nature communications

文章 https://doi.org/10.1038/s41467-023-36559-0

使用异构图转换器的单细胞生物网络推理

收稿日期:2022年10月16日

收稿日期:2023年2月6日

Published online: 21 February 2023

查看最新消息

马安军[1,2,7],王小英，李敬贤，王灿坤[1]，肖彤，刘云涛[3]，程浩[1]，介新[4,5]，李阳，常雨舟，李金普，王多林[4,5]，姜月旭，李苏，辛刚，古少鹏，李子海，刘炳强[3,8]，徐东[4,5,6,8]，秦马[1,2,8]

单细胞多组学(scMulti-omics)允许同时对多种模式进行量化，以捕获复杂分子机制和细胞异质性的复杂性。现有工具无法有效推断不同细胞类型中活跃的生物网络以及这些网络对外部刺激的反应。在这里，我们提出了用于scMulti-omics生物网络推断的DeepMAPS。它在异质图中建模scMulti-omics，并使用多头图转换器以鲁棒的方式在局部和全局上下文中学习细胞和基因之间的关系。基准测试结果表明，DeepMAPS在细胞聚类和生物网络构建方面优于现有工具。它还展示了在肺肿瘤白细胞CITE-seq数据和匹配的弥漫性小淋巴细胞淋巴瘤scRNA-seq和scATAC-seq数据中获得细胞类型特异性生物网络的竞争能力。此外，我们还部署了一个具有多种功能和可视化的DeepMAPS web服务器，以提高scMulti-omics数据分析的可用性和可重复性。

单细胞测序，如单细胞RNA测序(scRNA-seq)和单细胞ATAC测序(scATAC-seq)，重塑了细胞异质性的研究，并在神经科学、癌症生物学、免疫肿瘤学和治疗反应性[1,2]方面产生了见解。然而，单个单细胞模态仅反映了遗传特征的快照，并部分描述了细胞的特殊性，导致复杂生物系统中的表征偏差[2,3]。单细胞多组学(scMulti-omics)允许同时定量多种模式，以充分捕捉复杂分子机制和细胞异质性的复杂性。当与强大的计算分析方法配对时，此类分析可以推进各种生物学研究[4]。

现有的scMulti-omics数据综合分析工具，如Seurat[5]、MOFA+[6]、Harmony[7]和totalVI[8]，可以可靠地预测细胞类型和状态，去除批处理效应，并揭示多种模式之间的关系或一致性。然而，大多数现有方法没有明确考虑细胞和模态之间的拓扑信息共享。因此，它们不能在细胞聚集的同时有效地推断出不同细胞类型的活跃生物网络，并且在阐明这些复杂网络对特定细胞类型的外部刺激的反应方面能力有限。

最近，图神经网络(graph neural networks, GNN)在学习单个细胞的低维表示方面表现出了强大的能力，它通过在全局细胞图中预分页相邻细胞特征和构建细胞-细胞关系[9,10]。例如，我们的内部工具scGNN，一个GNN

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA. [2]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [3]School of Mathematics, Shandong University, Jinan, Shandong, China. [4]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA. [5]Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA. [6]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, USA. [7]These authors contributed equally: Anjun Ma, Xiaoying Wang. [8]These authors jointly supervised this work: Bingqiang Liu, Dong Xu, Qin Ma. e-mail: bingqiang@sdu.edu.cn; xudong@missouri.edu; qin.ma@osumc.edu

模型，基于大规模scRNA-seq数据显示了优越的细胞聚类和基因植入性能[11]。此外，具有不同类型节点和边的异构图已被广泛用于建模多关系知识图[12]。它为整合scMulti-omics数据和学习潜在的细胞类型特异性生物网络提供了一个自然的表示框架。此外，最近在建模和集成异构关系的注意机制方面的发展使得深度学习模型具有可解释性，并使细胞类型特异性生物网络的推理成为可能[12,13]。

在这项工作中，我们开发了DeepMAPS(基于单细胞数据的深度学习的多组学分析平台)，这是一个异构图转换框架，用于从scMulti-omics数据中推断细胞类型特异性生物网络。该框架采用了一种先进的GNN模型，即异构图转换器(heterogeneous graph transformer, HGT)，该模型具有以下优点:(1)构建了一个以细胞和基因为节点，以细胞和基因之间的关系为边的一体化异构图。(ii)该模型捕获细胞和基因之间的邻居和全局拓扑特征，以构建细胞-细胞关系和基因-基因关系simultaneously[9,14–16]。(iii)该HGT模型中的注意机制能够估计基因对特定细胞的重要性，从而可用于区分基因贡献并增强生物学可解释性。(iv)该模型是无假设的，不依赖于基因共表达的约束，因此有可能推断出其他工具通常无法发现的基因调控关系。值得注意的是，DeepMAPS与Docker一起实现为无代码、交互式和非编程接口，以减轻scMulti-omics数据的编程负担。

结果

DeepMAPS概述

总的来说，DeepMAPS是一个端到端的、无假设的框架，可以从scMulti-omics数据中推断出细胞类型特异性的生物网络。在DeepMAPS框架中有五个主要步骤(图1和方法)。(i)通过去除低质量细胞和低表达基因对数据进行预处理，然后根据具体的数据类型应用不同的归一化方法。集成细胞-基因矩阵生成，以表示每个细胞中每个基因的组合活性。不同的scMulti-omics数据采用不同的数据集成方法types[5–8]。(ii)从集成矩阵构建异构图，将细胞和基因作为节点，将细胞中存在的基因作为边缘。(iii)建立HGT模型，共同学习细胞和基因的低维嵌入，并生成一个注意力分数来表示基因对细胞的重要性。(iv)基于hgt学习的嵌入和注意分数预测细胞聚类和功能基因模块。(v)在每种细胞类型中推断出多种生物网络，例如基因调控网络(GRN)和基因关联网络。

为了学习细胞和基因的联合表示，我们首先生成一个整合输入scMulti-omics数据信息的细胞-基因矩阵。然后构建具有细胞节点和基因节点的异构图，其中未加权的细胞-基因边缘表示细胞中基因活性的存在，并且通过双层GNN图自编码器从基因-细胞集成矩阵中学习每个节点的初始嵌入(方法)。这种非均匀图为清晰地表示和有机地集成scMulti-omics数据提供了机会，从而可以协同学习具有生物学意义的特征。然后将整个异质图发送到图自编码器，以学习细胞和基因之间的关系，并更新每个节点的嵌入。在此，DeepMAPS采用异构多头关注机制对异构图上的整体拓扑信息(全局关系)和邻居消息传递(局部关系)进行建模。异构图表示学习提供了一种利用DeepMAPS中的转换器同时嵌入细胞和基因的方法。初始图决定了消息传递的路径以及如何在DeepMAPS中计算注意力分数。

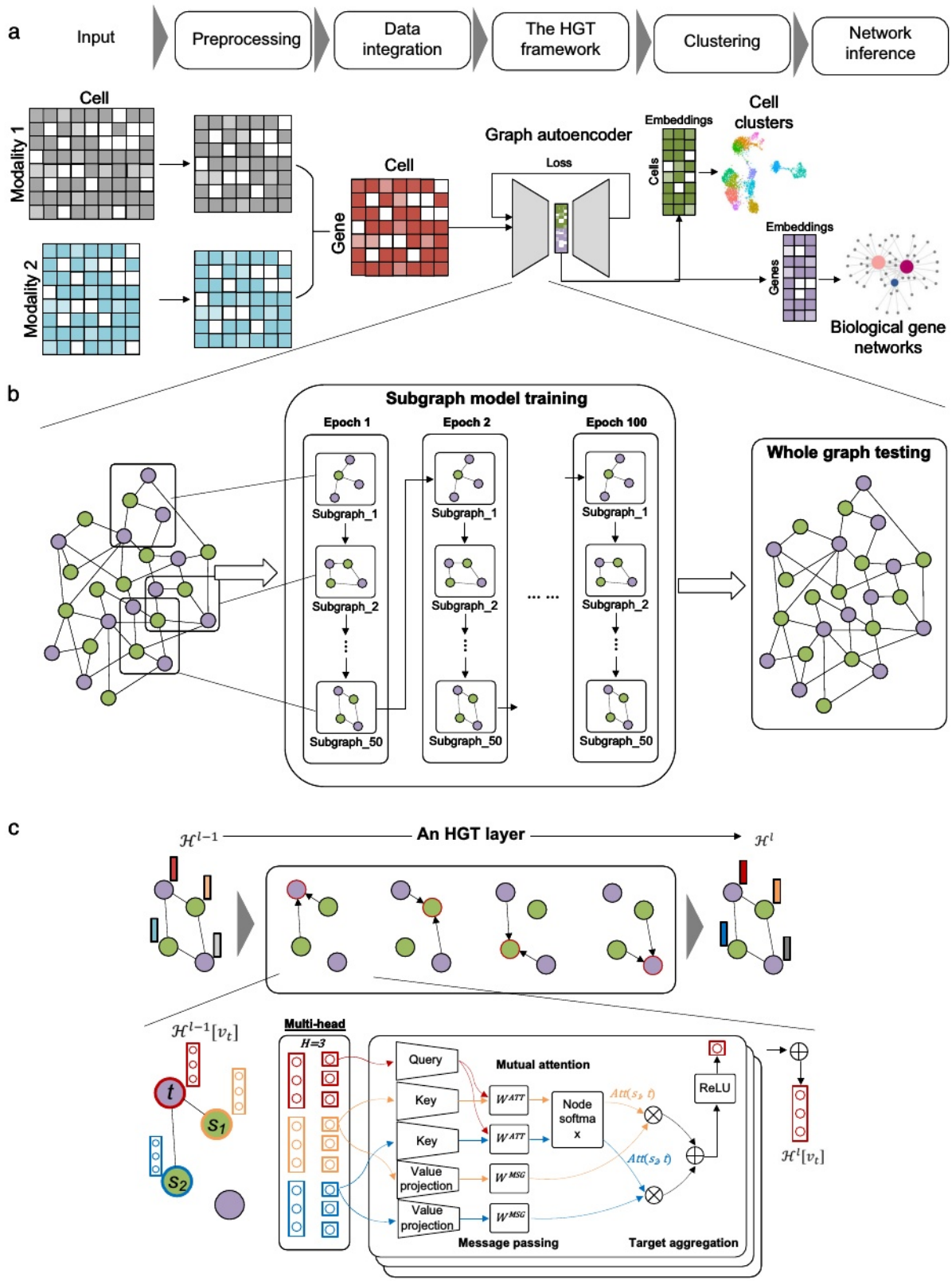
在每个HGT层中，每个节点(细胞或基因)被视为目标，其1跳邻居被视为源。DeepMAPS根据节点嵌入的协同作用(即注意力分数)评估其邻居节点的重要性和可以传递给目标的信息量。因此，具有高度正相关嵌入的细胞和基因更有可能在彼此内部传递信息，从而最大限度地提高嵌入的相似性和差异性。为了使无监督训练过程在大型异构图上可行，对从异构图中采样的50个子图进行DeepMAPS，覆盖至少30%的细胞和基因，以训练不同节点之间的共享参数，这些信息随后用于整个图的测试。作为一个重要的训练结果，给出一个注意力分数来表示一个基因对一个细胞的重要性。一个基因对细胞的高关注分数意味着该基因对于定义细胞身份和表征细胞异质性具有相对重要的意义。这种区分允许在每个细胞簇中构建可靠的基因关联网络，作为DeepMAPS的最终输出。然后，我们建立了一个斯坦纳森林问题(SFP)模型[17]，以识别具有较高注意力得分和与细胞集群相似嵌入特征的基因。SFP模型优化方案中的基因-基因关系和基因-细胞关系反映了基因的嵌入相似性和基因对每个细胞簇的关注重要性。基于其关注得分和嵌入相似性，在表征该细胞簇的身份方面具有最高重要性的基因可以建立一个基因关联网络，这些基因被认为是细胞类型活性的。

DeepMAPS在sc多组学数据的细胞聚类和生物网络推断方面取得了优异的成绩

我们在10个sc多组学数据集上对DeepMAPS的细胞聚类性能进行了基准测试，包括3个多个scRNA-seq数据集(R-bench 1、2和3)，3个ite -seq数据集(C-bench 1、2和3)，以及4个匹配的scRNA-seq和scATAC-seq (scRNA-ATAC-seq)数据集(a -bench 1、2、3和4)(补充数据1)。具体来说，6个R-bench和C-bench数据集在其原始论文中提供了基准注释。而4个A-bench数据集则没有。这些数据集涵盖的细胞数量从3,009到32,029不等;平均读取深度(仅考虑scRNA-seq数据)范围为2,885至11,127;零表达率(仅考虑scRNA-seq数据)为82% ~ 96%(补充数据1)。

我们将DeepMAPS与四种基准测试工具(Seurat v3和v4[5,18], MOFA + [6], TotalVI[8], Harmony[7]和GLUE[19](方法))在平均轮廓宽度(ASW)， Calinski-Harabasz (CH)， 戴维斯- bouldin指数(DBI)和调整后的Rand指数(ARI)方面进行了比较，以评估细胞聚类性能。对于每个数据集，我们在36个参数组合上训练DeepMAPS，包括头的数量、学习率和训练epoch的数量。为了确保公平性，每个基准测试工具也用不同的参数组合进行了调优(Methods)。在ARI (r -bench和c -bench)和ASW (a -bench)方面，比较所有基准工具在所有测试数据集中的最佳性能(图2a，补充图1-3和源数据1-3)。我们还注意到，Seurat是性能第二好的工具，在所有基准数据集中，不同参数选择的差异很小。我们根据参数组合在网格搜索基准测试中的表现选择了每个数据类型的默认参数。将所有基准数据集中平均ARI/ASW分数中位数最高的参数组合作为相应数据类型的默认参数。

https://doi.org/10.1038/s41467-023-36559-0



进行了额外的基准测试实验，以证明在DeepMAPS中选择不同的集成方法是正确的。具体而言，对于scRNA-ATAC-seq数据的分析，我们设计了一种使用基因速度的整合方法，以平衡基因表达和染色质可及性之间的权重，以表征细胞活性和状态(方法)。这种整合过程可以确保数据集的一致性(特别是对于多个scRNA-seq数据)，并生成一个集成矩阵(以基因为行，细胞为列)作为HGT的输入。我们的研究结果表明，对于基准数据1和2 (a -bench 1和- 2)，基于速度的方法在所有网格搜索参数组合上的ASW得分显著(p值<0.05)高于Seurat v 4.0中的加权最近邻(WNN)方法(补充图4和源数据4)。我们认为，包含速度信息后，基因之间的模态权重

https://doi.org/10.1038/s41467-023-36559-0

图1 | DeepMAPS与HGT图解的工作流程。a DeepMAPS的总体框架。从输入的scMulti-omics数据进行细胞聚类和生物基因网络推断包括五个主要步骤。b用HGT模型插入图形自编码器。使用集成的细胞-基因矩阵构建包括所有细胞(绿色)和基因(紫色)为节点的异构图。HGT模型在多个子图(以50个子图为例)上进行训练，这些子图尽可能多地覆盖整个图中的节点。每个子图用100个epoch来训练模型;这样，整个训练过程迭代5000次。然后将训练好的模型应用到整个图中进行学习和更新每个节点的嵌入。c单个HGT层中目标节点的嵌入更新过程示意图。上图中红色圆圈表示目标节点，黑色圆圈表示源节点。箭头表示目标节点和源节点之间的连接。彩色矩形表示不同节点的嵌入。底部面板中细化过程的放大显示了按摩传递过程和注意机制。一个HGT层的最终输出是所有节点的节点嵌入更新。HGT异构图形转换器。

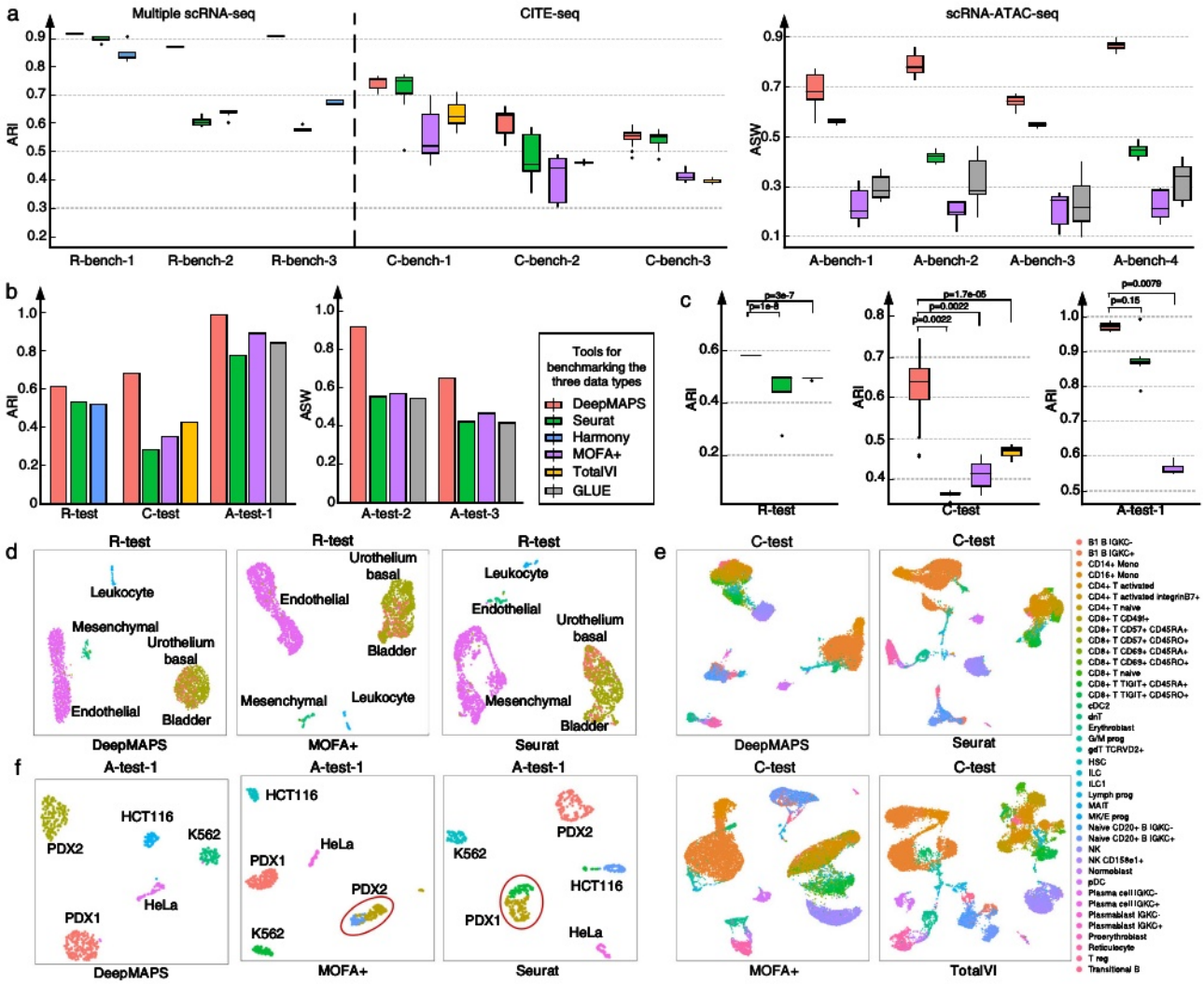


图2 | DeepMAPS在细胞聚类方面的基准测试。a 3个多scRNA-seq数据和3个带基准标签的CITE-seq数据在ARI中对10个数据集的基准细胞聚类结果，以及4个未带基准标签的scRNA-ATAC-seq数据的ASW。每个框显示了使用不同参数设置的工具的最小值、第一四分位数、中位数、第三四分位数和最大ARI或AWS结果(DeepMAPS: n = 96, RNA-RNA和CITE-seq的Seurat: n = 16, RNA-ATAC的36, Harmony: n = 36, MOFA +: n = 36, TotalVI: n = 48, GLUE: n = 72)。点表示异常值。b五个独立数据集的结果比较。不进行重复实验。c DeepMAPS的稳健性检验对三个独立的测试数据集使用基准单元标记的单元簇省略方法。p值的计算基于双尾t检验。每个框显示了对不同数据子集(R-test: n = 5, C-test: n = 20, A-test-1: n = 5)执行的工具的最小值、第一四分位数、中位数、第三四分位数和最大ARI结果。点表示异常值。d-f使用原始细胞标签的DeepMAPS与其他工具之间R-test、C-test和A-test-1数据集的UMAP比较。源数据作为源数据文件提供。ASW平均廓形宽度，ARI调整后的兰德指数。

有助于识别细胞类型的表达和染色质可及性得到了更好的平衡(补充图5)。使用或不使用速度加权平衡法对不同细胞簇中scATAC-seq的模态权重的比较。此外，我们比较了DeepMAPS中不同的聚类方法(即Leiden, Louvain和SLM)，并比较了聚类分辨率(即0.4,0.8,1.2和1.6)对细胞聚类结果的影响。我们发现这些聚类方法之间没有显著差异，Louvain的性能略好于其他两种(补充图6和源数据5)。最后，在选择相同的聚类分辨率时，DeepMAPS的得分高于其他工具。我们还发现，在大多数情况下，更高的分辨率会降低单元格聚类预测得分;因此，我们选择0.4的分辨率作为DeepMAPS的默认参数(补充图7和源数据6-8)。

我们进一步在五个独立的数据集(R-test、C-test、A-test-1、- 2和-3)上独立测试了我们的默认参数选择，通过将我们的结果与使用默认参数的相同基准测试工具进行比较。对于三个带有基准单元标签的测试数据集，DeepMAPS在ARI得分方面表现最好，而对于两个没有单元标签的scRNA-ATAC-seq数据集，比较中的基准测试工具实现了相似的性能(图2b和源数据9)。为了评估DeepMAPS的鲁棒性，对三个带有基准标签的独立测试数据集进行了一个“留下一个”测试(图2c和源数据10)。我们首先基于基准标签去除细胞簇中的所有细胞，然后对剩余的细胞应用DeepMAPS和其他工具。对于每个数据集，DeepMAPS的leave- out结果优于其他具有较高ARI分数的工具，这表明DeepMAPS中使用的消息传递和注意机制以稳健的方式维持了细胞-细胞关系。

在三个具有基准标记的独立数据集上的细胞聚类UMAP显示，在DeepMAPS中获得的潜在表征可以更好地保留scRNA-seq数据的异质性(图2d-f)。对于r检验数据集，所有工具都显示出分离间充质细胞、白细胞和内皮细胞的能力，但未能分离尿路上皮基底细胞和膀胱细胞。然而，DeepMAPS UMAP上的细胞更紧凑，膀胱细胞(红点)的分组优于MOFA +和Seurat(图2)。2 d)。对于c检验数据集，同一簇中的细胞更加有序和紧凑(例如，B细胞簇和NK细胞簇)，而来自不同簇的细胞在DeepMAPS UMAP上更加彼此分开(例如，CD8细胞簇和CD4细胞簇)。(图2 e)。对于A-test-1数据集，DeepMAPS是唯一能够准确分离每种细胞类型的工具。相比之下，Seurat和MOFA +错误地将PDX1或PDX2群体分为两个簇，并包含更多的不匹配(图2f)。

DeepMAPS可以从scMulti-omics数据中推断出具有统计学意义和生物学意义的基因关联网络

我们从中心性评分和功能富集两方面评估了DeepMAPS可以推断的两种生物网络，基因关联网络和GRN。对于r检验数据集(图3a)和c检验数据集(图3b)，我们使用了两个中心性分数，即接近中心性(CC)和特征向量中心性(EC)，这在之前已经使用过

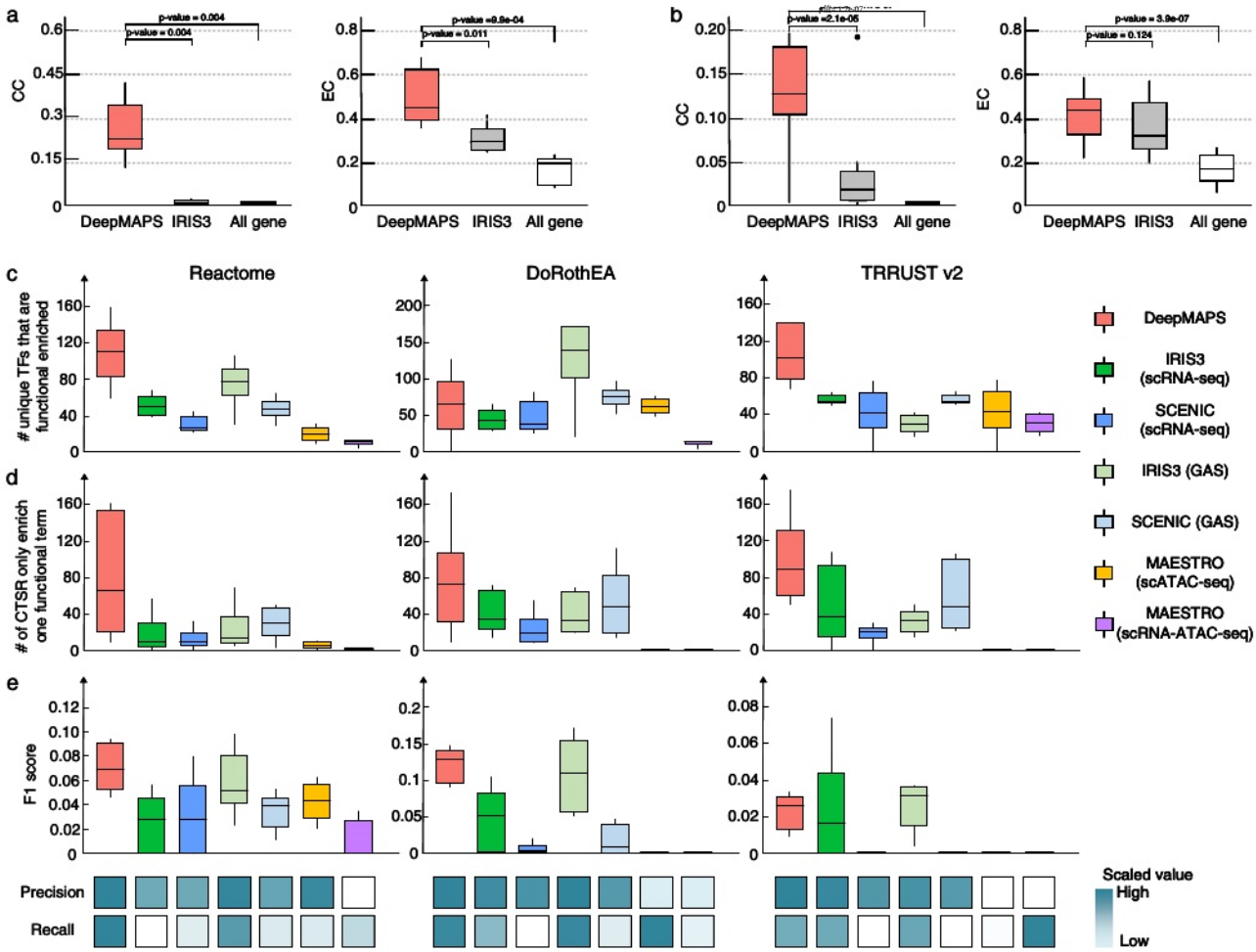


图3 | DeepMAPS基因关联网络推断的评价与比较。a, b使用接近中心性(CC)和特征向量中心性(EC)来表示基因对网络的紧密性和重要性。我们将结果与IRIS3和使用r检验数据集(n = 5) a和c检验数据集(n = 14) b的所有基因的背景网络进行比较，p值使用双尾t检验计算。c三个公共数据库中grn中具有显著生物功能的唯一tf数量的比较。每个方框包含6个scRNA-ATAC-seq数据集的结果(n = 6)。d grn中仅在一个中显著富集的细胞类型特异性调控数的比较e三个数据库中只富集了一个功能/通路的调控子的F1评分比较(n = 6)。选取的6个scRNA-ATAC-seq数据集的精密度和召回率的平均值按最大最小比例缩放，在热图中显示，深蓝色表示高值，浅蓝色表示低值。源数据作为源数据文件提供。图3中的每个方框显示了相应标准的最小值、第一四分位数、中位数、第三四分位数和最大值。CC接近中心性，EC特征向量中心性，CTSR细胞类型特异性调控。

单细胞基因关联网络评估[20]，以比较本次比较中所有工具中鉴定出的基因关联网络。CC反映了网络中节点与所有其他节点的平均连通性，EC反映了基于其连接节点的节点的重要性。CC和EC都可以解释节点在识别可能在网络中发挥更关键作用的基因方面的影响。节点中心性较高的基因关联网络表明，检测到的基因更有可能参与关键和功能性的生物系统。我们还通过计算细胞簇中基因表达的Pearson相关系数，使用数据集中的所有基因构建了基因共表达网络作为背景。将p值= 0.05设置为边缘截止值。我们比较了DeepMAPS中产生的细胞类型活跃的基因关联网络与IRIS3[15]和背景共表达网络中产生的基因关联网络。DeepMAPS构建的网络在r检验和c检验数据集上的平均CC和EC得分明显高于IRIS3和背景共表达网络(源数据11)。我们认为在DeepMAPS中产生的基因关联网络不仅是共表达的，而且对细胞有很大的关注影响;因此，网络中的基因往往对细胞类型更为重要。

为了评估DeepMAPS是否可以识别特定细胞类型中具有生物学意义的grn，我们使用三个公共功能数据库，Reactome[21]、DoRothEA[22]和TRRUST v2[23]，对基本基因调控模块(即regulons[14])进行了富集测试。为了避免比较中的任何偏差，我们将DeepMAPS推断的细胞类型特异性grn与(i) IRIS3和SCENIC在scRNA-seq矩阵上进行了比较，(ii) IRIS3和SCENIC在记录基因活性分数(GAS)的基因细胞矩阵上进行了比较，该基因细胞矩阵记录了DeepMAPS基于基于速度的积分方法计算的基因活性分数(GAS)，(iii) MAESTRO在scRNA-seq和scATAC-seq矩阵上，(iv) MAESTRO在原始scRNA-seq和scATAC-seq矩阵上。使用从人体组织中收集的6个数据集(即A-test-1、A-bench-2、A-bench-3、A-bench-4、A-test-1、A-test-2)。我们首先展示了在DeepMAPS中识别的grn比其他工具包含更多独特的转录因子(TF)调控，除了对DoRothEA数据库的富集(图3c和源数据12)。我们认为一个高度细胞类型特异性的调控子(CTSR)可能只代表一个重要的富集功能;或者，一个通用的调节子可能不正确地包含涉及几个途径的基因。因此，我们比较了不同工具中富集于一种功能/途径的ctsr的数量。在6个scRNA-ATAC-seq数据集的大多数数据集上，DeepMAPS在只富集一种功能/通路的调控子数量和富集F1分数方面优于其他工具(p值<0.05)(图3d、e和源数据12)。对于TRRUST v2数据库的富集测试F1分数，DeepMAPS (F1分数中位数为0.026)使用GAS矩阵(F1分数中位数为0.031)略低于IRIS3。我们还注意到，所有工具在TRRUST v2数据库中都没有实现很好的富集，主要原因是基因数量少(平均而言，一个TF调控10个基因;共计795个tf)。SCENIC还显示出具有竞争力的缩放精度分数(缩放平均值:Reactome为0.47,DoRothEA为0.66,TRRUST v2为0.61)，同时获得较低的缩放召回分数，使得F1分数小于大多数数据集的DeepMAPS。IRIS3和SCENIC在GAS矩阵上的富集结果优于仅使用scRNA-seq数据，这表明整合scRNA-ATAC-seq数据的信息比单独使用scRNA-seq数据对GRN推断更有用。

在PBMC和肺肿瘤免疫CITE-seq数据中，DeepMAPS准确地识别细胞类型并推断细胞间通讯

我们提出了一个案例研究，将DeepMAPS应用于已发表的混合外周血单个核细胞(PBMC)和肺肿瘤白细胞CITE-seq数据集(10x Genomics在线资源，补充数据1)，以证明scMulti-omics在表征细胞身份方面的建模能力。该数据集包括在3485个细胞上测量的rna和蛋白质。DeepMAPS识别了13个细胞群，包括4个CD4+T细胞群(初始、中枢记忆(CM)、组织驻留记忆(TRM)和调节性(Treg))、2个CD8+T细胞群(CM和TRM)、1个自然杀伤细胞群、1个记忆B细胞群、1个浆细胞群、2个单核细胞群、1个肿瘤相关巨噬细胞(TAM)群和1个树突状细胞(DC)群。我们通过可视化标记的maker基因和蛋白质的表达水平来注释每个簇(图4a, b和补充数据2)。与仅使用蛋白质或RNA鉴定的细胞类型相比，我们分离或准确地注释了无法使用个体模态分析表征的细胞群。例如，只有使用整合的蛋白质和RNA才能成功识别DC簇。通过结合从RNA和蛋白质中捕获的信号，DeepMAPS成功地在CITE-seq数据中识别出生物学上合理且有意义的细胞类型。

然后，我们比较了两种细胞类型之间的模态相关性。我们使用记忆B细胞和浆细胞之间的顶级差异表达基因和蛋白质，并对相关矩阵进行分层聚类。结果清楚地将这些特征分层为两个反相关模块:一个与记忆B细胞相关，另一个与浆细胞相关(图4c)。此外，我们发现两个模块中的特征与我们的HGT包埋所捕获的成熟轴显著相关(补充图8和补充数据3)。例如，一个HGT包埋(51st))显示浆细胞和记忆B细胞之间存在显著差异(图4d, e)。在比较EMCD8+T细胞和TRMCD8+T细胞时也观察到类似的结果(图4f)。然而，有可能确定一个具有代表性的HGT嵌入(56th)，该嵌入维持了两组明确分离的嵌入信号(图4,h)。这些结果指向任何两个由多个基因和蛋白质的协调激活和抑制组成的细胞簇，导致细胞状态的逐渐转变，可以通过DeepMAPS潜在HGT空间的特定维度捕获。另一方面，我们生成了具有EMCD8+T细胞、TRMCD8+T细胞、记忆B细胞和浆细胞高注意力得分基因的基因相关网络，并观察到不同的模式(补充图9)。

基于细胞类型和基因和蛋白质表达的原始数据，我们使用CellChat[25]推断细胞-细胞之间的通信，并在多个信号通路内构建不同细胞类型之间的通信网络(图4i)。例如，我们在肺癌肿瘤微环境(TME)中观察到DC(源)和TRMCD4+T细胞(靶)之间存在CD6-ALCAM信号通路。先前的研究表明，抗原呈递dc上的ALCAM与T细胞表面的CD6相互作用，有助于T细胞活化和proliferation[26–28]。作为另一个例子，我们发现在TAM(源)和TRMCD8+T细胞(靶)之间的相互作用过程中参与了NECTIN-TIGIT信号通路，这得到了先前报道的支持，即TAM上表达的NECTIN (CD155)在肺癌TME[29,30]中与CD8+T细胞上的表面受体TIGIT相互作用时可能具有免疫抑制作用。

DeepMAPS在弥漫性小淋巴细胞淋巴瘤scRNA-seq和scATAC-seq数据中识别特异性grn

为了进一步扩展DeepMAPS对GRN推理的能力，我们使用了10x基因组学网站(10x基因组学在线资源)上提供的单细胞多组ATAC +基因表达数据集。原始数据来源于一名诊断为淋巴结淋巴弥漫性小淋巴细胞淋巴瘤(DSLL)的患者的14566个快速冷冻的腹内淋巴结肿瘤细胞。我们通过基于RNA速度平衡细胞中基因的每种模式的重量来整合基因表达和染色质可及性(图5a和方法)。为了构建tf -基因连接，我们考虑了基因表达、基因可及性、TF-motif结合亲和力、峰-基因距离和tf编码基因表达。发现的基因

https://doi.org/10.1038/s41467-023-36559-0

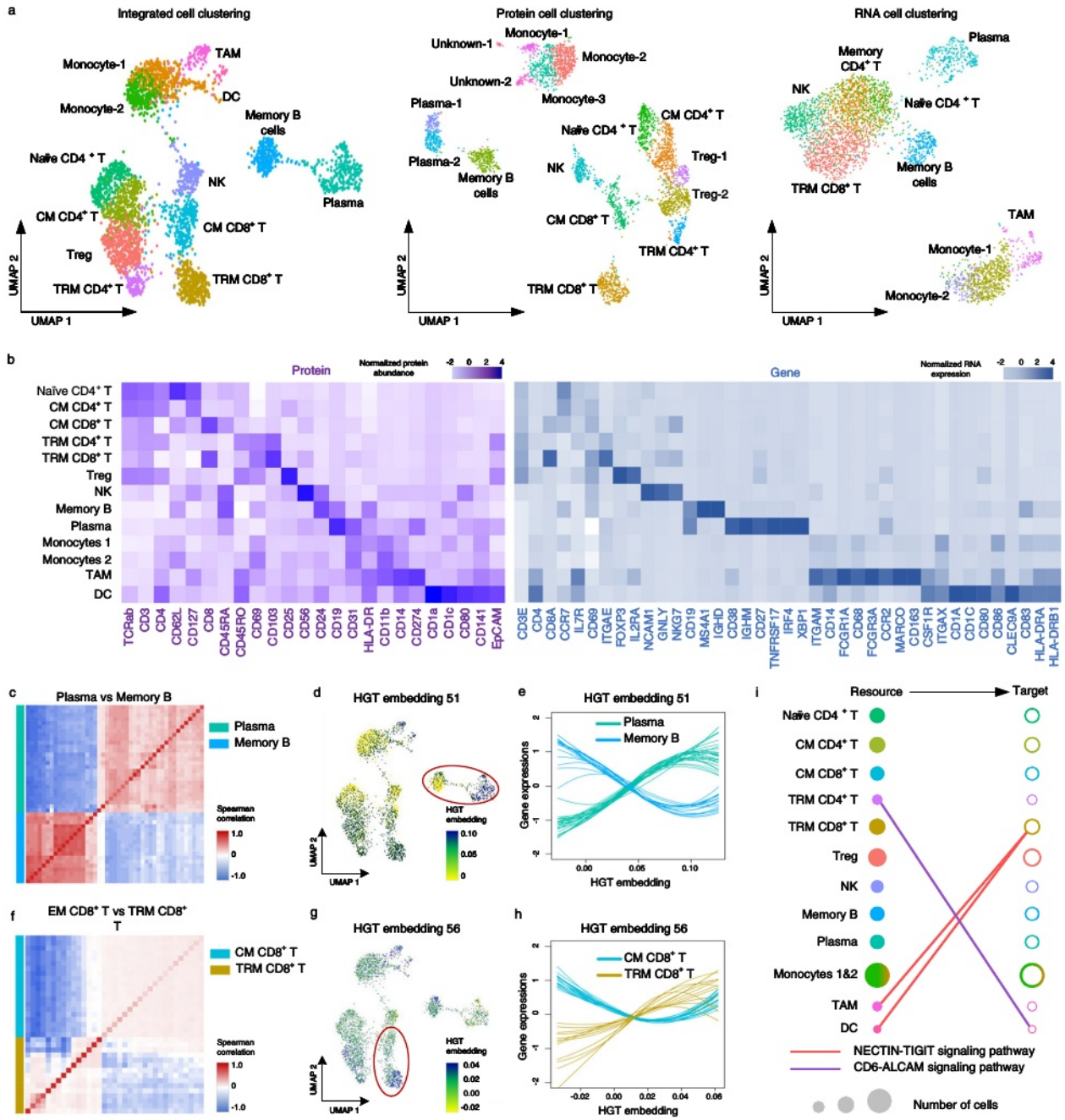


图4 | DeepMAPS识别pbmc和肺肿瘤白细胞的CITE-seq数据的异质性。a DeepMAPS细胞聚类的UMAPs结果来自集成的RNA和蛋白质数据，仅蛋白质数据和仅RNA数据。基于筛选的标记蛋白和基因对细胞聚类进行注释。b固化标记蛋白和决定细胞聚类和注释的基因的热图。c浆细胞和记忆B细胞中顶级差异表达基因和蛋白的Spearman相关比较热图。d UMAP用第51次包埋着色，表示在浆细胞和记忆B细胞中的不同包埋表征。e c中顶级差异表达基因和蛋白的表达作为第51次包埋的函数，观察浆细胞和记忆B细胞之间的模式关系。每条线代表一个基因/蛋白，用细胞类型着色。对于每个基因，使用黄土平滑函数绘制一条线

在细胞中相应的嵌入和缩放基因表达。f-h对56th包埋进行类似的可视化，比较EMCD8+T细胞和TRMCD8+T细胞c-e。i两个信号通路，NECTIN和ALCAM，显示了两个细胞簇之间预测的细胞-细胞通信。填充的圆圈(资源簇具有高表达的配体编码基因)和未填充的圆圈(靶簇具有高表达的受体编码基因)之间的联系表明信号通路的潜在细胞-细胞通信。圆圈的颜色代表不同的细胞簇，大小代表细胞的数量。两个单核细胞组合并。TRM组织驻留记忆，CM中枢记忆，TAM肿瘤相关巨噬细胞，HGT异质图转换器。

在细胞簇中受相同TF调控的基因被归为一个调控子。我们认为中心性得分较高的调控子对细胞簇的表征有更显著的影响。同一TF在不同细胞群中调控的调控子被比较不同的调控活性。那些具有显著较高的调节子活性评分(RAS)的被认为是细胞簇中细胞类型特异性的调节子。

DeepMAPS在DSLL数据中识别出11个细胞簇。基于筛选的基因标记对所有细胞簇进行人工注释(图5b和补充数据4)。两个类DSLL细胞簇(DSLL state-1和

https://doi.org/10.1038/s41467-023-36559-0

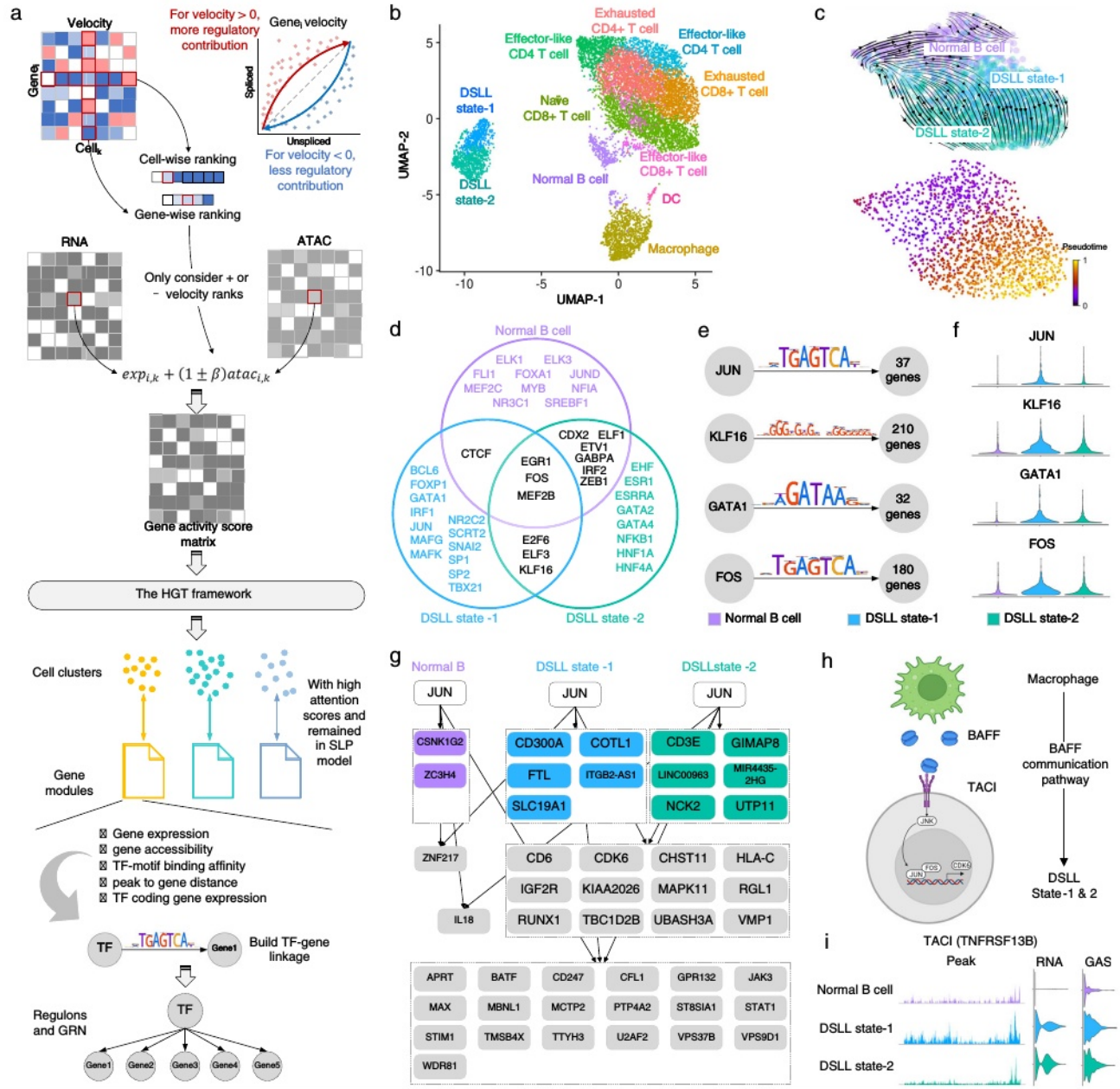


图5 | DeepMAPS识别DSLL子网中的特定gm。scRNA-ATAC-seq数据DeepMAPS分析的概念说明。模式首先基于速度加权平衡进行整合。然后使用集成的GAS矩阵构建异构图作为HGT框架的输入。然后使用具有高关注分数的细胞簇和基因模块来构建tf -基因连接并确定每个细胞簇中的规则。b UMAP为DeepMAPS聚类结果。基于策划的标记基因，对细胞簇进行了手动注释。c显示了基于正常B细胞的RNA速度和两种DSLL状态的观察和推断的未来状态(箭头)(上图)。基于速度的轨迹分析显示了从上到右下的伪dotime(下图)。d在三个聚类中各选择20个TF，代表中心性得分最高的前20个规则。颜色表示在每个聚类中唯一识别或共享的规则在不同的集群之间。e与其他集群相比，DSLL状态1的调控子活性存在显著差异。Motif的形状和调控基因的数量也被显示出来。f四种调控子在三个簇间比较的调节子活动的小提琴图。g三个集群中JUN (DSLL状态-1中差异最活跃的调控)的下游调控基因。h使用CellChat从基于gas的细胞-细胞通信预测中识别出的BAFF信号通路的说明。发现巨噬细胞与两种DSLL状态之间均存在BAFF信号通路。它进一步激活JUN调控，使CDK6等基因的转录成为可能。图由BioRender.com制作。i BAFF信号通路受体TACI编码基因TNFRSF13B的ATAC峰值、RNA表达和GAS水平。源数据以Source data文件形式提供。

状态2)被观察到。对三个B细胞簇(正常B细胞和两种DSLL状态)进行的基于RNA速度的伪时间分析假设这两种DSLL状态来自正常B细胞，并且状态1的衍生时间早于状态2，尽管这两种状态似乎部分混合(图5c)。我们进一步选择了三个细胞群中每个细胞群中调节中心性得分最高的前20个TFs(图5d和源数据13)。有趣的是，这些TFs显示了正常和两种DSLL状态之间的差异，并推断了两种DSLL状态下的不同调节模式。对于所有三种B细胞簇共有的调控，EGR1、MEF2B和FOS在正常B细胞和DSLL细胞中都具有转录活性，并负责调节B细胞的发育、增殖和生发

中心formation[31–34]。E2F6, ELF3和KLF16被确定为仅在两种DSLL状态中共享，并在tumorigenesis[35–40]中报告了其作用。此外，编码活化蛋白-1 (AP-1)区室的JUN、MAFK和MAFG[34,41,42]被发现在dsl状态1中具有活性，而编码NF-κB蛋白复合物亚基[43,44]的NFKB1被发现在dsl状态2中具有活性。

我们构建了一个由四种细胞类型特异性调控(JUN, KLF16, GATA1和FOS)组成的GRN(图5e和补充图10)，在dsl状态1中RAS显著高于正常B细胞和dsl状态2(图5f)。据报道，KLF16促进前列腺[39]和胃癌细胞[40]的增殖。FOS和JUN是AP-1家族中的转录因子，调节多种类型淋巴瘤的肿瘤发生[34,41,42,45]，GATA1对造血至关重要，涉及多种血液系统疾病和恶性肿瘤[46,47]的失调。当我们放大单个调控子时，也观察到不同的调控模式(图5和补充图11-12)。作为DSLL state-1中最活跃的调控子，JUN调控了5个独特的下游基因和12个与DSLL state-2共有的基因。下游基因，包括CDK6[33,34]、IGF2R[48]和RUNX1[49]，在DSLL中对细胞增殖、存活和发育功能至关重要。

此外，我们进一步在DSLL细胞中建立了上游细胞-细胞通信信号通路和下游调节机制之间的联系。我们发现巨噬细胞与两种DSLL状态之间通过B细胞活化因子(BAFF)信号通路进行细胞间通信，基于CellChat[25]的集成GAS矩阵，其中包括BAFF作为巨噬细胞上的配体，TACI(跨膜激活剂和钙调节剂以及亲环蛋白配体相互作用物)作为DSLL细胞上的受体(图5h)。BAFF信号对正常B细胞的存活和成熟至关重要[50,51]，而异常有助于恶性B细胞抵抗凋亡[52,53]。我们观察到TACI编码基因TNFRSF13B在两种DSLL状态下的表达明显更高，而相应的染色质可及性在状态1下保持高峰(图5i)。据报道，在与配体结合后，TACI转导信号并最终激活AP-1[54,55]和NF-κB[56,57]转录复合物，用于B细胞的下游信号传导。JUN (AP-1的一个亚基)被确定为state-1中最特异性和最关键的调节因子，负责细胞增殖和调节下游癌基因，如CDK6，据报道在多种类型的dsll以及其他血液学malignancies[58–60]中促进癌细胞的增殖。很明显，BAFF信号首先出现在DSLL state-1中，并触发JUN调控机制的激活，导致JUN的高调控活性。JUN调控子在DSLL中明确加速增殖和肿瘤发生，导致DSLL进入更终末期的分化阶段(state-2)。因此，state-1包括经历快速细胞增殖和分化的细胞，从正常B细胞过渡到成熟的DSLL。简而言之，DeepMAPS可以构建grn并识别细胞类型特异性调控模式，从而更好地了解患病亚群中的细胞状态和发育顺序。

DeepMAPS为分析scMulti-omics数据提供了一个多功能和用户友好的门户网站

由于单细胞测序数据的复杂性，在过去三年中开发了更多的web服务器和dockersyears[61–73](supplementary data 5)。然而，这些工具大多只提供细胞聚类和差异基因分析等最小的功能。它们不支持scMulti-omics数据的联合分析，尤其缺乏对生物网络推断的足够支持。另一方面，我们记录了DeepMAPS和基准工具在不同数据集上的运行时间，单元格数从1000到160,000(补充数据6)。深度学习模型(DeepMAPS和TotalVI)比Seurat和MOFA +的运行时间更长。为此，我们提供了一个无代码、交互式和非编程的接口，以减轻scMulti-omics数据的编程负担(图6a)。web服务器支持使用DeepMAPS分析多个RNA-seq数据、CITE-seq数据和scRNA-ATAC-seq数据(图6b)。一些其他方法，例如Seurat，也被纳入作为用户方便的替代方法。服务器包含三个主要步骤——数据预处理、单元聚类和注释以及网络构建。此外，DeepMAPS服务器支持实时计算和交互式图形表示。用户可以注册一个账户，拥有自己的工作空间来存储和分享分析结果。除了提到的这些进步之外，DeepMAPS网络服务器还强调了一个额外的功能，用于阐明特定细胞类型对外部刺激的复杂网络。用户可以上传包含表型信息的元数据文件(例如，经过处理和未经过处理的细胞)，选择并重新标记相应的细胞(例如，经过处理的CD8+ T细胞和未经处理的CD8+ T细胞)。通过这种方式，DeepMAPS将预测CD8+ T细胞中的治疗相关网络。在https://bmblx.bmi.osumc的在线教程中给出了示例。edu/tutorial。

讨论

DeepMAPS是一个深度学习框架，实现了异构图表示学习和图转换器，用于从scMulti-omics数据中研究生物网络。通过构建包含细胞和基因的异构图，DeepMAPS可以同时识别它们的联合嵌入，并能够在完整的框架中推断细胞类型特异性生物网络以及细胞类型。此外，异质图转换器的应用在可解释的统一多关系中对细胞-基因关系进行建模。通过这种方式，可以大大缩短图中的训练和学习过程，从而从更远的距离考虑细胞的影响。

通过联合分析基因表达和蛋白质丰度，DeepMAPS在PBMC和肺肿瘤白细胞的混合CITE-seq数据中准确地鉴定和注释了13种细胞类型，这些数据基于无法使用单一模式完全阐明的标记物。我们还证明了在DeepMAPS中识别的嵌入特征捕获了统计上显著的信号，并在原始信号有噪声时将其放大。此外，我们基于在两个集群中推断出的基因关联网络，确定了DC和TRMCD4+T细胞之间具有生物学意义的细胞-细胞通信途径。对于scRNA-ATAC-seq，我们采用基于RNA速度的方法来动态整合基因表达和染色质可及性，从而增强了对细胞簇的预测。利用这种方法，我们确定了正常B细胞和两种DSLL发育状态之间不同的基因调控模式。我们进一步阐明了细胞间通讯和下游grn之间的深层生物学联系，这有助于表征和定义DSLL状态。所鉴定的tf和基因可以作为进一步验证的潜在标记物和DSLL治疗的免疫治疗靶点。

虽然在分析scMulti-omics数据方面有优势和改进的性能，但DeepMAPS的能力仍有进一步提高的空间。首先，考虑到异构图表示(可能包含数十亿条边)的复杂性，超大数据集(例如，超过100万个细胞)的计算效率可能是一个实际问题。此外，建议在gpu上运行DeepMAPS，这会导致潜在的再现性问题。不同的GPU型号有不同的浮点数，这可能会影响训练过程中损失函数的精度。对于不同的GPU模型，DeepMAPS可能会产生略有不同的单元聚类和网络结果。最后，当前版本的DeepMAPS是基于基因和细胞的二部异构图。需要单独的预处理和整合步骤，才能将不同的模态转移到基因中，以便整合到细胞-基因矩阵中。要完全实现端到端的框架

https://doi.org/10.1038/s41467-023-36559-0

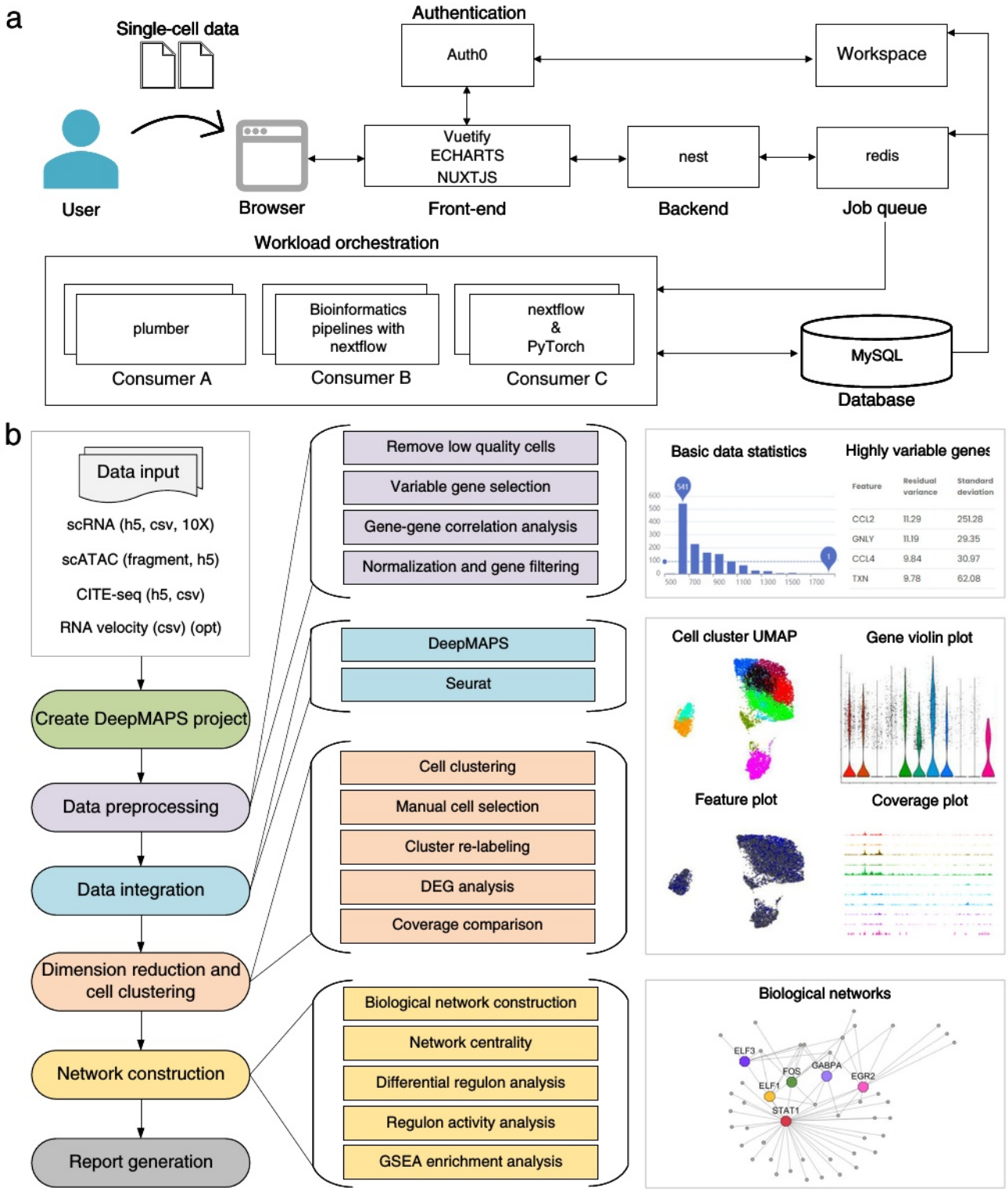


图6 | DeepMAPS门户网站的组织结构。DeepMAPS的软件工程图和框架概述。b服务器的流水线示意图，包括主要步骤(左;颜色表示不同的步骤)，详细的分析(中)，特色的图形和表格(右)。

scMulti-omics分析，二部图可以扩展到多部图，其中不同的模式可以包括为不相交的节点类型(例如，基因，蛋白质或峰区)。这样的多部异质图还可以包括基于知识的生物信息，例如已知的分子调控和一个图中的两个以上模态。然而，通过包含更多的节点类型，计算负担将以几何方式增加，这需要在未来专门发现模型和参数优化。

总之，我们将DeepMAPS评估为scMulti-omics数据综合分析和细胞类型特异性生物网络推断的先驱研究。它可能会为单细胞生物学中的深度学习部署提供不同的愿景。随着DeepMAPS网络服务器的开发和维护，我们的长期目标是创建一个基于深度学习的生态社区，用于存档，分析，可视化和传播ai就绪的scMulti-omics数据。

方法

数据描述

我们包括10个公共数据集(即R-bench-1-3, C-bench-1-3和A-bench-1-4)用于DeepMAPS和现有工具之间的网格测试基准测试，另外5个数据集(即R-test-1, C-test-1和A-test-1-3)用于优化参数的独立测试。这两个病例研究分别使用了人PBMC和肺肿瘤白细胞CITE-seq数据和10×淋巴结scRNA-seq和scATAC-seq数据。所有数据均可公开获取(补充数据1和数据可用性)。

数据预处理与整合

DeepMAPS的分析以多个scRNA-seq(多基因表达矩阵)、CITE-seq(基因和表面蛋白表达矩阵)和scRNA-ATAC-seq(基因表达和染色质可及性矩阵)数据的原始计数矩阵为输入。对于每个数据矩阵，我们将模态表示定义为行(基因、蛋白质或峰区)，并将细胞定义为整篇论文的列，除非有例外情况。在每个数据矩阵中，如果一行或一列包含的非零值少于0.1%，则会被删除。数据质量控制由Seurat v3进行，包括但不限于总读取次数、线粒体基因比例、黑名单比例。下面展示了额外的数据预处理和集成方法。

多个scRNA-seq数据。基因表达矩阵是对数归一化的，使用Seurat v3[18]从每个矩阵中选择前2000个高度可变的基因。如果矩阵中少于2000个基因，则会选择所有基因进行整合。然后，我们应用Seurat中广泛使用的典型相关分析(CCA)来对齐这些矩阵并协调scRNA-seq数据，从而得到矩阵X =fx$_{ij}$ | i = 1,2，⋯,我;J = 1,2，..⋯， JÞ为J细胞中的I基因。

CITE-seq数据。基因和表面蛋白表达矩阵是对数归一化的。选择前I个$_1$高度可变的基因(I_1= 2000)和I个$_2$蛋白(I_2是矩阵中蛋白的总数)。然后将这两个矩阵垂直连接起来，得到矩阵X = x.x$_{ij}$ | i = 1,2，⋯，ðI $_1$ +I_2Þ;j = 1,2，⋯， JÞ为J细胞中的I_1基因和I_2蛋白(可处理为J细胞中的I_1+ I_2= I基因)。对X进行中心对数比(CLR)转化，如下所示:

$$CLR(x_{ij}) = \log\left(1 + \left(\frac{x_{ij}}{\exp\left(\frac{\sum_{i\in\mathcal{I}_j}\log(1+x_{ij})}{|\mathcal{I}_j|}\right)}\right)\right) \quad (1)$$

其中Z_j表示细胞j中非零基因的索引集，|⊠|表示集合中元素的个数。

scRNA-ATAC-seq数据。基因表达矩阵^{XR} = fxRij | i = 1,2，⋯,我;J = 1,2，⋯， I基因和J细胞的Jg是对数归一化的。然后使用左截断混合高斯(LTMG)模型，通过对细胞群体中潜在调控信号如何控制基因表达的建模，为所有细胞中的每个基因提供定性表示[74]。具体来说，如果基因i可以用所有J细胞上的G_i高斯分布来表示，这意味着可能存在G_i调节信号来调节该基因。可以生成与$_{XR}$具有相同维数的矩阵$_{XR'=}$ fxRij ' g，其中基因表达用$x_{R_{ij}}$’ =1,2，⋯的离散值标记。$G_{i。}$

染色质可接近性矩阵表示为X ^{A} = fxAkj | k = 1,2，⋯K;J = 1,2，⋯， Jg表示J个细胞的K个峰区。我们根据MAESTRO[24]中描述的方法，将X^{A}中的峰区标注为相应的基因。具体来说，峰值k到基因i的调控电位权重w_{ik}是根据基因组中峰值k到基因i的距离来计算的:

$$w_{ik} = \begin{cases} 0, d_{ik} > 150\text{ kb or peak k located in any nearby genes} \\ \frac{1}{Length(exon)}, \text{peak } k \text{ located at the exon regions of the gene } j \\ 2^{-\frac{d_{ik}}{d_0}}, else \end{cases} \quad (2)$$

其中d_{ik}为k峰中心与基因i转录起始位点之间的距离，d_0为距离的半衰减(设为10 kb)。k峰对基因j的调控电位权重w_{ik}为$_{dik}$通常用$_{2d0}$计算。对于d_{ik}> 150kb的峰值，w_{ik}将小于0.0005，因此为了方便，我们将其设置为0。在MAESTRO中，对于位于外显子区域的峰，d_0为0，因此根据公式w_{ik}应为1，w_{ik}用基因i的总外显子长度归一化。原因是，在大量ATAC-seq数据中，我们观察到许多高表达基因在外显子区域也会有ATAC-seq峰，这主要是由于时间PolII和其他转录机制的结合。基于这一观察结果，为了更好地将模型与基因表达相匹配，MAESTRO添加了来自外显子区域的信号。然而，由于reads往往比较短的外显子更容易位于较长的外显子上，因此为了使背景reads的可能性正常化，它将外显子上的总reads按每个基因的总外显子长度标准化。最终，细胞j中k峰对基因i的调控电位得分可计算为$r_{ik\,|\,j}$=w_{ik}×xAkj。然后，将调节同一基因的峰的调控电位得分相加，将scATAC-seq矩阵X A转化为基因调控电位矩阵:

$$x_{ij}^{A'} = 0 + \sum_k r_{ikj}, \quad (3)$$

给grisisetotheregulatorypotentialmatrixx $_{A\,0=}$ fxAij $_{0\,|\,i}$ = 1,2，⋯,我;J = 1,2，⋯， Jg在X ^{R}的J细胞中表达相同的I基因。

我们假设一个基因在细胞中的活性是由具有不同贡献的基因表达和基因调控活性决定的。与Seurat v4(加权近邻)[5]中直接基于表达和染色质可及性值确定的贡献权重不同，我们假设基因对细胞的表达和染色质可及性的相对贡献是动态的，而不是静态的，不能通过细胞快照准确确定。RNA速度由细胞中未剪接和剪接的mRNA的丰度决定。未剪接mRNA的数量由基因调控和基因转录率决定，剪接mRNA的数量由未切片mRNA和降解mRNA的差异决定。我们推断，对于细胞中具有正RNA速度的基因，具有更高的驱动基因转录的潜力。因此，它们与染色质可及性相关的调控活性在定义当前快照细胞中的整体转录活性方面比基因表达具有更大的影响。对于负速度的基因，转录速率趋于减速;因此，染色质可及性对转录活性的影响小于基因表达。

速度矩阵X ^{V} = fxVij | i = 1,2，⋯我;J = 1,2，⋯， Jg是使用scVelo生成的，默认参数为[75]。考虑到某些基因可能无法获得有效的速度或调节电位值，我们同时从X R、X $_{R'}$、$_{X\,A'}$、X $_{V}$四个矩阵中去除X $_{A'}$或X $_{V}$中所有行为零的基因:在不丧失一般性的情况下，我们仍然使用I和J来表示这些新矩阵的大小。此外，考虑到在解释细胞中基因的速度时可能存在的偏差，我们使用LTMG表示xijR ' 2 f1,2，⋯。，$_{G}$ G离散xVij。对于基因i，设$_{jg}$为基因i具有的细胞集

相同的LTMG信号g 2 f1, 2, ⋯,G$_i$g。对于J$_g$中的细胞，我们使用这些细胞中基因i的平均速度来代替原来的速度。为了计算基因i在细胞j中的速度权β，我们首先提取X$_{Vi}$ = × X$_{Vi1}$, xVi2, ⋯。xViJ Þfor基因i在所有细胞中的速度和X$_{Vj}$ = x.x$_{V1j}$, xV2j, ⋯xVI2j Þfor细胞j中所有基因的速度。然后，对于X$_{Vi}$，令X$_{Vi}$$^{+}$ = fxVij ⁞ X$_{Vij}$ > 0,j = 1, 2, ⋯，Jg对于所有具有基因i和X V正速度的细胞i = fxVij ⁞ X Vij < 0,j = 1,2, ⋯，对于基因i为负速度的所有细胞，Jg。同理，对于X Vj，令X Vj + = fxVij ⁞ X Vij > 0;I = 1,2, ⋯细胞i和X$_{V中所有正速度基因}$的Ig j = fxVij ⁞ X$_{Vij}$ < 0;I = 1,2, ⋯对于xVij > 0，根据速度值从高到低排序X Vi +和X Vj $_{+,}$ 从1开$_{始排}$^{序}，计算速度权值为:

$$\beta^{+} = \sqrt{\left(|x_i^{V+}| - a - 1\right)^2 + \left(|x_j^{V+}| - b - 1\right)^2} \quad (4)$$

其中a为xVij在X$_{Vi}$$^{+}$中的排名，b为xVij在X$_{Vj}$$^{+}$中的排名。

同理，对于xVij < 0，根据排名从0开始的速度绝对值从高到低排列X$_{Vi}$和X Vj$_{,\ 计算}$速度权值为:

$$\beta^{-} = \sqrt{\left(x_j^{V-}| - a - 1\right)^2 + \left(|x_j^{V-}| - b - 1\right)^2} \quad (5)$$

其中a是xVij在X$_{i\ V中的排名}$，b是xVij在X中的排名V$_{j}$。

我们现在生成一个基因活性矩阵X$_G$ = fxGijg，基于速度权重整合基因表达和染色质可及性。xGij为细胞j中基因i的基因活性评分(GAS):

$$x_{ij}^{G} = \begin{cases} x_{ij}^{R} + (1+\beta^{+})x_{ij}^{A'}, \text{for } x_{ij}^{V} > 0 \\ x_{ij}^{R} + (1-\beta^{-})x_{ij}^{A'}, \text{for } x_{ij}^{V} < 0 \\ x_{ij}^{R} + x_{ij}^{A'}, \text{for } x_{ij}^{V} = 0 \end{cases} \quad (6)$$

构建基因-细胞异质图

为了简化符号，我们现在将上一节生成的任何积分矩阵重新定义为X =fx$_{ij}$ ⁞ i = 1,2, ⋯。,我;J = 1,2, ⋯，Jg带有I基因和J细胞。x$_{ij}$表示细胞j中基因i的规范化表达(用于多个scRNA-seq和CITE-seq)或GAS(用于scRNA-ATAC-seq)。我们通过两个自编码器计算基因和细胞的初始嵌入。我们使用两个自编码器分别生成细胞和基因的初始嵌入。细胞自编码器将每个细胞的基因维数从1维降至512维，最终降至256维;基因自编码器将每个基因的细胞维度从J维降低到512和256维。因此，每个细胞和基因都具有256维的相同初始嵌入。较低维度的数量被优化为每个数据集不同的超参数。输出层是与X相同维数的重构矩阵X^，单元格自编码器的损失函数是X和X的均方误差(MSE):^

$$loss = MSE(X, \hat{X}) = \sum_i (X - \hat{X})^2 \quad (7)$$

基因自编码器从所有细胞中学习基因的低维特征，它有一个编码器、潜在空间和一个类似于细胞自编码器的解码器，而输入X^{T}是X的转置矩阵。基因自编码器的损失函数为

$$loss = MSE\left(X^T, \hat{X^T}\right) = \sum_j \left(X^T - \hat{X^T}\right)^2, \quad (8)$$

其中，X ^T为输出层中与X^{T}具有相同维数的重构矩阵。

定义1(异构图):异构图是具有多种类型节点和/或多种类型边的图。我们将异构图表示为G = ðV, E, a, RÞ，其中V表示节点，E表示边，a表示节点类型联合，R表示边类型联合。

定义2(节点类型和边类型映射函数):我们定义τðvÞ: V !A和⊠ðeÞ: E !R分别作为节点类型和边类型的映射函数。

定义3(节点元关系):对于由边e$_{1,2}$链接的节点对v$_1$和v$_2$，v$_i$和v$_j$之间的元关系表示为langleτðv$_i$Þ,⊠ðe$_{i,j}$ Þ，τðv$_j$ Þi。

给定积分矩阵X，我们构建了一个具有两种节点类型(细胞和基因)和一种G G边缘类型(基因-细胞边缘)的二部基因-细胞异质图$_G$。V = VC∪VG，其中V = vi ⁞ i = 1,2, .. .。,我表示所有基因，V^{C}=v$_{Cj}$ ⁞ j = 1,2, ⋯。，J表示所有细胞。E =e$_{i,j}$表示vGi和vCj之间的边。对于x$_{ij}$> 0，对应边的权值ωðe i,j Þ= 1，否则，ωðe$_{i,j}$ Þ= 0。

通过异质图转换器进行联合嵌入

我们提出了一个无监督HGT框架[12,13]来学习所有节点的图嵌入并挖掘基因和细胞之间的关系。HGT的输入是积分矩阵X，输出是细胞和基因的嵌入以及表示基因对细胞重要性的关注分数。

定义4(目标节点和源节点):当执行HGT对该节点进行信息聚合和嵌入更新时，V中的一个节点被视为目标节点，表示为v$_t$。如果E中的v$_s$和v$_t$之间有一条边，则将节点视为源节点，表示为v$_s$,v$_s$≠v$_t$，为方便起见，将其表示为e$_{s,t}$。

定义5(目标节点的邻域图):诱导出目标节点v$_t$的邻域图from G并表示为com-作为G = (V, E, ‘R’),V0 =fv$_t$G∪Nv$_t$,Nv$_t$的完整邻居集，E0 = fei, j2e ⁞ v$_{i,\ vj2v}$ ' g, A '标记目标和源节点类型，R '表示目标-源边。e$_{s,t}$∈E '表示v$_s$和v$_t$之间的边。由于G中只包含一种边类型，因此vs和vt的节点元关系表示为τv$_s$，⊠e$_{s,t}$，τv$_t$。

1. 多头注意机制与向量的线性映射。设H^{l}表示lthHGT层的嵌入(l = 1,2, ⋯)L)。将v$_t$和v$_s$在lth层上的嵌入记为H^{l}½v$_t$和H^{l}½v$_s$。采用多头机制将H^{l}½v$_t$和H^{l}½v$_s$等分为H头。多头注意允许模型共同关注来自不同嵌入子空间的信息，每个头部可以并行运行一个注意机制，以减少计算时间。对于lthHGT层中的hth头，H^{l}½v$_t$是从H^{l-1}½v$_t$和H^{l-1}½v$_s$更新的。其中H^{0}½v$_t$和H^{0}½v$_s$分别是v$_t$和v$_s$的初始嵌入。应用三个线性投影函数将节点嵌入映射到hth中

h

th

向量。具体来说，Qlinear$_{τ}$ðv$_{vt}$Þfunction将v$_t$映射到h查询向量Q^{h}v$_t$，维度为R^{d}!R^{d}，其中d为$_h$

Hl的维数1½2v$_t$和d$_H$是每个头部的向量维数。类似地，K linearτhðvsÞ和V linearτhðvsÞfunction映射

$_{第}$源节点vs映射到hkey向量Kh vs和第h值$_h$向量Vv$_s$。

$$Q^h(v_t) = Q_{linear^h_{\tau(v_t)}}\left(\mathcal{H}^{(l-1)}[v_t]\right) \quad (9)$$

$$K^h(v_s) = K\,linear^h_{\tau(v_s)}\left(\mathcal{H}^{(l-1)}[v_s]\right) \quad (10)$$

$$V^h(v_s) = Vlinear^h_{\tau(v_s)}\left(\mathcal{H}^{(l-1)}[v_s]\right) \quad (11)$$

每种类型的节点都有一个唯一的线性投影，以最大限度地模拟分布差异。

2. 异质相互关注

为了计算v_t和v_s之间的相互关注，我们引入了attention算子来估计每个v_s到v_t的重要性:

$$Attention(v_s, e_{s,t}, v_t) = \underset{\forall v \in \mathcal{N}(v_t)}{\mathrm{Softmax}}\left(\Big\|_H\left(ATT\_head^h(v_s, e_{s,t}, v_t)\right)\right) \quad (12)$$

注意函数可以描述为将查询向量和一组键值对映射到每个节点对的输出e = ðv$_s$，v_tÞ。v_t和v_s的整体注意力是所有头部的注意力权重的串联，然后是一个softmax函数。 ‖ ðÞ是拼接函数。H

ATT_head^hv_s，e_s，$_t$，v_t项是v_t和v_s之间的hth头部关注权重，可以通过以下方式计算:

$$ATThead^h(v_s, e_{s,t}, v_t) = \left(K^h(v_s) W^{ATT}_{\phi(e_{s,t})} Q^h(v_t)^T\right) \cdot \frac{\mu\langle \tau(v_s), \phi(e_{s,t}), \tau(v_t)\rangle}{\sqrt{d}}, \quad (13)$$

测量查询和关键字之间的相似性，其中W $_{att<e:1>}$ ðes,t Þ2 RdH×Hd是捕获元关系特征的转换矩阵。ðÞT是调换功能和μ是之前张量表示 the⊠significanεe∓v_s为每个节点元关系,⊠$e_{s,t}$,τv_t,作为一种自适应缩放的注意。注意头的连接产生v_s和v_t之间的注意系数，然后是Eq. 12中的Softmax函数。

3. 异构消息传递

Message操作符用于提取v_s中可以传递到v_t的消息。多头消息的定义是:

$$Message(v_s, e_{s,t}, v_t) = \Big\|_H\left(MSGhead^h(v_s, e_{s,t}, v_t)\right) \quad (14)$$

th

对于每条边v_s,v_t， h头部消息MSG_headh vs, es, $_t$, vt定义为:

$$MSGhead^h(v_s, e_{s,t}, v_t) = V^h(v_s) W^{MSG}_{\phi(e_{s,t})} \quad (15)$$

其中头部h中的每个源节点vs通过线性投影V $_h$ vs: Rd !×RH $_{d.}$ W $_{MSG}$⊠ð $_e$ Þ2 RdH $_{×H也是}$一个类似于W⊠^{ATT}ð$_e$Þ的变换矩阵。

$_{s\ t}$

4. 目标特异性聚合

为了更新v_t的嵌入，lthHGT层的最后一步是将该层获得的邻居信息Hfl vt聚合到嵌入Hl 1vt的目标节点中。

$$\widetilde{\mathcal{H}^l}[v_t] = \underset{\forall v_s \in \mathcal{N}(v_t)}{Aggregate}\left(Attention(v_s, e_{s,t}, v_t) \cdot Message(v_s, e_{s,t}, v_t)\right) \quad (16)$$

$$\mathcal{H}^l[v_t] = \theta\left(ReLU(\widetilde{\mathcal{H}^l}[v_t])\right) + (\theta - 1)\mathcal{H}^{l-1}[v_t], \quad (17)$$

其中θ为可训练参数，ReLU为激活函数。最终嵌入v_t是通过所有L个HGT层叠加信息得到的，在DeepMAPS中L设为2。

5. 基因对细胞注意力的测定

我们呼出HGT过程完成后，基因i对最后一层HGT细胞j的最终关注得分$a_{i,j}$:

$$a_{i,j} = \sqrt{\sum_h ATT\_head^h(i,j)^2} \quad (18)$$

子图的HGT训练

为了提高HGT模型在大型异构图(数万个节点和数百万条边)上的效率和能力，我们部署了一种改进的HGSampling方法，用于子图选择和多个小批HGT训练[12]。对于含有I个基因和J个细胞的图G，子图的并集应覆盖基因和细胞节点的%(设为30%)节点，以保证训练功率。因此，采样器从给定的异构图G构建许多小子图(在DeepMAPS中为50个)，并使用多个gpu以不同批次将子图馈送到HGT模型中。每个图应该包含一个% ×I=50个基因，和一个% × J=50个细胞。取一个细胞j作为目标节点v_t，它的邻居v_s2Nv_t，对应基因i，作为源节点，我们计算在边$e_{s,t}$上的概率为:

$$Prob(e_{s,t}) = \frac{x(v_s, v_t)}{\sum_{v \in \mathcal{N}(v_t)} x(v, v_t)}, \quad (19)$$

其中，xðv$_s$,v$_t$Þ=x$_{ij}$表示基因i在细胞j中在积分矩阵x中的表达或GAS值，因此，对于每个目标节点v_t，我们基于采样为vt随机选择$_{一个\% ×I=50}$ a% ×J=50个邻居基因

概率Prob es,t。HGT超参数，如W $_{atte}$ Þ， W ⊠ $_{msg}$ - e $_{s,t}$ s,tÞ,

和θ，将在一个历元中从子图1到50依次训练和继承。子图训练以无监督的方式使用图自编码器(GAE)进行。HGT是编码器层，嵌入的内积是解码器层。我们将GAE的损失函数计算为重构矩阵X^和积分矩阵X的Kullback-Leibler散度(KL):

$$loss = KL\left(softmax(\hat{X}), softmax(X)\right) \quad (20)$$

如果损失被抑制或达到100次epoch，则子图训练将完成，以先发生者为准。

细胞簇中活性基因模块的确定

预测细胞簇。我们部署了Louvain聚类(Seurat v3)来预测从最终HGT层生成的细胞簇细胞嵌入HL vc。Louvain聚类的分辨率由多个HGT超参数组合的网格搜索测试确定，我们将聚类分辨率设置为默认值0.4。

识别细胞聚类-活性基因关联网络。我们使用SFP模型[17]来选择对细胞簇表征有高度贡献的基因，并构建细胞簇活性基因关联

网络。定义一个新的异构图Ge = V, Ee, v2 V $_G$∪V $_C$,Ee2Ee1∪Ee2，其中Ee1表示基因-基因关系，Ee2表示基因-细胞关系。对应边的权值ωð ee$^1_{i1,i2}$Þof vGi,2 V $_G$和vGi 2 V $_G$是vGj$_1$和vGj$_2$之间HGT嵌入的Pearson关联。的权值

vGi 2v $_G$和vGj 2v $_C$对应的边ωee2i,j为

最终注意力得分$a_{i,j}$。只有ωee$_{i1,i2}$> 0:5和

1

ωee$^{2}_{i,j}$>μa$_{i,j}$+ sd - ða i,j Þ，其中μ()表示平均值，sdðÞ表示a$_{i,j}$的标准差，将保留在一个单元簇内。然后将剩余边的权重进行ma$_{x-m}$in归一化，以确保权重最大的边被重新缩放为0，权重最小的边被重新缩放为1。

设Z为通过Louvain聚类预测的聚类数量，V$^{C_{1/2}z}$= fvC½Z g为与Z = 1,2，…的聚类标签中的单元集对应的节点集。Z。然后，我们使用下面定义的组合优化模型来表述这个问题

$$\min_{\tilde{E}\subseteq\tilde{E}^{1}\cup\tilde{E}^{2}} \sum_{e\in\tilde{E}} \omega(e)$$

酸处理

$$\mathcal{L}\left(v_{i_1}^{C}, v_{i_2}^{C}\right)=1, \forall v_{i_1}^{C}, v_{i_2}^{C} \in V^{C[z]}, z=1,2,\ldots,Z \quad (21)$$

其中，1 / $_{vCj1}$,vCj2 Þis一个二进制指示函数，表示两个单元节点vCj1和vCj2在Ge中是否可以通过eELj1,$_{j2}$ = fe2i1,$_{j1}$, ee1i1,$_{i2}$, ee1i2,$_{i3}$…, ee1it1,$_{it}$,eei2t, $_{j2g}$路径。设EeL = fEeL $_{j1,j2g}$为连接vjC1和vjC2的$_{EeLj1,j2}$的完整集合。组合优化模型旨在以最小的边权总和识别连接vCj$_{1}$和vCj$_{2}$的路径。我们认为在集群z的SFP结果中保留的基因网络是集群活性基因关联网络。

利用scRNA-ATAC-seq数据构建grn

对于SFP产生的细胞簇活性基因关联网络中的基因，一组tf q = 1, 2, …，则可以将Q分配给基因。通过在scATAC-seq数据中查找TF结合位点与峰区之间的比对来检索TF-峰关系，并且在之前计算潜在调节分数r$_{ik\,|\,j}$时建立了峰基因关系(Eq. 3)。我们设计了一个调节强度(RI)分数s$_{i,j,q}$来量化细胞j中TF q对基因i的调节强度:

$$\vartheta_{ij|q}=\sum_{k} b_{qk}^{A}\cdot r_{ik|j} \quad (22)$$

其中b$^{A}_{qk}$为TF q到k峰的结合亲和力评分。结合亲和力评分分为三步计算:(a)我们从JASPAR中检索基因组浏览器跟踪文件，该文件存储了每个TF的所有已知TF结合位点。在JASPAR中，p值得分计算为-log$_{10}$(p) ×100，其中0对应于p值为1,1000对应于p值<10−10。我们去掉了p值小于500的TF结合位点。(b)如果TF结合位点与scATAC-seq谱中的任何峰区重叠，则保留该位点，否则将其移除。(c)将对应的p值得分除以100。我们认为受同一TF调控的一组基因是一个调控子。

我们计算了一个调控子活性评分(RAS) rðq,zÞof在细胞簇z中一个受TF q调控的基因的调控子:

$$\vartheta(q,z)=\frac{\sum_{i\in I_q}\sum_{j\in C[z]} x_{ij}\cdot\vartheta_{ij|q}}{I\cdot J} \quad (23)$$

其中I$_{q}$表示细胞簇z中受TF q调控的基因。我们使用Wilcoxon秩和检验来识别基于RAS的簇中的差异活性调控。如果不同细胞簇之间经bh调整的p值小于0.05，且log fold变化大于0.10，则我们认为该调控子在该簇中具有差异活性，并将其定义为细胞类型特异性调控子(CTSR)。

细胞簇中的GRN是通过合并细胞簇中的规则来构建的。GRN中TF节点v的特征向量中心性(c$_{v}$)定义为:

$$c_v=\alpha_{\max}(v) \quad (24)$$

其中α$_{max}$为GRN加权邻接矩阵的最大特征值所对应的特征向量。c$_{v}$排名越高的tf被视为master tf(默认前10名)。

标杆量化和统计

基于基准数据的网格搜索参数聚类测试。为了确定不同数据类型下HGT的默认参数，我们对HGT参数进行了网格搜索测试，包括嵌入数对和头数对(91/13、104/13、112/16和128/16)、学习率(0.0001、0.001和0.01)和训练epoch(50、75和100)。总共测试了36个参数组合。对于三种数据类型，分别在三个基准数据上进行HGT参数训练，并根据三个数据集的最高中位数得分(多个scRNA-seq数据和带基准标签的CITE-seq数据为ARI，未带基准标签的scRNA-ATAC-seq数据为AWS)选择默认参数组合。

为了评估DeepMAPS与其他提出的scMulti-omics基准工具的性能，我们将DeepMAPS与Seurat (v3.2.3和v4.0, https://github.com/satijalab/seurat)、MOFA + (v1.0.0, https://github.com/bioFAM/MOFA2)、Harmony (v0.1, https:// github.com/immunogenomics/harmony)、TotalVI (v0.10.0, https:// github.com/YosefLab/scvi-tools)和GLUE (v0.3.2, https://github)进行了比较。com/gao-lab/GLUE)。由于对不同数据类型的集成能力，DeepMAPS在多个scRNA-seq数据上与Seurat v 3.2.3和Harmony进行了比较，在CITE-seq数据上与Seurat v4.0.0、MOFA+和TotalVI进行了比较，在scRNA-ATAC-seq数据上与Seurat v4.0.0、MOFA+和GLUE进行了比较。对于每个基准测试工具，网格搜索测试也应用于组合参数，例如用于细胞聚类的维数和聚类分辨率。

然后将为每种数据类型选择的默认HGT参数组合应用于其他数据集(一个多个scRNA-seq，一个CITE-seq和三个scRNA-ATAC-seq数据)进行独立测试。所有基准测试工具都使用它们的默认参数。

为了展示在DeepMAPS中选择综合方法和细胞聚类方法的基本原理，我们通过替换其他几种方法来评估细胞聚类性能。具体来说，对于数据集成，我们用Harmony集成方法(多个scRNA-seq)代替CCA方法，用Seurat加权最近邻方法(CITE-seq)代替CLR方法，用Seurat加权最近邻方法代替速度加权方法，不使用速度(scRNA-ATAC-seq)。对于细胞聚类方法，我们用Leiden和smart local moving (SLM)方法代替了Louvain聚类。我们还比较了聚类分辨率(使用0.4、0.8、1.2和1.6)对Deep-MAPS中细胞聚类结果的影响。每次比较都是在网格搜索测试中使用的所有36个参数组合上进行的。对于没有velocity的DeepMAPS，我们简单地将来自scRNA-seq数据的基因表达矩阵与来自scATAC-seq数据的基因潜在活性矩阵相加，考虑velocity对细胞i中基因j引入的平衡权为1。

调整后的兰德指数(ARI)。ARI通过考虑当前和以前的随机排列调整的聚类中分配的所有对样本来计算相似度。构建一个关联表来总结具有b个元素(单元)的两个单元标签列表之间的重叠，以计算ARI。每个条目表示两个标签列表之间共有的对象数量。

ARI可以计算为:

$$ARI = \frac{\frac{J_a+J_b}{C_n^2} - E\left[\frac{J_a+J_b}{C_n^2}\right]}{\max\left(\frac{J_a+J_b}{C_n^2}\right) - E\left[\frac{J_a+J_b}{C_n^2}\right]} \quad (25)$$

E½:为期望，J_a为分配给与基准标签相同的单元簇的单元数;J_b是作为基准标签分配给不同细胞簇的细胞数;C^2_n是从集群中总共n个细胞中选择两个细胞的组合。

平均轮廓宽度(ASW)。与ARI不同，它需要已知的基础真值标签，剪影分数是指一种解释和验证数据簇内一致性的方法。剪影权重表示一个对象与其聚类(内聚)相比与其他聚类(分离)的相似程度。轮廓宽度的范围从−1到+1，其中高值表示对象与其聚类匹配良好。轮廓分数sil(j)可以通过以下方式计算:

$$sil(j) = \frac{|n(j) - m(j)|}{\max\{m(j), n(j)\}} \quad (26)$$

其中mðjÞ为细胞j与同一簇中所有其他细胞之间的平均距离，nðjÞis为I到j不属于的最近簇中所有细胞的平均距离。我们计算了所有细胞的平均剪影分数作为ASW来表示数据集的剪影分数。

Calinski-Harabasz指数。CH指数计算所有集群的集群间分散和集群间分散之和的比率。CH指数越高，性能越好。对于规模为E的数据集n_E，k个聚类，定义CH指数为:

$$CH = \frac{t(B_k)}{t(W_k)} \times \frac{n_E - k}{k - 1} \quad (27)$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} \left(x - c_q\right)(x - c_q)^T \quad (28)$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_q)^T \quad (29)$$

其中 t Bk为组间色散矩阵的轨迹，t W_k为簇内色散矩阵的轨迹。C_q为簇q中点的集合，c_q为簇q的中心，c_E为E的中心，n_q为簇q中点的个数，T为矩阵变换。

Davies-Bouldin指数。DB指数表示集群之间的平均“相似性”，其中相似性是将集群之间的距离与集群本身的大小进行比较的度量。DB指数越低，表示模型在簇之间的分离越好。对于k个簇，i 2k, j 2k的数据，DB索引定义为:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \quad (30)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (31)$$

其中s_i和s_j为聚类内每个点到聚类质心的平均距离。d_{ij}为I和j的聚类质心之间的距离。

基因关联网络评价。我们通过将DeepMAPS中识别的基因关联网络与IRIS3[15]和使用所有基因的正常基因共表达网络推断进行比较，评估了其性能。我们计算了每个工具中生成的网络的接近中心性和特征向量中心性。公式如下所示。

接近中心性(CC)。顶点u的接近中心性(CC)[76]是由无向加权图中所有其他顶点v的最短路径长度和的倒数来定义的。其公式定义为:

$$CC(u) = \frac{1}{\sum_{v \in V} d_w(u,v)} \quad (32)$$

其中$d_w$ðu，vÞis u和v之间的最短加权路径。如果顶点u和v之间没有路径，则公式中使用顶点总数而不是路径长度。CC越高，表示该节点在网络中的集中度越高，反映出该基因在网络中的作用越重要。使用函数igraph::betweenness使用igraph R包计算CC。我们取网络中所有节点的平均CC来表示网络CC。

特征向量中心性(EC)。特征向量中心性(EC)[77]得分对应于图邻接矩阵的第一个特征向量的值。u的EC分数定义为:

$$EC(u) = \lambda \sum_{v \in G} a_{uv} x_v \quad (33)$$

其中λ是特征向量x = ðx的特征值的逆$_1$,x_2，⋯,x_nÞ，a_{uv}是非直接图g的相邻加权矩阵。特征向量中心性得分高的节点意味着它与许多本身得分高的节点相连。EC是用igraph R包和函数igraph:: event来计算的。我们取网络中所有节点的平均EC来表示网络的EC。

GRN评价。对于scRNA-ATAC-seq数据，我们将DeepMAPS推断的细胞类型特异性GRN与(i) IRIS3和SCENIC在scRNA-seq矩阵上，(ii) IRIS3和SCENIC在GAS矩阵上，(iii) MAESTRO在scATAC-seq矩阵上，以及(iv) MAESTRO在原始scRNA-seq和scATAC-seq矩阵上进行比较。对于每个数据集比较，我们将基准测试工具中使用的单元簇设置为与DeepMAPS中生成的相同，以确保公平性。将每个工具生成的GRN与三个公共功能数据库进行比较，包括Reactome[21]、DoRothEA[22]和trust v2[23]。仅使用人类样本数据集进行比较，因为这些数据库都与人类相关。我们对GRN进行了超几何测试，得到了每个数据库的每种工具，并比较了富集GRN和功能术语的精度、召回率和F1分数。

细胞簇遗漏测试

对于具有真实单元格类型标签的基准数据集，我们删除了一种单元格类型中的所有单元格并运行DeepMAPS。然后，我们遍历所有单元格类型(一次一个)来评估ARI的鲁棒性。对于没有基准标签的数据，我们从DeepMAPS和其他基准工具中删除了预测的细胞簇中的细胞。

DeepMAPS服务器构建

DeepMAPS托管在HPE XL675d RHEL系统上，具有2个×128-core AMD EPYC 7H12 CPU, 64GB RAM和2×NVIDIA A100 40GB

GPU。后端服务器是使用NestJs框架用TypeScript编写的。Auth0作为一个独立的模块，提供用户认证和授权服务。Redis容纳了所有待处理分析作业的队列。DeepMAPS中有两种类型的作业:有状态作业由Plumber R包处理，提供实时交互式分析;无状态作业，如cpu绑定的生物信息学管道和GPU训练任务，可能需要很长时间，使用Nextflow构建。所有正在运行的作业都使用Nomad进行编排，允许为每个作业分配适当的内核和存储，并根据服务器负载保持作业的可扩展性。作业结果存储在MySQL数据库中。前端使用NUXT、Vuetify作为UI库、Apache ECharts和Cytoscape.js进行数据可视化构建。前端服务器和后端服务器使用REST API进行通信。

报告总结

关于研究设计的更多信息可在本文链接的自然组合报告摘要中获得。

数据可用性

本研究中使用的所有数据均来自公共领域。原始数据从GEO数据库下载，登录号为:人胰岛scRNA-seq数据GSE84133和健康骨髓单个核细胞CITE-seq数据:GSE194122。以下数据集来自figshare:人胰腺scRNA-seq数据[https://figshare.com/articles/dataset/Benchmarking_atlas-level_ data_integration_in_single-cell_ genomics_-_integration_task_datasets_ Immune_and_pancreas_/12420968/8]，小鼠膀胱scRNA-seq数据[https://doi.org/10.6084/m9.figshare]。5968960.v1]和人肺腺癌PBMC CITE-seq数据[https://doi.org/10.6084/m9.figshare.c.5018987.v1]]。以下配对的scRNA-seq和scATAC-seq数据集来自10X Genomics网站:3k健康PBMC数据 [https://www]。10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-no-cell-sorting-3-k-1-standard-2-0-0]， 10k 健康PBMC数据[https:// www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0]，冷冻人类健康大脑数据[https://www.10xgenomics]。10k人类PBMC数据[https://www.10xgenomics. cn]。健康的PBMC数据[https://www]。10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-3-k-1-standard-2-0-0]，新鲜胚胎数据[https://www.10xgenomics.com/resources/datasets/fresh-embryonic-e-18-mouse-brain-5-k-1-standard-2-0-0]]，以及淋巴结数据[https://www.10xgenomics.com/resources/datasets/fresh-frozen-lymph-node-with-b-cell-lymphoma-14-k-sorted-nuclei-1-standard-2-0-0]]。scRNA-seq和scATAC-seq癌细胞系数据从CNGB核苷酸序列档案下载，登录码为CNP0000213。所有数据集均可公开获取，不受限制。详细的数据信息可以在补充数据1中找到。本文提供了源数据。

代码的可用性

DeepMAPS Docker的python源代码可在https://github.com/OSU-BMBL/deepmaps免费获得，DeepMAPS web服务器可在https://bmblx.bmi.osumc.edu/免费获得。在Zenodo[78]上也可以获得源代码

参考文献

1. Stuart, T. & Satija, R. Integrative single-cell analysis. Nat. Rev. Genet. 20, 257–272 (2019).
2. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. 38, 1007–1022 (2020).
3. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Com-putational principles and challenges in single-cell data integration. Nat. Biotechnol. 39, 1202–1215 (2021).
4. S Teichmann, M. E. Method of the year 2019: single-cell multimodal omics. Nat. Methods 17, 1 (2020).
5. Hao, Y. et al. Integrated analysis of multimodal single-cell data. Cell 184, 3573–3587.e3529 (2021).
6. Argelaguet, R. et al. MOFA+: a statistical framework for compre-hensive integration of multi-modal single-cell data. Genome Biol. 21, 111 (2020).
7. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296 (2019).
8. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods 18, 272–282 (2021).
9. Li, Y. et al. Elucidation of biological networks across complex dis-eases using single-cell omics. Trends Genet. 36, 951–966 (2020).
10. Ma, Q. & Xu, D. Deep learning shapes single-cell data analysis. Nat. Rev. Mol. Cell Biol. 23, 303–304 (2022).
11. Wang, J. et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat. Commun. 12, 1882 (2021).
12. Hu, Z., Dong, Y., Wang, K. & Sun, Y. In Proceedings of The Web Conference 2020 2704–2710 (Association for Computing Machin-ery, Taipei, Taiwan; 2020).
13. Wang, X. et al. In The World Wide Web Conference 2022−2032 (Association for Computing Machinery, San Francisco, CA, USA; 2019).
14. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086 (2017).
15. Ma, A. et al. IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. Nucleic Acids Res. 48, W275−W286 (2020).
16. Han, P., Gopalakrishnan, C., Yu, H. & Wang, E. Gene regulatory network rewiring in the immune cells associated with cancer. Genes (Basel) 8, 308 (2017).
17. Gassner, E. The Steiner Forest Problem revisited. J. Discret. Algo-rithms 8, 154−163 (2010).
18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Inte-grating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411−420 (2018).
19. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat. Biotechnol. 40, 1458−1466 (2022).
20. Iacono, G., Massoni-Badosa, R. & Heyn, H. Single-cell tran-scriptomics unveils gene regulatory network plasticity. Genome Biol. 20, 110 (2019).
21. Joshi-Tope, G. et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33, D428−D432 (2005).
22. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the esti-mation of human transcription factor activities. Genome Res. 29, 1363–1375 (2019).
23. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 46, D380–D386 (2018).
24. Wang, C. et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol. 21, 198 (2020).
25. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. Nat. Commun. 12, 1088 (2021).
26. Ampudia, J. et al. CD6-ALCAM signaling regulates multiple effec-tor/memory T cell functions. J. Immunol. 204,

150.113−150.113 (2020).

27. Skonier, J. E. et al. Mutational analysis of the CD6 ligand binding domain. Protein Eng. Des. Selection 10, 943–947 (1997).

28. Gimferrer, I. et al. Relevance of CD6-mediated interactions in T cell activation and proliferation. J. Immunol. (Baltim., Md.: 1950) 173, 2262–2270 (2004).

29. Johnston, R. J., Lee, P. S., Strop, P. & Smyth, M. J. Cancer immunotherapy and the nectin family. Annu. Rev. Cancer Biol. 5, 203–219 (2021).

30. Li, X.-Y. et al. CD155 loss enhances tumor suppression via combined host and tumor-intrinsic mechanisms. J. Clin. Invest. 128, 2613–2625 (2018).

31. Gururajan, M. et al. Early growth response genes regulate B cell development, proliferation, and immune response. J. Immunol. (Baltim., Md.: 1950) 181, 4590–4602 (2008).

32. Oh, Y.-K., Jang, E., Paik, D.-J. & Youn, J. Early growth response-1 plays a non-redundant role in the differentiation of B cells into plasma cells. Immune Netw. 15, 161–166 (2015).

33. Brescia, P. et al. MEF2B instructs germinal center development and acts as an oncogene in B cell lymphomagenesis. Cancer Cell 34, 453–465.e459 (2018).

34. Trøen, G. et al. Constitutive expression of the AP-1 transcription factors c-jun, junD, junB, and c-fos and the marginal zone B-cell transcription factor Notch2 in splenic marginal zone lymphoma. J. Mol. Diagn. 6, 297–307 (2004).

35. Sánchez-Beato, M. et al. Abnormal PcG protein expression in Hodgkin's lymphoma. Relation with E2F6 and NFkappaB transcrip-tion factors. J. Pathol. 204, 528–537 (2004).

36. Saha, A., Robertson, E. S. & Goodrum, F. Mechanisms of B-cell oncogenesis induced by Epstein-Barr virus. J. Virol. 93, e00238–00219 (2019).

37. Yachida, S. et al. Genomic sequencing identifies ELF3 as a driver of ampullary carcinoma. Cancer Cell 29, 229–240 (2016).

38. Wang, H. et al. Overexpression of ELF3 facilitates cell growth and metastasis through PI3K/Akt and ERK signaling pathways in non-small cell lung cancer. Int. J. Biochem. Cell Biol. 94, 98–106 (2018).

39. Zhang, J. et al. KLF16 affects the MYC signature and tumor growth in prostate cancer. Onco Targets Ther. 13, 1303–1310 (2020).

40. Ma, P. et al. KLF16 promotes proliferation in gastric cancer cells via regulating p21 and CDK4. Am. J. Transl. Res. 9, 3027–3036 (2017).

41. Mathas, S. et al. Aberrantly expressed c-Jun and JunB are a hallmark of Hodgkin lymphoma cells, stimulate proliferation and synergize with NF-κB. EMBO J. 21, 4104–4113 (2002).

42. Eferl, R. & Wagner, E. F. AP-1: a double-edged sword in tumorigenesis. Nat. Rev. Cancer 3, 859–868 (2003).

43. Nagel, D., Vincendeau, M., Eitelhuber, A. C. & Krappmann, D. Mechanisms and consequences of constitutive NF-κB activation in B-cell lymphoid malignancies. Oncogene 33, 5655–5665 (2014).

44. Jost, P. J. & Ruland, J. R. Aberrant NF-κB signaling in lymphoma: mechanisms, consequences, and therapeutic implications. Blood 109, 2700–2707 (2006).

45. Garces de Los Fayos Alonso, I. et al. The role of activator protein-1 (AP-1) family members in CD30-positive lymphomas. Cancers (Basel) 10, 93 (2018).

46. Crispino, J. D. & Horwitz, M. S. GATA factor mutations in hematologic disease. Blood 129, 2103–2110 (2017).

47. Shimizu, R., Engel, J. D. & Yamamoto, M. GATA1-related leukaemias. Nat. Rev. Cancer 8, 279–287 (2008).

48. Mosquera Orgueira, A. et al. Detection of new drivers of frequent B-cell lymphoid neoplasms using an integrated analysis of whole genomes. PLoS ONE 16, e0248886 (2021).

49. Blyth, K. et al. Runx1 promotes B-cell survival and lymphoma development. Blood Cells Mol. Dis. 43, 12–19 (2009).

50. Mackay, F., Schneider, P., Rennert, P. & Browning, J. BAFF AND APRIL: a tutorial on B cell survival. Annu Rev. Immunol. 21, 231–264 (2003).

51. Smulski, C. R. & Eibel, H. BAFF and BAFF-receptor in B cell selection and survival. Front. Immunol. 9, 2285 (2018).

52. Yang, S., Li, J. Y. & Xu, W. Role of BAFF/BAFF-R axis in B-cell non-Hodgkin lymphoma. Crit. Rev. Oncol. Hematol. 91, 113–122 (2014).

53. He, B. et al. Lymphoma B cells evade apoptosis through the TNF family members BAFF/BLyS and APRIL. J. Immunol. (Baltim., Md.: 1950) 172, 3268–3279 (2004).

54. Xia, X. Z. et al. TACI is a TRAF-interacting receptor for TALL-1, a tumor necrosis factor family member involved in B cell regulation. J. Exp. Med 192, 137–143 (2000).

55. Laâbi, Y., Egle, A. & Strasser, A. TNF cytokine family: more BAFFling complexities. Curr. Biol. 11, R1013–R1016 (2001).

56. Mackay, F. & Schneider, P. TACI, an enigmatic BAFF/APRIL receptor, with new unappreciated biochemical and biological properties. Cytokine Growth Factor Rev. 19, 263–276 (2008).

57. Rihacek, M. et al. B-cell activating factor as a cancer biomarker and its implications in cancer-related Cachexia. Biomed. Res. Int. 2015, 792187–792187 (2015).

58. Su, H., Chang, J., Xu, M., Sun, R. & Wang, J. CDK6 overexpression resulted from microRNA-320d downregulation promotes cell proliferation in diffuse large B-cell lymphoma. Oncol. Rep. 42, 321–327 (2019).

59. Lee, C., Huang, X., Di Liberto, M., Martin, P. & Chen-Kiang, S. Targeting CDK4/6 in mantle cell lymphoma. Ann. Lymphoma 4, 1 (2020).

60. Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. Nat. Rev. Cancer 17, 93–115 (2017).

61. Li, K. et al. cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells. bioRxiv, 2020.2008.2028.270652 (2020).

62. Pereira, W. et al. Asc-Seurat –Analytical single-cell Seurat-based web application. bioRxiv, 2021.2003.2019.436196 (2021).

63. Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C. & Deplancke, B. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinformatics 33, 3123–3125 (2017).

64. Li, B. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat. Methods 17, 793–798 (2020).

65. Hillje, R., Pelicci, P. G. & Luzi, L. Cerebro: interactive visualization of scRNA-seq data. Bioinformatics 36, 2311–2313 (2019).

66. Prompsy, P. et al. Interactive analysis of single-cell epigenomic landscapes with ChromSCape. Nat. Commun. 11, 5702 (2020).

67. Bolisetty, M. T., Stitzel, M. L. & Robson, P. CellView: Interactive exploration of high dimensional single cell RNA-seq data. bioRxiv, 123810 (2017).

68. Mohanraj, S. et al. CReSCENT: CanceR Single Cell ExpressioN Toolkit. Nucleic Acids Res. 48, W372–W379 (2020).

69. Patel, M. V. iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. Bioinformatics 34, 4305–4306 (2018).

70. Yousif, A., Drou, N., Rowe, J., Khalfan, M. & Gunsalus, K. C. NASQAR: a web-based platform for high-throughput sequencing data ana-lysis and visualization. BMC Bioinforma. 21, 267 (2020).

71. Zhu, Q. et al. PIVOT: platform for interactive analysis and visualization of transcriptomics data. BMC Bioinforma. 19, 6 (2018).

72. Innes, B. & Bader, G. scClustViz - Single-cell RNAseq cluster assessment and visualization. F1000Res. 7, ISCB Comm J-1522 (2018).

73. Granja, J. M. et al. ArchR is a scalable software package for inte-grative single-cell chromatin accessibility analysis. Nat. Genet. 53, 403–411 (2021).

74. Wan, C. et al. LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic Acids Res. 47, e111 (2019).

75. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. 38, 1408–1414 (2020).

76. Li, G. S., Li, M., Wang, J. X., Li, Y. H. & Pan, Y. United Neighborhood Closeness Centrality and Orthology for Predicting Essential. Pro-teins Ieee Acm T Comput Bi 17, 1451–1458 (2020).

77. Parisutham, N. & Rethnasamy, N. Eigenvector centrality based algorithm for finding a maximal common connected vertex induced molecular substructure of two chemical graphs. J. Mol. Struct. 1244, 130980 (2021).

78. Ma, A. et al. Single-cell biological network inference using a heterogeneous graph transformer. Zenodo https://doi.org/10.5281/zenodo.7559037 (2023).

致谢

这项工作得到了美国国立卫生研究院R01-GM131399 (q.m)、R35-GM126985 (D.X.)和U54-AG075931 (q.m)的资助。这项工作还得到了美国国家科学基金会NSF1945971 (q.m)奖的支持。这项工作得到了Pelotonia免疫肿瘤研究所(PIIO)的支持。内容完全是作者的责任，并不一定代表PIIO的官方观点。另外，感谢东北师范大学何飞博士在框架构建、数据检验等方面提出的宝贵建议。

作者的贡献

Q.M, b.l.和D.X.构思了基本思想并设计了框架。X.W.编写了DeepMAPS的主干代码。C.W.和H.C.构建了后端和前端服务器。S.G.设计了服务器上的交互式图形。Y.Liu进行了RNA速度计算。Y.Li设计了SFP模型用于基因模块预测。a.m.、X.W和J.L.进行了基准实验。X.W, y.c., B.L.进行稳健性检验。a.m.、j.l.、X.W、广新、志亮及T.X.进行个案研究。j.w.、d.w.、y.j.、j.l.和L.S.进行了工具优化。A.M.和q.m负责人物设计和稿件撰写。所有作者都参与了稿件的解读和撰写。

相互竞争的利益

作者声明没有利益竞争。

额外的信息

补充信息在线版本包含补充资料，可在https://doi.org/10.1038/s41467-023-36559-0上获得。

联系、索取资料请联系刘炳强、徐东或马勤。

同行评议信息Nature Communications感谢Saugato Rah-man Dhruba和其他匿名审稿人对本文同行评议的贡献。可查阅同行评议报告。