据库 (<https://reactome.org/>) 进行了通路富集分析，并使用严格的多重测试校正 (<https://mathworld.wolfram.com/BonferroniCorrection.html> 和方法) 确定了高置信度的“通路命中”。在图 5d 中，我们将 scGPT 获得的结果与共表达网络获得的结果进行了比较。值得注意的是，scGPT 在所有聚类分辨率中始终表现出更高的富集通路数量。此外，我们检查了 scGPT 和共表达网络之间已确定通路的相似性和差异性，如图 5e 所示。两种方法都确定了 15 个常见途径，包括与细胞周期和免疫系统相关的途径。scGPT 唯一确定了另外 22 条通路，其中 14 条与免疫相关。值得注意的是，scGPT 特别强调了与适应性免疫系统相关的通路。

T 细胞受体信号传导、PD-1 信号传导和 MHC II 类表现。这与微调数据集中存在适应性免疫群体的事实一致。这些发现证明了 scGPT 在更广泛的生物学背景下捕获复杂的基因-基因连接并揭示特定机制的卓越能力。补充表 5 中提供了富集通路的详细列表。

除了使用基因嵌入进行数据集级基因网络推断外，scGPT 注意力机制还使其能够在单细胞水平上捕获基因-基因相互作用。scGPT 通过聚合注意力图中的单个细胞信号来提取细胞状态特异性网络激活数据。这提供了对单个细胞内环境特异性基因调控相互作用的见解，这些相互作用可能因不同的细胞状态和条件而异。例如，在扰动实验中，scGPT 检查扰动前后基因网络激活的变化，以推断哪些基因受每个扰动基因的影响最大 (图 6a 和方法)。在 Adamson CRISPR 干扰数据集中，scGPT 确定了受 DDIT3 (编码转录因子) 抑制影响最大的前 20 个基因，这些基因在 ChIP-Atlas 数据库中都被发现是 DDIT3 的信号靶标 (图 6b)。此外，scGPT 在对照与 DDIT3 敲除设置中受 DDIT3 影响最大的前 100 个基因中捕获了不同的通路激活模式。值得注意的是，已知在 DDIT3 敲除环境中鉴定的 ATF3 转录因子通路可介导未折叠的蛋白质反应并调节细胞凋亡。同样，在 BHLHE40 抑制的情况下，发现前 20 个影响最大的基因中有 19 个是该转录因子的靶标，通过染色质免疫沉淀和测序 (ChIP-seq) 预测 (图 6c)。突出 DNA 合成和有丝分裂的通路激活谱反映了转录因子 BHLHE40 在细胞周期调控中的作用。这些基于注意力的发现

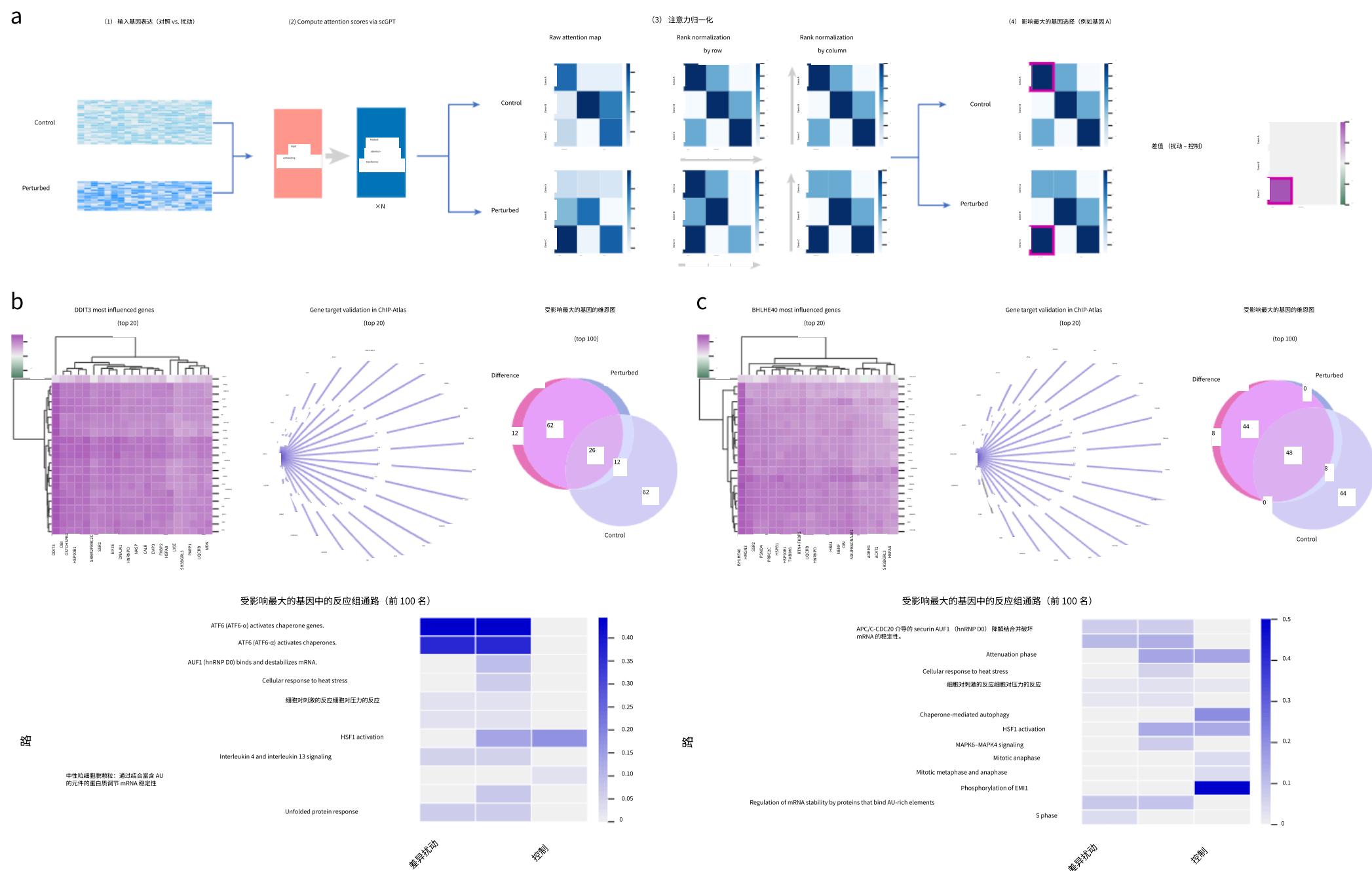


图 6 | 基于注意力的基因相互作用分析。a、基于注意力的 GRN-扰动数据的发现工作流程。获得对照和扰动细胞状态下的注意力分数，并按行和列连续排序。相应地选择对照、扰动和差异设置中受影响最大的基因。b，DDIT3 抑制的 GRN 分析。基因连接热图显示了受 DDIT3 抑制影响最大的前 20 个基因网络中的扰动后变化。基因靶标网络图展示了 ChIP-Atlas 数据库中验证的前 20 个基因，其中

ChIP-seq 预测的靶标以紫色突出显示。维恩图比较了在三种选择设置（即对照、扰动和扰动后差异）中鉴定出的前 100 个影响最大的基因集之间的重叠和差异。通路热图显示了在这三种选择设置中从前 100 个影响最大的基因中鉴定出的 Reactome 通路的差异。颜色表示基因重叠百分比表示每条通路的强度。c，BHLHE40 抑制的 GRN 分析，以类似的方式可视化。

在细胞状态水平上进一步验证 scGPT 学习的基因网络，为模型的学习生物学提供额外的可解释性。

迁移学习中的扩展和上下文效应

在前面的部分中，scGPT 通过以迁移学习方式进行微调，展示了巨大的潜力。我们通过将基础模型与无需预训练就为每个下游任务从头开始训练的类似 transformer 模型进行比较（表示为 scGPT（从头开始）），进一步证实了使用基础模型的好处。结果显示在补充表 2-4 中，其中微调的 scGPT 始终如一地展示了积分和细胞类型注释等任务的性能提升。鉴于观察到的基础模型对下游任务的贡献，我们进一步有兴趣探索影响迁移学习过程的因素。

首先，我们深入研究了预训练数据大小与微调模型性能之间的关系：对于某个分析任务，将更多的测序数据添加到图谱中进行预训练，可以获得多少改进？我们预训练了一系列具有相同参数数量但使用不同数量的数据的 scGPT 模型，从 30,000 到 3300 万个测序的正常人类细胞。补充图 13a 说明了使用这些不同的预训练模型在各种应用程序上进行微调的结果性能。我们观察到，随着预训练数据量的增加，微调模型的性能得到了提高（补充说明 4）。

这些结果表明了扩展效应，表明更大的预训练数据大小会导致更好的预训练嵌入和下游任务的性能提高。值得注意的是，我们的研究结果也与自然语言模型中报告的缩放定律一致，突出了数据大小在模型性能中的重要作用。预训练数据大小在微调结果中的关键作用表明，单细胞领域的预训练模型前景光明。随着更大、更多样化的数据集的出现，我们可以预期模型性能会进一步提高，从而促进我们对细胞过程的理解。

我们探索的第二个因素是特定于上下文的预训练的影响。在这里，上下文用法是指在特定细胞类型上进行预训练，然后针对类似细胞类型的下游任务进行微调的 scGPT 模型。为了探索这一因素的影响，我们在来自单个主要器官的正常人类细胞上预训练了七个器官特异性模型（图 1d）和另一个用于泛癌细胞的模型。我们通过可视化预训练数据的细胞嵌入来验证预训练的有效性：泛癌模型细胞嵌入准确区分了不同的癌症类型（补充图 2）。器官特异性模型能够揭示相应器官的细胞异质性（补充图 3）。接下来，我们在 COVID-19 数据集上微调了各个模型，以检查预训练上下文的影响。我们的分析揭示了模型上下文在预训练中的相关性与其

整合数据的后续性能（补充图 8）。数据集成任务中表现最好的是在全人、血液和肺数据集上预训练的模型，这与 COVID-19 数据集中存在的细胞类型密切相关。值得注意的是，即使是大脑预训练模型，尽管是在 1300 万个细胞的大量数据集上进行训练的，但与具有相似数据集大小的血液预训练模型相比，性能也落后了 8%。这强调了在预训练中将细胞环境与目标数据集保持一致的重要性，以便在下游任务中获得更好的结果。虽然考虑细胞环境是必不可少的，但全人类预训练模型已成为广泛运用的多功能且可靠的选择。

讨论

我们介绍了 scGPT，这是一个基础模型，它利用预训练转换器的强大功能来处理大量单细胞数据。在语言模型中自我监督预训练的成功基础上，我们在单细胞领域采用了类似的方法来解开复杂的生物相互作用。在 scGPT 中使用 transformer 可以同时学习基因和细胞嵌入，这有助于对细胞过程的各个方面进行建模。通过利用 transformer 的注意力机制，scGPT 在单细胞水平上捕获基因间的相互作用，提供额外的可解释性。

我们通过零镜头和微调设置的综合实验证明了预训练的好处。预训练模型展示了对看不见的数据集进行推断的强大能力，在零样本实验中根据细胞类型呈现有意义的聚类模式。此外，scGPT 中学习到的基因网络与已知的功能组表现出很强的比对性。此外，预训练模型的知识可以通过微调转移到多个下游任务。在细胞类型注释、扰动预测以及多批次和多组学整合等各种任务中，微调后的 scGPT 模型始终优于从头开始训练的模型。这证明了预训练模型对下游任务的价值，从而实现更准确和具有生物学意义的分析。值得注意的是，当前的预训练本身并不能减轻批处理效应，因此模型的零镜头性能可能会在技术变化很大的数据集上受到限制。鉴于经常缺乏明确的生物学基本事实和数据质量的差异，评估该模型也很复杂（详见补充说明 10）。

对于未来的方向，我们计划在具有更多多样性的更大规模数据集上进行预训练，包括多组学数据、空间组学和各种疾病。在预训练阶段结合扰动和时间数据也很有趣，使模型能够学习因果关系并推断基因和细胞如何响应随时间的变化。我们还旨在探索单细胞数据的上下文教学学习。这涉及开发技术，使预训练模型能够在零样本设置中理解和适应不同的任务和上下文，而无需微调。通过使 scGPT 能够掌握不同分析的细微差别和特定要求，我们可以增强其在各种研究场景中的可用性和适用性。我们设想预训练范式将很容易整合到单细胞研究中，并作为利用指数级增长的细胞图谱中的现有知识进行新发现的基础。

引用

1. Silverman, A. D., Karim, A. S. & Jewett, M. C. 无细胞基因表达：扩展的应用库。Nat. Rev. 基因。21, 151–170 (2020).
2. Preissl, S., Gaulton, K. J. & 任, B. 使用单细胞表观基因组学表征顺式调节元件。Nat. Rev. Genet. 24, 21–43 (2022).
3. Ding, J., Sharon, N. & Bar-Joseph, Z. 使用单细胞转录组学进行时间建模。Nat. Rev. Genet. 23, 355–368 (2022).
4. Wagner, D. E. & Klein, A. M. 谱系追踪遇见单细胞组学：机遇与挑战。Nat. Rev. Genet. 21, 410–427 (2020).
5. Regev, A. 科学论坛：人类细胞图谱。eLife 6, e27041 (2017).
6. Han, X. 通过 Microwell-seq 绘制小鼠细胞图谱。细胞 172, 1091–1107 (2018).
7. Angerer, P. 等人。单细胞制造大数据：转录组学的新挑战和机遇。电流。意见。系统生物学 4, 85–91 (2017).
8. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. 多组学数据集成、解释及其应用。生物信息。生物学见解 14, 1177932219899051 (2020)。
9. Miao, Z., Humphreys, B. D., McMahon, A. P. & Kim, J. 百万单细胞数据时代的多组学整合。Nat. Rev. Nephrol. 17, 710–724 (2021).
10. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen 预测单细胞扰动反应。Nat. 方法 16, 715–721 (2019).
11. Lotfollahi, M. 预测细胞对高通量筛选中复杂扰动的反应。分子系统生物学 19, e11517 (2023).
12. Lotfollahi, M. 通过迁移学习将单细胞数据映射到参考图谱。Nat. 生物技术。40, 121–130 (2022).
13. 曹 Z.-J. & Gao, G. 多组学单细胞数据集成和调控推断与图链接嵌入。国家生物技术。40, 1458–1466 (2022).
14. Zhang, Z. et al. scMoMat 联合进行单细胞镶嵌整合和多模式生物标志物检测。Nat. Commun. 14, 384 (2023).
15. Bommasani, R. 等人。关于基础模型的机遇和风险。<https://doi.org/10.48550/arXiv.2108.07258> 年预印本 (2021 年)。
16. Moor, M. 等人。通才医学人工智能的基础模型。自然 616, 259–265 (2023)。
17. Vaswani, A. 等人。你只需要关注。Adv. 神经 Inf. 过程。系统 6000-6010 (NeurIPS, 2017)。
18. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. 使用 CLIP 潜伏生成分层文本条件图像。<https://doi.org/10.48550/arXiv.2204.06125> 年预印本 (2022 年)。
19. Brown, T. 语言模型是少数机会的学习者。Adv. 神经。Inf. 过程。系统 1877-1901 (NeurIPS, 2020)。
20. OpenAI 团队。GPT-4 技术报告。预印本 <https://doi.org/10.48550/arXiv.2303.08774> (2023)。
21. Avsec, Z. 等人。通过整合远程相互作用从序列中有效预测基因表达。Nat. Methods 18, 1196–1203 (2021)。
22. Gururangan, S. 等人。不要停止预训练：使语言模型适应域和任务。计算语言学协会第 58 届年会论文集 8342–8360 (ACL, 2020)。
23. Qiu, X. 等人。自然语言处理的预训练模型：一项调查。中国科技科学 63, 1872–1897 (2020)。
24. 刘杰, 范, Z., 赵, W. & 周, X. 单细胞数据分析中的机器智能：进展和新挑战。前面。基因。12, 655536 (2021)。

在线内容

任何方法、其他参考文献、Nature Portfolio 报告摘要、源数据、扩展数据、补充信息、致谢、同行评审信息、作者贡献和利益争夺的详细信息；以及数据和代码可用性声明可在 <https://doi.org/10.1038/s41592-024-02201-0> 上获得。

25. Oller-Moreno, S., Kloiber, K., Machart, P. & Bonn, S. 机器学习的算法进展，用于单细胞表达分析。电流。*系统生物学* 25, 27–33 (2021)。
26. Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. 用于扰动单细胞组学的机器学习。*细胞系统* 12, 522–537 (2021)。
27. Theodoris, C. V. 等人。迁移学习支持在网络生物学中进行预测。*自然* 618, 616–624 (2023)。
28. McInnes, L., Healy, J. & Melville, J. UMAP：用于降维的均匀流形近似和投影。<https://doi.org/10.48550/arXiv.1802.03426> 年预印本 (2018 年)。
29. Schirmer, L. 多发性硬化症中的神经元脆弱性和多谱系多样性。*自然* 573, 75–82 (2019)。
30. 程 S. 肿瘤浸润髓细胞的泛癌种单细胞转录图谱。*单元格* 184, 792–809 (2021)。
31. Chen, J. 等人。用于一站式可解释像元类型注记的转换器。*Nat. Commun.* 14, 223 (2023)。
32. Yang, F. et al. scBERT 作为用于单细胞 RNA-seq 数据的细胞类型注释的大规模预训练深度语言模型。*Nat. Mach. Intell.* 4, 852–866 (2022)。
33. 亚当森 B. 多路复用单细胞 CRISPR 筛选平台能够系统地解剖未折叠的蛋白质反应。*细胞* 167, 1867–1882 (2016)。
34. Replogle, JM 使用基因组规模 Perturb-seq 绘制信息丰富的基因型-表型景观。*单元格* 185, 2559–2575 (2022)。
35. Norman, TM 等人。探索由丰富的单细胞表型构建的遗传相互作用歧管。*科学* 365, 786–793 (2019)。
36. Roohani, Y., Huang, K. & Leskovec, J. 用 GEARS 预测新型多基因扰动的转录结果。*Nat. 生物技术*。<https://doi.org/10.1038/s41587-023-01905-6> (2023 年)。
37. Traag, V. A., Waltman, L. & Van Eck, N. J. 从鲁汶到莱顿：保证社区的紧密联系。*科学代表* 9, 5233 (2019)。
38. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. 单细胞转录组学的深度生成建模。*Nat. Methods* 15, 1053–1058 (2018)。
39. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. 单细胞基因表达数据的空间重建。*Nat. 生物技术* 33, 495–502 (2015)。
40. Korsunsky, I. 等人。使用 Harmony 快速、灵敏、准确地集成单细胞数据。*Nat. Methods* 16, 1289–1296 (2019)。
41. 加约索, A. 用于单细胞组学数据概率分析的 Python 库。*Nat. 生物技术* 40, 163–166 (2022)。
42. Siletti, K. 成人大脑中细胞类型的转录组多样性。*科学* 382, eadd7046 (2023)。
43. 来自健康供体的 PBMC, Cell Ranger ARC 1.0.0 的单细胞多组 ATAC 基因表达演示数据。10X Genomics https://support.10xgenomics.com/single-cellmultiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k (2020 年)。
44. Hao, Y. 多模态单细胞数据的综合分析。*单元格* 184, 3573–3587 (2021)。
45. Luecken, M. 等人。用于预测和整合单细胞中 DNA、RNA 和蛋白质的沙盒。在神经信息处理系统会议记录中，数据集和基准跟踪 13 (NeurIPS, 2021)。
46. Mimitou, EP 单细胞中染色质可及性、基因表达和蛋白质水平的可扩展、多模式分析。*Nat. 生物技术* 39, 1246–1258 (2021)。
47. Pratapa, A., Jalihal, A. P., Law, JN, Bharadwaj, A. & Murali, T. M. 从单细胞转录组数据中进行基因调控网络推断的基准算法。*Nat. 方法* 17, 147–154 (2020)。
48. 周 S. Y.HLA 系统：遗传学、免疫学、临床检测和临床意义。*延世医学杂志* 48, 11–23 (2007)。
49. Norman, PS 免疫生物学：健康和疾病中的免疫系统。*J. 过敏临床。免疫学* 96, 274 (1995)。
50. Luecken, M. D. 单细胞基因组学中的图谱级数据集成基准测试。*Nat. 方法* 19, 41–50 (2022)。
51. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 更新：一个数据挖掘套件，通过完全整合 ChIP-seq、ATAC-seq 和 Bisulfite-seq 数据来探索表观基因组景观。*核酸研究* 50, W175–W182 (2022)。
52. Yang, H., Niemeijer, M., van de Water, B. & Beltman, J. B. ATF6 是 CHOP 动力学的关键决定因素。*iScience* 23, 100860 (2020)。
53. Yoshida, H. et al. 在 NF-Y (CBF) 存在下，由蛋白水解激活的 ATF6 直接与负责哺乳动物未折叠蛋白反应的顺式作用元件结合。摩尔。细胞。*生物学* 20, 6755–6767 (2000)。
54. Kaplan, J. 等人。神经语言模型的缩放定律。<https://doi.org/10.48550/arXiv.2001.08361> 年预印本 (2020 年)。

出版商注 Springer Nature 对已发布地图中的管辖权主张和机构隶属关系保持中立。

根据与作者或其他权利持有人签订的出版协议，Springer Nature 或其许可方（例如协会或其他合作伙伴）拥有本文的专有权；本文已接受的手稿版本的作者自行存档仅受此类出版协议的条款和适用法律的约束。

© 作者，经 Springer Nature America, Inc. 2024 独家许可

方法

输入嵌入

单细胞测序数据被处理成逐个细胞矩阵 $XXX \in R$, 其中每个元素 $X \in R$ 代表 scRNA-seq 数据的 RNA 分子的读取计数或 scATAC-seq 数据峰区域的染色质可及性。具体来说, 对于 scRNA-seq 数据, 该元素表示细胞 $i \in \{0, 1, \dots, N\}$ 中基因 $j \in \{0, 1, \dots, G\}$ 的 RNA 丰度。在后续部分中, 我们将此矩阵称为原始计数矩阵。scGPT 的输入由三个主要部分组成: (1) 基因 (或峰值) 标记, (2) 表达值和 (3) 条件标记。对于每个建模任务, 基因标记和表达值相应地从原始计数矩阵 X 进行预处理。

基因标记。在 scGPT 框架中, 每个基因都被认为是最小的信息单位, 类似于 NLG 中的一个单词。因此, 我们使用基因名称作为标记, 并为每个基因 g 分配一个唯一的整数标识符 $id(g)$ 。这些标识符构成了 scGPT 中使用的标记词汇表。这种方法提供了极大的灵活性, 可以协调具有不同基因集 (即由不同的测序技术或预处理流程生成的) 的多项研究。具体来说, 通过跨研究获取所有基因的联合集, 可以将不同的基因标记集整合到一个通用词汇表中。此外, 我们还在词汇表中加入了特殊标记, 例如用于将所有基因聚合到细胞表示中的 $<cls>$, 以及用于将输入填充到固定长度的 $<pad>$ 。从概念上讲, 我们在 NLG 中将基因标记和单词标记相提并论。因此, 每个细胞 i 的输入基因标记由向量 ttt 表示 $\in N$:

$$ttt = [id(g), id(g), \dots, id(g)], \quad (1)$$

其中 M 是预定义的最大输入长度。

表达式值。基因表达矩阵 X 在用作建模的输入之前需要额外的处理。基因表达建模的一个基本挑战是不同测序方案中绝对幅度的可变性。测序深度的变化和稀表达基因的存在导致不同批次测序样本之间的数据规模存在巨大差异。这些差异不能用常见的预处理技术 (如每百万转录本归一化和 log1p 转换) 来轻易缓解。即使在这些转换之后, 相同的绝对值也可以在测序批次中传达不同的“语义”含义。为了解决这种规模差异, 我们提出了值分箱技术, 将所有表达式计数转换为相对值。对于每个单元格中的每个非零表达式计数, 我们计算原始绝对值并将它们划分为 B 个连续的区间 $[b_k, b_{k+1}]$, 其中 $k \in \{1, 2, \dots, B\}$ 。每个区间代表所有表达基因的相等部分 ($1/B$)。请务必注意, 将为每个单元格计算一组新的 bin 边缘, 因此区间边缘 b 可能因单元格而异。单元格 i 的分箱表达式值 x 定义为:

$$x = \begin{cases} k, & \text{如果 } X > 0 \text{ 和 } X \in [b_k, b_{k+1}], \\ 0, & \text{如果 } X = 0. \end{cases} \quad (2)$$

通过这种分箱技术, x 的语义含义在各种测序批次的细胞中是一致的。例如, $x = B$ 的值始终表示基因中的最高表达。值得注意的是, 对于微调任务, 我们还在值分箱步骤之前执行了 log1p 转换和 HVG 选择。为了简化表示法, 我们在分箱之前使用 X 来表示原始数据矩阵和预处理过的数据矩阵。因此, 单元格 i 的分箱表达式值的最终输入向量表示为

$$x = [x_1, x_2, \dots, x_B]. \quad (3)$$

条件令牌。条件标记包含与单个基因相关的各种元信息, 例如扰动实验改变 (由扰动标记表示)。为了表示位置条件标记, 我们使用与输入基因共享相同维度的输入向量。此向量表示为:

$$t = [t_1, t_2, \dots, t_M], \quad (4)$$

其中 t 表示与条件对应的整数索引。

嵌入层。我们分别使用常规嵌入层 (即 PyTorch 嵌入层 (<https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html>)) emb 来表示基因标记和条件标记, 以方便将每个标记映射到维度 D 的固定长度嵌入向量。我们使用全连接层 (表示为 emb) 作为分箱表达式值, 以增强表现力。这种选择可以对基因表达值的序数关系进行建模。因此, 最终嵌入 hh

$\in R$ 对单元格 i 定义为:

$$hh^{(i)} = emb(t) + emb(x) + emb(t). \quad (5)$$

通过 transformer 进行细胞和基因表达建模

scGPT 转换器。我们使用自注意力转换器对完整的输入嵌入 hh 方程 (5) 进行编码。自我注意机制作用于 M 嵌入向量的序列, 使其特别适合捕获基因之间的相互作用。堆叠变压器块的输出可以定义如下:

$$\begin{aligned} h_0^{(i)} &= hh^{(i)} \\ h_l^{(i)} &= \text{transformer_block}(h_{l-1}^{(i)}) \quad \forall l \in [1, n]. \end{aligned} \quad (6)$$

用于基因水平的微调任务 hh 直接应用基因水平的微调目标 (Fine-tuning objectives)。示例包括基因表达预测 (GEP) 目标和扰动表达预测任务 (perturb-GEP)。对于 cell 级任务, 我们首先集成 hh

n 转换为单元格嵌入向量 (Cell 表示)。一个例子是细胞类型分配任务, 其中细胞嵌入用于通过细胞类型分类训练目标中添加的分类器来预测细胞类型标签。输入维度 M 可以达到数万个基因, 大大超过了 NLG 中常用的常规转换器的输入长度。为了应对这一挑战并确保有效的自我注意机制, 我们利用了 FlashAttention 的加速自我注意实现。此实现有效地增强了模型容量, 并支持对大型输入维度的有效处理。尽管采用了 FlashAttention, 但任何高效的转换器也有可能用于 scGPT, 例如具有线性复杂性的转换器 (Linerformer) 和内核化自注意力 (KSA)。

单元格表示。每个细胞都类似于一个由基因组成的“句子”, 其表示 hh

$c \in R^{n \times D}$ 是通过聚合获得的。因此, hh 很容易地使用各种池化操作, 例如元素均值池化或加权池化。在这项研究中, 我们选择使用一个特殊的 $<\text{cls}>$ 标记来表示单元, 使模型能够学习 transformer blocks 内的池化操作。 $<\text{cls}>$ 标记将附加到 input 标记的开头, 并将此位置的最终嵌入提取为单元表示形式。因此, 单元嵌入 hh

c 可以通过堆叠的最终层嵌入 hh 中的相应行提取 n [$<\text{cls}>$], 其中 [$<\text{cls}>$] 作检索 $<\text{cls}>$ 标记位置的索引处的行。

批次和模态的表示形式。我们使用额外的标记集来表示不同的测序批次和测序模式（来自 RNA-seq 的基因、来自 ATAC-seq 的峰等），特别是用于 scRNA-seq 和 scMultiomic 整合任务。这类似于 Input embeddings 中引入的条件标记，并且使用标准嵌入层类似地实现。模态标记 t_{tttare} 与单个输入特征 g 相关联（例如，指示它是基因、区域还是蛋白质）。批处理令牌最初位于单元级别，但也可以传播到单个单元的所有特征。换句话说，相同的批量标记 t 可以重复到单个单元格 i 的输入特征的长度 M ：

$$\text{batch}[t, t, \dots, t] = [t, t, \dots, t]. \quad (7)$$

Input embeddings 中描述的 token 与 batch 和 modality token 之间的区别在于，这些 batch 和 modality token 的嵌入不用作 transformer 块的 Importing。相反，在进入特定的微调目标之前，它们与特征或单元级别的变压器输出连接。这是为了防止转换器放大相同模态特征内的注意力，同时低估不同模态的特征。此外，了解模式和/或批次身份有助于在下游微调目标中进行基因表达建模。当模型学习预测基于模态和/或批次身份的表达值时，这种偏差会从基因和细胞表征本身中隐式消除。

这用作促进批量更正的技术。

例如，在 scMultiomic 集成任务中，我们将 transformer 输出与 batch 和 modality 嵌入之和连接起来。这用作表达建模的下游微调目标的输入：

$$h_n^{(-)} = \text{concat} (H \overset{(i)}{\text{batch}} \text{emb} (ttt) + \text{emb} (ttt)), \quad (8)$$

其中 emb 和 emb 分别表示 batch-embedding 层和 modality-embedding 层。 $h_n^{(i)}$ 表示变压器层的输出

(scGPT 转换器)。

或者，在 scRNA-seq 集成任务中，批量嵌入与细胞表示的串联产生以下表示作为输入：

$$h_c^{(-)} = \text{concat} (H \overset{(i)}{\text{batch}} \text{emb} (t)), \quad (9)$$

其中 H 表示单元格 i 的批次令牌的单个单元格表示。注意修改后的版本 $h_n^{(i)}$

c 仅与表达式建模目标相关，不适用于基于分类的目标，如微调目标中所述。

生成式预训练

基础模型预训练。基础模型被设计为一个可推广的特征提取器，可以使各种下游任务受益。预训练中使用的标记词汇包含人类基因组中的整组基因。在模型预训练 (Input embeddings) 之前对表达式值进行分箱。为了加快训练速度，我们将输入限制为每个输入细胞仅具有非零表达的基因。为了有效地训练模型来捕捉基因-基因关系和基因-细胞关系，我们引入了一种带有专用注意力掩码的生成训练策略，如下一节所述。

用于生成式预训练的注意力掩码。自我注意已被广泛用于捕获标记之间的共现模式。在自然语言处理中

这主要通过两种方式实现：(1) transformer 编码器模型（如 BERT 和 RoBERTa）中使用的掩码令牌预测，其中输入序列中随机掩码的令牌在模型的输出中被预测，以及 (2) 在因果 transformer 解码器模型中使用顺序预测生成自回归，如 OpenAI GPT 系列。OpenAI GPT-3（参考文献 19）和 GPT-4（参考文献 20）中使用的生成式预训练使用了一个统一的框架，在该框架中，模型从由已知输入标记组成的“提示”中预测最可能的下一个标记。此框架为各种 NLG 应用程序提供了极大的灵活性，并演示了 zero-shot 和 fine-tuned 设置中的上下文感知等功能。我们相信生成式训练可以以类似的方式对单细胞模型有益。具体来说，我们对两项任务感兴趣：(1) 根据已知基因表达生成未知基因表达值，即通过“基因提示”生成，以及 (2) 在给定输入细胞类型条件下生成全基因组表达，即通过“细胞提示”生成。

尽管标记和提示的想法相似，但由于数据的非顺序性质，建模遗传读取与自然语言本质上不同。与句子中的单词不同，细胞内的基因顺序是可以互换的，并且没有等效的“下一个基因”概念可以预测。这使得将 GPT 模型的因果掩码公式直接应用于单细胞数据变得具有挑战性。为了应对这一挑战，我们为 scGPT 开发了一种专门的注意力掩蔽机制，该机制根据注意力分数定义预测顺序。

注意力掩蔽通常可以应用于 transformer 模块中的自我注意力图：对于 M 个基因标记的输入 (方程 (1))，第 $(l+1)$ 个 transformer 模块对其输入 $h_n^{(i)}$ 应用多头自我注意

$\in M$ 个标记的 R (等式 (6))。 $\overset{(i)}{W}$ 具体来说，每个 self-attention 作的计算方式如下：

$$Q = h_n^{(i)} \overset{(i)}{W}, \quad K = h_n^{(i)} \overset{(i)}{W}, \quad V = h_n^{(i)} \overset{(i)}{W}, \\ \text{注意 } (Q, K, V) = \text{softmax} \left(\frac{Q^T K}{\sqrt{d}} + AAA \right) V, \quad (10)$$

其中 $Q, K, V \in R$ 表示查询、键和值向量。 $W, W, W \in R$ 表示稀有的可学习权重矩阵。术语 d 是特征维度，用作保持

数值稳定性。注意力掩码 $AAA \in \{0, -inf\}$ 通过修改查询和键之间的原始注意力权重来描述自我注意力的范围，如下所示。具体来说，将 $-inf$ 添加到矩阵中的位置 (i, j) 会使 softmax 之后的注意力权重无效，从而禁止第 i 个查询和第 j 个键之间的注意力。另一方面，添加 0 意味着注意力权重保持不变。这种注意力掩码技术使模型能够专注于特定的上下文元素。

我们专门设计了 scGPT 注意力掩码，以统一的方式支持基因提示和细胞提示生成。注意掩码 $AAA \in \{0, -inf\}$

在补充图 1a 中可视化，其中查询按行组织，键按列组织。如图底部注释的那样，输入嵌入向量 $h_n^{(i)}$ 中的每个标记

1 可以是以下三组之一：(1) 保留的 $\langle CLS \rangle$ 用于细胞嵌入的标记（在 Cell 表示中介绍），(2) 具有标记嵌入和表达值嵌入的已知基因，以及 (3) 要预测表达值的未知基因。scGPT 注意力掩蔽的经验法则是只允许“已知基因”的嵌入和查询基因本身之间的注意力计算。这是通过使用 AAs 中的元素来实现的，如下所示：

$$\text{一个} = \begin{cases} 0, & \text{如果 } j \notin \text{未知基因,} \\ 0, & \text{如果 } i=j \text{ 和 } j \in \text{未知基因,} \\ -inf, & \text{如果 } i \neq j \text{ 和 } j \in \text{未知基因.} \end{cases} \quad (11)$$

在每次生成迭代中，scGPT 都会预测一组新基因的基因表达值，这些基因反过来成为下一次迭代中的“已知基因”，用于注意力计算。这种方法

这种方法通过在非序列单单元数据中进行顺序预测，反映了传统 transformer 解码器中具有下一个标记预测的因果掩码设计。

如补充图 1a 所示，在训练过程中，我们随机选择一定比例的基因为未知，以便在输入中省略它们的表达值。注意力只应用于已知基因和查询未知基因本身之间，而不应用于其他未知基因的位置。例如，在位置 j 预测的基因具有与细胞嵌入、已知基因和自身的注意力分数，而没有其他未知基因，如注意力掩码的最后一行所示。scGPT 模型通过带有上述掩蔽注意力图的堆叠 transformer 块预测这些未知基因的表达。推理步骤如补充图 1b 所示。在细胞提示生成的推理过程中，scGPT 生成以特定细胞类型为条件的所有全基因组基因表达。在表示像元类型条件的第一个位置输入经过训练的细胞嵌入。数千个基因表达值的整个生成过程以 K 个迭代步骤进行（即补充图 1b 中的 $K = 3$ 个步骤）。例如，在一次迭代中， $i \in \{1, 2, \dots, K\}$ 的 intent 掩码机制允许对从之前的 0 到 $i - 1$ 次迭代的所有预测基因进行注意力。在每次迭代中，scGPT 从未知集中选择具有最高预测置信度的前 $1/K$ 基因作为已知基因包含在下一次迭代 $i + 1$ 中。直观地说，该工作流程以自回归方式简化了基因表达的生成，其中首先生成具有最高预测置信度的基因表达值，并将其用于帮助后续轮次生成。基因提示生成以迭代方式类似地工作。不同之处在于，它从一组具有观察到的表达值的已知基因开始，而不是细胞嵌入。

scGPT 注意力掩蔽统一了已知基因的编码过程和未知基因的产生。它也是最早对非序列数据进行自回归生成的 transformer 方案之一。

预训练的学习目标。我们使用基因表达预测目标来优化模型以预测未知基因的表达值。具体来说，我们使用多层感知器网络（MLP）来估计未知的表达值并计算均方损失 L ：

$$L = \frac{1}{|UU|} \sum_{j \in UU} (\text{MLP}_{\text{hh}}^{(i)} h_{j-} x)^2, \quad (12)$$

其中 UU 表示未知基因的输出位置集， x 表示要预测的实际基因表达值。 $| \cdot |$ 作检索集合的元素数。

如用于生成式预训练的注意力掩码中所述，支持基因提示生成和细胞提示生成。在训练过程中，这两种模式是连续进行的。在一个给定细胞的输入基因标记中，一部分基因被选择为“未知”基因，并且它们的表达值被省略。首先，在基因提示步骤中，模型的输入包含 $\langle \text{cls} \rangle$ 标记嵌入、已知基因嵌入和未知基因嵌入。损失（方程（12））使用模型的输出计算。其次，在 cell-prompt 步骤中，输出 cell 嵌入（即 hh ）

c 用于替换位置的嵌入。其他计算保持不变。最后，将两个步骤的损失值相加，用于计算梯度以优化模型参数。

微调目标

scGPT 利用各种微调目标来促进学习细胞和基因的生物学有效表示，以及用于正则化目的，例如批量校正。

基因表达预测。为了鼓励学习基因-基因相互作用，scGPT 结合了 GEP。此微调目标的工作方式类似于预训练中的目标（预训练的学习目标），但适用于蒙版位置。具体来说，对于每个输入单元格，基因标记的子集及其相应的表达值是随机屏蔽的。scGPT 经过优化，可准确预测掩蔽位置的表达值。这个微调目标有利于模型有效地编码数据集中基因之间的共表达。该物镜使掩蔽位置的均方误差最小，表示为 M 。GEP 的工作原理如下：

$$\tilde{x}_{xx} = \text{MLP}_{\text{hh}}^{(i)}, \quad (13)$$

$$L = \frac{1}{|M|} \sum_{j \in M} (\boxtimes_{x-} x)^2.$$

$\in N$ 表示单元格 i 的表达估计值行。值得注意的是，如果提供了测序批次或模式条件，我们使用 hh

等式（8）中的 n 而不是 HHH

GEP 提出了一个通用的自我监督微调目标，旨在预测基因表达值。在某些下游任务中，例如扰动预测，模型需要预测扰动的基因表达值，而不是原始值。我们将这种变化称为 perturb-GEP。我们将 MLP 估计器保留在等式（13）中，但使用扰动后基因表达作为目标 x 。在 perturb-GEP 中，该模型应该预测所有输入基因的扰动后表达。

用于细胞建模的基因表达预测。此微调目的与 GEP 类似，但根据细胞表示 HHH 预测基因表达值 c 显式促进单元格表示学习。对于输入单元格 i 中的每个基因 j ，我们创建一个查询向量 q ，并使用 q 的参数化内积和单元格表示 hh

c 作为预测的 expression 值：

$$\text{qq} = \text{MLP}_{\text{emb}}(\text{ttt}), \quad \boxtimes$$

$$x = \text{qqq} \cdot \text{WW}^{(i)} \text{Whhh} \quad (14)$$

$$L = \frac{1}{|M|} \sum_{j \in M} (\boxtimes_{x-} x)^2.$$

用于细胞建模的 GEP（GEPC）从方程（5）继承了基因标记嵌入 emb (tg)。在积分任务中，我们使用 $hhhc$ 来自方程（9）而不是 hh

c 。在我们的实验中，我们观察到，与单独使用任何一种方法相比，将 GEP 和 GEPC 相结合可以显着提高性能。

弹性单元相似性。这个微调目标通过利用相似性学习损失来增强单元格表示：

$$L = -(\text{sim}(hhc^{(i)}, h_{\text{refer}}^{(i)}) - \beta)^2, \quad (15)$$

其中 sim 表示余弦相似性函数，而 i 和 refer 表示小批量中的两个单元格。此外， β 表示预定义的阈值，而 ECS 表示弹性单元相似性。这种方法背后的基本思想是增强余弦相似度值高于 β 的对之间的相似性，从而使它们更加相似。相反，鼓励不同的对相距更远。

通过反向反向传播进行域自适应。细胞表征学习受到批次效应的阻碍，批次效应是由测序技术引入的非生物批次差异引起的。为了缓解这个问题，我们使用不同的 MLP 分类器来预测与每个输入细胞相关的测序批次

他^们的单元格表示模型⁽¹⁾内的梯度来修改反向传播过程。这种方法利用了 Ganin 和 Lempitsky 提出的稳健域适应方法的见解。

细胞类型分类。此微调目标旨在利用学习的 cell 表示来注释单个 cells。我们使用单独的 MLP 分类器根据它们的细胞表示 h_{hh} 来预测细胞类型

c .此微调目标通过预测的细胞类型概率和真值标签之间的交叉熵损失 ce 进行优化。

对下游任务进行微调

单元格类型注释。对于单元格类型注释任务，我们在带有真值标签的参考集上微调了模型，并在保留的查询集上验证了注释性能。保留了预训练基础模型和参考集之间的公共基因标记集。在模型微调之前，对基因表达值进行归一化、对数转换和分箱。所有预训练模型权重都用于初始化微调模型，但输出单元类型分类器除外，它是随机初始化的。在训练中使用了所有具有零和非零表达值的基因标记。使用细胞类型分类微调目标来最小化分类损失。

扰动响应预测。为了对扰动预测任务进行微调，我们选择了 HVG 并在模型训练之前对表达式值进行了预处理。预训练模型中的嵌入层和转换器层的参数用于初始化微调模型。在微调过程中，包括所有具有零和非零表达值的基因标记。扰动预测任务中的输入采用了两个显着的变化：首先，我们使用 log1p 转换的表达式值作为输入和目标值，而不是分箱值，以更好地预测该任务的绝对扰动后表达式。其次，我们在每个输入基因位置附加了一个二元条件标记，以指示该基因是否受到干扰。我们采用了 perturb-GEP 微调目标，并进一步修改了训练设置。我们没有使用同一单元格的掩蔽和未掩蔽的表达值作为输入和学习目标，而是使用对照单元格作为输入，并使用扰动单元格作为目标。这是通过将未扰动的对照细胞与每个受扰动的细胞随机配对以构建输入-靶标对来实现的。输入值由对照细胞中的所有基因表达值组成。因此，该模型学会了根据对照基因表达和扰动标记来预测扰动后反应。

集成多个 scRNA-seq 数据集的批量校正。当输入的原始计数矩阵包含来自不同测序批次或技术的多个数据集时，批次效应可能是细胞类型聚类中的主要混杂因素。因此，我们的目标是在整合多个 scRNA-seq 数据集时纠正批次效应，同时保留生物学差异。为了微调此集成任务，保留了预训练基础模型和当前数据集之间的通用基因标记集。我们进一步从公共集合中选择了 HVG 的子集作为输入。我们在模型训练之前对表达式值进行了预处理，类似于 cell type-annotation 任务。所有预训练模型权重都用于初始化微调后的模型。默认情况下，所有具有零和非零表达值的基因标记都用于训练。除了 GEP 和 GEPC 之外，ECS、通过反向反向传播 (DAR) 的域适应和 DSBN 微调目标同时进行了优化，以通过反向反向传播和域特异性归一化来增强单元对比学习和显式批量校正。

scMultiomic 数据的综合表示学习。scMultiomic 数据可能包含不同实验批次的不同测序模式。我们检查了 scMultiomic 数据的两种数据集成设置，配对和镶嵌。在配对设置中，所有样本（细胞）共享所有测序的数据模态。在镶嵌设置中，某些批次共享一些常见的数据模态，但不是全部。由于存在额外的 ATAC 和/或蛋白质标记，我们仅继承了 RNA 数据的训练基因嵌入，并从头开始训练额外的标记嵌入和模型的其余部分。如果数据集包含额外的蛋白质数据，则在训练中仅使用具有非零表达值的标记。否则，默认情况下会同时使用零和非零表达式值。我们使用了一组额外的模态标记来表示每个标记的数据类型（即基因、区域或蛋白质），并促进 GEP 和 GEPC 微调目标中的掩蔽基因和值预测（批次和模态的表示）。默认情况下，该模型使用 GEP 和 GEPC 微调目标进行优化。如果存在多个批次，则包括 DAR 以促进多模态批次校正。

基因调控网络推理。对于图 5 中基于基因嵌入的 GRN 推断，在零镜头设置中，我们基于 k 最近邻从 scGPT 的预训练基因嵌入构建了基因相似性网络。在微调设置中，我们以类似的方式从免疫人类数据集上微调的 scGPT 模型构建了基因相似性网络。继 Ceglia 等人之后，我们进一步在相似性图上进行了 Leiden 聚类，并从由 5 个或更多基因组成的基因簇中提取了基因程序。

对于图 6 中基于注意力的靶基因选择，我们在 Adamson 扰动数据集上微调了 scGPT 血液模型，该数据集由白血病细胞系的 87 个 CRISPR 干扰实验组成。我们在图 6a 中说明了靶基因选择管道。对于每个感兴趣的 perturbed 基因，我们首先通过分别给模型 perturbed 和对照细胞集来检索两组注意力图，perturbed 和 control。请注意，原始注意力分数是从模型最后一个注意力层的所有八个注意力头中获得的。然后，原始注意力分数进行两轮排名标准化，首先按行，然后按列。然后将 8 个注意力头的排名标准化注意力分数取平均值，以输出聚合注意力图。这就得出了用于影响最大的基因选择的最终注意力图。对于每个感兴趣的 perturbed 基因，我们通过在 perturbed 基因列中的最终注意力图中对分数进行排序来选择其影响最大的基因。这反映了一种直觉，即注意力图中的列表明了目标基因对其他基因的影响程度。我们提供了三种影响最大的基因选择设置，即来自控制注意力图谱的 'control'、来自扰动注意力图的 'perturbed' 和来自两者之间差异的 'difference'。从对照注意力图中选择的基因靶标应反映目标基因参与的基础途径，而扰动注意力图反映扰动后效应。这两个注意力图之间的差异应该突出基因网络中从扰动前到扰动后变化最大的边缘。

同样，对于涉及多个转录因子的基于注意力的扩展基因相互作用预测（补充注 7），我们在 Replogle 数据子集中微调了 scGPT 血液模型，并报告了来自“扰动”设置的影响最大的基因。

数据

CELLxGENE scRNA-seq 集合。我们使用 Census API（Census API 可在 <https://chanzuckerberg.github.io/cellxgene-census/python-api.html> 处访问）从 CELLxGENE 门户 (<https://cellxgene.cziscience.com/>) 收集了全人类基础模型预训练的数据。它定期托管和更新在线数据发布。我们使用了 2023 年 5 月 15 日的版本）。我们纳入了测序

我们包括 scRNA-seq 和 snRNA-seq 的测序方案，并在无病条件的样品中过滤。这导致了 3300 万个细胞的测序数据。特别是为了预训练 scGPT 血液模型，我们检索了超过 1030 万份人类血液和骨髓 scRNA-seq 样本 (<https://cellxgene.cziscience.com/>)。通过过滤生物体（即智人）、组织（即血液、骨髓）和疾病（即正常、COVID-19、流感），从 CELLxGENE 中收集了总共 65 个数据集。此外，我们收集了 570 万个各种癌症类型的细胞来训练泛癌模型。

多发性硬化症。MS 数据集是从 EMBL-EBI

(<https://www.ebi.ac.uk/gxa/sc/experiments/E-HCAD-35>) 访问的。数据集中包括 9 个健康对照样品和 12 个 MS 样品。我们将对照样本拆分为参考集以进行模型微调，并将 MS 样本作为评估的查询集。此设置用作 out-of-distribution 数据的示例。我们排除了三种细胞类型：B 细胞、T 细胞和少突胶质细胞 B 细胞，它们仅存在于查询数据集中。最终的单元格计数在训练参考集中为 7,844 个，在查询集中为 13,468 个。从原始出版物中提供的细胞类型标签用作评估的真值标签。数据处理方案涉及选择 HVG 以保留 3,000 个基因。

髓系。髓系数据集可以使用入藏号 GSE154763 从基因表达综合 (GEO) 数据库访问。该数据集由 9 种不同的癌症类型组成，但是，为了训练和评估模型，在参考集中选择了 6 种癌症类型进行训练，而 3 种癌症类型用于查询集。参考集包含髓系癌类型 UCEC、PAAD、THCA、LYM、cDC2 和肾脏，而查询集包含 MYE、OV-FTC 和 ESCA。该数据集也被随机子采样。最终的单元格计数在参考集中为 9,748 个，在查询集中为 3,430 个。

在数据处理过程中选择了 3000 个 HVG。

人类胰腺。人类胰腺数据集包含来自人类胰腺细胞的五项 scRNA-seq 研究的数据，由 Chen 等人再处理用于细胞类型注释任务。这 5 个数据集按数据源分为参考集和查询集。引用集由来自两个数据源的数据组成，查询集包含其他三个数据源。参考集和查询集都有 3000 个基因和从其原始出版物中保留的真值注释。该参考集包含 13 个细胞群（α、β、导管、腺泡、δ、胰腺星状、胰腺多肽、内皮细胞、巨噬细胞、肥大细胞、ε 细胞、雪旺细胞和 T 细胞）的 10,600 个细胞。该查询集包含 11 个细胞组（α、β、导管、胰腺多肽、腺泡、δ、胰腺星状、内皮、ε、肥大和 MHC II 类）的 4,218 个细胞。

PBMC 10k。PBMC 10k 数据集包括从健康供体获得的两批人类 PBMC 的 scRNA-seq 批次。该数据集由 Gayoso 等人再处理，得到 3,346 个差异表达基因。第一批包含 7,982 个单元格，而第二批包含 4,008 个单元格。使用 Seurat 注释的细胞群包括九类，即 B 细胞、CD4T 细胞、CD8T 细胞、CD14 单核细胞、树突状细胞、NK 细胞、FCGR3 单核细胞、巨核细胞等。

免疫人类。免疫人类数据集包含五个 scRNA-seq 数据集：一个来自人类骨髓，四个来自人类外周血。使用了多种测序技术，包括 10x Genomics、10x Genomics (v.2)、10x Genomics (v.3) 和 Smart-seq2。该数据集总共包含 33,506 个细胞和 12,303 个基因。这 10 个不同的批次是根据供体的来源定义的。协调数据包含 16 个单元组。我们使用了经过再处理的数据和 Luecken 等人提供的注释。

鼻周皮层。鼻周皮层数据集包括两个不同的样本，来自 Siletti 等人的一项更大规模的研究，该研究最初包含 606 个高质量样本，涵盖 10 个不同的大脑区域。从鼻周皮层数据集中选择的两个批次中，每个批次都包含大量细胞，第一批由 8,465 个细胞组成，第二批次包含 9,070 个细胞。这些数据集中包含了 59,357 个基因的广泛范围。我们利用了原始研究中提供的 10 种独特细胞类型的注释。COVID-19 的。COVID-19 数据集源自 Lotfollahi 等人的工作，分为 18 个不同的批次，并提供了来自肺组织、PBMC 和骨髓的细胞的多样化表示。该数据集最初包含 274,346 个细胞和 18,474 个基因，为了本研究的目的，该数据集已被子采样为总共包含 20,000 个细胞。我们利用了原始研究提供的注释。对于参考映射评估，我们随机选择了 12 个样本批次作为参考数据集，其他 6 个批次作为查询数据集。生成的参考数据集由 15,997 个单元格组成，查询数据集包含 4,003 个单元格。

亚当森。Adamson 扰动数据集包含来自受 Perturb-seq 扰动的 K562 白血病细胞系的基因表达数据。该数据集包括 87 个由 CRISPR 干扰引起的独特单基因扰动，每个扰动在大约 100 个细胞中复制。

诺曼。Norman 扰动数据集包含来自 Perturb-seq 扰动的 K562 白血病细胞系的基因表达数据。该数据集有 131 个双基因扰动和 105 个单基因扰动。每个扰动在大约 300-700 个单元格中复制。

Reprogle 的。Reprogle 扰动数据集包含具有 CRISPR 干扰的 K562 白血病细胞系中的全基因组扰动。考虑到数据质量，我们保留了与原始研究中确定的具有强转录表型的 1,973 个扰动相匹配的数据子集。我们还删除了 150 个在测序数据中没有扰动基因表达记录的扰动。我们进一步保留了每个扰动的 100 个样品和 2,500 个对照样品。处理后的整个数据集由来自 1,823 次扰动的 171,542 个样本组成，其中 99 个是转录因子的扰动。该测试集由 456 个扰动组成，其中 25 个是转录因子的扰动。

10x 多组 PBMC。10x Multiome PBMC 数据集 (https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k) 包含通过 10x Single Cell Multiome 方案测序的人 PBMC 细胞的配对单细胞 RNA 和 ATAC 数据。在这个数据集中，所有样本都来自同一个健康供体。每个细胞都有基因表达和染色质可及性测量。处理后的数据来自 9,631 个细胞，具有来自 29,095 个基因和 107,194 个染色质区域的读取计数。注释包括 19 个细胞组（CD14 单核细胞、CD16 单核细胞、CD4 幼稚、CD4 TCM、CD4TEM、CD8 幼稚、CD8TEM 1、CD8TEM 2、HSPCs、中间体 B、MAIT、记忆 B、NK、幼稚 B、血浆、T、cDC、gdT 和 pDC）。

BMMC.BMMC 数据集包含通过 CITE-seq 协议测序的 BMMC 的配对单细胞 RNA 和蛋白质丰度测量值。这些细胞来自 12 个健康的人类供体，由该数据集中的 12 个批次组成。处理后的数据代表 90,261 个细胞，来自 13,953 个基因和 134 个表面蛋白的测量值。注释由 45 种详细的免疫细胞亚型组成。

尽快 PBMC。ASAP PBMC 数据集包含四个测序批次，具有三种数据模式（基因表达、染色质可及性和蛋白质丰度）。这四个批次分别代表

这四个批次分别代表 5,023、3,666、3,517 和 4,849 个细胞。在第 1 批和第 2 批中，所有样品都有来自 CITE-seq 的 4,768 个基因和 216 个蛋白质测量值。在第 3 批和第 4 批中，所有样品都有 17,742 个区域和来自 ASAP-seq 的相同 216 个蛋白质测量值。注释包含四个细胞组（B 细胞、髓系、NK 和 T 细胞）。

金隆。该数据集包括 14 个原代人肺腺癌样本，总计 32,493 个细胞。该数据集可通过 Curated Cancer Cell Atlas (<https://www.weizmann.ac.il/sites/3CA/lung>) 公开访问。在我们的研究中，我们将其分为一个包含 10 个样本的参考集和一个包含 4 个样本的查询集。最初，数据集包括 11 个细胞组：B 细胞、树突状细胞、内皮细胞、上皮细胞、成纤维细胞、巨噬细胞、恶性细胞、肥大细胞、NK 细胞、T 细胞和一个未确定的类别。对于预处理，我们删除了未确定的类别，并选择了 3,000 个 HVG 进行下游评估。最终数据集包含 30,472 个单元格。对于参考映射评估，我们随机选择了 10 个患者样本作为参考数据集，其他 4 个患者样本作为查询数据集。生成的参考数据集由 24,746 个单元格组成，查询数据集表示 7,747 个单元格。

基准测试实验设置

scRNA-seq 细胞类型注释。我们在骨髓、MS 和人胰腺数据集上将 scGPT 与两种最近的基于 transformer 的细胞类型注释方法 scBERT 和 TOSICA 进行了基准测试。对于每个数据集，如上一节所述，我们使用参考数据分区进行模型训练和验证。检索查询集上的预测单元格类型标签以进行评估。我们根据四个标准分类指标评估了细胞类型分配性能：准确度、精密度、召回率和宏 F1。准确度、精密度和召回率是针对整体性能进行全局计算的，而宏 F1 是每个类别的平均值，以增加稀有细胞类型的权重。我们还报告了一个按细胞类型划分的具有

“精确率”的标准化混淆矩阵，以获取更多详细信息。有关度量计算的详细信息，请参阅补充说明 12。

scRNA-seq 扰动。我们将 scGPT 与最近的扰动预测方法 GEARS 和线性回归模型进行了比较。线性回归模型采用来自对照细胞的基因表达值和每个基因上扰动状态的二进制编码作为输入特征。该模型使用输入特征的线性组合，通过最小化回归误差来估计每个基因的扰动后表达。为了确保一致性，我们遵循了

Roohani 等人在其基准测试中概述 al.in 预处理步骤。25% 的扰动被分成测试集，在训练期间保持不可见。虽然 CPA 方法在 GEARS 研究中被报告为次优，但我们在实验中证实了这些发现。然而，我们将这种性能差距主要归因于实验设置的差异。CPA 主要旨在学习适用于看不见的细胞类型的共享扰动嵌入，这与我们专注于完全看不见的扰动不同。为了确保公平的基准测试，我们将 CPA 排除在最终比较之外。我们在与以下相同的设置中训练了所有模型，并报告了评估结果。最初，使用所有基因的总数对每个细胞的基因表达值进行标准化，并应用对数转换。随后，我们选择了 5,000 个 HVGs，并将最初未考虑的任何受干扰基因纳入基因集。在实验中，对于所有三个数据集中的单基因扰动，扰动被分割以确保在训练中看不到测试扰动，即训练集中没有细胞经历任何测试扰动。对于 Norman 等人的双基因扰动。数据集中，训练-测试拆分由三种难度递增的场景组成：(1) 训练集中两个看不见的基因中的 0，(2) 两个看不见的基因中的一个，以及 (3) 两个看不见的基因中的两个。评估扰动预测的准确性

检查扰动后细胞状态和对照细胞状态之间的表达变化 ('delta')。我们计算了数据中预测和观察到的表达变化之间的 Pearson 相关性，表示为 Pearson。我们还报告了差异表达最多的基因的前 20 个基因的这些 Pearson 指标。因此，我们提出了差异表达条件的附加评价指标，即 Pearsonon 差异表达基因。有关度量计算的详细信息，请参阅补充说明 12。此外，我们还测试了预测基因表达和真实表达值之间的 Pearson 相关性 (Pearson) 的评估。在补充说明 9 中提出的比较分析中，我们发现基于 delta 的评估比仅仅与真实表达水平的相关性更忠实地表示模型性能。

对于基于聚类的生物学验证，我们首先从 scGPT 模型中检索了每个扰动条件的代表性基因表达谱。scGPT 从样本对照基因表达的单个向量（即大小为 $1 \times M$ 基因）预测每个扰动条件的代表性扰动响应，该向量通过平均数据集中所有对照细胞的基因表达获得。Norman 等人的数据集包含 105 个独特的扰动基因，总共产生 5,565 个独特的扰动组合可供预测。我们将高维预测的扰动响应投影到二维 UMAP 上。我们首先将 UMAP 图与 Norman 等人在原始出版物中发现的功能团进行了比较，其中 236 个扰动实验根据基本事实扰动反应进行了聚类，并通过标记基因表达对其功能作用进行了注释。我们检查了 scGPT 预测的 UMAP 投影与原始论文中发现的功能分组之间的一致性。然后，我们分析了 scGPT 预测的 UMAP 中存在的子集群。当 Leiden 聚类分辨率为 0.5 时，在预测扰动响应的 UMAP 中确定了 54 个子聚类。我们用出现次数最多的 perturbed 基因作为其显性基因对每个簇进行注释。

对于反向扰动预测任务，我们从 Norman 数据集中选择了 20 个基因来构建一个扰动用例，并微调和测试新的扰动模型。这个 20 个基因的子集是通过基于 scGPT 训练-测试拆分最大化训练和测试用例的真值扰动数据的比例来选择的。这个选定的子空间包含 210 个独特扰动组合的 39 个训练案例、3 个验证案例和 7 个测试用例。其余的都是没有实验结果的未见过的案例。反向扰动预测遵循 top-K 检索任务设置：我们使用所有 210 个扰动条件的预测响应作为参考数据，使用来自 7 个测试用例的真值响应作为查询集。目标是检索生成与查询结果最相似的主要扰动条件。对于参考数据，我们不是为每个扰动条件提供单一的代表性基因表达谱，而是从 30 个随机采样的对照细胞中获得预测响应，以增加多样性。这产生了一个包含 6,300 个预测的扰动后基因表达谱的参考数据库。对于 X + Y 扰动的每个测试用例，我们使用来自所有经历 X + Y 扰动的细胞的真值基因表达谱作为查询集。对于 top-K 检索，我们设计了一个涉及两轮选择的集成投票策略。在第一轮中，每个单独的查询单元格通过与参考数据集的欧几里得距离选择其前 K 个最相似的表达谱。在第二轮中，我们根据从所有查询单元格收到的票数对候选扰动条件进行排名。我们报告了前 K 名得票最多的扰动条件，作为集合投票后第二轮排名的预测源扰动条件。我们通过修改后的 top-K 准确率指标评估了正确检索（即精确匹配）和相关检索（即匹配实际扰动组合中的至少一个基因）的检索性能，如补充说明 12 中所述。我们报告了 scGPT 在五次不同运行中的平均分数

我们报告了 scGPT 在不同随机种子的五次运行中的平均分数。我们将 scGPT 的检索性能与 GEARS 和差异基因的检索性能进行了基准测试。为了使用 GEARS 进行基准测试，我们使用 GEARS 预测的表达谱作为参考数据库，并报告了相同的指标，直到前 K 检索。对于差异基因，我们通过扰动细胞和对照细胞之间的 Wilcoxon 秩和检验确定了每个测试扰动的前两个差异表达基因。我们将这两个差异表达基因视为前 1 个预测的双基因扰动组合。我们仅报告了差异表达的前 1 个检索指标，因为随着差异表达基因列表的扩展，扰动组合变得模糊。

scRNA-seq 批量集成。在这项工作中，我们将 scGPT 的性能与其他三种方法（即 Seurat、Harmony 和 scVI）的性能进行了比较。评估涵盖三个集成数据集上的批量校正和细胞类型聚类：COVID-19（参考文献 12）、PBMC 10k 和鼻周皮层。Harmony 和 scVI 在最近的集成基准测试中被强调为性能最好的方法。为确保公平比较，所有方法都提供了相同数量的 1,200 HVG 作为输入。通过考虑所有基因的总计数，对每个细胞的基因表达值进行标准化，然后进行对数转换。整合的细胞包埋是在完成训练后获得的，并用于评估。

使用生物保护指标对整合的细胞包埋进行评估。这些指标包括 NMI、ARI 和 ASW。这些分数衡量派生的像元类型聚类与真实标签之间的一致性。为了便于比较，我们还计算了这些指标的平均值，称为 AvgBIO。此外，我们还报告了批次校正指标以评估批次混合。使用用于批量聚类的平均轮廓宽度的倒数（表示为 ASW）和图形连接度量（表示为 GraphConn）来量化批量校正性能。我们将 AvgBATCH 计算为 ASW 和 GraphConn 的平均值，以总结批量混合性能。此外，我们引入了一个总分，即 AvgBIO 和 AvgBATCH 的加权总和，与最近基准研究中采用的方法一致。有关公制计算的详细信息，请参阅补充说明 12。

scMultiomic 集成。我们在配对和马赛克两种积分设置中对 scGPT 进行了基准测试，并与最近的 scMultiomic 整合方法 Seurat (v.4)、scGLUE 和 scMoMat 进行了基准测试。在配对数据集成实验中，我们在 10x Multiome PBMC 数据集上用 scGLUE 和 Seurat (v.4) 对 scGPT 进行了基准测试，其中 RNA 和 ATAC-seq 数据作为第一个示例。使用相同的 1,200 个 HVG 和 4,000 个高度可变峰的设置作为所有方法的输入。我们进一步在 BMMC 数据集上将 scGPT 与 Seurat (v.4) 进行了基准测试，其中包含配对的 RNA 和蛋白质读数。在这种情况下，我们没有对 scGLUE 进行基准测试以进行公平比较，因为该方法不是专门为模拟蛋白质数据而设计的。同样，相同的 1,200 个 HVG 和所有 134 个蛋白质被用作输入。在马赛克数据集成实验中，我们在 ASAP PBMC 数据集上用 scMoMat 对 scGPT 进行了基准测试。总共使用了 1,200 个 HVG、4,000 个高度可变的峰和所有 216 个蛋白质特征作为两种方法的起始量。在保持输入特征集一致的同时，我们使用了每种方法的自定义预处理管道来规范化表达式值。训练后检索整合的细胞嵌入进行评估。

在配对和镶嵌数据集成设置的所有三个数据集中，我们使用四个生物守恒指标 NMI、ARI、ASW 和 AvgBIO 评估了细胞包埋质量。由于三个数据集中的两个，BMMC（配对）和 ASAP PBMC（马赛克）包含多个批次，因此我们使用三个批次校正指标 ASW、GraphConn 和 AvgBATCH 进一步评估了不同组学批次的混合。还报告了马赛克整合实验的总分。有关度量计算的详细信息，请参阅补充说明 12。

基因调控网络推理。我们针对已知的 HLA 和 CD 基因网络验证了 scGPT 基因嵌入相似性网络。对于每个网络，我们首先通过过滤带有设置前缀的基因名称（即 HLA- 和 CD-）来定义相关基因集。然后，我们从 Reactome 2022 数据库 (<https://reactome.org/>) 中过滤了参与免疫系统 R-HSA-168256 通路的基因。对于 CD 基因，我们使用了来自免疫人类数据集的 HVG 的共同基因集，以便于在预训练模型和微调模型之间进行比较。然后，我们从 scGPT 模型中提取这些选定基因的基因嵌入，并构建了一个 kNN 相似性网络。我们通过选择余弦相似性大于某个阈值的边缘（即 HLA 为 0.5, CD 基因网络为 0.4）来突出强连接的子网络。然后，我们将子网络与免疫系统的已知功能组进行了比较。

此外，我们通过通路富集分析验证了 scGPT 模型提取的基因程序的质量。我们使用每个基因程序作为输入基因列表，并选择具有统计学意义的通路作为“通路命中”。根据执行的测试总数，即基因程序的数量乘以通路测试的数量，使用 Bonferroni 校正 (<https://mathworld.wolfram.com/BonferroniCorrection.html>) 将 P 值阈值调整为 0.05。我们在 Reactome 2022 数据库 (<https://reactome.org/>) 中报告了通路命中的数量。作为基准，我们将结果与从基线共表达图中提取的基因程序进行了比较。共表达图由免疫人类数据集中标准化基因表达的基因之间的 Pearson 相关性定义。为了确保与 scGPT 网络相似的模块化，我们将此图稀疏化为 kNN 相似性网络 ($k = 15$)。遵循与 scGPT 相同的管道，我们通过 Leiden 聚类从基因簇中鉴定了基因程序。作为敏感性分析，我们报告了 scGPT 和共表达方法在 1、5、10、20、30、40、50 和 60 的不同 Leiden 分辨率下的通路命中。我们进一步检查了每种方法在 Leiden 分辨率为 40 时鉴定的常见和独特通路，以更深入地了解性能差异。

我们在 ChIP-Atlas 数据库中验证了基于 scGPT 注意力影响最大的基因选择方法，其中包含经过实验验证的已知转录因子的基因靶标。我们首先通过使用 ChIP-Atlas 交叉检查来自 Adamson 扰动数据集的扰动基因列表，选择了两个由 DDXIT3 和 BHLHE40 编码的示例转录因子。对于每个转录因子，我们通过将注意力在“差异”设置中选择的前 20 个影响最大的基因与经过验证的基因靶标进行比较来验证它们。请注意，在 'difference' 设置中，通过检查扰动注意力图和控制注意力图之间的差异，根据扰动后的变化选择前 20 个基因。通过过滤转录起始位点位于转录因子峰值调用区间 10 kbp 距离内的人类基因 (hg38)，从 ChIP-Atlas 获得真实基因靶标列表。我们报告了前 20 个关注选择的基因靶标与 ground truth 靶基因的重叠数量。

随后，我们通过检查所选前 100 个基因之间的重叠，比较了三种影响最大的基因选择方法（即 control、perturbed 和 difference）。这三个前 100 个基因集的重叠和差异在维恩图中可视化。我们进一步验证了这些顶级基因与 Reactome 数据库中的转录因子一起参与的途径。通路命中和基因重叠的百分比在热图中可视化。

我们进一步验证了对功能相关转录因子组影响最大的基因选择分析。我们使用了 Replogle 等人鉴定了两个示例扰动组，它们分别与 mRNA 聚腺苷酸化和组蛋白乙酰化有关。

然后，我们通过与 ChIP-Atlas 数据库进行交叉检查，从扰动的基因列表中选择转录因子。因此，我们获得了以下具有功能注释的转录因子基因组：(1) 用于 mRNA 多聚腺苷酸化的 CPSF2、CPSF3、CPSF4 和 CSTF3 和 (2) 用于组蛋白乙酰化的 KAT8、MCRS1 和 YEATS4。请注意，KANSL3 和 CPSF1 也被 Replogle 等人包含在 al.as 这两个官能团的一部分。然而，与这些基因对应的转录因子在用于构建注意力图的 1,200 个 HVG 中具有十多个注释基因靶标，因此被从后续分析中删除。与之前对单个转录因子的分析一致，我们首先使用“扰动”设置为每个转录因子选择前 20 个影响最大的基因，并验证了 ChIP-Atlas 数据库中影响最大的 20 个基因的基因靶标。对于每个功能组，我们随后报告了由所有转录因子组合的前 100 个影响最大的基因（以及转录因子）富集的 Reactome 通路。更具体地说，对于 mRNA 聚腺苷酸化，我们从 CPSF2、CPSF3、CPSF4 和 CSTF3 的前 25 个影响最大的基因的联合集中获得了前 100 个影响最大的基因，从组蛋白乙酰化分别从 KAT8、MCRS 和 YEATS4 的前 33 个影响最大的基因的联合集中获得了。为了进行验证，我们比较了丰富的 Reactome 通路，以确定特定于 Replogle 等人的功能注释的术语。我们通过文献检索进一步验证了任何相关途径。如果一个术语满足以下两个标准之一，则根据文献检索认为该术语相关：

- (1) 它包含一个或多个转录因子，或
- (2) 现有文献支持它与功能注释的联系。

实现细节

预训练的基础模型的嵌入大小为 512。它由 12 个堆叠的变压器块组成，每个变压器块有 8 个关注头。全连接层的隐藏大小为 512。在使用 3300 万个细胞对全人模型进行预训练时，我们随机拆分数据，并使用 99.7% 的数据进行训练，使用 0.3% 的数据进行验证。对于其他模型的预训练，包括器官特异性模型和泛癌模型，我们随机拆分数据，并使用 97% 的数据进行训练，3% 的数据用于验证。请注意，在预训练中，只有非零表达的基因才会被输入到模型中。我们将最大输入长度设置为 1,200。对于非零基因数量大于最大输入长度的细胞，每次迭代将随机采样 1,200 个输入基因。我们将要生成的基因比例设置为从 0.25、0.50 和 0.75 三个选项中均匀采样。该模型由 Adam 优化器优化，使用小批量大小 32，起始学习率为 0.0001，每个 epoch 后的权重衰减为 0.9。该模型总共训练了 6 个 epoch。

对于 scRNA-seq 批量集成、细胞类型注释和扰动预测任务，我们使用了从预训练模型继承的相同模型层配置。在微调过程中，我们从 0.0001 的学习率开始，在每个 epoch 后衰减到 90%。对于集成任务，GEP 和 GEPC 的掩码率设置为 0.4，而 ECS 中的参数 β 设置为 0.6。当与其他损失合并时，ECS 的权重为 10。为了将数据集分为训练集和验证集，我们使用了 9: 1 的比率。该模型训练了 15 个 epoch 的固定持续时间，在每个 epoch 之后，在验证集上评估 GEP 损失值。报告的结果与具有最佳验证分数的模型相对应。

对于多组学整合任务，我们从预训练模型中加载了基因嵌入，并对任何新标记（即基因、ATAC 峰或蛋白质）使用相同的嵌入大小 512。主模型被设置为有四个堆叠的变压器块，每个块有 8 个注意力头，隐藏层大小为 512。除预训练的嵌入权重外，所有层都已重新初始化。每个数据集以 9: 1 的比例分为训练集和评估集。我们使用 1 的 DAR 称量

0 表示批量集成。我们在每个 epoch 后使用了 0.001 的起始学习率和 0.95 的权重衰减。我们对模型进行了 25 个 epoch 的固定持续时间训练，批次大小为 16，并同样报告了验证最佳的模型。

我们使用 PyTorch 实现了 scGPT 神经网络模型。

Scanpy Python 库用于基因表达预处理，包括归一化、对数转换和 HVG 选择。我们使用 EpiScanpy Python 库研究染色质可及性数据进行高度可变的峰选择。在 scRNA-seq 批量集成和 scMultiomic 集成任务中，使用 scib.metrics 中的实现计算评估指标。在 cell-annotation 任务中，评估指标是使用 scikit-learn 包实现的。在 GRN 推理任务中，使用 Scanpy 库执行相似性图构建和 Leiden 聚类。使用 GSEApY 软件包实现通路富集分析。其他依赖项包括 torchtext 0.14.0、torch-geometric 2.3.0、flash-attn 1.0.1、pandas 1.3.5、cell-gears 0.0.1、umap-learn 0.5.3、leidenalg 0.8.10 和 wandb 0.12.3。

报告摘要

有关研究设计的更多信息，请参阅本文链接的 Nature Portfolio Reporting Summary。

数据可用性

数据集中报告了所有已使用数据集的来源。预训练数据集可以从 CELLxGENE 普查 2023 年 5 月 15 日发布版 (<https://chanzuckerberg.github.io/cellxgene-census/python-api.html>, <https://cellxgene.cziscience.com/>) 中检索。对于注释任务，MS 数据集是从 <https://www.ebi.ac.uk/gxa/sc/experiments/E-HCAD-35> 访问的。骨髓数据集可通过 GEO 数据库使用入藏号 GSE154763 公开访问。处理后的人胰腺数据集取自 <https://github.com/JackieHanLab/TOSICA>。对于参考映射，Lung-Kim 数据集可通过 Curated Cancer Cell Atlas (<https://www.weizmann.ac.il/sites/3CA/lung>) 公开访问。处理后的 COVID-19 数据集在 <https://github.com/theislab/scarches-reproducibility> 访问。对于扰动预测任务，从以下链接检索了 Norman 和 Adamson 数据集：<https://dataverse.harvard.edu/api/access/datafile/6154020> 和 <https://dataverse.harvard.edu/api/access/datafile/6154417>。Replogle 数据集是从 <https://gwpw.mit.edu> 中检索的。对于批量集成任务，使用 API scvi.data.pbmc_dataset 从 scVI 工具 (<https://scvi-tools.org/>) 中检索 PBMC 10k 数据集。鼻周皮层数据集取自 CELLxGENE 人脑细胞图谱 1.0 版 (<https://cellxgene.cziscience.com/collections/283d65eb-dd53496d-adb7-7570c7caa443>)。对于多组学整合任务，从 <https://scglue.readthedocs.io/en/latest/data.html> 中检索 10x Multiome PBMC 数据集。BMMC 数据集可通过 GEO 数据库的入藏号 GSE194122 访问。ASAP PBMC 数据集是从 tree/main/data/real/ASAP-PBMC <https://github.com/PeterZZQ/scMoMaT> 检索的。对于 GRN 分析，从 <https://doi.org/10.6084/m9.figshare.12420968.v8> 访问处理后的 Immune Human 数据集。所有处理过的数据集都可以在 <https://github.com/bowang-lab/scGPT> 和 <https://doi.org/10.6084/m9.figshare.24954519.v1> (参考文献 73) 访问。

代码可用性

scGPT 的代码库可在 <https://github.com/bowang-lab/scGPT> 和 Zenodo 存储库 (<https://doi.org/10.5281/zenodo.10466117>) 上使用 MIT 许可证公开获得。

引用

55. Sarkar, A. & Stephens, M. 分离测量和表达模型澄清了单细胞 RNA 测序分析中的混淆。Nat. Genet. 53, 770–777 (2021).

56. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. 用于生物医学研究和临床应用的单细胞 RNA 测序实用指南。基因组医学 9, 1-12 (2017)。
57. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: 用于语言理解的深度双向转换器的预训练。计算语言学协会北美分会 2019 年会议论文集 4171–4186 (ACL, 2019)。
58. Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. FlashAttention: 快速且记忆高效的精确注意与 IO-Awareness。Adv. 神经 Inf. 进程。系统 16344–16359 (NeurIPS, 2022)。
59. Wang, S., Li, B. Z., Khabsa, M., Fang, H. & 马, H. Linformer: 线性复杂性的自我注意。预印本 <https://doi.org/10.48550/arXiv.2006.04768> (2020)。
60. Katharopoulos, A., Vyas, A., Pappas, N. & Fleuret, F. 变压器是 RNN: 具有线性注意力的快速自回归变压器。第 37 届机器学习国际会议 5156–5165 (PMLR, 2020)。
61. Liu, Y. RoBERTa: 一种稳健优化的 BERT 预训练方法。<https://doi.org/10.48550/arXiv.1907.11692> 年预印本 (2019 年)。
62. Bubeck, S. 等人。通用人工智能的火花: GPT-4 的早期实验。预印本 <https://doi.org/10.48550/arXiv.2303.12712> (2023)。
63. Liu, C. 等人。用于图像检索的引导式相似性分离。Adv. 神经 Inf. 进程。系统 1556–1566 (NeurIPS, 2019)。
64. Eisenstein, M. 单细胞 RNA-seq 分析软件提供商争先恐后地提供解决方案。Nat. 生物技术。38, 254–257 (2020)。
65. Tran, HTN 等人。单细胞 RNA 测序数据的批次效应校正方法的基准。基因组生物学 21, 12 (2020)。
66. Ganin, Y. & Lempitsky, V. 无监督域适应反向传播。第 32 届机器学习国际会议 1180–1189 (PMLR, 2015)。
67. Ceglia, N. 使用 GeneVector 的互信息定义的密集载体表示鉴定转录程序。Nat. Commun. 14, 4400 (2023)。
68. 单细胞 RNA 测序证明了转移性肺腺癌的分子和细胞重编程。Nat. Commun. 11, 2285 (2020)。
69. Paszke, A. PyTorch: 命令式、高性能深度学习库。Adv. 神经 Inf. 进程。系统 1–12 (NeurIPS, 2019 年)。
70. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: 大规模单细胞基因表达数据分析。基因组生物学 19, 15 (2018)。
71. Danese, A. 等人。EpiScanpy: 集成的单细胞表观基因组分析。Nat. Commun. 12, 5228 (2021)。
72. Fang, Z., Liu, X. & Peltz, G. GSEAp: 一个用于在 Python 中进行基因集富集分析的综合包。生物信息学 39, btac757 (2023)。
73. Wang, C. scGPT 基础模型中使用的处理数据集。Figshare <https://doi.org/10.6084/m9.figshare.24954519.v1> (2024 年)。
74. Cui, H., Wang, C. & Pang, K. scGPT 代码库: 使用生成式人工智能为单细胞多组学构建基础模型。芝诺多 <https://doi.org/10.5281/zenodo.10466117> (2024)。

确认

我们感谢 L. Zhang 在撰写手稿期间的宝贵反馈。图 1a 中的 UMAP 插图是使用 CELLxGENE Annotate (<https://github.com/chanzuckerberg/cellxgene>) 创建的。图 1d 是使用 BioRender (<https://www.biorender.com>) 创建的。这项工作得到了加拿大自然科学与工程研究委员会 (RGPIN-2020-06189 和 DGECR-2020-00294, BW)、CIFAR 人工智能主席计划 (BW) 和大学健康网络 (BW) 的彼得蒙克心脏中心 AI 基金的资助。这项研究的进行, 部分归功于加拿大研究主席计划的资助。H.M. 得到了自然科学与工程研究的博士奖学金的支持

加拿大委员会。

作者贡献

H.C. 开发了这项工作的概念, 并为算法的设计和实施做出了贡献。C.W. 和 K.P. 为算法的设计和实现做出了贡献。H.C.、C.W.、H.M.、K.P. 和 F.L. 为计算实验的分析做出了贡献。H.C. 和 C.W. 起草了手稿的初始版本。H.C.、C.W.、H.M.、K.P.、F.L. 和 B.W. 为该作品的修订做出了贡献。N.D. 为算法的设计做出了贡献。B.W. 为这项工作的构思和设计做出了贡献。

利益争夺

B.W. 是 Vevo Therapeutics 的顾问委员会成员。N.D. 是 Microsoft 的员工, 持有该公司的股权。其余作者声明没有利益冲突。

其他信息:

补充资料 网上版
包含 <https://doi.org/10.1038/s41592-024-02201-0> 提供的补充材料。

信件和材料请求应发送至 Bo Wang。

同行评审信息 Nature Methods 感谢匿名审稿人对这项工作的同行评审做出的贡献。Primary Handling 编辑: Lin Tang, 与 Nature Methods 团队合作。

重印本和权限信息可在 www.nature.com/reprints 上获得。

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection python 3.9, pandas 1.3.5, scanpy 1.9.1, datasets 2.3.0, cell-gears 0.0.1, numba 0.55.1

Data analysis python 3.9, pandas 1.3.5, scanpy 1.9.1, cell-gears 0.0.1, torch 1.13.0, torchtext 0.14.0, umap-learn 0.5.3, leidenalg 0.8.10, scib 1.0.1, flash-attn 1.0.1, torch-geometric 2.3.0, wandb 0.12.3, episcanpy 0.4.0, scikit-learn 1.0.2, gseapy 1.0.6, scGPT codebase (<https://github.com/bowang-lab/scGPT>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All datasets used are described in the Methods section in the manuscript and also as follows.

The pre-training datasets can be retrieved from the CELLxGENE census, release version May 15, 2023 (<https://chanzuckerberg.github.io/cellxgene-census/python->

api.html).

For the annotation task, the multiple sclerosis (M.S.) dataset was accessed from EMBL-EBI (<https://www.ebi.ac.uk/gxa/sc/experiments/E-HCAD-35>). The myeloid (Mye.) dataset is publicly accessible from the Gene Expression Omnibus (GEO) database using the accession number GSE154763. The processed hPancreas dataset was retrieved from <https://github.com/JackieHanLab/TOSICA>. For reference mapping, the Lung-Kim dataset is publicly accessible via the Curated Cancer Cell Atlas (<https://www.weizmann.ac.il/sites/3CA/lung>). The processed COVID-19 dataset was accessed at <https://github.com/theislab/scraches-reproducibility>.

For the perturbation prediction task, the Norman and Adamson datasets were retrieved from the GEARS API via the following links (<https://dataverse.harvard.edu/api/access/datafile/6154020>; <https://dataverse.harvard.edu/api/access/datafile/6154417>). The Replogle dataset was retrieved from the original study at <https://gwpss.wi.mit.edu/>.

For the batch integration task, the PBMC 10K dataset was retrieved from the SCVI's API at scvi.data.pbmc_dataset. The two data batches Perirhinal Cortex dataset were retrieved from the CELLxGENE Human Brain Cell Atlas v1.0 (<https://cellxgene.cziscience.com/collections/283d65eb-dd53-496d-adb7-7570c7caa443>). For the multiomic integration task, the 10X Multiome PBMC dataset was retrieved from scGLUE's processed datasets repository at <https://scglue.readthedocs.io/en/latest/data.html>. The BMMC dataset is accessible from the NeurIPS 2021 Multimodal Single-Cell Data Integration benchmark via accession number GSE194122. The ASAP PBMC dataset was retrieved at scMoMat's github repository at <https://github.com/PeterZZQ/scMoMaT/tree/main/data/real/ASAP-PBMC>.

For GRN analysis, the processed Immune Human dataset was accessed from <https://doi.org/10.6084/m9.figshare.1242096852>.

All processed datasets can be accessed at <https://github.com/bowang-lab/scGPT> and <https://doi.org/10.6084/m9.figshare.24954519.v1>

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<input type="button" value="Not applicable"/>
Population characteristics	<input type="button" value="Not applicable"/>
Recruitment	<input type="button" value="Not applicable"/>
Ethics oversight	<input type="button" value="Not applicable"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The scGPT whole-human model was pre-trained on over 33 million single-cell sequencing samples.
Data exclusions	We included sequencing protocols of scRNA-seq and snRNA-seq and filtered in samples without disease conditions. It aims to maintain data heterogeneity.
Replication	Not applicable
Randomization	The training, validation, and test single-cell samples are randomly assigned.
Blinding	The authors did not access group allocation, since the training, validation, and test groups were randomly assigned by software.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- n/a Involved in the study
- Antibodies
 - Eukaryotic cell lines
 - Palaeontology and archaeology
 - Animals and other organisms
 - Clinical data
 - Dual use research of concern

Methods

- n/a Involved in the study
- ChIP-seq
 - Flow cytometry
 - MRI-based neuroimaging