Supervised Adversarial Alignment of Single-Cell RNA-seq Data

监督对抗对齐单细胞 RNA-seq 数据

SONGWEI GE, [1] HAOHAN WANG, [2] AMIR ALAVI, ERIC XING, [2,3] and ZIV BAR-JOSEPH [1,3]
SONGWEI GE, [1] HAOHAN WANG, [2] AMIR ALAVI, ERIC XING, [2,3] 和 ZIV BAR-JOSEPH [1,3]

ABSTRACT

摘要

Dimensionality reduction is an important first step in the analysis of single-cell RNA-sequencing (scRNA-seq) data. In addition to enabling the visualization of the profiled cells, such representations are used by many downstream analyses methods ranging from pseudo-time reconstruction to clustering to alignment of scRNA-seq data from different experiments, platforms, and laboratories. Both supervised and unsupervised methods have been proposed to reduce the dimension of scRNA-seq. However, all methods to date are sensitive to batch effects. When batches correlate with cell types, as is often the case, their impact can lead to representations that are batch rather than cell-type specific. To overcome this, we developed a domain adversarial neural network model for learning a reduced dimension representation of scRNA-seq data. The adversarial model tries to simultaneously optimize two objectives. The first is the accuracy of cell-type assignment and the second is the inability to distinguish the batch (domain). We tested the method by using the resulting representation to align several different data sets. As we show, by overcoming batch effects our method was able to correctly separate cell types, improving on several prior methods suggested for this task. Analysis of the top features used by the network indicates that by taking the batch impact into account, the reduced representation is much better able to focus on key genes for each cell type.

降维是单细胞 RNA 测序 (scRNA-seq) 数据分析中的一个重要第一步。除了能够可视化所分析的细胞外，这种表示方法还被许多下游分析方法使用，从伪时间重建到聚类，再到来自不同实验、平台和实验室的 scRNA-seq 数据的对齐。目前已经提出了监督和无监督方法来降低 scRNA-seq 的维度。然而，迄今为止的所有方法都对批次效应敏感。当批次与细胞类型相关时，通常会出现这种情况，其影响可能导致表示方法更倾向于批次而非细胞类型特异性。为了解决这个问题，我们开发了一种领域对抗神经网络模型，用于学习 scRNA-seq 数据的降维表示。对抗模型试图同时优化两个目标。第一个是细胞类型分配的准确性，第二个是无法区分批次 (领域)。我们通过使用生成的表示对齐几个不同的数据集来测试该方法。正如我们所展示的，通过克服批次效应，我们的方法能够正确分离细胞类型，优于为此任务建议的几个先前方法。对网络使用的主要特征的分析表明，通过考虑批次影响，降维表示能够更好地关注每种细胞类型的关键基因。

Keywords: batch effect removal, data integration, dimensionality reduction, domain adversarial training, single-cell RNA-seq.

关键词: 批次效应去除，数据整合，降维，领域对抗训练，单细胞 RNA-seq。

1. INTRODUCTION

1. 引言

INGLE-CELL RNA SEQUENCING (scRNA-seq) has revolutionized the study of gene expression programs (Hwang et al., 2018; Papalexi and Satija, 2018). The ability to profile genes at the single-cell level has revealed novel specific interactions and pathways within cells (Yu et al., 2016), differences in the proportions of cell types between samples (Jaitin et al., 2014; Zeisel et al., 2015), and the identity and characterization of new cell types (Villani et al., 2017). Several biological tissues, systems, and processes have recently been studied using this technology (Jaitin et al., 2014; Zeisel et al., 2015; Yu et al., 2016).

单细胞 RNA 测序 (scRNA-seq) 革新了基因表达程序的研究 (Hwang et al., 2018; Papalexi and Satija, 2018)。在单细胞水平上分析基因的能力揭示了细胞内的新特定相互作用和通路 (Yu et al., 2016)、样本之间细胞类型比例的差异 (Jaitin et al., 2014; Zeisel et al., 2015)，以及新细胞类型的身份和特征 (Villani et al., 2017)。最近，多个生物组织、系统和过程已使用该技术进行了研究 (Jaitin et al., 2014; Zeisel et al., 2015; Yu et al., 2016)。

Although studies using scRNA-seq provide many insights, they also raise new computational challenges. One of the major challenges involves the ability to integrate and compare results from multiple scRNA-seq studies. There are several different commercial platforms for performing such experiments, each with their own biases. Furthermore, similar to other high-throughput genomic assays, scRNA-seq suffers from batch effects that can make cells profiled in one laboratory look very different from the same cells profiled at another laboratory (Tung et al., 2017; Stuart and Satija, 2019). This is a key issue for consortium-scale analysis, such as the Human Cell Atlas (Rozenblatt-Rosen et al., 2017; Regev et al., 2017) and HUBMaP Consortium (2019), where researchers across the globe are profiling single cells in their own laboratories and seeking to perform large-scale analysis that integrates data across the entire consortia. Even for cell profiles in the same laboratory, we often cannot avoid batch effects, for example, in studies where samples are collected at different times or across a large set of individuals (Nowotschin et al., 2019; Pijuan-Sala et al., 2019). Moreover, other types of high-throughput transcriptomics profiling, including microscopy-based techniques, are also generating single-cell expression data sets (Wang et al., 2018; Eng et al., 2019). The goal of fully utilizing these spatial data sets motivates the development of methods that can combine them with scRNA-seq when studying specific biological tissues and processes.

尽管使用 scRNA-seq 的研究提供了许多见解，但它们也带来了新的计算挑战。主要挑战之一是整合和比较来自多个 scRNA-seq 研究的结果的能力。进行此类实验的商业平台有多种，每种平台都有其自身的偏见。此外，类似于其他高通量基因组检测，scRNA-seq 也受到批次效应的影响，这可能导致在一个实验室中分析的细胞与在另一个实验室中分析的相同细胞看起来非常不同 (Tung et al., 2017; Stuart and Satija, 2019)。这是联盟规模分析的关键问题，例如人类细胞图谱 (Rozenblatt-Rosen et al., 2017; Regev et al., 2017) 和 HUBMaP 联盟 (2019)，在这些研究中，全球的研究人员在自己的实验室中分析单细胞，并寻求进行跨整个联盟的数据整合的大规模分析。即使是在同一实验室中的细胞分析中，我们也常常无法避免批次效应，例如在样本在不同时间收集或跨一大组个体的研究中 (Nowotschin et al., 2019; Pijuan-Sala et al., 2019)。此外，其他类型的高通量转录组分析，包括基于显微镜的技术，也正在生成单细胞表达数据集 (Wang et al., 2018; Eng et al., 2019)。充分利用这些空间数据集的目标推动了在研究特定生物组织和过程时能够将其与 scRNA-seq 结合的方法的发展。

A number of recent methods have attempted to address this challenge by developing methods for aligning scRNA-seq data from multiple studies of the same biological system. Many of these methods rely on identifying nearest neighbors between the different data sets and using them as anchors. Methods that use this approach include mutual nearest neighbors (MNNs) (Haghverdi et al., 2018) and Seurat (Stuart et al., 2019). Others including scVI and scAlign first embed all data sets into a common lower dimensional space. scVI encodes the scRNA-seq data with a deep generative model conditioning on the batch identifiers (Lopez et al., 2018), whereas scAlign regularizes the representation between two data sets by minimizing the random walk probability differences between the original and embedding spaces. Although these methods were successful for some data sets, here we show that they are not always able to correctly match all cell types. A key problem with these methods is the fact that they are unsupervised and rely on the assumption that cell types profiled by the different studies overlap. Although this works for some data sets, it may fail for studies in which cells do not fully overlap or for those containing rare cell types. Unsupervised methods tend to group rare types with the larger types, making it hard to identify them in a joint space.

最近有多种方法试图通过开发对齐来自同一生物系统的多个研究的 scRNA-seq 数据的方法来应对这一挑战。这些方法中的许多依赖于识别不同数据集之间的最近邻并将其用作锚点。使用这种方法的方法包括互最近邻 (MNNs) (Haghverdi et al., 2018) 和 Seurat (Stuart et al., 2019)。其他方法如 scVI 和 scAlign 首先将所有数据集嵌入到一个共同的低维空间中。scVI 使用深度生成模型对 scRNA-seq 数据进行编码，并以批次标识符为条件 (Lopez et al., 2018)，而 scAlign 通过最小化原始空间和嵌入空间之间的随机游走概率差异来规范两个数据集之间的表示。尽管这些方法在某些数据集上取得了成功，但我们在此表明，它们并不总能正确匹配所有细胞类型。这些方法的一个关键问题在于它们是无监督的，并依赖于不同研究所描绘的细胞类型重叠的假设。尽管这对某些数据集有效，但对于细胞未完全重叠的研究或包含稀有细胞类型的研究，它可能会失败。无监督方法倾向于将稀有类型与较大类型分组，从而使其在联合空间中难以识别。

Recent machine learning work has focused on a related problem termed "domain adaptation/ generalization." Methods developed for these problems attempt to learn representations of diverse data that

[1] Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
[1] 计算生物学系，卡内基梅隆大学，匹兹堡，宾夕法尼亚州，美国。
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
[2] 语言技术研究所，卡内基梅隆大学，匹兹堡，宾夕法尼亚州，美国。
[3] Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
[3] 机器学习系，卡内基梅隆大学，匹兹堡，宾夕法尼亚州，美国。

are invariant to technical confounders (Csurka, 2017; Motiian et al., 2017; Wang et al., 2019b). These methods have been used for multiple applications such as machine translation for domain-specific corpus (Chu and Wang, 2018) and face detection (Patel et al., 2015). Several methods proposed for domain adaptation rely on the use of adversarial methods (Ganin et al., 2016; Csurka, 2017; Li et al., 2018; Wang et al., 2019a), which has been proved effective to align latent distributions. In addition to the original task such as classification, these methods apply a domain classifier upon the learned representations. The encoder network is used for improving accurate classification while at the same time reducing the impact of the domain (by "fooling" a domain classifier). This is achieved by learning encoder weights that simultaneously perform gradient descent on the label classification task and gradient ascent on the domain classification task.

最近的机器学习研究集中在一个相关的问题上，称为“领域适应/泛化”。为这些问题开发的方法试图学习对技术干扰不变的多样数据表示 (Csurka, 2017; Motiian et al., 2017; Wang et al., 2019b)。这些方法已被用于多个应用，例如针对特定领域语料库的机器翻译 (Chu 和 Wang, 2018) 和人脸检测 (Patel et al., 2015)。为领域适应提出的几种方法依赖于对抗性方法的使用 (Ganin et al., 2016; Csurka, 2017; Li et al., 2018; Wang et al., 2019a)，这些方法已被证明在对齐潜在分布方面有效。除了原始任务 (如分类) 外，这些方法还在学习到的表示上应用领域分类器。编码器网络用于提高准确分类，同时减少领域的影响 (通过“欺骗” 领域分类器)。这通过学习编码器权重来实现，这些权重同时对标签分类任务执行梯度下降，对领域分类任务执行梯度上升。

Here we extend these approaches, coupling them with Siamese network learning (Koch et al., 2015) for overcoming batch effects in scRNA-seq analysis. We define a "domain" in this article as a standalone data set profiled at a single laboratory using a single platform. We define "label" as the cell type for each cell in the data set. Considering the specificity of the cell types in the scRNA-seq data sets, we propose a conditional pair sampling strategy that constrains input pair selection when training the adversarial network. We discuss how to formulate a domain adaptation network for scRNA-seq data, how to learn the parameters for the network, and how to train it using available data.

在这里，我们扩展了这些方法，将其与西阿摩斯网络学习 (Koch et al., 2015) 结合，以克服 scRNA-seq 分析中的批次效应。我们在本文中将“领域” 定义为在单一实验室使用单一平台分析的独立数据集。我们将“标签” 定义为数据集中每个细胞的细胞类型。考虑到 scRNA-seq 数据集中细胞类型的特异性，我们提出了一种条件对采样策略，该策略在训练对抗网络时限制输入对的选择。我们讨论了如何为 scRNA-seq 数据制定领域适应网络，如何学习网络的参数，以及如何使用可用数据对其进行训练。

We tested our method on several data sets ranging in size from 10 to 39 cell types and from 4 to 155 batches. As we show, for all of the data sets, our domain adversarial method improves on previous methods, in some cases significantly. Visualization of the learned representation from several different methods helps highlight the advantages of the domain adversarial framework. As we show, the framework is able to accurately mitigate the batch effects while maintaining the grouping of cells from the same type across different batches. Biological analysis of the resulting model identifies key genes that can correctly distinguish between cell types across different experiments. Such batch invariant genes are promising candidates for a cell-type specific signature that can be used across different studies to annotate cells.

我们在多个数据集上测试了我们的方法，这些数据集的细胞类型数量从 10 到 39 不等，批次数量从 4 到 155 不等。正如我们所展示的，对于所有数据集，我们的领域对抗方法在某些情况下显著优于以前的方法。对几种不同方法学习到的表示进行可视化，有助于突出领域对抗框架的优势。正如我们所展示的，该框架能够准确减轻批次效应，同时保持来自不同批次的相同类型细胞的分组。对结果模型的生物学分析识别出能够正确区分不同实验中细胞类型的关键基因。这些批次不变基因是细胞类型特异性标记的有希望的候选者，可以在不同研究中用于注释细胞。

2. METHODS

2. 方法

2.1. Problem formulation

2.1. 问题表述

To formulate the problem we start with a few notation definitions. We assume that the single-cell RNA-seq data are drawn from the input space $\mathbf{X} \in \mathbb{R}^p$, where each sample (a cell) $\mathbf{x}$ has p features corresponding to the gene expression values. Cells are also associated with the label $y \in \mathbf{Y} = \{1, 2, \ldots, K\}$, which represents their cell types. We associate each sample with a specific domain/batch $d \in \mathcal{D}$ that represents

any standalone data set profiled at a single laboratory using a single platform. Note that we will use domain and batch interchangeably in this article for convenience. The data are divided into a training set and a test set that are drawn from multiple studies. The domains used to collect training data are not used for the test set and so batch effects can vary between the training and test data. In practice, each of the domains only contains a small subset of the cell types. This means that the distribution of cell types is correlated with the distribution of domains. Thus, the methods that naively learn cell types based on expression profile (Alavi et al., 2018; Kiselev et al., 2018; Lieberman et al., 2018) may instead fit domain information and not generalize well to the unobserved studies.

为了制定问题，我们首先定义一些符号。我们假设单细胞 RNA 测序数据来自输入空间 $\mathbf{X} \in \mathbb{R}^p$ ，其中每个样本 (一个细胞) $\mathbf{x}$ 具有 p 个特征，对应于基因表达值。细胞还与标签 $y \in \mathbf{Y} = \{1, 2, \ldots, K\}$ 相关联，表示它们的细胞类型。我们将每个样本与一个特定的领域/批次 $d \in \mathcal{D}$ 相关联，该领域/批次代表在单一实验室使用单一平台分析的任何独立数据集。请注意，为了方便起见，我们将在本文中交替使用领域和批次。数据被分为训练集和测试集，这些数据来自多个研究。用于收集训练数据的领域不用于测试集，因此批次效应在训练数据和测试数据之间可能会有所不同。在实践中，每个领域仅包含少量细胞类型。这意味着细胞类型的分布与领域的分布相关。因此，基于表达谱天真学习细胞类型的方法 (Alavi et al., 2018; Kiselev et al., 2018; Lieberman et al., 2018) 可能会拟合领域信息，而无法很好地推广到未观察到的研究中。

2.2. Domain adversarial training with Siamese network

2.2. 使用孪生网络的领域对抗训练

To overcome this problem and remove the domain impact when learning a cell-type representation, we propose a neural network (NN) framework that includes three modules as shown in Figure 1: scRNA encoder, label classifier, and domain discriminator. The encoder module $f_e(\mathbf{x}; \theta_e)$ is used to reduce the dimensions of the data and contains fully connected layers that produce the hidden features, where θ_e represents the parameters in these layers. The label classifier $f_l(f_e; \theta_l)$ attempts to predict the label of input $\mathbf{x_1}$, whereas the goal of the domain discriminator $f_d(f_e; \theta_d)$ is to determine whether a pair of inputs $\mathbf{x_1}$ and $\mathbf{x_2}$ is from the same domain or not. Past work for classifying scRNA-seq data only attempted to minimize the loss function for the label classifier $\mathcal{L}_l(f_i(f_e; \theta_l))$ (Lin et al.,2017; Alavi et al.,2018). Here, we extend these methods by adding a regularization term based on the adversarial loss of the domain discriminator $\mathcal{L}_d(f_d(f_e; \theta_d))$, which we will elaborate later. The overall loss E on a pair of samples $\mathbf{x_1}$ and $\mathbf{x_2}$ is denoted by:

为了克服这个问题并消除在学习细胞类型表示时的领域影响，我们提出了一种神经网络 (NN) 框架，包括如图 1 所示的三个模块:scRNA 编码器、标签分类器和领域鉴别器。编码器模块 $f_e(\mathbf{x}; \theta_e)$ 用于降低数据的维度，并包含生成隐藏特征的全连接层，其中 θ_e 表示这些层中的参数。标签分类器 $f_l(f_e; \theta_l)$ 尝试预测输入 $\mathbf{x_1}$ 的标签，而领域鉴别器 $f_d(f_e; \theta_d)$ 的目标是确定一对输入 $\mathbf{x_1}$ 和 $\mathbf{x_2}$ 是否来自同一领域。过去对 scRNA-seq 数据的分类工作仅试图最小化标签分类器的损失函数 $\mathcal{L}_l(f_i(f_e; \theta_l))$ (Lin et al., 2017; Alavi et al., 2018)。在这里，我们通过添加基于领域鉴别器的对抗损失 $\mathcal{L}_d(f_d(f_e; \theta_d))$ 的正则化项来扩展这些方法，稍后我们将详细阐述。样本对 $\mathbf{x_1}$ 和 $\mathbf{x_2}$ 的整体损失 E 表示为:

$$E(\theta_e, \theta_l, \theta_d) = \mathcal{L}_l(f_l(f_e(\mathbf{x_1}; \theta_e); \theta_l)) - \lambda \mathcal{L}_d(f_d(f_e(\mathbf{x_1}; \theta_e); \theta_d), f_d(f_e(\mathbf{x_2}; \theta_e); \theta_d)),$$

where λ can control the trade-off between the goals of domain invariance and higher classification accuracy. For convenience, we use $\mathbf{z_1}$ and $\mathbf{z_2}$ to denote the hidden representations of $\mathbf{x_1}$ and $\mathbf{x_2}$ calculated from $f_e(\mathbf{x}; \theta_e)$. Inspired by Siamese networks (Koch et al.,2015), we implement our domain discriminator by adopting a contrastive loss (Hadsell et al., 2006):

其中 λ 可以控制领域不变性和更高分类准确性目标之间的权衡。为了方便起见，我们用 $\mathbf{z_1}$ 和 $\mathbf{z_2}$ 表示从 $f_e(\mathbf{x}; \theta_e)$ 计算得到的 $\mathbf{x_1}$ 和 $\mathbf{x_2}$ 的隐藏表示。受到孪生网络 (Koch et al., 2015) 的启发，我们通过采用对比损失 (Hadsell et al., 2006) 来实现我们的领域鉴别器:

$$\mathcal{L}_d(f_d(\mathbf{z_1}; \theta_d), f_d(\mathbf{z_2}; \theta_d)) = U\frac{1}{2}D(f_d(\mathbf{z_1}), f_d(\mathbf{z_2}))^2$$

$$+ (1 - U)\frac{1}{2}(\max\{0, m - D(f_d(\mathbf{z_1}), f_d(\mathbf{z_2}))\})^2,$$

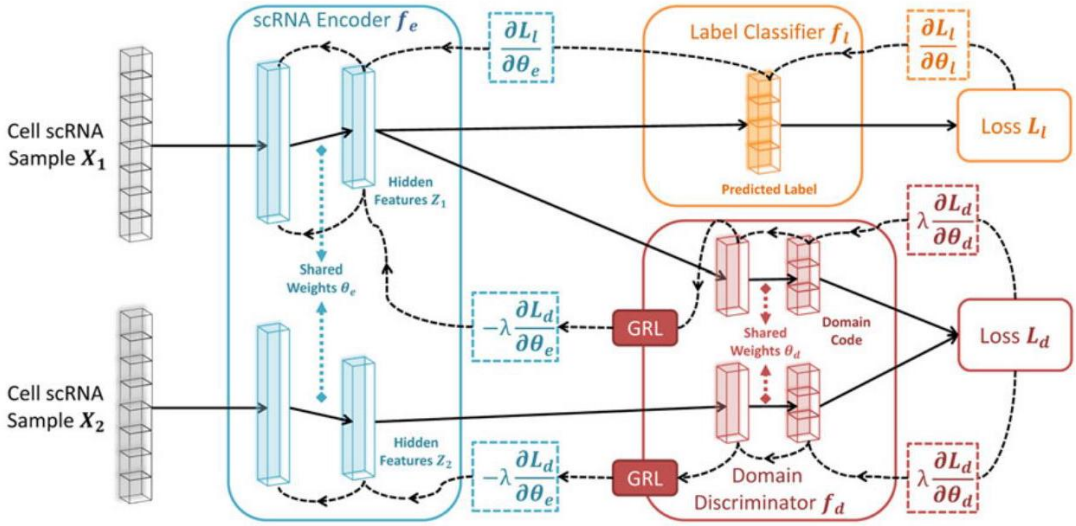

FIG. 1. Architecture of scDGN. The network includes three modules: scRNA encoder f_e (blue), label classifier f_l (orange), and domain discriminator f_d (red). Note that the red and orange networks use the same encoding as input. Solid lines represent the forward direction of the NN, whereas the dashed lines represent the backpropagation direction with the corresponding gradient it passes. GRLs have no effect in forward propagation, but flip the sign of the gradients that flow through them during backpropagation. This allows the combined network to simultaneously optimize label classification and attempt to "fool" the domain discriminator. Thus, the encoder leads to representations that are invariant to the different domains while still distinguishing cell types. GRLs, gradient reversal layers; NN, neural network; scDGN, single-cell domain generalization network; scRNA, single-cell RNA.

图 1. scDGN 的架构。该网络包括三个模块:scRNA 编码器 f_e (蓝色)、标签分类器 f_l (橙色) 和领域判别器 f_d (红色)。请注意，红色和橙色网络使用相同的编码作为输入。实线表示神经网络的前向传播方向，而虚线表示反向传播方向及其所经过的相应梯度。GRL 在前向传播中没有影响，但在反向传播过程中会翻转流经它们的梯度符号。这使得组合网络能够同时优化标签分类并尝试“欺骗”领域判别器。因此，编码器生成的表示对不同领域是不变的，同时仍能区分细胞类型。GRL，梯度反转层；NN，神经网络；scDGN，单细胞领域泛化网络；scRNA，单细胞 RNA。

where $U = 1$ indicates that two samples are from the same domain d and $U = 0$ indicates that they are not, $D(\cdot)$ is the euclidean distance, and m is the margin that indicates the prediction boundary. The domain discriminator parameters, θ_d , are updated using back propagation to maximize the total loss E , whereas the encoder and classifier parameters, θ_e and θ_l , are updated to minimize E . To allow all three modules to be updated together end-to-end, we use a gradient reversal layer (GRL, Fig. 1) (Ganin et al., 2016; Pei et al., 2018). Specifically, GRLs have no effect in forward propagation, but flip the sign of the gradients that flow through them during backpropagation. The following provides the overall optimization problems solved for the network parameters:

其中 $U = 1$ 表示两个样本来自同一领域 d ，而 $U = 0$ 表示它们不来自同一领域，$D(\cdot)$ 是欧几里得距离，m 是表示预测边界的边际。领域判别器参数 θ_d 通过反向传播进行更新，以最大化总损失 E ，而编码器和分类器参数 θ_e 和 θ_l 则被更新以最小化 E 。为了使所有三个模块能够端到端地一起更新，我们使用了一个梯度反转层 (GRL，图 1)(Ganin 等，2016；Pei 等，2018)。具体而言，GRL 在前向传播中没有影响，但在反向传播过程中会翻转流经它们的梯度符号。以下提供了为网络参数解决的整体优化问题:

$$\left(\widehat{\theta}_e, \widehat{\theta}_l\right) = \arg\min_{\theta_e, \theta_l} E\left(\theta_e, \theta_l, \widehat{\theta}_d\right)$$

$$\left(\widehat{\theta}_d\right) = \arg\max_{\theta_d} E\left(\widehat{\theta}_e, \widehat{\theta}_l, \theta_d\right)$$

In other words, the goal of the domain discriminator is to tell whether two samples are drawn from the same or different batches. By optimizing the scRNA encoder adversarially against the domain discriminator, we attempt to make sure that the network representation cannot be used to classify based on domain knowledge. During the training, the maximization and minimization tasks compete with each other, which is achieved by adjusting the representations to improve the accuracy of the label classifier and simultaneously fool the domain discriminator.

换句话说，领域鉴别器的目标是判断两个样本是否来自同一批次或不同批次。通过对抗性地优化 scRNA 编码器以对抗领域鉴别器，我们试图确保网络表示不能基于领域知识进行分类。在训练过程中，最大化和最小化任务相互竞争，这通过调整表示来提高标签分类器的准确性，同时欺骗领域鉴别器来实现。

2.3. Conditional domain generalization strategy

2.3. 条件领域泛化策略

Most prior domain adaption or generalization methods focused on the cases wherein the distribution of labels is independent of the domains (Csurka, 2017; Motiian et al., 2017). In contrast, as we show in Results section, for scRNA-seq experiments different studies tend to focus on certain cell types (Jaitin et al., 2014; Zeisel et al. 2015; Yu et al. 2016). Consequently, it is not reasonable to completely merge the scRNA-seq data from different batches. To be specific, aligning the scRNA-seq data from two batches with different sets of cell types would sacrifice its biological significance and prevent the cell classifier from predicting effectively. To overcome this issue, instead of arbitrarily choosing positive pairs (samples from the same domain) and negative pairs (samples from different domains), we constrain the selection as follows: (1) for positive pairs, only the samples with different labels from the same domain are selected, (2) for negative pairs, only the samples with the same label from different domains are selected. Figure 2 provides a visual interpretation of this strategy. Formally, letting y_i and z_i represent the i -th sample' s cell-type label and domain label respectively, we have the following equations to define the value of U for sample pairs:

大多数先前的领域适应或泛化方法集中在标签分布与领域独立的情况 (Csurka, 2017; Motiian et al., 2017)。相比之下，正如我们在结果部分所示，对于 scRNA-seq 实验，不同研究往往集中在某些细胞类型上 (Jaitin et al., 2014; Zeisel et al. 2015; Yu et al. 2016)。因此，完全合并来自不同批次的 scRNA-seq 数据是不合理的。具体而言，将来自不同细胞类型集合的两个批次的 scRNA-seq 数据对齐将牺牲其生物学意义，并阻碍细胞分类器的有效预测。为了解决这个问题，我们不再任意选择正样本对 (来自同一领域的样本) 和负样本对 (来自不同领域的样本)，而是将选择限制为以下内容:(1) 对于正样本对，仅选择来自同一领域但标签不同的样本；(2) 对于负样本对，仅选择来自不同领域但标签相同的样本。图 2 提供了该策略的可视化解释。形式上，设 y_i 和 z_i 分别表示第 i 个样本的细胞类型标签和领域标签，我们有以下方程来定义样本对的 U 值:

$$U = \begin{cases} 0, & z_1 \neq z_2 \text{ and } y_1 = y_2 \\ 1, & z_1 = z_2 \text{ and } y_1 \neq y_2 \end{cases}$$

This strategy prevents the domain adversarial training from aligning samples with different labels or separating samples with same labels. For example, to fool the discriminator with a positive pair, the encoder must implicitly increase the distance of two samples with different cell types. Therefore, combining this strategy with domain adversarial training allows the network to learn cell-type specific focused representations. We term our model single-cell domain generalization network (scDGN).

该策略防止领域对抗训练将不同标签的样本对齐或将相同标签的样本分离。例如，为了欺骗带正样本对的判别器，编码器必须隐式地增加不同细胞类型的两个样本之间的距离。因此，将该策略与领域对抗训练结合使用，使网络能够学习细胞类型特定的聚焦表示。我们将我们的模型称为单细胞领域泛化网络 (scDGN)。

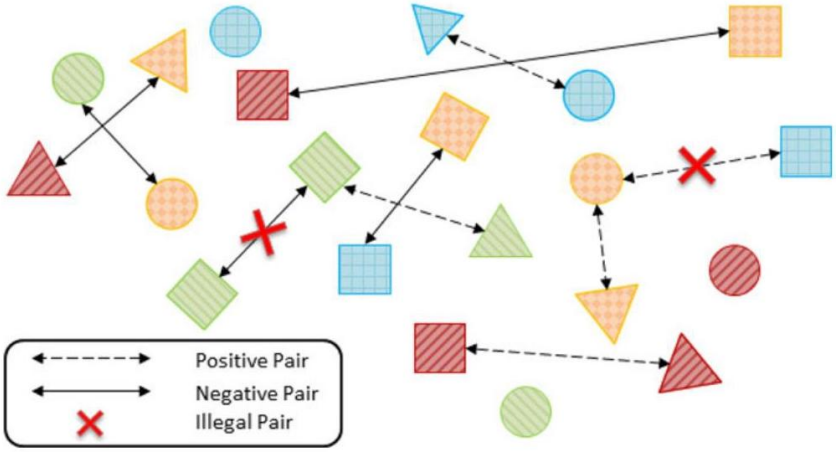

FIG. 2. Conditional domain generalization strategy: shapes represent different labels, and colors (or patterns) represent different domains. For negative pairs from different domains, we only select those samples with the same label. For positive pairs from the same domain, we only select the samples with different labels.

图 2. 条件领域泛化策略: 形状代表不同的标签，颜色 (或图案) 代表不同的领域。对于来自不同领域的负样本对，我们仅选择那些具有相同标签的样本。对于来自同一领域的正样本对，我们仅选择具有不同标签的样本。

3. RESULTS

3. 结果

3.1. Experiment setups

3.1. 实验设置

3.1.1. Data sets. To test our method and to compare it with previous methods for aligning and classifying scRNA-seq data, we used several recent data sets. These data sets contain between 6000 and 45,000 cells, and all include cells profiled in multiple experiments by different laboratories and on different platforms.

3.1.1. 数据集。为了测试我们的方法并与之前用于对齐和分类 scRNA-seq 数据的方法进行比较，我们使用了几个最近的数据集。这些数据集包含 6000 到 45000 个细胞，所有数据集均包含在不同实验和不同平台上由不同实验室分析的细胞。

The evaluation data sets include a subset of the data from scQuery (Alavi et al., 2018), which contains 44,490 samples from 155 different experiments, including a broad range of cell types. In addition, we use a peripheral blood mononuclear cell (PBMC) data set with nine batches (sequencing technologies) and 28,969 cells (Ding et al., 2019). Finally we also test on a data set of human pancreatic islet cells from Seurat (Stuart et al., 2019), which we artificially split into 6 smaller data sets to simulate cases wherein cell types and domains are highly correlated. See Table 1 and Section A. 1 in Supplementary Material for details on batches, cell-type distributions, train and test split information, and normalization for these three data sets.

评估数据集包括来自 scQuery(Alavi 等，2018) 的数据子集，其中包含来自 155 个不同实验的 44,490 个样本，涵盖广泛的细胞类型。此外，我们使用一个外周血单核细胞 (PBMC) 数据集，包含九个批次 (测序技术) 和 28,969 个细胞 (Ding 等，2019)。最后，我们还在 Seurat(Stuart 等，2019) 的人胰腺岛细胞数据集上进行测试，我们将其人为地拆分为 6 个较小的数据集，以模拟细胞类型和领域高度相关的情况。有关这些数据集的批次、细胞类型分布、训练和测试拆分信息以及归一化的详细信息，请参见补充材料中的表 1 和 A. 1 节。

3.1.2. Model configurations. We used the network of Lin et al. (2017) as the components for the encoder and the label classifier in our model. The encoder contains two hidden layers with 1136 and 100 units. The label classifier is directly connected to the 100 unit layer and makes predictions based on these values. The domain discriminator contains an additional hidden layer with 64 units and is also connected to the 100 unit layer of the encoder (Fig. 1). For each layer, tanh( ) is used as the nonlinear

activation function. We test several other possible configurations but did not observe improvement in performance. As is commonly done, we use a validation set to tune the hyperparameters for learning including learning rates, decay, momentum, and the adversarial weight and margin parameters λ and m . In general, our analysis indicates that for larger data sets, a lower weight λ and larger margin m for the adversarial training are preferred and vice versa. More details about the hyperparameters and training are provided in Section A. 3 in Supplementary Material.

3.1.2. 模型配置。我们使用 Lin 等人 (2017) 的网络作为我们模型中编码器和标签分类器的组件。编码器包含两个隐藏层，分别有 1136 和 100 个单元。标签分类器直接连接到 100 单元层，并基于这些值进行预测。领域判别器包含一个额外的隐藏层，具有 64 个单元，并且也连接到编码器的 100 单元层 (图 1)。对于每一层，使用 tanh( ) 作为非线性激活函数。我们测试了几种其他可能的配置，但没有观察到性能的改善。通常，我们使用验证集来调整学习的超参数，包括学习率、衰减、动量，以及对抗权重和边际参数 λ 和 m 。一般来说，我们的分析表明，对于较大的数据集，更低的权重 λ 和更大的边际 m 更适合对抗训练，反之亦然。有关超参数和训练的更多细节见补充材料的 A.3 节。

3.1.3. Baselines. We compared scDGN with several prior methods for classifying and aligning scRNA-seq data. These included the NN model of Lin et al. (2017), which is developed for classifying scRNA-seq data, CaSTLe (Lieberman et al., 2018), which performs cell-type classification based on transfer learning, and several state-of-the-art alignment methods. For alignment, we compared with MNN (Haghverdi et al., 2018), which utilizes MNNs to align data from different batches, scVI (Lopez et al., 2018), which trains a deep generative model on the scRNA-seq data and uses an explicit batch identifier to retain conditional independence property of the representation, and Seurat (Stuart et al., 2019), which first identifies the anchors among different batches and then projects different data sets using a correction vector based on the order defined by hierarchical clustering with pairwise distances.

3.1.3. 基线。我们将 scDGN 与几种先前的方法进行了比较，以对 scRNA-seq 数据进行分类和对齐。这些方法包括 Lin 等人 (2017) 开发的 NN 模型，该模型用于分类 scRNA-seq 数据，CaSTLe(Lieberman 等人，2018)，该模型基于迁移学习进行细胞类型分类，以及几种最先进的对齐方法。在对齐方面，我们与 MNN(Haghverdi 等人，2018) 进行了比较，该方法利用 MNN 对来自不同批次的数据进行对齐，scVI(Lopez 等人，2018)，该方法在 scRNA-seq 数据上训练深度生成模型，并使用显式批次标识符以保持表示的条件独立性，以及 Seurat(Stuart 等人，2019)，该方法首先识别不同批次之间的锚点，然后使用基于层次聚类和成对距离定义的顺序的修正向量对不同数据集进行投影。

TABLE 1. BASIC STATISTICS FOR SCQUERY, SUERAT PANCREAS, AND PERIPHERAL BLOOD MONONUCLEAR CELL DATA SETS

表 1. SCQUERY、SUERAT 胰腺和外周血单核细胞数据集的基本统计信息

	scQuery			Seurat pancreas			Seurat PBMC		
	Data	Cell type	Domain	Data	Cell type	Domain	Data	Cell type	Domain
Training	37,697	39	99	6321	13	3	25,977	10	8
Validation	3023	19	26	—	—	—	—	—	—
Test	3770	23	30	638	13	1	2992	10	1

	scQuery			Seurat 胰腺			Seurat PBMC		
	数据	细胞类型	领域	数据	细胞类型	领域	数据	细胞类型	领域
训练	37,697	39	99	6321	13	3	25,977	10	8
验证	3023	19	26	—	—	—	—	—	—
测试	3770	23	30	638	13	1	2992	10	1

PBMC, peripheral blood mononuclear cell.

PBMC，外周血单核细胞。

Our comparisons include both visual projection of the learned alignment (Fig. 5 and 6) and quantitative analysis of the accuracy of the predicted test cell types (Table 2). For the latter, to enable comparisons of the supervised and unsupervised methods, we used the resulting aligned data from the unsupervised methods to train a NN that has the same configuration as Lin et al. (2017). For scVI, which results in a much lower dimensional representation, we used a smaller input vector and a smaller hidden layer. Note that these alignment methods actually use the scRNA-seq test data to determine the final dimensionality reduction function, whereas our method does not utilize the test data for any model decision or parameter learning. To effectively apply Seurat to scQuery, we remove the batches that have < 100 samples. Also, for those data sets that the assumption of overlapped cell types is not guaranteed such as scQuery, we find that the performance of MNNs highly depends on the order of alignment. Therefore, for MNNs on the scQuery data set, we use 10 random permutations of batch orders and report the average accuracy.

我们的比较包括学习到的对齐的可视化投影 (图 5 和 6) 以及对预测测试细胞类型准确性的定量分析 (表 2)。对于后者，为了能够比较监督和无监督方法，我们使用无监督方法得到的对齐数据来训练一个与 Lin 等人 (2017) 相同配置的神经网络。对于 scVI，它生成了一个维度更低的表示，我们使用了一个更小的输入向量和一个更小的隐藏层。请注意，这些对齐方法实际上使用 scRNA-seq 测试数据来确定最终的降维函数，而我们的方法并不利用测试数据进行任何模型决策或参数学习。为了有效地将 Seurat 应用于 scQuery，我们移除了具有 < 100 样本的批次。此外，对于那些不保证重叠细胞类型假设的数据集，例如 scQuery，我们发现 MNN 的性能高度依赖于对齐的顺序。因此，对于 scQuery 数据集上的 MNN，我们使用 10 个随机的批次顺序排列，并报告平均准确性。

3.2. Overall performance

3.2. 整体性能

As already mentioned, we use the validation set to select the best model when using the scQuery data set. For the smaller data sets, we use the model obtained after 250 epochs (all models converged after this number of epochs). Test accuracy for the different methods is presented in Table 2. We show both mean and standard deviation of the accuracy for 10 randomly initialized experiments. An example for the performance for all methods we tested on the two data sets is shown in Figure 3. As can be seen, on average scDGN outperforms the other methods in terms of test accuracy, although for a particular cell type, we sometimes see other methods perform better. Performance comparisons for all cell types are presented in Section C (Tables C1-C8) in Supplementary Material.

如前所述，我们使用验证集来选择在使用 scQuery 数据集时的最佳模型。对于较小的数据集，我们使用在 250 个周期后获得的模型 (所有模型在这个周期数后收敛)。不同方法的测试准确率见表 2。我们展示了 10 次随机初始化实验的准确率的均值和标准差。所有我们在两个数据集上测试的方法的性能示例见图 3。可以看出，平均而言，scDGN 在测试准确率方面优于其他方法，尽管对于特定的细胞类型，我们有时会看到其他方法表现更好。所有细胞类型的性能比较在补充材料的 C 节 (表 C1-C8) 中呈现。

In addition, Table 2 presents the mutual information (MI) between labels and domains that corresponds to the difficulty of the data set. A larger MI indicates that models that do not account for the domain are likely to fit the domain information rather than the cell type. For the scQuery data set, we find the accuracy is low for all methods, indicating that this data set is relatively difficult. This is corroborated by the large MI value. For such data, we see a clear advantage for the scDGN: scDGN improves by $> 10\%$ over all other methods ($p = 5.069 \times 10^{-5}$ based on Student' s t -test when compared with the NN baseline that is tied for second best). The improvements over other single-cell alignment methods are even more significant. scDGN also achieves the best performance on the second largest data set, the PBMC data set. However,

此外，表 2 展示了标签与领域之间的互信息 (MI)，这与数据集的难度相关。较大的 MI 表示不考虑领域的模型可能更倾向于拟合领域信息而非细胞类型。对于 scQuery 数据集，我们发现所有方法的准确率都较低，表明该数据集相对困难。这一点得到了较大 MI 值的证实。对于这样的数据，我们看到 scDGN 有明显优势: 与并列第二好的 NN 基线相比，scDGN 在所有其他方法 ($p = 5.069 \times 10^{-5}$ 的基础上提高了 $> 10\%$ ，基于学生 t -检验。与其他单细胞对齐方法的改进更为显著。scDGN 在第二大数据集 PBMC 数据集上也取得了最佳性能。然而，

Table 2. Overall Performances of Different Methods

表 2. 不同方法的整体性能

Experiments	MI	NN	CaSTLe	MNNs	*scVI*	Seurat	scDGN
scQuery	3.025	0.255	0.156	0.200	0.257	0.144	0.286
PBMC	0.112	0.861	0.865	0.859	0.808	0.830	0.868
Pancreas 1	0.902	0.720	0.705	0.591	0.855	0.812	0.856
Pancreas 2	0.733	0.891	0.764	0.764	0.852	0.825	0.918
Pancreas 3	0.931	0.545	0.722	0.722	0.651	0.751	0.663
Pancreas 4	0.458	0.927	0.914	0.914	0.925	0.881	0.925
Pancreas 5	0.849	0.928	0.882	0.932	0.895	0.865	0.923
Pancreas 6	0.670	0.944	0.917	0.946	0.893	0.907	0.950
Average	—	0.826	0.817	0.842	0.845	0.840	0.872

实验	MI	NN	CaSTLe	MNNs	*scVI*	Seurat	scDGN
scQuery	3.025	0.255	0.156	0.200	0.257	0.144	0.286
PBMC	0.112	0.861	0.865	0.859	0.808	0.830	0.868
胰腺 1	0.902	0.720	0.705	0.591	0.855	0.812	0.856
胰腺 2	0.733	0.891	0.764	0.764	0.852	0.825	0.918
胰腺 3	0.931	0.545	0.722	0.722	0.651	0.751	0.663
胰腺 4	0.458	0.927	0.914	0.914	0.925	0.881	0.925
胰腺 5	0.849	0.928	0.882	0.932	0.895	0.865	0.923
胰腺 6	0.670	0.944	0.917	0.946	0.893	0.907	0.950
平均值	—	0.826	0.817	0.842	0.845	0.840	0.872

MI represents the mutual information between batch and cell type in the corresponding data set. The highest test accuracy for each data set is bolded.

MI 表示对应数据集中批次与细胞类型之间的互信息。每个数据集的最高测试准确率以粗体显示。

MI, mutual information; MNNs, mutual nearest neighbors; NN, neural network; scDGN, single-cell domain generalization network.

MI，互信息；MNNs，互最近邻；NN，神经网络；scDGN，单细胞领域泛化网络。

# train	#test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
330	25	gamma	0.832	0.56	0.82	0.716	0.724	0.794
462	21	ductal	1	1	1	1	1	1
323	18	mast	0.95	0.83333	0.9556	0.9833	0.8833	0.8833
21	14	endothelial	0.4786	0.428571	0.4786	0.75	0.0286	0.5
6	3	beta	0	0	0.0333	0	0.3333	0.0333
15	1	quiescent_stellate	1	1	1	1	0	0.9
5	1	macrophage	0.6	1	0.6	0.7	0.1	0.4
1	1	activated_stellate	0.5	1	0.4	1	0	0
327	36	schwann	0.7806	0.805556	0.7639	0.8556	0.1667	0.866
3	5	epsilon	0.16	0	0.18	0.34	1	0
1199	239	alpha	0.8695	0.648536	0.8833	0.777	0.9063	0.9381
74	16	delta	0.975	0.625	0.9687	0.9	1	0.9812
469	258	acinar	0.9647	0.910853	0.9725	0.8	0.8725	0.9667
3235	638	Average	0.891	0.764	0.899	0.852	0.825	0.918

# 训练	# 测试	细胞类型	NN	CaSTLe	MNN	scVI	Seurat	scDGN
330	25	gamma	0.832	0.56	0.82	0.716	0.724	0.794
462	21	导管	1	1	1	1	1	1
323	18	乳腺	0.95	0.83333	0.9556	0.9833	0.8833	0.8833
21	14	内皮	0.4786	0.428571	0.4786	0.75	0.0286	0.5
6	3	beta	0	0	0.0333	0	0.3333	0.0333
15	1	静息星形胶质细胞	1	1	1	1	0	0.9
5	1	巨噬细胞	0.6	1	0.6	0.7	0.1	0.4
1	1	激活星形胶质细胞	0.5	1	0.4	1	0	0
327	36	施旺细胞	0.7806	0.805556	0.7639	0.8556	0.1667	0.866
3	5		0.16	0	0.18	0.34	1	0
1199	239		0.8695	0.648536	0.8833	0.777	0.9063	0.9381
74	16		0.975	0.625	0.9687	0.9	1	0.9812
469	258	腺泡	0.9647	0.910853	0.9725	0.8	0.8725	0.9667
3235	638	平均	0.891	0.764	0.899	0.852	0.825	0.918

FIG. 3. Test accuracy of each model on different cell types from Pancreas2 data set. The darker shades represent better performance.

图 3. Pancreas2 数据集上不同细胞类型的每个模型的测试准确率。较深的阴影表示更好的性能。

given the very low MI for this data set, the performance of the other methods, including the baseline NN, is almost as good as the performance of scDGN. The third data set we test on is the Securat pancreas data set. This is the smallest data set and so it has the least number of training samples. Still, of the six settings we tested (which differed in the subset of cells that were excluded from training), we find that scDGN is the top performer in four of them, comparable with the top performer for another one and in only one setting (Pancreas 3, with the highest MI) is significantly outperformed by Seurat. Note

that even for the Pancreas 3 data, the domain adversarial training helps: using this the scDGN is able to improve by $> 20\%$ over the baseline NN used for the label classifier.

鉴于该数据集的 MI 非常低，其他方法的性能，包括基线 NN，几乎与 scDGN 的性能相当。我们测试的第三个数据集是 Securat 胰腺数据集。这是最小的数据集，因此它的训练样本数量最少。尽管如此，在我们测试的六个设置中 (这些设置在排除的细胞子集上有所不同)，我们发现 scDGN 在其中四个设置中表现最佳，与另一个设置的最佳表现者相当，只有在一个设置 (胰腺 3，具有最高的 MI) 中被 Seurat 显著超越。请注意，即使对于胰腺 3 数据，领域对抗训练也有所帮助: 使用这一方法，scDGN 能够在用于标签分类器的基线 NN 上提高 $> 20\%$ 。

3.3. Visualization of the representation learned by alignment and classification methods

3.3. 对齐和分类方法学习的表示的可视化

To further explore the effectiveness of the batch removal provided by our proposed domain adversarial training with conditional domain generalization strategy, we visualize the 100-dimensional hidden representations learned by NN and scDGN: Figure 4 presents both principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) plots for several different cell types across the three data sets. Note that we only used the top two components for visualization, although the actual classification for all methods was performed using all raw feature values without any dimensionality reduction. Points are colored using their batch IDs to evaluate batch effects. As can be seen, using scDGN we obtain results that are much better at mixing cells from the different batches when compared with the baseline NN model. The impact is larger for the pancreas data sets that have larger MI compared with the PBMC data set, which helps explain the large increase in performance for these two data sets.

为了进一步探讨我们提出的具有条件领域泛化策略的领域对抗训练所提供的批次去除的有效性，我们可视化了 NN 和 scDGN 学习到的 100 维隐藏表示: 图 4 展示了在三个数据集上针对几种不同细胞类型的主成分分析 (PCA) 和 t-分布随机邻居嵌入 (t-SNE) 图。请注意，我们仅使用前两个成分进行可视化，尽管所有方法的实际分类是使用所有原始特征值而没有进行任何降维。点的颜色使用它们的批次 ID 来评估批次效应。可以看出，使用 scDGN 时，我们获得的结果在混合来自不同批次的细胞方面比基线 NN 模型要好得多。对于胰腺数据集，由于其 MI 较大，相较于 PBMC 数据集，影响更为显著，这有助于解释这两个数据集性能的大幅提升。

We next extended this comparison and visualized the learned (aligned) representations for all methods using data from both the Pancreas2 and scQuery data sets (Figs. 5 and 6). For the Pancreas2 data set, we visualize the entire data set. For scQuery, given the large number of cell types and domains, we present PCA visualization of a subset of cell types and domains. As can be seen, in addition to scDGN, Seurat is also able to successfully mix the data from different batches. However, as the results in Table 2 indicate, this may come at the expense of not correctly separating cell types. MNNs and scVI are not always effective at removing batch effects for the cell types. In contrast, scDGN is able to do both domain mixing and cell-type assignment, leading to its better performance overall. For example, for the acinar and alpha cell types in the pancreas data set (Fig. 5), only scDGN, MNNs, and Seurat are able to align the data from different domains. However, MNNs and Seurat overcorrect the representation by aligning different cell types from different domains, mixing acinar and gamma cells. Additional visualizations for other cell types and domains can be found in Section D in Supplementary Material, where the same advantages of scDGN over other methods can be consistently observed.

我们接下来扩展了这种比较，并使用来自 Pancreas2 和 scQuery 数据集的数据可视化了所有方法学习到的 (对齐的) 表示 (图 5 和 6)。对于 Pancreas2 数据集，我们可视化了整个数据集。对于 scQuery，由于细胞类型和领域的数量庞大，我们展示了细胞类型和领域子集的 PCA 可视化。如所示，除了 scDGN，Seurat 也能够成功混合来自不同批次的数据。然而，正如表 2 中的结果所示，这可能以未能正确分离细胞类型为代价。MNNs 和 scVI 并不总是有效地去除细胞类型的批次效应。相比之下，scDGN 能够同时进行领域混合和细胞类型分配，从而整体上表现更好。例如，对于胰腺数据集中的腺泡细胞和 细胞类型 (图 5)，只有 scDGN、MNNs 和 Seurat 能够对齐来自不同领域的数据。然而，MNNs 和 Seurat 通过对齐来自不同领域的不同细胞类型而过度校正了表示，混合了腺泡细胞和 细胞。其他细胞类型和领域的附加可视化可以在补充材料的 D 节中找到，在那里可以一致观察到 scDGN 相较于其他方法的相同优势。

3.4. Analysis of key genes

3.4. 关键基因分析

Although NNs are often treated as black boxes, recent methods provide useful directions for making them more interpretable (Ribeiro et al., 2016). Here we use activation maximization, which relies on the

尽管神经网络通常被视为黑箱，但最近的方法为使其更具可解释性提供了有用的方向 (Ribeiro et al., 2016)。在这里，我们使用激活最大化，这依赖于

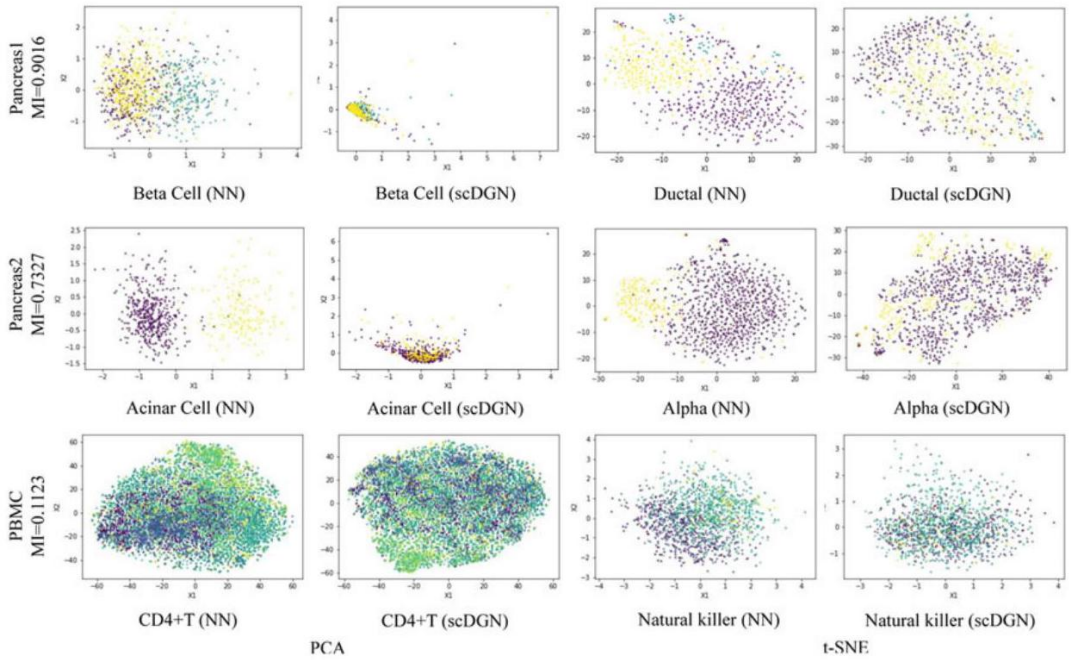

FIG. 4. Visualization of learned representations for NN and scDGN: using PCA and t-SNE rows: the three data sets we tested the method on. Columns: Methods and cell types. For each row, data from different batches are distinguished using different colors.

图 4. NN 和 scDGN 学习表示的可视化: 使用 PCA 和 t-SNE 行: 我们测试该方法的三个数据集。列: 方法和细胞类型。对于每一行，来自不同批次的数据用不同颜色区分。

gradient of the correct category logit with respect to the input vector to select the key inputs for each of the models (Erhan et al., 2009; Simonyan et al., 2013; Springenberg et al., 2014). Formally, given a particular cell type i and a trained NN ϕ , activation maximization looks for important input genes x' by solving the following optimization problem:

关于输入向量的正确类别对数几率的梯度，以选择每个模型的关键输入 (Erhan et al., 2009; Simonyan et al., 2013; Springenberg et al., 2014)。形式上，给定特定的细胞类型 i 和训练好的神经网络 ϕ ，激活最大化通过解决以下优化问题来寻找重要的输入基因 x' :

$$x' = \max_x \left(\phi\left(x\right) \cdot e_i\right)$$

where e_i is the natural basis vector associated with the i -th category. This can be solved through back-propagation, where the gradient of $\phi(x)$ with respect to x , which can be viewed as the weight of the first-order Taylor expansion of the NN, is calculated to iteratively update the input. We follow a previous method (Simonyan et al., 2013) and initialize the optimization with a zero vector. Given this setting, we ran the optimization for 100 iterations with learning rate set to 1 . The important genes are selected as those inputs leading to the largest changes compared with the initialization values. To compare scDGN and NN for certain cell types, we select the top k genes with the largest changes and perform gene ontology (GO) analysis on these selected genes.

其中 e_i 是与 i 类别相关的自然基向量。这可以通过反向传播来解决，其中 $\phi(x)$ 相对于 x 的梯度可以视为神经网络的一阶泰勒展开的权重，通过计算该梯度来迭代更新输入。我们遵循之前的方法 (Simonyan et al., 2013)，并用零向量初始化优化。在这种设置下，我们进行了 100 次迭代的优化，学习率设置为 1。

重要基因被选为与初始化值相比导致最大变化的输入。为了比较某些细胞类型的 scDGN 和神经网络，我们选择变化最大的前 k 个基因，并对这些选定基因进行基因本体 (GO) 分析。

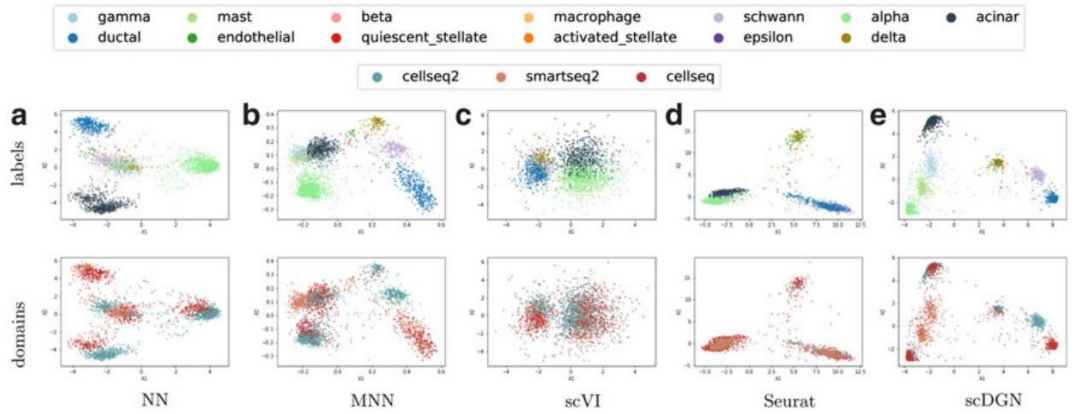

FIG. 5. PCA visualizations of the representations learned by different models on the full Pancreas2 data set. Colors for different cell types and domains are shown in the legend at the top.

图 5. 不同模型在完整的 Pancreas2 数据集上学习到的表示的 PCA 可视化。不同细胞类型和领域的颜色在顶部的图例中显示。

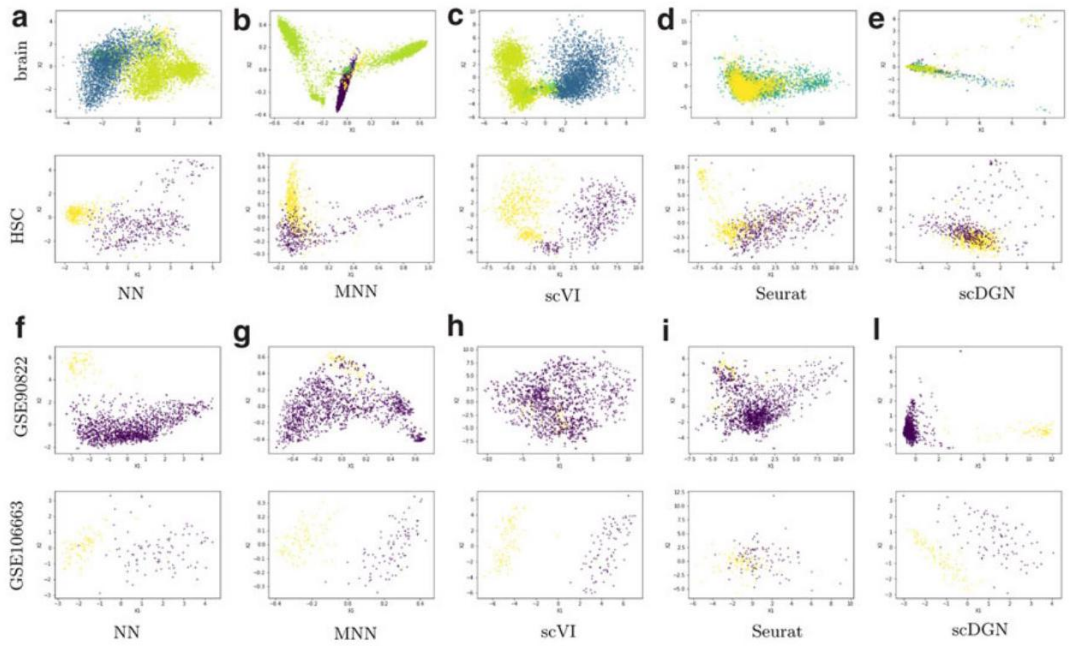

FIG. 6. PCA visualizations of the representations of certain cell types and batches by different models for the scQuery data set. Top two rows: Cell types. Colors represent different batches. Bottom two rows: Batches. Colors represent different cell types. HSC, hematopoietic stem cell.

图 6. 不同模型在 scQuery 数据集上对某些细胞类型和批次的表示的 PCA 可视化。前两行: 细胞类型。颜色代表不同的批次。后两行: 批次。颜色代表不同的细胞类型。HSC，造血干细胞。

As an example, consider the genes identified for the liver cell type using the scQuery data set. We select the top 100 genes for this cell type from NN and scDGN and present the enriched GO categories on biological process with adjusted p-value $< 1.0 \times 10^{-4}$ in Tables 3 and 4 . We also list these genes by order in Section A. 3 in Supplementary Material. As can be seen, although a number of significant GO categories are identified for the top 100NN genes, these are generic and not liver specific. They include general terms related to interactions between organs and immune response categories that are active in multiple organs and cell types. In sharp contrast, the categories identified for scDGN are much more specific and highlight key pathways that are mainly utilized in the liver.

作为一个例子，考虑使用 scQuery 数据集识别的肝细胞类型的基因。我们从 NN 和 scDGN 中选择该细胞类型的前 100 个基因，并在表 3 和表 4 中展示调整后的 p 值 $< 1.0 \times 10^{-4}$ 的生物过程富集 GO 类别。我们还在补充材料的 A.3 节中按顺序列出了这些基因。可以看出，尽管为前 100NN 个基因识别出了一些显著的 GO 类别，但这些类别是通用的，并非特定于肝脏。它们包括与器官之间相互作用和在多个器官及细胞类型中活跃的免疫反应类别相关的一般术语。相比之下，为 scDGN 识别的类别则要具体得多，突出了主要在肝脏中利用的关键通路。

For example, the top category for the scDGN genes, "chylomicron remodeling," refers to the main physiological purpose of chlyomicron remnants: to facilitate the return of bile lipoproteins and cholesterol to the liver (Redgrave, 2004). Specifically, in this pathway, chylomicrons (lipoproteins) are broken down (remodeled through hydrolysis) and converted to a form called "chlyomicron remnant" that is taken up by specific receptors that exist primarily on the surface of liver cells (Hara et al., 1997). The second term, "pos. regulation of cholesterol esterification," refers to cholesterol esterification, a critical step in reverse cholesterol transport, the process in which excess cholesterol is sent to the liver to be removed from the body (Murakami et al., 1995; Komoda, 2010). Furthermore, cholesteryl ester transfer protein (CETP) is a key enzyme involved in this process and is highly expressed in liver cells, and variants of CETP are associated with increased risk of atherosclerosis (Komoda, 2010; Seidman et al., 2014). The fifth most significant term, "lipoprotein remodeling," is part of the two aforementioned processes. The top 100 genes identified by the scDGN include apoal (main protein component of high-density lipoprotein cholesterol), apoa2, and apoc1, all of which encode lipoproteins that are primarily expressed in the liver (Ko et al., 2014; Domingo-Espın et al., 2018). These genes were not included in the top 100 genes by the NN. We present the results of GO analysis comparison for several additional cell types in Section E. 2 in Supplementary Material.

例如，scDGN 基因的顶级类别 “乳糜微粒重塑” 指的是乳糜微粒残余物的主要生理目的: 促进胆汁脂蛋白和胆固醇返回肝脏 (Redgrave, 2004)。具体而言，在这一途径中，乳糜微粒 (脂蛋白) 被分解 (通过水解重塑) 并转化为一种称为 “乳糜微粒残余物” 的形式，该形式被主要存在于肝细胞表面的特定受体所摄取 (Hara et al., 1997)。第二个术语 “胆固醇酯化的正调节” 指的是胆固醇酯化，这是反向胆固醇转运中的一个关键步骤，该过程将多余的胆固醇送往肝脏以从体内清除 (Murakami et al., 1995； Komoda, 2010)。此外，胆固醇酯转移蛋白 (CETP) 是参与这一过程的关键酶，并在肝细胞中高度表达，CETP 的变异与动脉粥样硬化的风险增加相关 (Komoda, 2010； Seidman et al., 2014)。第五个最重要的术语 “脂蛋白重塑” 是上述两个过程的一部分。scDGN 识别的前 100 个基因包括 apoal(高密度脂蛋白胆固醇的主要蛋白成分)、apoa2 和 apoc1，所有这些基因编码的脂蛋白主要在肝脏中表达 (Ko et al., 2014； Domingo-Espın et al., 2018)。这些基因未被 NN 的前 100 个基因所包含。我们在补充材料的 E.2 节中展示了对几种额外细胞类型的 GO 分析比较结果。

Table 3. GO Analysis Results for Top 100 scQuery Liver Genes in the Neural Network Metwork Method

表 3. 神经网络方法中前 100 个 scQuery 肝脏基因的 GO 分析结果

Term_name	Term_id	p adj	- lo g_{10}p$_{adj}$
Symbiotic process	GO:0044403	1.16E-08	7.935246875
Interspecies interaction between organisms	GO:0044419	3.14E-08	7.503093471
Viral process	GO:0016032	3.69E-08	7.433145019
Immune response	GO:0006955	2.5491E-06	5.593613105
Multiorganism process	GO:0051704	1.40837E-05	4.851282542
Immune effector process	GO:0002252	4.53533E-05	4.34339136
Response to stress	GO:0006950	5.56335E-05	4.254663785
Defense response	GO:0006952	6.18759E-05	4.208478308

术语名称	术语 ID	p adj	- lo g_{10}p$_{adj}$
共生过程	GO:0044403	1.16E-08	7.935246875
生物体之间的物种间相互作用	GO:0044419	3.14E-08	7.503093471
病毒过程	GO:0016032	3.69E-08	7.433145019
免疫反应	GO:0006955	2.5491E-06	5.593613105
多生物过程	GO:0051704	1.40837E-05	4.851282542
免疫效应过程	GO:0002252	4.53533E-05	4.34339136
应激反应	GO:0006950	5.56335E-05	4.254663785
防御反应	GO:0006952	6.18759E-05	4.208478308

Table 4. GO Analysis Results for Top 100 scQuery Liver Genes IN THE SINGLE-CELL DOMAIN GENERALIZATION NETWORK METHOD

表 4. 单细胞领域泛化网络方法中前 100 个 scQuery 肝脏基因的 GO 分析结果

Term_name	Term_id	p adj	- lo g_{10}p adj
Chylomicron remodeling	GO:0034371	3.04042E-05	4.517066786
Positive reg. of cholesterol esterification	GO:0010873	3.04042E-05	4.517066786
Negative reg. of cellular component organization	GO:0051129	3.94437E-05	4.404022507
Protein-lipid complex remodeling	GO:0034368	7.34551E-05	4.133978335
Plasma lipoprotein particle remodeling	GO:0034369	7.34551E-05	4.133978335
Protein-containing complex remodeling	GO:0034367	8.8522E-05	4.052948555

术语名称	术语 ID	p adj	- lo g_{10}p adj
乳糜微粒重塑	GO:0034371	3.04042E-05	4.517066786
胆固醇酯化的正调节	GO:0010873	3.04042E-05	4.517066786
细胞组分组织的负调节	GO:0051129	3.94437E-05	4.404022507
蛋白质-脂质复合物重塑	GO:0034368	7.34551E-05	4.133978335
血浆脂蛋白颗粒重塑	GO:0034369	7.34551E-05	4.133978335
含蛋白质复合物重塑	GO:0034367	8.8522E-05	4.052948555

reg., regularization.
reg.，正则化。

4. DISCUSSION

4. 讨论

Single-cell computational methods that do not account for batch effects are likely to fit the noise introduced by the batches. Several recent methods have been proposed for aligning scRNA-seq from multiple studies of the same tissues or processes. Most of these methods are unsupervised and assume that the cell types among different batches overlap. However, we show that these methods would fail on the studies in which cell types do not fully overlap, which is often the case when dealing with multiple data sets. To overcome this problem, we extend a supervised scRNA-seq cell-type assignment method based on NN and regularize its prediction to be invariant to batch effects.

不考虑批次效应的单细胞计算方法可能会拟合批次引入的噪声。最近提出了几种方法，用于对齐来自同一组织或过程的多个研究的 scRNA-seq 数据。这些方法大多是无监督的，并假设不同批次之间的细胞类型重叠。然而，我们表明，当细胞类型不完全重叠时，这些方法会失败，这在处理多个数据集时往往是常见的。为了解决这个问题，我们扩展了一种基于神经网络的监督 scRNA-seq 细胞类型分配方法，并对其预测进行了正则化，使其对批次效应不变。

Our method is based on the ideas of domain adversarial training. In such training, two competing tasks are used to optimize the representation of scRNA-seq data. The first focuses on the traditional goal of cell-type identification, whereas the second attempts to construct representations that are not affected by specific batch or experimental artifacts. This is accomplished by jointly minimizing a loss function that takes into account both goals, accounting for the weight of each of the goals using a GRL. We also proposed a conditional strategy to avoid overcorrection. We presented efficient learning methods for this setting and tested it on three large scale scRNA-seq data sets containing experiments from several different platforms for partially overlapping cell types.

我们的方法基于领域对抗训练的思想。在这种训练中，使用两个竞争任务来优化 scRNA-seq 数据的表示。第一个任务专注于细胞类型识别的传统目标，而第二个任务则试图构建不受特定批次或实验伪影影响的表示。这是通过共同最小化一个考虑到两个目标的损失函数来实现的，使用 GRL 考虑每个目标的权重。我们还提出了一种条件策略以避免过度校正。我们为这种设置提出了有效的学习方法，并在包含来自多个不同平台的实验的三个大规模 scRNA-seq 数据集上进行了测试，以处理部分重叠的细胞类型。

As we show, our scDGN method is able to correctly identify cell types in the test data sets. For the largest data set we tested on which contained close to 40 different cell types, scDGN significantly outperformed all prior methods. It also ranked first for the second largest data set, and for all but one of the six tests on the third data set. Importantly, it always outperformed the supervised learning-based method, indicating that batch effects should be addressed when designing such methods for cell-type assignments. In addition to accurately assigning cell types, further analysis of significant genes indicates that by overcoming batch effects, scDGN is better able to focus on relevant sets of genes when compared with prior supervised methods, explaining its improvement in accuracy.

我们展示了，我们的 scDGN 方法能够正确识别测试数据集中的细胞类型。在我们测试的最大数据集中，包含近 40 种不同的细胞类型，scDGN 显著优于所有先前的方法。它在第二大数据集中排名第一，并

且在第三数据集的六个测试中除了一个外均表现最佳。重要的是，它始终优于基于监督学习的方法，这表明在设计细胞类型分配的方法时，应解决批次效应。除了准确分配细胞类型外，对显著基因的进一步分析表明，通过克服批次效应，scDGN 能够更好地关注相关基因集，相较于先前的监督方法，这解释了其准确性的提升。

Although scDGN performed best on the data we analyzed, there are a number of possible issues with this approach. First, it learns a large number of parameters that require large input data sets. However, as we showed, scDGN is able to perform well even for data sets with a few thousand cells that match current sizes of scRNA-seq data sets. Second, scDGN is based on NNs that are often seen as a black box, making it hard to interpret the resulting model and its biological relevance. Recent work provides a number of directions that can be used to overcome this issue. As we showed, using activation maximization, we were able to identify several relevant cell-type specific genes in the learned network. Future work would include using

尽管 scDGN 在我们分析的数据上表现最佳，但这种方法仍存在一些可能的问题。首先，它学习了大量的参数，这需要大量的输入数据集。然而，正如我们所展示的，scDGN 即使在与当前 scRNA-seq 数据集大小相匹配的几千个细胞的数据集上也能够表现良好。其次，scDGN 基于神经网络，通常被视为黑箱，这使得解释所得到的模型及其生物学相关性变得困难。最近的研究提供了多种方向，可以用来克服这个问题。正如我们所展示的，通过使用激活最大化，我们能够在学习的网络中识别出几个相关的细胞类型特异性基因。未来的工作将包括使用

additional NN interpretation methods, including Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) or Remove and Retrain (ROAR) and Keep and Retrain (KAR) (Hooker et al., 2018), to further identify the set of genes that play the largest role in the decisions the network makes. Third, as shown in Section D. 3 in Supplementary Material, scDGN sometimes does not mix up the representations from different batches for all cell types. Considering the visualization results for NN in Section D. 8 in Supplementary Material and its competitive performance in Table 2 together, it may indicate that it is not always necessary to remove batch effects for the model to achieve high test accuracy. Therefore, it is worthwhile to further study when the alignment is imperative.

额外的神经网络解释方法，包括局部可解释模型无关解释 (LIME)(Ribeiro et al., 2016) 或移除并重新训练 (ROAR) 和保留并重新训练 (KAR)(Hooker et al., 2018)，以进一步识别在网络决策中发挥最大作用的基因集合。第三，如补充材料中的 D.3 节所示，scDGN 有时不会混合来自不同批次的所有细胞类型的表示。考虑到补充材料中 D.8 节的神经网络可视化结果及其在表 2 中的竞争表现，这可能表明并不总是需要去除批次效应，以使模型达到高测试准确率。因此，进一步研究何时对齐是必要的，具有重要意义。

Finally, unlike prior scRNA-seq alignment methods, scDGN is supervised. Although this is an advantage when it comes to accuracy, as we have shown, it may be a problem for the new data. We believe that as more scRNA-seq and other high-throughput single-cell data accumulate, we would have labeled data for most cell types, which would enable training an scDGN for even more cell types. As we have shown with the scQuery data set, for which scDGN significantly outperformed all other methods, when such data exist, scDGN is able to correctly align experiments and platforms not seen in the training set. More generally, this article presents a method that connects the batch effect removal problem to domain adaptation tasks in machine learning. Recent developments in this direction in the machine learning community may lead to even better results for batch removal problems. For instance, it has been recently shown that self-supervision with domain knowledge, for instance rotation prediction (Sun et al., 2020), can greatly improve the learned features and generalization to unseen data. It would be interesting to consider whether similar biological information, for example, knowledge about gene interactions, can be used to further improve the solution for alignment problems.

最后，与之前的 scRNA-seq 对齐方法不同，scDGN 是有监督的。虽然这在准确性方面是一个优势，但正如我们所示，这可能对新数据造成问题。我们相信，随着更多的 scRNA-seq 和其他高通量单细胞数据的积累，我们将拥有大多数细胞类型的标记数据，这将使得训练一个针对更多细胞类型的 scDGN 成为可能。正如我们在 scQuery 数据集上所展示的，scDGN 在该数据集上显著优于所有其他方法，当这样的数据存在时，scDGN 能够正确对齐训练集中未见过的实验和平台。更一般而言，本文提出了一种将批次效应去除问题与机器学习中的领域适应任务相连接的方法。机器学习社区在这一方向上的最新进展可能会为批次去除问题带来更好的结果。例如，最近的研究表明，结合领域知识的自我监督，例如旋转预测 (Sun et al., 2020)，可以大大改善学习到的特征和对未见数据的泛化。考虑是否可以利用类似的生物信息，例如关于基因相互作用的知识，以进一步改善对齐问题的解决方案，将是非常有趣的。

scDGN is implemented in Python with the PyTorch API (Steiner et al., 2019), and users can obtain the code and sampled data from (https://github.com/SongweiGe/scDGN).

scDGN 是用 Python 和 PyTorch API 实现的 (Steiner et al., 2019)，用户可以从 (https://github.com/SongweiGe/scDGN 获取代码和样本数据。

AUTHOR DISCLOSURE STATEMENT

作者声明

The authors declare they have no conflicting financial interests.

作者声明他们没有冲突的财务利益。

FUNDING INFORMATION

资助信息

This study was partially supported by National Institute of Health grants 1R01GM122096 and OT2OD026682 to Z.B.J. and by a Scholars Award in Studying Complex Systems from the James S. McDonnell Foundation to Z.B.J. H.W. was supported by the National Institutes of Health grants R01- GM093156 and P30-DA035778.

本研究部分得到了国家卫生研究院的资助，拨款编号为 1R01GM122096 和 OT2OD026682，资助人 Z.B.J.，以及詹姆斯 · S · 麦克唐纳基金会的复杂系统研究学者奖，资助人 Z.B.J. H.W. 得到了国家卫生研究院的资助，拨款编号为 R01-GM093156 和 P30-DA035778。

SUPPLEMENTARY MATERIAL

附加材料

Supplementary Material

附加材料

REFERENCES

参考文献

Alavi, A., Ruffalo, M., Parvangada, A., et al. 2018. A web server for comparative analysis of single-cell RNA-seq data. Nat. Commun. 9, 4768.

Chu, C., and Wang, R. 2018. A survey of domain adaptation for neural machine translation. arXiv 1806.00258.

Csurka, G. 2017. Domain adaptation for visual applications: A comprehensive survey. arXiv 1702.05374.

Ding, J., Adiconis, X., Simmons, S.K., et al. 2019. Systematic comparative analysis of single cell RNA-seq methods. BioRxiv 632216.

Domingo-Espın, J., Nilsson, O., Bernfur, K., et al. 2018. Site-specific glycations of apolipoprotein A-I lead to differentiated functional effects on lipid-binding and on glucose metabolism. Biochim. Biophys. Acta -Mol. Basis Dis. 1864(9, Part B), 2822-2834.

Eng, C.-H.L., Lawson, M., Zhu, Q., et al. 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature 568, 235-239.

Erhan, D., Bengio, Y., Courville, A., et al. 2009. Visualizing higher-layer features of a deep network. Techreport 1341. University of Montreal.

Ganin, Y., Ustinova, E., Ajakan, H., et al. 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 2096-2030.

Hadsell, R., Chopra, S., and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping, 2, 1735- 1742. Presented at the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, New York, NY, USA.

Haghverdi, L., Lun, A.T., Morgan, M.D., et al. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36, 421-427.

Hara, T., Tan, Y., and Huang, L. 1997. In vivo gene delivery to the liver using reconstituted chylomicron remnants as a novel nonviralvector. Proc. Natl Acad. Sci. U. S. A. 94, 14547-14552.

Hooker, S., Erhan, D., Kindermans, P.-J., et al. 2018. Evaluating feature importance estimates. arXiv 1806.10758.

HuBMAP Consortium. 2019. The human body at cellular resolution: The NIH human biomolecular atlas program. Nature 574, 187-192.

Hwang, B., Lee, J.H., and Bang, D. 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. 50, 1-14.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., et al. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776-779.

Kiselev, V.Y., Yiu, A., and Hemberg, M. 2018. scmap: Projection of single-cell RNA-seq data across data sets. Nat. Methods 15, 359-362.

Ko, H.-L., Wang, Y.-S., Fong, W.-L., et al. 2014. Apolipoprotein C1 (APOC1) as a novel diagnostic and prognostic biomarker for lung cancer: A marker phase I trial. Thorac. Cancer 5, 500-508.

Koch, G., Zemel, R., and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop 2, Lille, France.

Komoda, T., ed. 2010. Chapter 3, 35-59. In The HDL Handbook: Biological Functions and Clinical Implications. Academic Press, Boston.

Li, H., Pan, S.J., Wang, S., et al. 2018. Domain generalization with adversarial feature learning. In CVPR, Salt Lake City, Utah, USA.

Lieberman, Y., Rokach, L., and Shay, T. 2018. Castle-classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. PLoS One 13, e0205499.

Lin, C., Jain, S., Kim, H., et al. 2017. Using neural networks for reducing the dimensions of single-cell RNA-seq data. Nucleic Acids Res. 45:e156.

Lopez, R., Regier, J., Cole, M.B., et al. 2018. Deep generative modeling for single-cell transcriptomics. Nat. Methods 15, 1053-1058.

Motiian, S., Piccirilli, M., Adjeroh, D.A., et al. 2017. Unified deep supervised domain adaptation and generalization, vol. 2, 3. In ICCV, Venice, Italy.

Murakami, T., Michelagnoli, S., Longhi, R., et al. 1995. Triglycerides are major determinants of cholesterol esterification/ transfer and HDL remodeling in human plasma. Arterioscler. Thromb. Vasc. Biol. 15, 1819-1828.

Nowotschin, S., Setty, M., Kuo, Y.Y., et al. 2019. The emergent landscape of the mouse gut endoderm at single-cell resolution. Nature 569, 361-367.

Papalexi, E., and Satija, R. 2018. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. 18, 35-45.

Patel, V.M., Gopalan, R., Li, R., et al. 2015. Visual domain adaptation: A survey of recent advances. IEEE Signal Process. Mag. 32, 53-69.

Pei, Z., Cao, Z., Long, M., et al. 2018. Multi-adversarial domain adaptation. Presented at AAAI Conference on Artificial Intelligence, New Orleans, Couisiana, USA.

Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., et al. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature 566, 490-495.

Redgrave, T. 2004. Chylomicron metabolism. Biochem. Soc. Trans. 32, 79-82.

Regev, A., Teichmann, S.A., Lander, E.S., et al. 2017. Science forum: the human cell atlas. Elife 6, e27041.

Ribeiro, M.T., Singh, S., and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier, 1135-1144. In SIGKDD. ACM, San Francisco, California, USA.

Rozenblatt-Rosen, O., Stubbington, M.J., Regev, A., et al. 2017. The human cell atlas: From vision to reality. Nature 550, 451-453.

Seidman, M.A., Mitchell, R.N., and Stone, J.R. 2014. Chapter 12-Pathophysiology of atherosclerosis, 221-237. In Willis M.S., Homeister J.W., and Stone J.R., eds. Cellular and Molecular Pathobiology of Cardiovascular Disease. Academic Press, SanDiego.

Simonyan, K., Vedaldi, A., and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv 1312.6034.

Springenberg, J.T., Dosovitskiy, A., Brox, T., et al. 2014. Striving for simplicity: The all convolutional net. arXiv 1412.6806.

Steiner, B., DeVito, Z., Chintala, S., et al. 2019. Pytorch: An imperative style, high-performance deep learning library. NeurIPS 32, 1-4.

Stuart, T., Butler, A., Hoffman, P., et al. 2019. Comprehensive integration of single-cell data. Cell 177:1888.e21- 1902.e21.

Stuart, T., and Satija, R. 2019. Integrative single-cell analysis. Nat. Rev. Genet. 20:257-272.

Sun, Y., Wang, X., Liu, Z., et al. 2020. Test-time training with self-supervision for generalization under distribution shifts. Presented at International Conference on Machine Learning (ICML) online.

Tung, P.-Y., Blischak, J.D., Hsiao, C.J., et al. 2017. Batch effects and the effective design of single-cell gene expression studies. Sci. Rep. 7, 39921.

Villani, A.-C., Satija, R., Reynolds, G., et al. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, eaah4573.

Wang, G., Moffitt, J.R., and Zhuang, X. 2018. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. Sci. Rep. 8, 4847.

Wang, H., Ge, S., Xing, E.P., et al. 2019a. Learning robust global representations by penalizing local predictive power. arXiv 1905.13549.

Wang, H., He, Z., Lipton, Z.C., et al. 2019b. Learning robust representations by projecting superficial statistics out. arXiv 1903.06256.

Yu, Y., Tsang, J.C., Wang, C., et al. 2016. Single-cell RNA-seq identifies a PD-1 hi ILC progenitor and defines its development pathway. Nature 539, 102-106.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., et al. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138-1142.

Address correspondence to:
通信地址:
Mr. Songwei Ge
格松伟先生
Computational Biology Department
计算生物学系
Carnegie Mellon University
卡内基梅隆大学
5000 Forbes Avenue
福布斯大道 5000 号
Pittsburgh, PA 15213
宾夕法尼亚州匹兹堡 15213
USA
E-mail: songweig@cs.cmu.edu
电子邮件:songweig@cs.cmu.edu
Dr. Ziv Bar-Joseph
巴尔-约瑟夫博士
Computational Biology Department
计算生物学系
Carnegie Mellon University
卡内基梅隆大学
5000 Forbes Avenue
5000 福布斯大道
Pittsburgh, PA 15213
宾夕法尼亚州匹兹堡 15213
USA
E-mail: zivbj@cs.cmu.edu
电子邮件: zivbj@cs.cmu.edu