文章编号:1674-2869(2026)01-0103-08

基于自监督预训练的单细胞类型注释方法

张 晴[1],吴晓晓[1],李 想[1],马 威[1],吴通权[2],谢诒诚[*3],吴兴隆[*1]
1. 武汉工程大学计算机科学与工程学院,湖北 武汉 430205;
2. 合肥工业大学机械工程学院,安徽 合肥 230009;
3. 浙江大学医学院附属儿童医院皮肤科,浙江 杭州 310052

摘 要:为了应对单细胞RNA测序中准确注释细胞类型的挑战,提出了基于迁移学习和Transformer的深度学习网络单细胞标签注释网络(ScLabel-Net),旨在对小鼠肺部的大规模单细胞数据集进行高效、准确的细胞类型注释。ScLabel-Net首先在约10万个细胞的单细胞肺部数据集上进行预训练,通过自监督学习捕捉基因间的相似性,然后将模型迁移到相对较少的数据集上,对特定细胞类型注释任务进行微调。考虑到单细胞数据中常见的细胞类型分布不平衡现象,微调数据集时应用了随机上采样技术,以减轻分布不平衡对注释结果的影响。实验结果表明,ScLabel-Net在GSE267861、GSE264032和Quake等3个小鼠肺部数据集上的细胞类型注释准确率分别达到0.955、0.922和0.986。此外,ScLabel-Net在小鼠其他器官(如气管、肾脏、胰腺)的单细胞数据集上也表现出优异的泛化能力,准确率分别达到0.981、0.951和0.987,验证了ScLabel-Net跨器官的适用性,进一步证明了ScLabel-Net在复杂生物系统和疾病研究中的广泛应用潜力。

关键词:细胞类型注释;自监督预训练;深度学习;单细胞RNA测序
中图分类号:TP391.1 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202412007

A single-cell type annotation method based on self-supervised pretraining

ZHANG Qing[1], ***WU Xiaoxiao***[1], ***LI Xiang***[1], ***MA Wei***[1], ***WU Tongquan***[2], ***XIE Yicheng***[*3], ***WU Xinglong***[*1]
1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China;
2. School of Mechanical Engineering, Hefei University of Technology, Hefei 230009, China;
3. Department of Dermatology, Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310052, China

Abstract: To address the challenges in accurate cell type annotation in single-cell RNA sequencing, such as efficiently and accurately annotate cell types in large-scale single-cell datasets from mouse lung, a deep learning single-cell label annotation network (ScLabel-Net) based on transfer learning and Transformer was proposed. ScLabel-Net was first pre-trained on a single-cell lung dataset containing approximately 100 000 cells, leveraging self-supervised learning to capture inter-gene relationships. It was then transferred to smaller datasets for fine-tuning, enabling precise cell type annotation. To mitigate the impact of class imbalance (a common issue in single-cell data), a random up-sampling technique was applied during fine-tuning. Experimental results showed that ScLabel-Net achieved cell type annotation accuracies of 0.955, 0.922, and 0.986 on three mouse lung datasets (GSE267861, GSE264032, and Quake). Moreover, ScLabel-Net also demonstrated excellent generalization ability on single-cell datasets from other mouse organs (trachea,

收稿日期:2024-12-06
基金项目:国家自然科学基金(82302085);新疆维吾尔自治区人工智能辅助影像诊断重点实验室基金(XJRG ZN2024008);湖北省自然科学基金(2020BCB002)
作者简介:张 晴,硕士研究生。Email:1714436732@qq.com
***通信作者**:谢诒诚,博士,研究员。Email: ycxie@zju.edu.cn
吴兴隆,博士,副教授。Email: xwu@wit.edu.cn
引文格式:张晴,吴晓晓,李想,等.基于自监督预训练的单细胞类型注释方法[J].武汉工程大学学报,2026,48(1):103-110.

kidney, and pancreas) with accuracies of 0.981, 0.951, and 0.987 respectively, verifying cross-organ applicability of ScLabel-Net and its potential for application in fields of complex biological systems and diseases.

Keywords: cell type annotation; self-supervised pretraining; deep learning; single-cell RNA sequencing

单细胞RNA测序技术为研究人员提供了深入分析单个细胞的能力[1],该技术的核心目标是精确识别每个细胞的具体细胞类型。然而,细胞类型的标注过程不仅耗时,而且高度依赖专家的专业知识[2]。面对庞大的单细胞RNA测序数据集,传统的基于专家知识的人工注释方法面临着巨大挑战,并且注释结果往往带有一定的主观性[3]。

目前关于细胞类型注释的研究大致可以分为基于标记基因的方法、基于相关性的方法和基于深度学习的方法[4]。首先,基于标记基因的方法在单细胞RNA测序的初期阶段发挥了重要作用。然而,这种方法的有效性取决于标记基因数据库的准确性。尽管有如文献[5-6]等可公开访问的数据库,但一些标记基因仍依赖于现有的研究知识。其次,基于相关性的方法也被广泛采用,但其易受批次效应的影响,需要高质量的参考数据集以确保类型注释的准确性。最后,基于深度学习的方法对噪声和数据的可变性具有鲁棒性,不依赖于人工选择的标记基因[7],近年来得到了广泛应用。例如,自动编码器[8]、变分自编码器[9]和生成对抗网络[10]等技术,常被用于细胞聚类、细胞类型注释和推断基因调控网络等任务。此外,图模型也常常用于细胞类型注释[11],利用基因共表达关系或细胞相似性来提高细胞类型注释的准确性[12]。随着单细胞组学数据量的激增,开发自动化、快速且准确的细胞类型注释方法已成为该领域的研究热点。

在将Transformer架构应用于单细胞数据的细胞类型注释的研究中,Yang等[13]展现了在大规模无标记单细胞RNA测序数据上进行自监督学习的潜力,提供了稳健、准确的细胞类型注释。然而,当面对细胞类型分布不平衡的数据集时,其性能受到了影响,这可能是由于存在稀有细胞类型或偏采样造成的[14]。因此,本研究提出了基于迁移学习和Transformer的深度学习网络单细胞标签注释网络(single-cell label annotation network, ScLabel-Net),用于单细胞数据的细胞类型注释,该网络引入了位置嵌入和数据增强等技术。位置嵌入被整合到每个基因表达嵌入中,以建立基因与基因之间的空间相对位置关系[15]。随机上采样这一数据增强技术对小样本数据进行复制扩充,达到与最多数据的样本量一致的数据量。这两种技术分别解决了自注意力机制不考虑输入序列的顺序和小样本数据得不到充分训练的难题。通过上述改进,ScLabel-Net能够更有效地处理不平衡数据集,从而在单细胞数据的细胞类型注释任务中表现出更高的准确性和鲁棒性。

1 实验部分

整体的研究流程如图1所示,主要分为数据准备、数据预处理、ScLabel-Net的构建以及性能评估4个部分。

首先,将多个数据汇集并整合到一个文件中[16]。接着,模型接收基因表达矩阵作为输入,并使用Python工具包SCANPY[17]对数据进行预处理,包括去除低质量细胞、数据归一化、对数转换

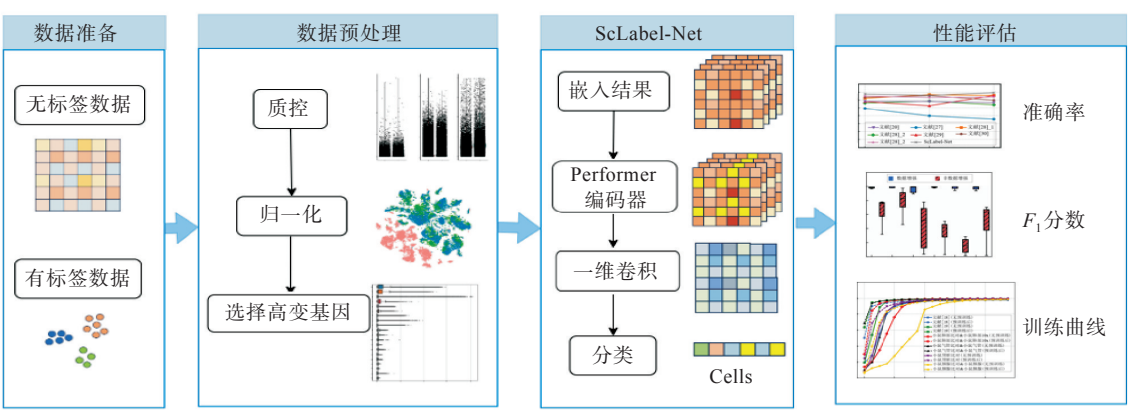


图1 整体研究流程图

Fig. 1 Overall Research Workflow Diagram

以及筛选高变基因。然后，ScLabel-Net在未标记的大规模单细胞数据集上进行预训练，旨在识别细胞类型的潜在特征，以减少不同数据集之间的批次效应[13]，随后在特定的有标签的目标数据集上进行微调。最后，对模型性能进行全面评估，以验证模型的实用性和通用性。

1.1　数据获取

研究所涉及的数据涵盖了4个小鼠的不同器官，包括肺部、气管、肾脏和胰腺，共10个单细胞数据集，并且包含了5种不同的测序技术。如表1所示，这些数据集被分为3个类别：用于预训练的无标签数据集、用于微调训练和性能评估的有标签数据集以及用于模型性能评估的无标签数据集。

预训练阶段使用的若干个小鼠肺部数据集来源于PanglaoDB[5]，一共包含了100 024个细胞，涵盖了25种不同的细胞类型。在数据预处理阶段，将这些小鼠肺部数据集合并成一个大的数据集（Panglao）用于预训练任务。微调训练涉及的小鼠肺部数据集包括来源于基因表达数据库的GSE267861[18]和GSE264032[19]，以及4个来自动物细胞图谱（animal cell atlas，ACA）数据库中小鼠其他器官的数据集[20]，分别是ALIGNED_Lung、ALIGNED_Trachea、ALIGNED_Kidney和ALIGNED_Pancreas数据集。为了进一步验证模型在真实场景中的泛化能力，研究中还特别选用了3个来自ACA数据库中的Quake、Plasschaert、Baron数据集[20]，并且这3个数据集未在预训练和微调训练过程中使用。

表1　研究中所有的单细胞数据集详情

Table 1　Details of all single-cell datasets in the study

类别	数据集	类型数量	细胞数量	基因数量
预训练	Panglao	25	100 024	45 549
微调训练	GSE267861	12	24 520	32 285
	GSE264032	16	31 928	32 285
	ALIGNED_Lung	17	16 599	25 174
	ALIGNED_Trachea	18	12 619	33 948
	ALIGNED_Kidney	37	63 659	35 210
	ALIGNED_Pancreas	22	3 450	25 410
性能评估	Quake	13	5 449	23 341
	Plasschaert	8	6 977	28 205
	Baron	13	1 886	14 877

1.2　数据预处理

在单细胞RNA测序数据分析中，基因表达矩阵通常具有高稀疏性[21]，因此对数据集进行预处理至关重要。首先，利用SCANPY将每个数据集的基因表达矩阵转换为有注释的数据对象，并进行质量控制，包括细胞过滤（移除基因数少于200且线粒体含量超过25%的细胞）和归一化。接着，对每个细胞的基因表达值进行对数转换，并筛选出变化较大的基因。

用于预训练和微调训练的数据集按8∶2的比例分为训练集和测试集，其中80%的数据通过五折交叉验证策略用于模型训练，其余20%用于性能评估。此外，在预训练和微调训练过程中未见的数据集没有被划分，而是直接用于评估模型在真实场景中的预测能力。

1.3　ScLabel-Net网络架构

ScLabel-Net是一种基于自监督预训练的深度学习网络，由位置嵌入、基因表达嵌入、Performer[22]编码器和分类层组成。

首先，ScLabel-Net根据单细胞数据的属性引入分箱技术，用来表示不同的基因表达水平。基因表达嵌入通常以矩阵形式表示，而位置嵌入的运用有助于模型识别并利用基因的相对位置关系，从而更深入地理解数据的结构[23]。其次，在预训练阶段，嵌入结果被输入到编码器中。编码器首先对输入数据进行编码将原始数据转换为抽象表示，随后对编码后的数据进行表征学习，以生成预测结果，然后经过维度转换，重建原始数据，并利用原始标签计算交叉熵损失。如图2所示，ScLabel-Net在微调训练过程中，针对特定任务的数据输入到预训练好的编码器中，表征学习的结

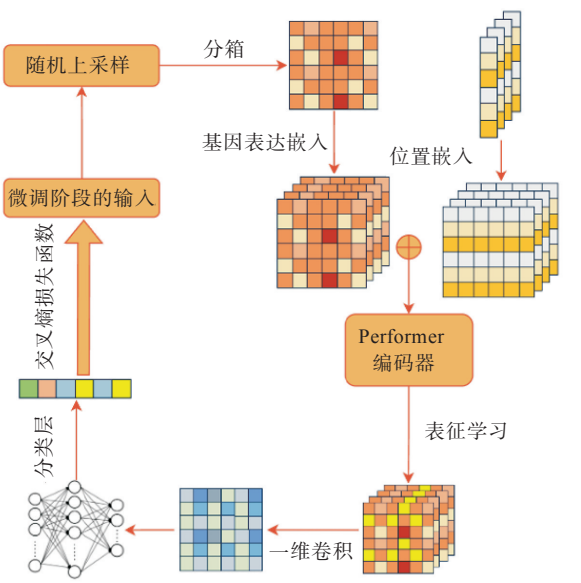


图2　ScLabel-Net微调训练网络架构

Fig. 2　ScLabel-Net fine-tuning training network architecture

果经过一维卷积进行特征提取，然后将一维卷积结果送入到分类层进行分类，分类层输出每个类别的概率。最终通过归一化指数生成正确的细胞类型，从而完成分类任务。

1.3.1 基因表达嵌入 基因表达嵌入面临的主要挑战是不同测序技术之间存在的绝对表达水平变异性。由于测序深度的不同，不同批次的测序样本在数据量级上可能会表现出显著的差异[24]。为此，基于分箱的技术通过将所有表达值转换为相对值，从而在不同测序批次之间建立一个统一的比较基准。这种方法能够最大限度地减少不同测序方案的影响，降低数据噪声。具体来说，将基因表达矩阵 $\boldsymbol{Y}$ 中每个细胞的非零表达值划分为 B 个连续区间 $[b_k, b_{k+1}]$，其中 $k \in \{1, 2, \cdots, B\}$，$b_k$ 表示某个细胞的区间边界表达值。每个细胞都计算了一套新的分箱边界，因此区间边界 b_k 可能在不同细胞之间有所不同。每个区间包含一部分具有相似表达值的基因，$Y_{i,j}$ 表示第 j 个基因在第 i 个细胞中的表达值，$y_{i,j}$ 则表示第 i 个细胞的第 j 个基因的分箱表达值，如式(1)所示。

$$y_{i,j} = \begin{cases} k, & Y_{i,j} > 0 \text{ 且 } Y_{i,j} \in \left[b_k, b_{k+1}\right] \\ 0, & Y_{i,j} = 0 \end{cases} \tag{1}$$

分箱技术确保了 $y_{i,j}$ 的语义在不同的批次中保持一致。例如，$y_{i,j} = B$ 的值始终表示基因的最高表达量。此外，在分箱步骤之前还进行了 log1p 转换，以尽量减少高基因表达值的影响，并使数据符合正态分布。每个基因表达值 $Y_{i,j}$ 的 log1p 变换可表示为：

$$Y'_{i,j} = \ln\left(1 + Y_{i,j}\right) \tag{2}$$

因此，第 i 个细胞的分箱表达值的最终输入向量表示为：

$$\boldsymbol{y}_i = \left[y_{i,1}, y_{i,2}, \cdots, y_{i,N}\right] \tag{3}$$

其中，N 表示基因的数量。

综上所述，分箱技术将基因表达值这一连续变量离散化，并将分箱结果转换为一个200维的向量，该向量随后用作ScLabel-Net的基因表达嵌入。

1.3.2 损失函数 在预训练过程中，采用基于Transformer的双向编码器表示(bidirectional encoder representations from Transformers)[25]的掩码语言模型方法，随机屏蔽非零基因表达数据，然后使用剩余的表达值重建原始输入数据。因此，预训练阶段利用交叉熵损失作为重建损失，具体如下：

$$L_{\text{Rec}} = -\sum_{i=1}^{C}\sum_{j=1}^{G} x_{i,j} \ln\left(p_{i,j}\right) \tag{4}$$

其中：C 表示细胞数，G 表示被掩盖的基因表达值的数量，$x_{i,j}$ 和 $p_{i,j}$ 分别表示第 j 个基因在第 i 个细胞中的真实表达和预测表达。

此外，在微调训练阶段使用交叉熵损失量化真实细胞类型标签和预测细胞类型标签之间的差异，具体如下：

$$L_{\text{Pred}} = -\sum_{i=1}^{M} z_i \ln\left(q_i\right) \tag{5}$$

其中，z_i 和 q_i 分别表示第 i 个细胞的真实标签和预测标签。

1.4 性能评估指标

由于需要解决数据分布不平衡的多分类问题，因此采用以下4个指标来评估模型性能，包括准确率(A)、宏观精确率(P_{m})、宏观召回率(R_{m})和宏观 F_1 分数(F_1)[26]，计算公式如下：

$$A = \frac{T + N}{T + N + F + M} \tag{6}$$

$$P_{\text{m}} = \frac{1}{n}\sum_{l=1}^{n}\frac{T_l}{T_l + F_l} \tag{7}$$

$$R_{\text{m}} = \frac{1}{n}\sum_{l=1}^{n}\frac{T_l}{T_l + M_l} \tag{8}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

其中：n 是细胞类型的数量，l 表示每个细胞类型的索引，T 和 N 分别表示正确分类的阳性数量和阴性数量，F 和 M 分别表示被错误分类的阴性数量和阳性数量。

2 结果与讨论

2.1 实验设备

所有实验均在一台搭载Ubuntu 18.04 LTS操作系统的服务器上执行，该服务器配备了英特尔酷睿i9-10920X处理器、2块英伟达GeForce GTX 3090 GPU以及188 GB的运行内存。ScLabel-Net采用Python 3.8和PyTorch 1.8.1深度学习框架构建。同时还使用SCANPY、R语言等用于数据预处理、分析和可视化。

ScLabel-Net在Panglao小鼠肺部数据集上进行了20轮的预训练，每个批次大小设置为2，使用Adam优化器，初始学习率大小为 10^{-4}，并使用余弦退火热重启学习率调整策略。预训练结束后，模型在目标数据集上继续进行20轮的微调训练，采用五折交叉验证策略。微调训练得出的最优模型权重被保存起来，以便在测试集上进行性能评估。

2.2　ScLabel-Net在小鼠肺部细胞类型预测中的应用

ScLabel-Net 在 GSE267861、GSE264032 和 ALIGNED_Lung小鼠肺部数据集上完成了微调。如图3(a)所示,GSE267861和GSE264032的数据按照8∶2的比例划分为训练集和测试集。因此,ScLabel-Net在这两个数据集的测试集以及Quake数据集上进行了性能评估。如表2所示,ScLabel-Net与基于相关性的方法(Cell BLAST[20]、SingleR[27]、Scmap_cell[28]、Scmap_cluster[28]和

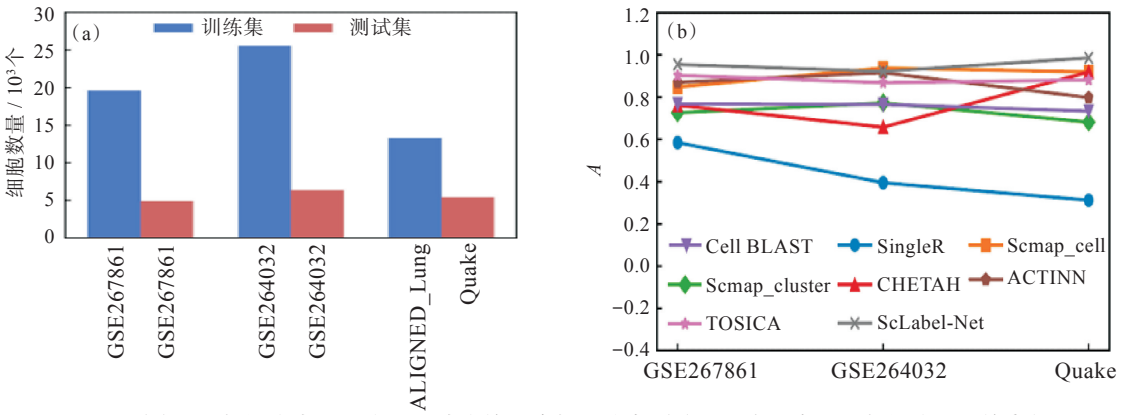


图3　(a)不同数据集中训练集和测试集的细胞数量分布;(b)不同方法在不同数据集上*A*的变化

Fig. 3　(a) Cell distribution in the training and testing sets of different datasets; (b) Changes in *A* of different methods across different datasets

表2　ScLabel-Net与7种现有方法在3个小鼠肺部数据集上的性能比较

Table 2　Comparison of the performance of ScLabel-Net with that of seven existing methods on three mouse lung datasets

肺部数据集		方法	*A*	F_1分数(95%置信区间)	P_m	R_m
微调训练	性能评估					
GSE267861	GSE267861	Cell BLAST	0.768	0.770(0.761~0.778)	0.827	0.733
		SingleR	0.585	0.284(0.282~0.285)	0.271	0.304
		Scmap_cell	0.848	0.839(0.829~0.846)	0.889	0.801
		Scmap_cluster	0.726	0.775(0.765~0.783)	0.804	0.775
		CHETAH	0.763	0.392(0.387~0.411)	0.402	0.422
		ACTINN	0.920	0.932(0.925~0.938)	**0.930**	**0.966**
		TOSICA	0.915	0.924(0.896~0.950)	0.910	0.923
		ScLabel-Net	0.955	**0.943(0.936~0.949)**	0.928	0.962
GSE264032	GSE264032	Cell BLAST	0.766	0.529(0.517~0.540)	0.603	0.490
		SingleR	0.396	0.169(0.163~0.182)	0.189	0.171
		Scmap_cell	0.939	**0.783(0.772~0.847)**	**0.852**	0.732
		Scmap_cluster	0.773	0.567(0.556~0.575)	0.664	0.586
		CHETAH	0.658	0.274(0.268~0.279)	0.324	0.284
		ACTINN	0.917	0.692(0.680~0.703)	0.757	0.738
		TOSICA	0.869	0.649(0.637~0.661)	0.692	0.727
		ScLabel-Net	0.922	0.706(0.695~0.716)	0.803	**0.754**
ALIGNED_Lung	Quake	Cell BLAST	0.735	0.460(0.440~0.502)	0.702	0.368
		SingleR	0.313	0.208(0.206~0.208)	0.209	0.206
		Scmap_cell	0.920	0.802(0.784~0.816)	0.922	0.733
		Scmap_cluster	0.683	0.642(0.620~0.702)	0.847	0.548
		CHETAH	0.919	0.442(0.435~0.481)	0.474	0.418
		ACTINN	0.878	0.725(0.713~0.773)	0.749	0.704
		TOSICA	0.912	0.726(0.717~0.833)	0.724	0.732
		ScLabel-Net	**0.986**	**0.847(0.834~0.974)**	0.848	**0.836**

CHETAH[29]），以及基于深度学习的方法（ACTINN[30]、TOSICA[31]）进行了对比。结果表明，ScLabel-Net在A和F_1方面表现出更优越的性能。特别是，在Quake数据集上进行评估时，ScLabel-Net的性能最为突出，A为0.986，F_1为0.847。此外，如图3（b）所示，ScLabel-Net在GSE264032数据集上的表现略逊于其他两个数据集，这可能是由于该数据集的复杂度较高且规模较大（包含16种细胞类型，其中4种类型占总数据的比例不到1%），尽管如此，ScLabel-Net在GSE264032数据集上的A为0.922，F_1为0.706，仍优于其他方法。

2.3 ScLabel-Net的跨器官类型注释

在实际应用中，可用的数据集往往来自不同器官，因此需要评估模型在跨器官数据集上的性能。如表3所示，ScLabel-Net在小鼠不同器官上表现出色，平均准确率达到97.4%，而两种基于深度学习的方法的平均准确率分别为93.7%和95.7%，基于相关性的方法的性能更是明显下降。其中，SingleR表现最差，在3个器官上的A分别为0、0.047和0.094，这可能是因为其附带的7个参考数据集（5个人类数据集和2个小鼠数据集，其中小鼠参考数据集为小鼠免疫遗传学数据库和小鼠RNA测序数据库）没有很好地覆盖所有细胞类型。与基于相关性的方法和基于深度学习的方法相比，ScLabel-Net在跨器官和跨平台场景中表现出更优越和稳定的性能。

2.4 随机上采样

组学分析中普遍存在数据分布不平衡的问题，通常由选择性采样和罕见数据的稀缺性引起。随机上采样技术可以有效缓解微调数据集中的不平衡分布。如图4所示，ScLabel-Net在4个器官的6个数据集上都取得了显著的性能提升。在经过上采样处理的数据集上，ScLabel-Net的A分别提升了1.6%、2.2%、2.8%和8.3% Quake和ALIGNED_Kidney数据集的A分别提升了约

表3 ScLabel-Net在小鼠不同器官中的性能

Table 3 Performance of ScLabel-Net across different mouse organs

器官数据集		方法	A	F_1分数（95%置信区间）	P_m	R_m
微调训练	性能评估					
ALIGNED_Trachea	Plasschaert	Cell BLAST	0.799	0.456（0.439~0.510）	0.532	0.426
		SingleR	0	0	0	0
		Scmap_cell	0.888	0.493（0.453~0.534）	**0.884**	0.422
		Scmap_cluster	0.527	0.341（0.317~0.395）	0.517	0.435
		CHETAH	0.815	0.366（0.359~0.415）	0.369	0.399
		ACTINN	0.978	**0.783（0.775~0.790）**	0.792	**0.775**
		TOSICA	**0.989**	0.706（0.698~0.873）	0.690	0.725
		ScLabel-Net	0.981	0.650（0.633~0.734）	0.718	0.614
ALIGNED_Kidney	ALIGNED_Kidney	Cell BLAST	0.624	0.468（0.459~0.476）	0.712	0.407
		SingleR	0.047	0.042（0.041~0.043）	0.033	0.075
		Scmap_cell	0.733	0.446（0.434~0.455）	0.802	0.372
		Scmap_cluster	0.145	0.062（0.055~0.068）	0.282	0.044
		CHETAH	0.794	0.351（0.345~0.360）	0.420	0.342
		ACTINN	0.915	0.636（0.627~0.645）	0.787	0.597
		TOSICA	0.927	**0.768（0.756~0.778）**	0.795	**0.757**
		ScLabel-Net	**0.951**	0.745（0.733~0.765）	**0.839**	0.724
ALIGNED_Pancreas	Baron	Cell BLAST	0.880	0.612（0.558~0.651）	0.781	0.540
		SingleR	0.094	0.166（0.148~0.179）	0.156	0.191
		Scmap_cell	0.819	0.425（0.399~0.453）	0.634	0.377
		Scmap_cluster	0.178	0.222（0.194~0.253）	0.489	0.168
		CHETAH	0.907	0.476（0.472~0.521）	0.483	0.471
		ACTINN	0.918	**0.961（0.930~0.981）**	0.936	0.912
		TOSICA	0.956	0.910（0.891~0.945）	0.921	0.906
		ScLabel-Net	**0.987**	0.922（0.922~0.996）	**0.941**	**0.924**

10.9%和18.8%。在F_1分数方面，ALIGNED_Kidney数据集的提升最为显著，达到214.5%。这表明，在缺乏数据增强的情况下，某些类别的预测性能并不理想。这些发现表明，基于随机上采样的数据增强技术能显著提升ScLabel-Net在细胞类型标注任务中的性能。

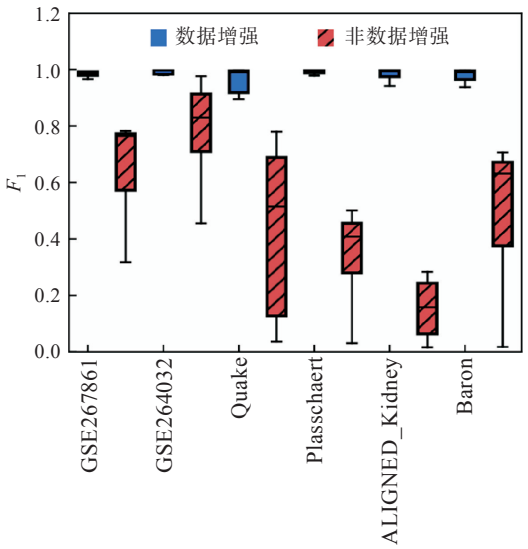


图4　ScLabel-Net在数据增强后的性能表现

Fig. 4　Performance of ScLabel-Net with data enhancement

2.5　自监督预训练对ScLabel-Net性能和训练效率的影响

在单细胞数据集的细胞类型标注任务中，采用了在大型未标记数据集上进行自监督预训练的迁移学习策略，有效提升了本文模型在小鼠肺部及其他器官上的性能。如图5所示，经过预训练的ScLabel-Net在GSE267861、GSE264032和Quake等3个数据集上取得了更好的性能且在肾脏等其他器官数据集上的性能也有所增强。以GSE267861的数据集为例，预训练后的A和F_1分别达到了0.955和0.943，相较于未进行预训练的A和F_1有了显著的提升。同时，经过预训练的ScLabel-Net在目标数据集上进行微调后，收敛速度明显加快。

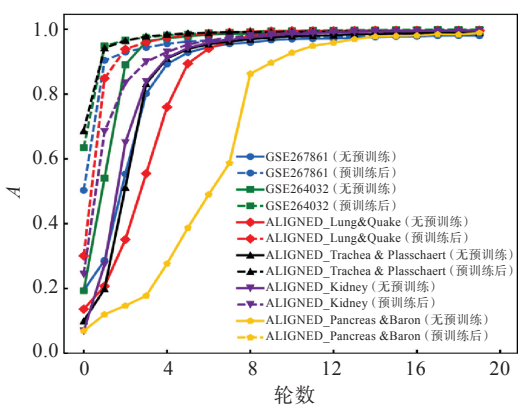


图5　ScLabel-Net在6个微调数据集上有无自监督预训练的学习效果对比曲线

Fig. 5　Comparison curves of the learning effects of ScLabel-Net with and without self-supervised pre-training on six fine-tuned datasets

3　结　论

本研究通过ScLabel-Net框架，采用自监督预训练策略对单细胞数据中的细胞类型进行注释，涵盖10个数据集。结果表明，ScLabel-Net在小鼠肺部数据集上的平均准确率比现有最佳方法提升了约2%，且在处理复杂数据集时显著优于基于相关性和深度学习的方法。研究还发现，随机上采样的数据增强技术能有效缓解单细胞数据集中常见的类型分布不平衡问题。ScLabel-Net不仅减少了训练时间，还有效提升了小鼠细胞类型注释的准确性和效率，加速了生物学领域的研究进程，解决了大规模数据处理中的挑战，提升了细胞类型注释的可行性和有效性。

参考文献

[1] JOVIC D, LIANG X, ZENG H, et al. Single-cell RNA sequencing technologies and applications: a brief overview [J]. Clinical and Translational Medicine, 2022, 12(3): e694.

[2] 丁宁，张然然，范广轩，等. 基于单细胞转录组测序的鹿茸生长中心细胞异质性研究[J]. 黑龙江畜牧兽医，2024(19)：108-112，123-127.

[3] CUI H T, WANG C, MAAN H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI[J]. Nature Methods, 2024, 21(8): 1470-1480.

[4] 袁佳欣，刘宏德. scTransformer：一种基于深度学习的单细胞类型识别方法[J]. 生物信息学，2025，23(2)：96-110.

[5] FRANZÉN O, GAN L M, BJÖRKEGREN J L M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data [J]. Database, 2019, 2019: baz046.

[6] HU C X, LI T Y, XU Y Q, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data [J]. Nucleic Acids Research, 2023, 51(D1): D870-D876.

[7] 熊怡绚. 面向单细胞转录组数据的细胞身份识别方法研究[D]. 武汉：华中师范大学，2024.

[8] 李霖，沈永健，张鹏宇，等. 基于CNN-Transformer的自编码器红外和可见光图像融合方法[J]. 遥测遥

控,2024,45(5):109-119.
[9] KINGMA D P, WELLING M. Auto-encoding variational bayes [Z/OL]. (2022-12-10) [2025-01-10].https://doi.org/10.48550/arXiv.1312.6114.
[10] GOODFELLOW I,POUGET-ABADIE J,MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.
[11] 郭兴君,李晓红,史婉嫄,等.融合模体感知和图Transformer编码的社区检测[J].计算机工程与科学,2024,46(11):2081-2090.
[12] 杨文瀚,胡之浩,郭伟壮,等.基于加权基因共表达网络分析和机器学习筛选椎间盘退行性变中的标志线粒体基因[J].脊柱外科杂志,2024,22(5):321-327.
[13] YANG F,WANG W C,WANG F,et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data [J]. Nature Machine Intelligence,2022,4(10):852-866.
[14] 郑建华,李小敏,刘双印,等.融合级联上采样与下采样的改进随机森林不平衡数据分类算法[J].计算机科学,2021,48(7):145-154.
[15] JIAO L F, WANG G, DAI H H, et al. scTransSort: Transformers for intelligent annotation of cell types by gene embeddings [J]. Biomolecules,2023,13(4):611.
[16] 刘德真,李圆媛.基于深度学习和多组学数据的肺腺癌分期预测研究[J].武汉工程大学学报,2024,46(2):190-196.
[17] WOLF F A, ANGERER P, THEIS F J. SCANPY: large-scale single-cell gene expression data analysis [J]. Genome Biology,2018,19:15.
[18] PAN C Z, WEI H, CHEN B, et al. Inhalation of itraconazole mitigates bleomycin-induced lung fibrosis via regulating SPP1 and C3 signaling pathway pivotal in the interaction between phagocytic macrophages and diseased fibroblasts[J]. Journal of Translational Medicine,2024,22(1):1058.
[19] KANG Z Y, HUANG Q Y, ZHEN N X, et al. Heterogeneity of immune cells and their communications unveiled by transcriptome profiling in acute inflammatory lung injury [J]. Frontiers in Immunology,2024,15:1382449.
[20] CAO Z J, WEI L, LU S, et al. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST[J]. Nature Communications,2020,11:3458.
[21] ZENG Y S, LUO M, SHANGGUAN N Y, et al. Deciphering cell types by integrating scATAC-seq data with genome sequences [J]. Nature Computational Science,2024,4(4):285-298.
[22] WANG H C, FU T F, DU Y Q, et al. Scientific discovery in the age of artificial intelligence [J]. Nature,2023,620:47-60.
[23] 宋鑫伟,李瑞芳,高姗,等.线粒体核糖体蛋白基因内含子间最佳匹配片段的相对位置分布[J].内蒙古师范大学学报(自然科学版),2024,53(3):306-312.
[24] WAN H, YUAN M S, FU Y W, et al. Continually adapting pre-trained language model to universal annotation of single-cell RNA-seq data[J]. Briefings in Bioinformatics,2024,25(2):bbae047.
[25] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional Transformers for language understanding[Z/OL]. (2019-05-24)[2025-01-10]. https://doi.org/10.48550/arXiv.1810.04805.
[26] 张渊,姚峰.基于知识嵌入式预训练语言模型的文本分类方法研究[J].武汉工程大学学报,2023,45(6):674-679.
[27] ARAN D,LOONEY A P,LIU L Q,et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage [J]. Nature Immunology,2019,20(2):163-172.
[28] KISELEV V Y, YIU A, HEMBERG M. Scmap: projection of single-cell RNA-seq data across data sets [J]. Nature Methods,2018,15(5):359-362.
[29] de KANTER J K,LIJNZAAD P,CANDELLI T,et al. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing [J]. Nucleic Acids Research,2019,47(16):e95.
[30] MA F Y, PELLEGRINI M. ACTINN: automated identification of cell types in single cell RNA sequencing [J]. Bioinformatics, 2020, 36(2): 533-538.
[31] CHEN J W,XU H,TAO W Y,et al. Transformer for one stop interpretable cell type annotation[J]. Nature Communications,2023,14:223.
[32] JAMIL A, AHMAD A, MOEEN-UD-DIN M ,et al. Unveiling the mechanism of micro-and-nano plastic phytotoxicity on terrestrial plants: a comprehensive review of omics approaches [J]. Environment International,2025,195:109257.

本文编辑:霍淑利