Distribution-Independent Cell Type Identification for Single-Cell RNA-seq Data

Yuyao Zhai[1] , **Liang Chen**[4] and **Minghua Deng**[1,2,3]
[1]School of Mathematical Sciences, Peking University
[2]Center for Statistical Science, Peking University
[3]Center for Quantitative Biology, Peking University
[4]Huawei Technologies Co., Ltd.
zhaiyuyao@stu.pku.edu.cn, chenliang260@huawei.com, dengmh@pku.edu.cn

Abstract

Automatic cell type annotation aims to transfer the label knowledge from label-abundant reference data to label-scarce target data, which makes encouraging progress in single-cell RNA-seq data analysis. While previous works have focused on classifying close-set cells and detecting open-set cells during testing, it is still essential to be able to classify unknown cell types as human beings. Additionally, few efforts have been devoted to addressing the challenge of common long-tail dilemma in cell type annotation data. Therefore, in this paper, we propose an innovative distribution-independent universal cell type identification framework called scDET from the perspective of autonomously equilibrated dual-consultative contrastive learning. Our model can generate fine-grained predictions for both close-set and open-set cell types in a long-tailed open-world environment. scDET consists of a contrastive-learning branch and a pseudo-labeling branch, which work collaboratively to provide interactive supervision. Specifically, the contrastive-learning branch provides reliable distribution estimation to regularize the predictions of the pseudo-labeling branch, which in turn guides itself through self-balanced knowledge transfer and a designed novel soft contrastive loss. Extensive experimental results on various evaluation datasets demonstrate the superior performance of scDET over other state-of-the-art single-cell clustering and annotation methods.

1 Introduction

Since being recognized as the yearly technology by Nature Methods in 2013, single-cell RNA sequencing (scRNA-seq) technology has seen substantial and swift development. The scale of the sequencing data has expanded, encompassing from a small group of dozens to hundreds of thousands, even millions [Regev *et al.*, 2017]. Using gene expression profiling, scRNA-seq enables researchers to inspect the individual cellular-level variability of disease tumors. The crucial process of identifying cell types in the analysis of scRNA-seq data aids in comprehending the source of tissue heterogeneity. The standard cell type annotation method initially clusters the cell population before identifying cluster-specific marker genes. The cells are then classified according to the ontological functions of their genes. However, as the scale of sequencing data rapidly expands, identifying marker genes to annotate cells turns into an increasingly daunting and time-intensive task [Kiselev *et al.*, 2019].

Given the abundance of extensively annotated scRNA-seq datasets, researchers have started to leverage classification and retrieval machine learning approaches to automate the process of transferring cell type labels from reference data to target data [Cao *et al.*, 2020]. During the early stages, the majority of automated annotation methods were implemented within closed domains, meaning that every cell type present in the target data also existed in the reference data. However, in realistic scenarios, this limitation can be excessively stringent. Thus, researchers eventually introduced automated annotation tasks applicable to open domains [Xu *et al.*, 2021]. To simplify terminologies, we shall denote cell types that are shared with both the reference data and the target data as seen cell types. Conversely, cell types absent in the reference data but present in the target data will be referred to as novel cell types. For tasks operating in open domains, it's a common practice among most annotation methods to classify these novel cell types under a category named "unassigned".

Given the demand for more detailed cellular and gene-level scrutiny within the "unassigned" group, the recent proposition of scGAD highlights the need to integrate cell classification and cell clustering within a unified framework [Zhai *et al.*, 2023]. However, certain limitations reside in scGAD's algorithm. Firstly, its mutual nearest neighbor retrieval process, an essential component of model training, is noted for its time-intensiveness which may significantly hinder the algorithm's computational efficiency. Secondly, scGAD overlooks a crucial characteristic of scRNA-seq data: the imbalance of cell types [Lähnemann *et al.*, 2020]. The differing frequencies of diverse cell types follow a long-tailed distribution where a minor fraction of classes heavily influence the overall data distribution, and many classes only link with a minuscule amount of instances. If we avoid formulating an algorithm specifically tailored to the peculiarities of these long-tailed distribution datasets, any derived solution would unquestionably be subpar.

To address the challenges discussed earlier, this study introduces a cutting-edge, distribution-independent framework for universal cell type identification termed scDET, premised on the concept of autonomously equilibrated dual-consultative contrastive supervision. The scDET framework encompasses two synergistic components: a contrastive learning branch and a pseudo-labeling branch. These branches work in tandem to offer mutual guidance, particularly for the imbalanced cell type annotation challenge within open-world settings. More precisely, the contrastive learning branch undertakes the role of distributional estimation throughout the model's training phase. This estimation acts as a means of regularizing the output of a linear classifier, thereby optimizing the generation of pseudo-labels. Conversely, the pseudo-labels thus obtained are selectively resampled and adjusted to counterbalance and augment the supervision provided to the contrastive learning branch. To amalgamate the insights from both branches and cultivate an improved representational space, we have engineered an innovative contrastive loss function predicated on pseudo-labels. This function is designed to coalesce samples in the feature space by leveraging their respective positiveness scores, thereby enhancing the clustering process.

We highlight the main contributions as follows:

- We integrate the paradigms of long-tail and open-world learning within the context of scRNA-seq data annotation, pinpointing distinct challenges in this domain that remain insufficiently tackled by current methodologies.
- We design a novel distribution-independent cell type annotation paradigm that comprises a contrastive learning branch and pseudo-labeling branch, which work collaboratively to tackle the data imbalance and classify novel cell types simultaneously.
- We carry out comprehensive experimental evaluations across diverse datasets and engage in detailed comparative analyses with leading-edge benchmarks, thereby affirming the effectiveness of scDET.

2 Related Work

2.1 Cell Type Identification for scRNA-Seq Data

Unsupervised clustering and supervised classification represent the two primary methodological approaches for determining cell types within scRNA-seq data [Petegrosso *et al.*, 2020]. With the accelerated advancement of deep learning technologies, a plethora of statistical learning approaches have been developed in both directions. In the realm of clustering, scziDesk identifies distinct cell populations by employing a soft self-training k-means algorithm within a low-dimensional feature space [Chen *et al.*, 2020a]. scNAME enhances clustering accuracy by integrating a mask estimation strategy alongside a neighborhood contrastive learning framework [Wan *et al.*, 2022]. As a semi-supervised clustering technique, scCNC embeds expert knowledge into the process using a capsule network framework [Wang *et al.*, 2022]. On the classification front, scNym leverages a combination of semi-supervised and adversarial learning approaches to effectively incorporate gene expression information [Kimmel and Kelley, 2021]. scArches adopt transfer learning accompanied by fine-tuned parameter optimization to contextualize target datasets appropriately [Lotfollahi *et al.*, 2022]. MARS utilizes meta-learning principles to reinforce the feature similarity amongst identical cell types [Brbić *et al.*, 2020]. To streamline and integrate the processes of cell clustering and classification, scGAD has been introduced [Zhai *et al.*, 2023]. This self-supervised learning architecture establishes a coherent linkage between the reference and target data, thus creating a unified framework for cell type identification.

2.2 Long-Tail Learning and Contrastive Learning

Long-tail learning is a typical machine learning scenario where the majority of samples belong to a few classes, while the minority classes have only a small number of samples [Zhang *et al.*, 2023]. Existing research on long-tail learning mainly focused on fully/semi/self-supervised scenarios, with representative methods that highlight the minority samples via re-sampling [Shen *et al.*, 2016], re-weighting [Cao *et al.*, 2019], or knowledge transferring [Wang *et al.*, 2017]. However, the requirement of the data distribution obstructs the application of fully/semi-supervised methods in our task, while self-supervised methods would disregard the available label information for known samples. Contrastive learning provides distinctive representations by controlling the instance similarity in feature space, achieving significant progress in recent years [He *et al.*, 2020]. The existing literature generalizes the idea to different scenarios [Khosla *et al.*, 2020], implements it for solving various downstream tasks [Chaitanya *et al.*, 2020], and provides theoretical support [Wang and Isola, 2020]. However, few attempts have been made at contrastive-based open-world cell type identification. scGAD generates pseudo-positive pairs for closely aligned representations, but this paradigm could lead to another dilemma: the representations and pseudo-labels are interdependent, which means an inferior feature space could lead to false positive pairs, and in turn deteriorate the learning of feature space.

3 Method

3.1 Problem Formulation

To commence, we'll clarify some notations. Our examination of a distribution-independent cell type identification task provides us with certain labeled reference data, denoted as $\mathcal{D}_r = \{(x_i^r, y_i^r)_{i=1}^{n_r}\}$, and unlabeled target data represented as $\mathcal{D}_t = \{(x_i^t)_{i=1}^{n_t}\}$. Both datasets may originate from either the same or different scRNA-seq datasets. The labels for $\mathcal{D}_r$ and $\mathcal{D}_t$ are represented respectively as $\mathcal{C}_r$ and $\mathcal{C}_t$. In the problem under consideration, we assume that $\mathcal{C}_r$ is a subset of $\mathcal{C}_t$. Moreover, the set of labels we have seen, or the seen label set, is defined as $\mathcal{C}_s = \mathcal{C}_r \cap \mathcal{C}_t$. Additionally, the novel label set is demarcated as $\mathcal{C}_n = \mathcal{C}_t \backslash \mathcal{C}_r$. The purpose of our work is to assign either seen cell type labels or clustering labels to cells within the target data. It is generally accepted that the total number of cell types within the whole dataset represented as $\mathcal{D}_r \cup \mathcal{D}_t$, is known prior to the assignment, since we can estimate it efficaciously using existing methods.

Given the high-dimensional and sparse characteristics inherent in scRNA-seq data, we have developed a model com-

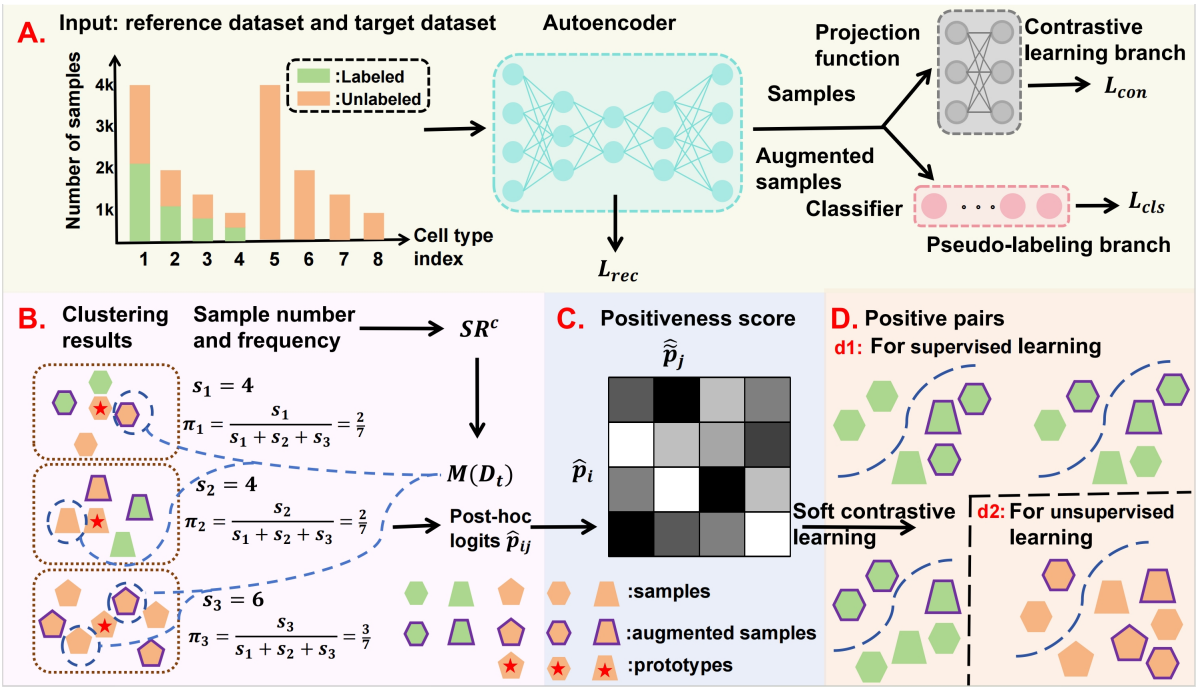

Figure 1: An overview of scDET. (A) The long-tailed reference and target datasets are inputs to the autoencoder, which outputs the embeddings of samples. Then two branches are connected to the embedding space to complete our cell type identification task. (B) $M(\mathcal{D}_t)$ is constructed by selecting samples with high prediction confidence in a ratio of SR^c. (C) The positiveness score is calculated by the similarity of the rectified cell type assignment distribution. (D) The illustration of unsupervised, supervised, soft contrastive learning paradigms.

posed of three primary components: a denoising autoencoder, a contrastive branch, and a pseudo-labeling branch (see Figure 1). The autoencoder compresses the input, represented as x, into an embedded feature that we signify as z. Following the compression, the input is then reconstructed through the use of z. The contrastive branch transforms the embedding, denoted as z, into a renewed representation symbolized as v via a projection layer. Simultaneously, the pseudo-labeling branch categorizes the same embedding z into one of the classes within the set $|\mathcal{C}_r \cup \mathcal{C}_t|$. This assignment is done with a probability p through classifier layer implementation. Motivated by the advancements in self-supervised learning [Liu *et al.*, 2021], we have employed a data augmentation approach to produce an alternate view, $\tilde{x}$, of gene expression, enhancing the ability to capture inter-gene correlations. Comprehensive details are available in the supplementary.

3.2 Revisiting the Contrastive Learning

Contrastive Learning provides distinctive and transferable representations by controlling the instance similarity in feature space, achieving significant progress in recent years [Chen *et al.*, 2020b]. Specifically, it aims to find a projection function to acquire optimal feature representation v of the gene expression input x, such that v retains the discriminative semantic information of the input cell. The general contrastive learning loss function can be defined as,

$$\mathcal{L}_{cl}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\left(\frac{1}{|P(i)|}\sum_{j\in P(i)} -\log\frac{\exp(v_i\cdot v_j/\tau)}{\sum_{a\in A(i)}\exp(v_i\cdot v_a/\tau)}\right), \tag{1}$$

where $\mathcal{D}$ is the training set and $A(i) = \mathcal{D}\setminus i$. $P(i)$ is the set of indices of positive pairs in $A(i)$, $|\cdot|$ denotes the operation to compute the inner product similarity, and τ is the temperature parameter. For the unsupervised learning scenario, the positive set is formulated as the two views of the same cell. For the supervised learning scenario, the positive set can be the cells that share the same ground truth labels.

3.3 Dynamic Distribution Estimation

To enhance and re-balance the contrastive representation learning, we propose to learn an auxiliary classifier for pseudo-labeling. Our pseudo-labeling branch includes a cross-entropy loss $\mathcal{L}_r$ on reference data, a self-training objective $\mathcal{L}_t$ on target data, and a regularization term $\mathcal{L}_{reg}$ on the whole data. Formally, we use the prototype parameterized classifier and randomly initialize a set of prototypes $\{\mu_i\}_{i=1}^{|\mathcal{C}_r\cup\mathcal{C}_t|}$, each standing for one cell type. Then we can calculate the assignment probability p_i for each cell x_i by softmax on cosine similarity between the hidden feature h_i and the prototypes $\{\mu_i\}_{i=1}^{|\mathcal{C}_r\cup\mathcal{C}_t|}$, i.e.,

$$p_{ij} = \frac{\exp(h_i\cdot\mu_j/\tau)}{\sum_{l=1}^{|\mathcal{C}_r\cup\mathcal{C}_t|}\exp(h_i\cdot\mu_l/\tau)}, \tag{2}$$

and the soft pseudo-label q_i is produced with a sharper temperature $\hat{\tau}$ in a similar fashion. Considering that we have two different views of cells, the classification objectives are then simply cross-entropy (CE) loss between the predictions and

pseudo-labels or ground-truth labels,

$$\mathcal{L}_r = \frac{1}{2n_r}\sum_{i=1}^{n_r}(CE(y_i^r, p_i^r) + CE(y_i^r, \tilde{p}_i^r)), \tag{3}$$

$$\mathcal{L}_t = \frac{1}{2n_r}\sum_{i=1}^{n_r}(CE(\tilde{q}_i^t, p_i^t) + CE(q_i^t, \tilde{p}_i^t)). \tag{4}$$

However, these two objectives do not guarantee the nontrivial solution of the model due to the prediction bias induced by weak supervision of novel cell types. So we propose to align the predictions with the data distribution to avoid non-activated classifiers. The difficulty lies in the fact that the distribution of the training set is independent in our task, and simply using the distribution of reference set $\pi_{\mathcal{D}^r}$ or a balance prior results in inferior performance. Furthermore, we observed that estimating the distribution from the pseudo-labeling branch itself could accumulate the estimation error and deteriorate model performance.

To this end, we perform estimation on the contrastive-learning branch as an alternative to avoid bias accumulation. Concretely, we perform k-means clustering on all samples in $\mathcal{D}_r \cup \mathcal{D}_t$ to obtain $|\mathcal{C}_r \cup \mathcal{C}_t|$ clusters with size $\{s_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|}$, and normalize the sample number $\{s_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|}$ to frequency $\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|}$, i.e., $\pi_i = s_i / \sum_{j=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|} s_j$. Moreover, we need to determine the corresponding relationship between clusters and cell types. So we use the Hungarian optimal assignment algorithm [Kuhn, 1955] to map $|\mathcal{C}_s|$ clusters to each known cell type. For the remaining $|\mathcal{C}_n|$ clusters, we sort them by cluster size and assign them sequentially to novel cell types. Finally, we regularize the mean prediction of the pseudo-labeling branch with the aligned distribution $\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|}$,

$$\begin{aligned}\mathcal{L}_{reg} = KL(\frac{1}{n_r + n_t}\sum_{i \in \mathcal{D}_r \cup \mathcal{D}_t} p_i || align(\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|})^\rho) \\ + KL(\frac{1}{n_r + n_t}\sum_{i \in \mathcal{D}_r \cup \mathcal{D}_t} \tilde{p}_i || align(\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|})^\rho),\end{aligned} \tag{5}$$

where $KL(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence between the two distributions, and $\rho \in [0, 1]$ is a hyper-parameter used to smooth the target long-tailed distribution. To balance estimation accuracy and computation overhead, we re-estimate the dataset distribution every m epoch. The training objective of the pseudo-labeling branch is as follows,

$$\mathcal{L}_{cls} = \mathcal{L}_r + \mathcal{L}_t + \mathcal{L}_{reg}. \tag{6}$$

3.4 Debiasing and Sampling Process

In the last section, we regularize the pseudo-labeling branch with the estimated data distribution $\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|}$ acquired from the contrastive-learning branch. In this step, we aim to transfer the knowledge from the pseudo-labeling branch to the contrastive-learning branch in turn and further enhance the representation learning, which is also beneficial to the estimation of data distribution. However, the long-tail distribution of the underlying dataset places additional requirements on knowledge transferring. On one hand, we should ensure that the knowledge to be transmitted is not affected by the long-tailed distribution. Meanwhile, this process should help the contrastive-learning branch cope with the issue of data imbalance and the lack of supervision for novel cell types.

We design a debiasing and sampling step of the pseudo-labels to meet the two aforementioned requirements for knowledge transfer. First, similar to the previous work [Menon *et al.*, 2020], we apply post-hoc logits adjustment based on the estimated distribution to the predicted logits to eliminate the bias caused by long-tailed distribution,

$$\hat{p}_{ij} = \frac{\exp(h_i \cdot \mu_j / \tau - \gamma \log \pi_j)}{\sum_{l=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|} \exp(h_i \cdot \mu_l / \tau - \gamma \log \pi_l)}, \tag{7}$$

where $\hat{p}_i$ denotes the rectified cell type probability prediction of cell x_i and γ is a hyper-parameter. Next, to re-balance the learning process and filter low-precision pseudo-labels, we propose to sample the unlabeled cells. Specifically, we sample the pseudo-labels of unlabeled cells in a training batch $\hat{p}_B = \{\hat{p}_i : i = 1, 2, ..., B\}$ according to their prediction cell type c, where $c_i = \arg\max \hat{p}_i$. Formally, the cell type-wise sampling rate SR^c can be defined as,

$$SR^c = \begin{cases} (\frac{\pi_c}{\min(\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|})})^{-\alpha}, & c \in \mathcal{C}_B, \\ (\frac{\pi_c}{\min(\{\pi_i\}_{i=1}^{|\mathcal{C}_r \cup \mathcal{C}_t|})})^{-\beta}, & otherwise, \end{cases} \tag{8}$$

where $\mathcal{C}_B \subset \mathcal{C}_r$ denotes the set of labels that are involved in the current batch, and $\alpha, \beta \in [0, 1]$ are two hyperparameters. Setting $\alpha = \beta = 0$ means sampling all pseudo-labels, while $\alpha = \beta = 1$ means setting the sampling rate to be inversely proportional to the estimated number of samples in its cell types. Note that we prioritize selecting samples with higher prediction confidence in each cell type to remove the potentially false pseudo-labels. The sampled $M(\mathcal{D}_t)$ complements the original long-tailed distribution to serve as a re-balance term for $\mathcal{D}_r$ and also provides additional supervision for novel cell types. With this debiasing and sampling step, the sampled cells $\{\mathcal{D}_r \cup M(\mathcal{D}_t)\}$ and their corresponding rectified pseudo-labels mitigate the impact of the imbalance issue and compensate the unlabeled open-set cells simultaneously.

3.5 Soft Contrastive Learning

In the above two sections, we generate pseudo-labels via the pseudo-labeling branch with help from the contrastive-learning branch. Those pseudo-labels could provide additional information to improve the contrastive learning branch as well. In particular, we adopt a soft contrastive strategy to fully leverage the probabilistic information in pseudo-labels. We design a pair-wise positiveness score to adjust the contribution of different samples to the anchor cell. For cell pair $(x_i, \tilde{x}_j)$, the positiveness score $w_{ij} = Sim(\hat{p}_i, \hat{\tilde{p}}_j)$. w_{ij} is obtained by calculating the similarity of the rectified cell type assignment distribution between x_i and $\tilde{x}_j$. In practice, we implement the similarity metric with the dot product operation, which can be interpreted as the probability of cell x_i and $\tilde{x}_j$ belonging to the same cell type. Finally, we formulate the soft contrastive loss by incorporating w into Equation (1),

$$\begin{aligned}\mathcal{L}_{cl}^{soft}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i \in \mathcal{D}}(\frac{1}{\sum_{j \in A(i)} w_{ij}}\sum_{j \in A(i)} -w_{ij} \\ \log \frac{\exp(\upsilon_i \cdot \tilde{\upsilon}_j / \tau)}{\sum_{a \in A(i)} \exp(\upsilon_i \cdot \tilde{\upsilon}_a / \tau)}).\end{aligned} \tag{9}$$

Minimizing the above equation encourages the similarity of features between two samples to be proportional to the corresponding positiveness score. In this way, we effectively transfer knowledge from the pseudo-labeling branch to contrastive

	Cao			Quake 10x			Quake Smart-seq2			Wagner			Zheng		
	seen	novel	overall	seen	novel	overall	seen	novel	overall	seen	novel	overall	seen	novel	overall
scziDesk	85.2	74.1	63.8	84.1	58.5	73.3	76.7	72.5	70.7	72.1	48.2	54.6	57.7	52.0	45.7
scNAME	79.1	78.5	75.1	82.2	62.0	69.8	76.5	61.2	63.5	74.4	48.4	54.8	57.7	52.0	45.7
scCNC	50.2	60.9	52.7	85.0	49.8	61.3	65.0	40.8	39.0	85.8	51.4	55.0	61.5	56.6	48.6
MARS	88.6	75.8	64.3	92.1	52.8	68.9	80.3	70.6	69.2	81.6	42.6	50.9	72.5	59.5	50.6
scNym	**99.2**	69.4	66.2	**98.4**	52.8	60.8	**96.9**	59.2	56.4	**96.5**	42.3	44.2	**98.8**	56.5	51.4
scArches	73.4	46.5	52.2	88.3	56.6	69.1	72.3	54.7	57.2	58.1	35.9	41.7	60.4	72.9	68.4
scGAD	92.4	**81.0**	78.3	95.8	62.1	83.7	91.3	76.3	75.7	92.1	49.6	56.4	97.6	**74.8**	66.3
scDET	97.2	79.7	**83.3**	98.2	**65.9**	**86.7**	95.3	**78.6**	**79.3**	96.0	**53.1**	**60.9**	94.7	72.8	**74.5**

Table 1: Performance comparison between various baselines on ten real datasets in intra-data annotation experiments.

	Lawlor (R) Baron_human (T)			Xin (R) Baron_human (T)			Vento _10x (R) Vento_Smart-seq2 (T)			Plasschaert (R) Montoro 10x (T)			Haber region (R) Haber largecell (T)		
	seen	novel	overall	seen	novel	overall	seen	novel	overall	seen	novel	overall	seen	novel	overall
scziDesk	81.3	80.3	81.2	75.1	84.2	81.3	81.7	79.0	81.5	67.9	74.6	68.3	85.3	80.8	71.0
scNAME	80.7	79.4	79.9	73.6	85.3	77.7	87.4	80.3	86.0	95.1	90.2	96.0	89.1	80.9	71.6
scCNC	54.0	43.9	40.9	46.6	54.7	36.5	92.1	63.4	84.8	79.7	73.1	73.0	75.7	50.4	51.6
MARS	80.9	**90.7**	80.3	93.6	78.0	88.6	71.3	78.6	70.3	88.6	94.5	89.1	83.8	64.1	67.1
scNym	90.2	52.2	82.8	**97.9**	40.0	52.3	98.7	66.5	75.9	**96.1**	77.7	83.1	84.2	53.7	53.0
scArches	47.3	66.8	52.5	61.5	52.2	52.7	87.6	52.9	78.2	91.4	67.4	85.3	71.9	45.4	50.4
scGAD	**96.6**	82.6	90.3	93.6	86.0	91.3	**98.8**	80.5	92.4	93.6	94.0	96.2	**89.8**	**81.5**	72.2
scDET	94.2	79.0	**90.9**	96.0	**94.4**	**95.2**	96.7	**84.7**	**94.1**	93.0	**98.3**	**96.2**	86.3	79.2	**80.5**

Table 2: Performance comparison between various baselines in inter-data annotation experiments. "R": reference data; "T": target data.

learning. The training objective of the contrastive learning (CL) branch consists of an unsupervised CL loss on all cells, a supervised CL loss on labeled cells, and the proposed soft CL loss on both labeled and sampled unlabeled subset,

$$\mathcal{L}_{con} = \mathcal{L}_{CL}^{u}(\mathcal{D}_r \cup \mathcal{D}_t) + \mathcal{L}_{CL}^{s}(\mathcal{D}_r) + \mathcal{L}_{CL}^{soft}(\mathcal{D}_r \cup M(\mathcal{L}_t)). \tag{10}$$

It is worth noting that the backbone of the two branches is shared. So the two-branch structure only slightly increases computational overheads. During inference, we utilize the pseudo-label learning branch and obtain predictions by finding the maximum component of the classification probability.

Overall loss. Together with the reconstruction loss $\mathcal{L}_{rec}$ (see supplementary), we give the overall training objective as,

$$\mathcal{L}_{tol} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{con}, \tag{11}$$

where λ_1 and λ_2 are two weight hyper-parameters.

4 Experiment

4.1 Setup

Data preparation. Our experiments encompass two types of annotation scenarios: intra-data annotation and inter-data annotation. In the process of intra-data annotation, we meticulously assembled a collection of 10 distinct datasets, each obtained from various organisms. The number of cells within these datasets exhibits a considerable range, with a minimum of 6,462 and a maximum of 110,704 cells. Additionally, the diversity of cell types is notable, fluctuating between 9 and 45 different types. In the absence of specific indications to the contrary, our standard procedure involves bifurcating all the cell types into two equal cohorts: 50% are classified as "seen" and the remaining 50% as "novel". After this classification, we randomly select 50% of the samples from the seen cell types to constitute the reference set, denoted as $\mathcal{D}_r$, while the remainder of the samples is designated to form the testing set, $\mathcal{D}_t$. Regarding inter-data annotation, our approach entails the selection of 10 paired groups of datasets. Each pair is composed of a reference dataset and a corresponding target dataset, between which batch effects are observed. The foundational attributes and details about these datasets are comprehensively enumerated in the supplementary.

Comparison baseline. Our research investigates a new task, wherein scGAD emerges as the optimal baseline method for our comparative analysis. Furthermore, we evaluate the performance of our method against an array of specialized techniques, including three clustering algorithms, i.e., scziDesk, scNAME, and scCNC, and three annotation methodologies, i.e., MARS, scNym, and scArches, each tailored for scRNA-seq data. In the context of the clustering approaches, it is notable that only scCNC is trained using both datasets $\mathcal{D}_r$ and $\mathcal{D}_t$, whereas scziDesk and scNAME are exposed solely to $\mathcal{D}_t$ during their training phase. We present their clustering efficacy for both known and novel cell types, offering a comprehensive assessment of their performance. Turning to the annotation strategies, we initially deploy these methods to categorize the target cells into established cell types and segregate an "unassigned" category. Subsequently, we implement k-means clustering on the cells within this "unassigned" classification to delineate novel cellular clusters. Lastly, we run each compared method in their default settings.

Evaluation metrics. Similar to scGAD, our study presents the classification accuracy for recognized cell types, denoted as $\mathcal{C}_s$, alongside the clustering accuracy for previously

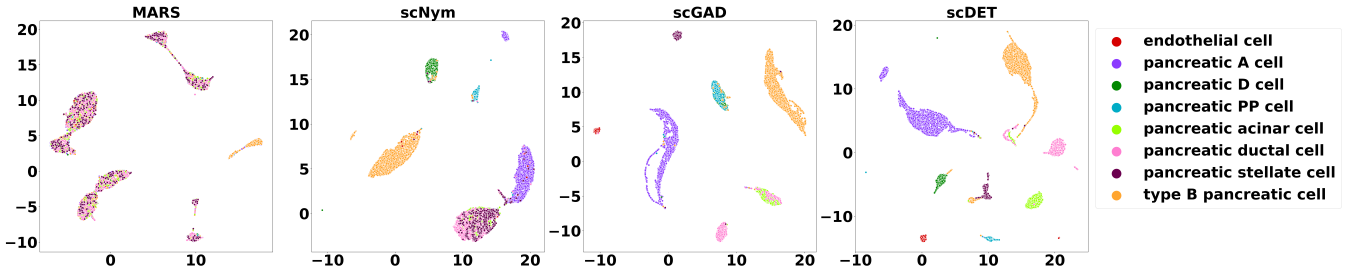

Figure 2: UMAP visualization of four methods on one inter-data annotation task from Xin dataset to Baron_human dataset, where "pancreatic A cell", "pancreatic D cell", "pancreatic PP cell" and "type B pancreatic cell" are seen cell types and other four cell types are novel ones.

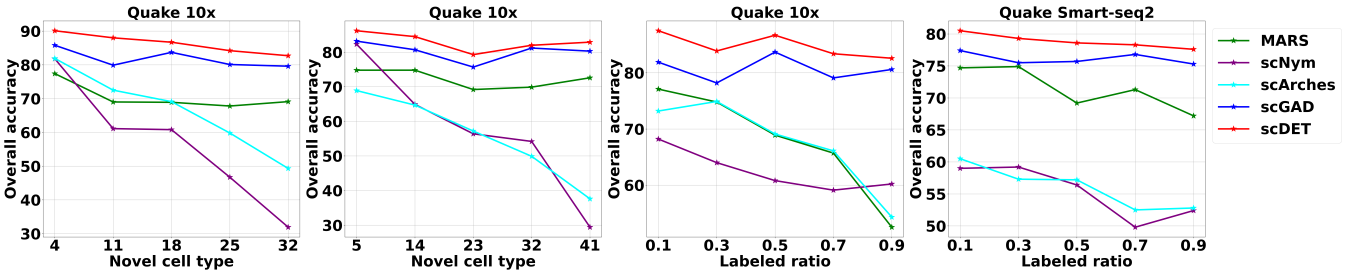

Figure 3: The variation of overall accuracy in the face of changing the novel cell type number and the labeled ratio value on Quake 10x and Quake Smart-seq2 datasets.

unidentified cell types, symbolized as $\mathcal{C}_n$. These metrics are used for evaluating the performance of annotation baselines. Conversely, the performance of clustering baselines is assessed by examining clustering accuracy across both known and novel cell types. To determine the clustering accuracy more precisely, the Hungarian algorithm is employed to address the optimal assignment problem, as detailed in [Kuhn, 1955]. In scenarios where accuracy across the collective set of cell types, $\mathcal{C}_s \cup \mathcal{C}_n$, is being evaluated, the Hungarian algorithm is applied to ascertain the best assignment strategy for both seen and novel cell categories. It is important to note that the accuracies reported are not singular measurements but rather the mean of three independent experimental runs, ensuring the reliability of the results.

Implementation details. We implement all our techniques using PyTorch and conduct the experiments using 2 Tesla A100 GPUs. Following scGAD, the encoder has two layers with sizes 512 and 256, the decoder has the reverse structure of the encoder, and the latent space has a dimension of 128. The pseudo-labeling branch is implemented by two fully connected layers with size 128 and a number of the whole cell types. The contrastive-learning branch also consists of two linear layers sized 128. The training mini-batch size is set to 256, and the optimizer is Adam with a learning rate of 1e-4. The temperature τ is set to 1.0, and the loss weight parameters λ_1 and λ_2 are both set to 1.0. The other hyperparameters ρ, γ, α and β are all set to 0.5. Lastly, the whole model is trained by 500 epochs for each dataset and the data distribution is re-estimated every 10 epochs.

4.2 Results

Intra-data experiments. We first explore the performance of scDET under the intra-data annotation setting without access to the batch effect. From the results in Table 1, scDET gives consistently superior performance than other methods under the overall accuracy on almost all datasets. Specially, compared with scGAD, our method achieves higher novel accuracy and better trade-off between classifying the seen cell types and clustering the novel cell types. It is not surprising that scDET gets such an excellent performance since the autonomously equilibrated dual-consultative contrastive learning framework can collaboratively tackle the scRNA-seq data imbalance issue and discover the out-of-distribution novel cell types. Besides, although scNym can obtain relatively high annotation accuracy on seen cell types, it has a sharp drop in clustering accuracy on novel cell types. MARS and scArches occasionally achieve competitive novel accuracy on some datasets, but they can only provide sub-optimal results for both classification and clustering accuracy. This evidence fully shows that the two-step strategy that first detects novel cells with "unassigned" labels and then clusters them is not an appropriate solution to this open-world task. As clustering methods, scziDesk and scNAME cannot provide satisfactory results for the reason that they do not utilize the information in reference datasets, which makes them less competitive. In general, scDET outperforms other baselines and achieves remarkable progress in this setting.

Inter-data experiments. Then we turn to study a more challenging setting with the batch effect, i.e., inter-data annotation. From the results in Table 2, scDET still achieves better results than other algorithms on most mixed datasets, especially for novel accuracy and overall accuracy. For those

groups where scGAD performs well, scDET can also perform well, and when the scGAD's performance is less than satisfactory on some datasets, scDET can show relatively stable and excellent performance. Moreover, compared with the intra-data setting, there is no significant decline in the performance of scDET, indicating that it could resist the effect of batch effect to some extent. In comparison, MARS and scArches are easily susceptible to batch effect and result in inferior precision on cell type identification, because they separate the learning process on reference data from that on target data. Similarly, scNym is easy to overfit to seen cell types and lacks strong discriminative power on novel cell types. Although the accuracy of scziDesk and scNAME seems competitive, that is because they do not use reference sets, thereby avoiding the batch effect. However, the high clustering accuracy on seen cell types does not facilitate the annotation process in practicality. In summary, scDET can work well under the challenging inter-data annotation scenario.

Feature visualization. To observe the cell type identification results more intuitively, we extract the low-dimensional embedding features of four methods and use the UMAP approach to visualize them in Figure 2. We can see that MARS confuses known cell types with novel cell types, and scNym fails to separate groups of novel cell types. Although scGAD performs better than them, it still does not recognize the known pancreatic D cells well and mixes some novel pancreatic ductal cells with pancreatic acinar cells. By contrast, our method does a good job of separating each known cell type from each novel cell type.

	Quake 10x			Quake Smart-seq2		
	seen	novel	overall	seen	novel	overall
reg with $\pi_{\mathcal{D}_r}$	89.3	34.7	56.1	83.8	42.4	57.6
balanced prior	87.5	48.2	67.4	80.7	58.2	69.5
cls k-means	95.1	60.8	82.0	92.9	72.3	74.7
ours	98.2	65.9	86.7	95.3	78.6	79.3
oracle	98.6	68.3	87.9	96.6	80.1	80.5

Table 3: Ablation study for regularization term.

	Quake 10x			Quake Smart-seq2		
	seen	novel	overall	seen	novel	overall
baseline	93.4	52.5	76.9	91.6	62.8	68.0
hard	95.3	60.4	83.1	91.7	73.8	75.2
ours	98.2	65.9	86.7	95.3	78.6	79.3

Table 4: Ablation study for contrastive loss design.

4.3 Ablation Study

Robustness analysis. We first investigate the effect of novel cell type numbers on the performance of each method and conduct control experiments on Quake 10x and Quake Smart-seq2 datasets that hold massive cell types. From the results in Figure 3, we can see that the overall accuracy of almost all tested methods declines with the increase in the number of novel cell types. This is reasonable because the number of new cell types determines the difficulty of discovering them. However, scDET always outperformed other methods regardless of the number of novel cell types, demonstrating the stability of our method. Then we vary the ratio of labeled data to study its impact on the results of five methods. Figure 3 shows the variation in overall accuracy with the changing of labeled ratio on Quake 10x and Quake Smart-seq2 datasets. We find that scDET still achieves consistently better results than the other baselines and maintains its superior performance without being affected by the ratio of labeled data.

Effectiveness of $\mathcal{L}_{reg}$. Recall in Section 3.3, we propose to regularize the predictions of the pseudo-labeling branch by the estimated train set distribution. In Table 3, we show the performance of the pseudo-labeling branch on Quake 10x and Quake Smart-seq2 with different estimation strategies. "Oracle" denotes we use the true distribution π of $\mathcal{D}_r \cup \mathcal{D}_t$ (unknown in practice) as the target distribution in Equation 5, and it serves as an upper bound of the performance. Compared to using a balance prior, regularizing the predictions with oracle distribution significantly improves the performance on both seen and novel cell types, showing the importance of the distribution estimation. Meanwhile, the similar results achieved by our estimation strategy imply it could be a reliable proxy to π. Furthermore, we investigate whether two alternative estimation strategies could help the pseudo-labeling branch: 1) only regularize seen cell types prediction with $\pi_{\mathcal{D}_r}$, 2) perform k-means clustering on the feature of the pseudo-labeling branch. Both of them result in inferior accuracy and could in turn deteriorate the contrastive learning process.

Effectiveness of $\mathcal{L}_{CL}^{soft}$. In Section 3.5, we design a novel soft contrastive loss based on pseudo-labels to transfer the knowledge of the pseudo-labeling branch into the contrastive learning branch. As an opposite, we could also construct the loss in a hard manner where we formulate the positive pairs on top of the predictive cell type with the largest logit and further perform the supervised contrastive loss. Intuitively, the hard design discards the probability distribution information and is more susceptible to the false pseudo-labels, while the soft contrastive loss utilized in our method could help alleviate the influence of erroneous pseudo-labels. The results also support the intuition, as shown in Table 4, that the proposed $\mathcal{L}_{CL}^{soft}$ outperforms the supervised CL loss by a large margin. This phenomenon is also observed in knowledge distillation that transferring knowledge by using soft labels rather than one-hot predictions can achieve better performance.

5 Conclusion

In this paper, we formulate a realistic distribution-independent cell type identification task that unifies long-tailed and open-set learning and design a novel framework scDET for this task to fight against scRNA-seq data imbalance and classify novel cell types simultaneously. Extensive experiments on various datasets verify the significance of scDET, and deeper analyses show the effectiveness of its proposed individual components.

Contribution Statement

Yuyao Zhai and Liang Chen made the same contribution to this paper, and Minghua Deng is the corresponding author.

References

[Brbić *et al.*, 2020] Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods*, 17(12):1200–1206, 2020.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[Cao *et al.*, 2020] Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, and Ge Gao. Searching large-scale scrna-seq databases via unbiased cell embedding with cell blast. *Nature communications*, 11(1):3458, 2020.

[Chaitanya *et al.*, 2020] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in neural information processing systems*, 33:12546–12558, 2020.

[Chen *et al.*, 2020a] Liang Chen, Weinan Wang, Yuyao Zhai, and Minghua Deng. Deep soft k-means clustering with self-training for single-cell rna sequence data. *NAR genomics and bioinformatics*, 2(2):lqaa039, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[Kimmel and Kelley, 2021] Jacob C Kimmel and David R Kelley. Semisupervised adversarial neural networks for single-cell classification. *Genome research*, 31(10):1781–1793, 2021.

[Kiselev *et al.*, 2019] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[Kuhn, 1955] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[Lähnemann *et al.*, 2020] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[Liu *et al.*, 2021] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[Lotfollahi *et al.*, 2022] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology*, 40(1):121–130, 2022.

[Menon *et al.*, 2020] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[Petegrosso *et al.*, 2020] Raphael Petegrosso, Zhuliu Li, and Rui Kuang. Machine learning and statistical methods for clustering single-cell rna-sequencing data. *Briefings in bioinformatics*, 21(4):1209–1223, 2020.

[Regev *et al.*, 2017] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

[Shen *et al.*, 2016] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 467–482. Springer, 2016.

[Wan *et al.*, 2022] Hui Wan, Liang Chen, and Minghua Deng. scname: neighborhood contrastive clustering with ancillary mask estimation for scrna-seq data. *Bioinformatics*, 38(6):1575–1583, 2022.

[Wang and Isola, 2020] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[Wang *et al.*, 2017] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2022] Hai-Yun Wang, Jian-Ping Zhao, Chun-Hou Zheng, and Yan-Sen Su. sccnc: a method based on capsule network for clustering scrna-seq data. *Bioinformatics*, 38(15):3703–3709, 2022.

[Xu *et al.*, 2021] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620, 2021.

[Zhai *et al.*, 2023] Yuyao Zhai, Liang Chen, and Minghua Deng. scgad: a new task and end-to-end framework for

generalized cell type annotation and discovery. *Briefings in Bioinformatics*, 24(2):bbad045, 2023.

[Zhang *et al.*, 2023] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.