PLOS Computational Biology

RESEARCH ARTICLE

Partial domain adaptation enables cross domain cell type annotation between scRNA-seq and snRNA-seq

Xiran Chen[1,2,3], **Quan Zou**[2], **Qinyu Cai**[4], **Xiaofeng Chen**[3], **Weikai Li**[1]*, **Yansu Wang**[2]*

1 School of Computer and Artificial Intelligence, Shandong Jianzhu University, Shandong, China, **2** Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, China, **3** School of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing, China, **4** School of Life Sciences, Westlake University, Hangzhou, China

* leeweikai@outlook.com (WL); wangyansu@uestc.edu.cn (YW)

Check for updates

OPEN ACCESS

Citation: Chen X, Zou Q, Cai Q, Chen X, Li W, Wang Y (2026) Partial domain adaptation enables cross domain cell type annotation between scRNA-seq and snRNA-seq. PLoS Comput Biol 22(5): e1014223. https://doi.org/10.1371/journal.pcbi.1014223

Editor: Wei Li, University of Maryland School of Medicine, UNITED STATES OF AMERICA

Received: November 12, 2025

Accepted: April 9, 2026

Published: May 6, 2026

Copyright: © 2026 Chen et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The source code is located at https://github.com/OPUS-Lightphenexx/ScNucAdapt. The data used in this study are available in Gene Expression Omnibus (GEO) with accession numbers GSE267964, GSE140989, GSE121862, GSE123454, GSE140819.

Abstract

Accurate cell type annotation across datasets is a key challenge in single-cell analysis. snRNA-seq enables profiling of frozen or difficult-to-dissociate tissues, complementing scRNA-seq by capturing fragile or rare cell types. However, cross-annotation between these two datasets remains largely unexplored, as existing methods treat them independently. We introduce ScNucAdapt, a method designed for cross-annotation between paired and unpaired scRNA-seq and snRNA-seq datasets. To address distributional and cell composition differences, ScNucAdapt employs partial domain adaptation. Experiments across both unpaired and paired scRNA-seq and snRNA-seq show that ScNucAdapt achieves robust and accurate cell type annotation, outperforming existing approaches. Therefore, ScNucAdapt provides a practical framework for the cross-domain cell type annotation between scRNA-seq and snRNA-seq data.

Author summary

Single-cell and single-nucleus RNA sequencing are two powerful technologies that allow scientists to study gene activity in individual cells. However, comparing data between these methods remains challenging because they capture different parts of the cell and are often collected under different conditions. This makes it difficult to consistently identify cell types across experiments, hindering our understanding of health and disease. We developed ScNucAdapt, a computational framework that can automatically transfer cell type knowledge between these two types of datasets, even when they come from different laboratories or tissue conditions. Our method learns to recognize shared patterns while ignoring dataset-specific differences. Through testing on diverse tissues, including bladder, kidney, tumors, and brain, we show that ScNucAdapt consistently outperforms existing approaches. By enabling reliable integration of single-cell and

Funding: This work was supported by the National Natural Science Foundation of China (Grant No.62131004 to QZ, Grant No.62531002 to YW, Grant No.62306051 to WL, Grant No.62481540175 to WL, Grant No.62276035 XFC), the Taishan Scholars Foundation of Shandong Province (Grant No.tsqn202507225 to WL), and the Natural Science Foundation of Chongqing (Grant No.CSTB2025NSCQ-GPX0857 to WL). The Fundamental Research and the Scientific and the Technological Research Program of Chongqing Municipal Education Commission (Grant No.KJQN202300718 to WL). The Science and Technology Research Program of Chongqing Municipal Education Commission, China (Grant No. KJZD-K202400703 to XFC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

single-nucleus data, our work helps researchers build more complete pictures of cellular diversity across tissues and disease states. This capability is particularly valuable for studying archived frozen samples or fragile cell types that are difficult to analyze with conventional methods, potentially accelerating discoveries in various fields.

Introduction

Cells are the fundamentals of life [1]. It is vital to annotate each cell correctly in terms of its transcriptomic profiles [2], enabling the identification of distinct cellular populations, comparison across samples, and linkage of molecular profiles to biological function or disease [3]. Most published methods on automatic cell type annotations are based on scRNA-seq, including SingleCellNet [4], which uses an ensemble of Random Forest classifiers for the cell type annotation of scRNA-seq datasets, and ScMap [5] works by comparing the gene-expression profile of each new cell to reference data and labeling the cell with the type that shows the highest similarity.

However, when confronted with frozen samples or tissues that are difficult to dissociate, snRNA-seq offers a practical alternative to scRNA-seq [6], capturing nuclear transcripts without viable whole cells and enabling detection of fragile or rare cell types that single-cell methods often underrepresent [7]. Many studies have integrated scRNA-seq and snRNA-seq, for instance, neurodegenerative diseases [8], skeletal muscle [9], frozen and fresh tumor samples [10], and even PBMC for a disease progression study [11]. This shows that cross-domain annotation between scRNA-seq and snRNA-seq is essential for unifying cellular identities across two datasets and ensuring consistent interpretation of data generated from different tissue conditions or experimental protocols [12]. Previous research on annotating cell types for snRNA-seq and scRNA-seq data relies on traditional machine learning methods [13], particularly for kidney cell types [14]. However, these methods overlooked the relationships between scRNA-seq and snRNA-seq, treating them as separate datasets. Therefore, there's an urgent need for development in cross-domain cell type annotation between scRNA-seq and snRNA-seq.

However, distributional differences often occur between scRNA-seq and snRNA-seq [15], including paired and unpaired ones, as shown in Fig 1. Moreover, in a real automatic annotation situation, the cell type composition of target datasets is unknown, which could cause cell type compositions to differ between the two datasets, making it challenging to achieve robust annotation between them. Therefore, inspired by partial domain adaptation, which can simultaneously address both the distribution differences between the two datasets and the mismatch in their label spaces, we developed a framework called ScNucAdapt that selectively transfers knowledge from the source dataset to the target dataset.

Partial domain adaptation [16] addresses the problem of transferring knowledge from a labeled source domain to an unlabeled target domain when the target label space is a subset of the source label space, as shown in Fig 2. Unlike

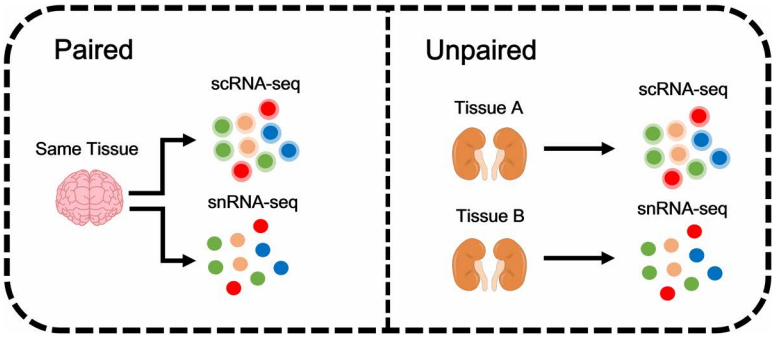

Fig 1. Concept of pair and unpaired scRNA-seq and snRNA-seq.

https://doi.org/10.1371/journal.pcbi.1014223.g001

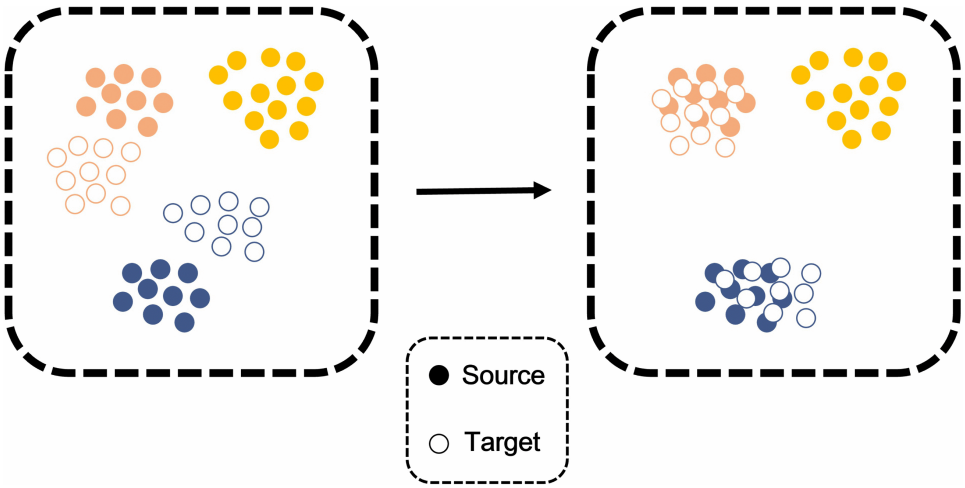

Fig 2. Concept of partial domain adaptation.

https://doi.org/10.1371/journal.pcbi.1014223.g002

traditional domain adaptation, which assumes identical label spaces across domains, Partial domain adaptation mitigates the negative transfer caused by irrelevant source classes. Partial domain adaptation has been applied to various fields, including fault diagnosis [17], pneumonia diagnosis from chest x-ray images [18], and cross-session neural decoding [19].

This design enables ScNucAdapt to focus on cell types that are shared across datasets while minimizing the negative impact of non-overlapping or dataset-specific cell types. Moreover, ScJoint [20], ScNCL [21], ScCobra [22], and ScCorrect [23], which utilize transfer learning and domain adaptation methods for label transfer from unpaired scRNA-seq to ScATAC-seq datasets, also provided inspiration for developing ScNucAdapt.

To the best of our knowledge, Our study is the first to focus on cross-annotation between paired or unpaired scRNA-seq and snRNA-seq datasets. In summary, the contributions of our proposed method can be listed as follows:

- ScNucAdapt is a cross-domain annotation framework that enables robust label transfer between paired or unpaired scRNA-seq and snRNA-seq datasets.
- ScNucAdapt also considers distributional differences between scRNA-seq and snRNA-seq, making it more robust when annotating cell types in the target datasets.

- ScNucAdapt also could handle cell type compositional differences between scRNA-seq and snRNA-seq, where only a subset of cell types are shared across these two datasets.

The following passages are organized as follows. In Section 2, the methods of ScNucAdapt are presented, and the descriptions of the datasets used and the evaluation metrics are included. In Section 3, the experimental results on classification accuracy between a set of scRNA-seq and snRNA-seq are presented, including an ablation experiment that demonstrates the effectiveness of each component in ScNucAdapt, as well as a sensitivity analysis and a runtime and memory scaling test. In Section 4, the discussion is presented. Finally, in Section 5, the conclusion is presented.

Materials and methods

Shared source and target encoder

To extract features from both source and target datasets into a common label space, a shared encoder is used. The encoder is composed of two fully connected layers. The first layer transforms the input features into hidden units. The second layer reduces these features into a latent space, creating a compact representation that captures the most important patterns in the gene expression data.

Dynamic clustering in target data

In this section, we will introduce the concept of dynamic clustering in the target dataset without the prior knowledge of the number of clusters, which was inspired by DeepDPM [24] for an unknown number of clusters for deep clustering, and PRAGA [25] for spatial multi-omics clustering.

Given the target datasets $X_t \in R^{n\times m}$, the target representations $Z_t \in R^{n\times m'}$ are obtained in Eq. (1). Where n represents the number of samples. m, m' represent the original number of features and the number of features after passing the shared encoder.

$$Z_t = MLP(X_t) \tag{1}$$

After obtaining the representations of the target dataset, we set the initial cluster C for the Gaussian mixture model to assign the target representations into C clusters; note that C doesn't represent the true number of clusters in the target representations or the true cell type labels in the source dataset, and would further be adjusted through a split and merge framework, which is performed through the Metropolis-Hastings framework [26].

The sample count and the target representations of each cluster are denoted by N_c and Z_c^t, respectively, where the subscript $c = 1, 2, \ldots, C$ denotes the cluster index. To enable dynamic adjustment of the total number of clusters, each cluster is further divided into two sub-clusters using Gaussian mixture. The corresponding sample count of the sub-clusters is represented as $N_{c,s}$, where $s \in \{1, 2\}$ denotes the sub-cluster index. A splitting criterion is then defined for each cluster to determine whether it should be further partitioned, where the split is accepted with probability $\min(1, H_s)$. The hasting ratio H_s is defined in Eq. (2).

$$H_s = \frac{\Gamma(N_{c,1})L(Z_{c,1}^t; \nu, \kappa, m, \psi)\Gamma(N_{c,2})L(Z_{c,2}^t; \nu, \kappa, m, \psi)}{\Gamma(N_c)L(Z_c^t; \nu, \kappa, m, \psi)} \tag{2}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $L(\cdot; \nu, \kappa, m, \psi)$ corresponds to the marginal likelihood evaluated under a Normal–Inverse–Wishart (NIW) prior, parameterized by the hyperparameters ν, κ, m, ψ. When $H_s > 1$, the original cluster is substituted with one of its derived subclusters, and the remaining subcluster is incorporated as an additional, distinct cluster shown in Eq. (3).

$$Z_c^t := Z_{c,1}^t; Z_C^t := Z_{c,2}^t (C := C + 1) \tag{3}$$

After splitting the clusters, we introduce a decision for merging criterion based on the clusters after splitting, where the merging is accepted with probability $\min(1, H_m)$, as shown in Eq. (4).

$$H_{m(i,j)} = \frac{1}{H_{s(i,j)}} = \frac{\Gamma(N_i + N_j)L(Z_i^t \cup Z_j^t; \nu, \kappa, m, \psi)}{\Gamma(N_i)L(Z_i^t; \nu, \kappa, m, \psi)\Gamma(N_j)L(Z_j^t; \nu, \kappa, m, \psi)} \tag{4}$$

Rather than exhaustively considering all possible merges in sequence, we limit the merge candidates for each cluster N_i to its nearest neighbors N_j. If $H_{m(i,j)} > 1$, then the new merge cluster will replace the original two clusters, shown in Eq. (5).

$$Z_i^t := \varnothing; Z_j^t := \varnothing; Z_C^t := Z_i^t \cup Z_j^t (C := C - 1) \tag{5}$$

Cauchy-Schwarz Divergence

In this section, we introduce Cauchy-Schwarz Divergence [27] and the emperical estimator of $D_{CS}(p_s(z); p_t(z))$. Given the source representations of a certain class $Z_s \in R^{N \times m'}$ and target representations of a certain cluster $Z_t \in R^{n \times m'}$, where the source representations vectors $\{z_i^s\}_{i=1}^N \in Z_s$ and target representations vectors $\{z_i^t\}_{i=1}^n \in Z_t$, the Cauchy-Schwarz Divergence (CS Divergence) is defined in Eq. (6).

$$D_{CS}(p_s; p_t) = -\log\left(\frac{(\int p_s(z)p_t(z)\mathrm{d}z)^2}{\int p_s^2(z)\mathrm{d}z \int p_t^2(z)\mathrm{d}z}\right) \tag{6}$$

Using kernel density estimation, we estimate the densities from finite samples defined in Eq. (7) and Eq. (8).

$$\hat{p}_s(z) = \frac{1}{M}\sum_{i=1}^{M} \kappa_\sigma(z, z_i^s) \tag{7}$$

$$\hat{p}_t(z) = \frac{1}{N}\sum_{i=1}^{M} \kappa_\sigma(z, z_i^t) \tag{8}$$

Here we chose gaussian kernel function as the kernel estimator $\kappa_\sigma(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{2\sigma^2}\right)$, where σ is the bandwidth parameter that controls the smoothness of the kernel. The calculation results of $\int \hat{p}_s^2(z)\mathrm{d}z$, $\int \hat{p}_t^2(z)\mathrm{d}z$, $\int \hat{p}_s(z)\hat{p}_t(z)\mathrm{d}z$ are shown in Eq. (9), Eq. (10) and Eq. (11).

$$\int \hat{p}_s^2(z)\mathrm{d}z = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} \kappa_{\sqrt{2}\sigma}(z_j^s, z_i^s) \tag{9}$$

$$\int \hat{p}_t^2(z)\mathrm{d}z = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \kappa_{\sqrt{2}\sigma}(z_j^t, x_i^t) \tag{10}$$

$$\int \hat{p}_s(x_s)\hat{p}_t(x_t)\mathrm{d}z = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\kappa_{\sqrt{2}\sigma}(z_j^t, z_i^t) \tag{11}$$

By substituting Eq. (9)-Eq. (11) into CS Divergence, we measure the divergence between source and target datasets as follows:

$$D_{CS}(p_s; p_t) = \log(\frac{1}{M^2}\sum_{i,j=1}^{M}\kappa_{\sqrt{2}\sigma}(z_j^s - z_i^s))+$$

$$\log(\frac{1}{N^2}\sum_{i,j=1}^{N}\kappa_{\sqrt{2}\sigma}(z_j^t - z_i^t))$$

$$-2\log(\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\kappa_{\sqrt{2}\sigma}(z_j^t, z_i^s))$$

Source class-target cluster matching

In this section, we will be introducing the Source Class-Target Cluster Matching after obtaining the CS divergence between source known cell classes and predicted target clusters. And further assign the matched pairs' CS Divergence to the final training loss.

Given the source representations vectors Z_s and with p labels, we could divide the source representations into p subsets, noting as $Z_s = \{Z_{1,s}, Z_{2,s}, \ldots, Z_{p,s}\}$. After performing dynamic clustering on the target representations on target representations vectors Z_t, we obtain $\hat{C}$ clusters $Z_t = \{Z_{1,t}, Z_{2,t}, \ldots, Z_{\hat{C},t}\}$. Then, in terms of CS divergence described in Section 2.3, we calculate the CS divergence of each pair from the source classes and target clusters $Z_{i,s}$ and $Z_{j,t}$, which is $D_{CS}(p_{i,s}; p_{j,t})$, where $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, \hat{C}$. To identify which subsets in the source representations best correspond to each cluster in the target representations, we selected those with the lowest CS divergence paired with each target cluster. Therefore, we could obtain the global merge decision, the total loss function for minimization on the mini-batched source and target dataset are shown in Eq. (12), where a_i represents the corresponding subsets that match the i$-th target cluster.

$$L_{cs} = \sum_{i=1}^{\hat{C}} D_{CS}(p_{a_i,s}^b; p_{i,t}^b) \tag{12}$$

Shared source and target classifier

The shared source and target classifier operates on the latent representations produced by the encoder to predict the corresponding cell type. It consists of a single fully connected layer that maps directly to the output nodes, with the number of nodes equal to the number of cell types in the source dataset. This design allows the classifier to effectively translate the compact latent representations into accurate cell-type predictions while maintaining computational efficiency. The same classifier structure is consistently applied across all experiments without modification.

The overview of ScNucAdapt

In conclusion, ScNucAdapt consists of three parts. The framework is shown in Fig 3. The first is the shared encoder for the source and target datasets, aiming to extract representations in the same latent space. The second is the dynamic clustering in target representations. In this part, ScNucAdapt is responsible for clustering cell types in target datasets without giving prior knowledge on the number of clusters, and further adjusting through a split and merge framework. Then, we introduce CS Divergence and the rule for merging between the predicted target clusters and the source datasets.

ScNucAdapt employs a two-stage training strategy, and the pseudocode is shown in supplementary S5 Fig. In the first stage, the encoder is trained for T warm-up epochs using only minibatch-based representations learning without clustering, thereby learning meaningful initial feature spaces. In the second stage, each epoch begins by applying GMM clustering and split/merge operation to the full learned representations of the target dataset, followed by source-target matching via argmin of Cauchy-Schwarz divergence between cluster distributions. The encoder is then updated via backpropagation on minibatches using the combined loss shown in Eq. (13) while treating the current cluster assignments as fixed. GMM clustering, split/merge operation, and source-target matching are recomputed every epoch to refine alignments as representations improve. To reduce computational time, we could also recompute every n epochs, but in our experiments, the operation is computed every epoch. Moreover, the total training loss on the selected source and target batch dataset consists of classification loss L_{cls}, which is the weighted cross-entropy loss, and the L_{cs} described in section 2.4. λ represents the trade-off hyperparameter.

$$L = L_{cls} + \lambda * L_{cs} \tag{13}$$

Datasets

Various scRNA-seq and snRNA-seq data are compiled from previous publications. Most datasets are preprocessed; datasets that require preprocessing are preprocessed using Scanpy [28]. We gathered cells from the bladder, kidney, the mouse cortex, and the frozen and fresh tumor tissues.

For bladder cell types, the dataset is GSE267964 [29], which contains two subsets, Immune and Stromal. The datasets are preprocessed in advance and paired.

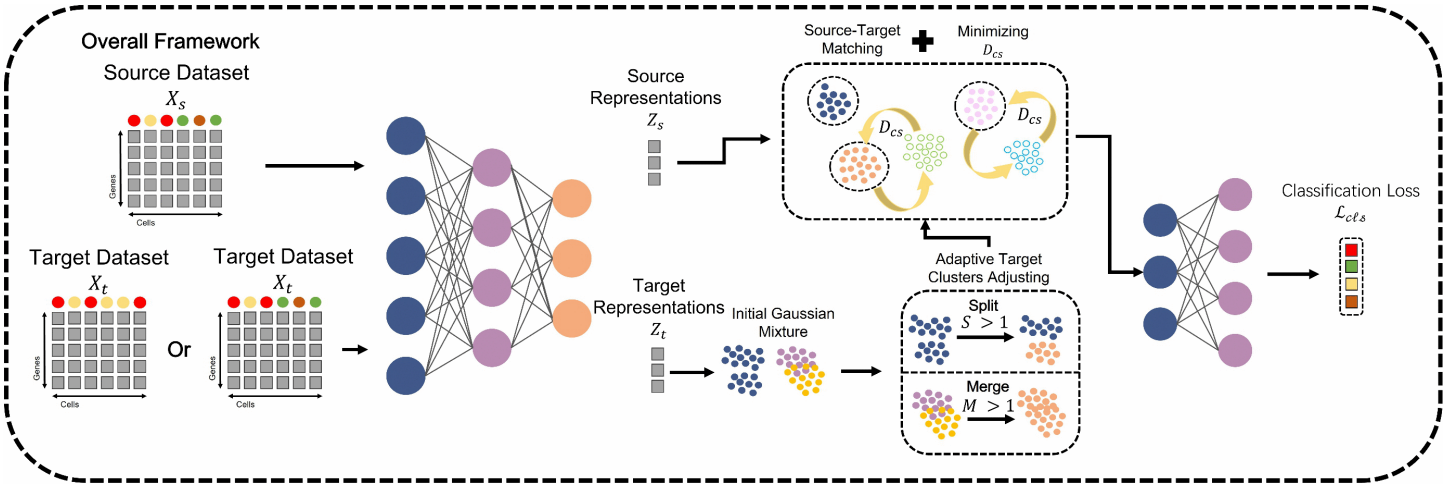

Fig 3. Overall framework of ScNucAdapt.

https://doi.org/10.1371/journal.pcbi.1014223.g003

Moreover, we also collected unpaired scRNA-seq and snRNA-seq of kidney cell types from different publications, GSE140989 [30], which is the scRNA-seq, and GSE121862 [31], which is the snRNA-seq. The cell type labels are gathered from a previous study on annotating cell types in kidneys from scRNA-seq and snRNA-seq using traditional machine learning methods.

For frozen and fresh tumor cell types, the datasets were gathered from the GEO database under accession number GSE140819, which contains many types of frozen tumors of scRNA-seq and snRNA-seq. We collected the cell types from metastatic breast cancer (MBC) and Chronic lymphocytic leukemia (CLL), then we further preprocessed them by filtering cells and genes that have low counts.

For mouse cortical cell types, the datasets are gathered from the GEO database under accession number GSE123454 [32]. All the cell type labels are collected from previous publications, with each cell annotated.

The detailed statistics of the datasets are shown in Table 1, including the number of samples, genes, and the number of unique cell types in each dataset.

Moreover, we provide the detailed adaptation settings between each dataset in Table 2. These include a total of eight adaptation scenarios: four partial settings and four closed-set settings. Each configuration specifies the source–target dataset pairs, the shared and non-shared label spaces.

Table 1. Statistical results of the datasets, including bladder cell types, kidney cell types, frozen and fresh tumor cell types, and mouse cortical cell types.

Datasets	Cells	Genes	Cell Types
GSE267964-Immune (Sc)	1725	36387	9
GSE267964-Immune (Sn)	369	36387	7
GSE267964-Stromal (Sc)	7227	36387	8
GSE267964-Stromal (Sn)	5737	36387	8
GSE140989 (Sc)	20927	18743	13
GSE121862 (Sn)	11684	18743	11
GSE123454 (Sc)	463	40023	2
GSE123454 (Sn)	463	40023	2
GSE140819-CLL (Sc)	2562	33538	3
GSE140819-CLL (Sn)	2297	33538	2
GSE140819-MBC (Sc)	5163	30316	8
GSE140819-MBC (Sn)	7260	30316	7

https://doi.org/10.1371/journal.pcbi.1014223.t001

Table 2. Adaptation settings between datasets.

Datasets	Setting
GSE267964-Immune (Sc)→ GSE267964-Immune (Sn)	Partial
GSE267964-Stromal (Sc)→GSE267964-Stromal (Sn)	Closed Set
GSE267964-Stromal (Sn)→GSE267964-Stromal (Sc)	Closed Set
GSE140989 (Sc)→GSE121862 (Sn)	Partial
GSE123454 (Sc)→ GSE123454 (Sn)	Closed Set
GSE123454 (Sn)→GSE123454 (Sc)	Closed Set
GSE140819-CLL (Sc)→ GSE140819-CLL (Sn)	Partial
GSE140819-MBC (Sc)→GSE140819-MBC (Sn)	Partial

https://doi.org/10.1371/journal.pcbi.1014223.t002

Evaluation metrics

To assess the performance of ScNucAdapt in cell type classification tasks, we evaluated its classification accuracy on datasets with known cell type annotations. The accuracy score quantifies the proportion of correctly predicted cell types among all predictions, providing a straightforward yet informative measure of model performance. Formally, accuracy is defined as shown in Eq. (14), where y_i denotes the true label of the i-th sample, $\hat{y}_i$ represents the predicted class label for the same sample, and n is the total number of samples in the dataset. The indicator function $1(y_i = \hat{y}_i)$ returns 1 if the predicted label matches the true label and 0 otherwise.

$$Acc(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=0}^{n-1} 1(y_i = \hat{y}_i) \tag{14}$$

Moreover, we also included the Macro-F1 score to evaluate model performance. Unlike overall accuracy, which can be biased toward majority classes, the Macro-F1 score provides a balanced assessment by treating all classes equally.

First, for each class $c \in \mathcal{C}$ (where $\mathcal{C}$ is the set of all classes), we compute class-specific precision and recall based on the true positives (TP_c), false positives (FP_c), and false negatives (FN_c):

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{15}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{16}$$

Precision measures the proportion of cells predicted to be class c that are correctly assigned, reflecting the model's exactness. Recall measures the proportion of true class c cells that were successfully retrieved, reflecting the model's completeness. The $F1$ score for each class is then defined as the harmonic mean of precision and recall, balancing the trade-off between the two:

$$\text{F1}c = 2 \cdot \frac{\text{Precision}c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{17}$$

Finally, the Macro-F1 score is calculated as the arithmetic mean of these per-class F1 scores across all classes:

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{F1}_c \tag{18}$$

Experimental setup

The proposed method was evaluated against several existing classifiers that have been widely applied in single-cell transcriptomic analyses. Including SingleCellNet and ScMap. And also a popular domain adaptation method for the cross-batches cell type annotation method in scRNA-seq, ScAdapt [33]. All the comparison methods are tuned to achieve the best performance and are performed on a research server equipped with an NVIDIA GeForce RTX 4090 GPU.

For initial clustering, we applied Gaussian Mixture Models with diagonal covariance matrices set to 'diag' in scikit-learn, and C as the number of mixture components, which would be further adjusted.

In each experiment, the specific hyperparameters, including the width of the first hiddenlayer, the latent representations width, the early stopping, the trade-off hyperparameter, the learning rate, batchsize, and the epoch before performing

initial Gaussian mixture clustering, and the initial cluster number K, are shown in Supplementary S1 Table. Moreover, the optimizer Adam and the hyperparameter $\sigma = 5$ are fixed across datasets.

Given a set of counts $\{c_1, c_2, \ldots, c_n\}$ for n classes, the weight w_i for class i is computed as:

$$w_i = \frac{1}{c_i} \tag{19}$$

Then, the weights in cross-entropy loss are calculated as:

$$\hat{w}_i = \frac{\frac{1}{c_i}}{\sum_{j=1}^{n} \frac{1}{c_j}} \times n \tag{20}$$

Where c_i is the count for class i, n is the total number of classes, and $\hat{w}_i$ is the final normalized weight for class i. Note that we've set a slightly higher weight for monocyte and plasma cell in bladder immune scRNA-seq to snRNA-seq. And on CLL, we set a slightly higher weight for the T cell.

The NIW hyperparameters are fixed across experiments and follow the settings from PRAGA and DeepDPM, where we used $\kappa = 0.0001$, set m to be the data mean, ν to be $K+2$, K represents the initial clustering hyperparameter, and $\psi = I \times 0.005$, where I denotes the identity matrix.

Experimental results

Simulation experiments

We first generated simulated datasets using the R package Splatter [34], including both imbalanced and balanced datasets. Each dataset comprises two batches and five cell types. Following this, we performed two target-controlled experiments. In the first, we varied the batch effect strength by adjusting the batch.facScale parameter while keeping the number of cell types in the target dataset fixed at three. In the second experiment, we held the batch.facScale constant at one and varied the number of cell types present in the target data.

We provide UMAP visualizations shown in Fig 4 to illustrate batch effects and cell type distributions under three conditions: before correction, at the onset of merging, and after complete merging.

The results show that across all conditions and evaluation metrics, ScNucAdapt consistently achieves robust performance. Furthermore, the UMAP visualizations shown in Fig 4 on imbalanced simulated datasets with batch.facScale set to 1, which clearly demonstrates that ScNucAdapt effectively merges the correct cell types while exhibiting minimal negative transfer.

ScNucAdapt enables accurate cross-annotation of bladder and kidney cell types across scRNA-seq and snRNA-seq

This section presents the classification performance of bladder and kidney cell types across domains between scRNA-seq and snRNA-seq data.

The experimental results shown in Table 3 and in Supplementary S2 Table indicate that ScNucAdapt outperforms existing classifiers, which focus solely on scRNA-seq datasets. Moreover, ScAdapt. On the immune subset, which is a partial domain adaptation problem where scRNA-seq is the source data and snRNA-seq is the target data, ScNucAdapt achieves an accuracy of 91.05 and a macro-F1 Score 84.69, which performed better than ScAdapt's 90.24 (accuracy) and 82.37 (macro-f1), and outperformed SingleCellNet's 81.02 (accuracy) and 55.51 (macro-f1). On the Stromal datasets where scRNA-seq is the source data, and snRNA-seq is the target data, ScNucAdapt achieves an accuracy of 97.80 and 89.42 of macro-f1 under a closed set setting, performing better than ScAdapt's 96.95 (accuracy) and 80.00 (macro-f1),

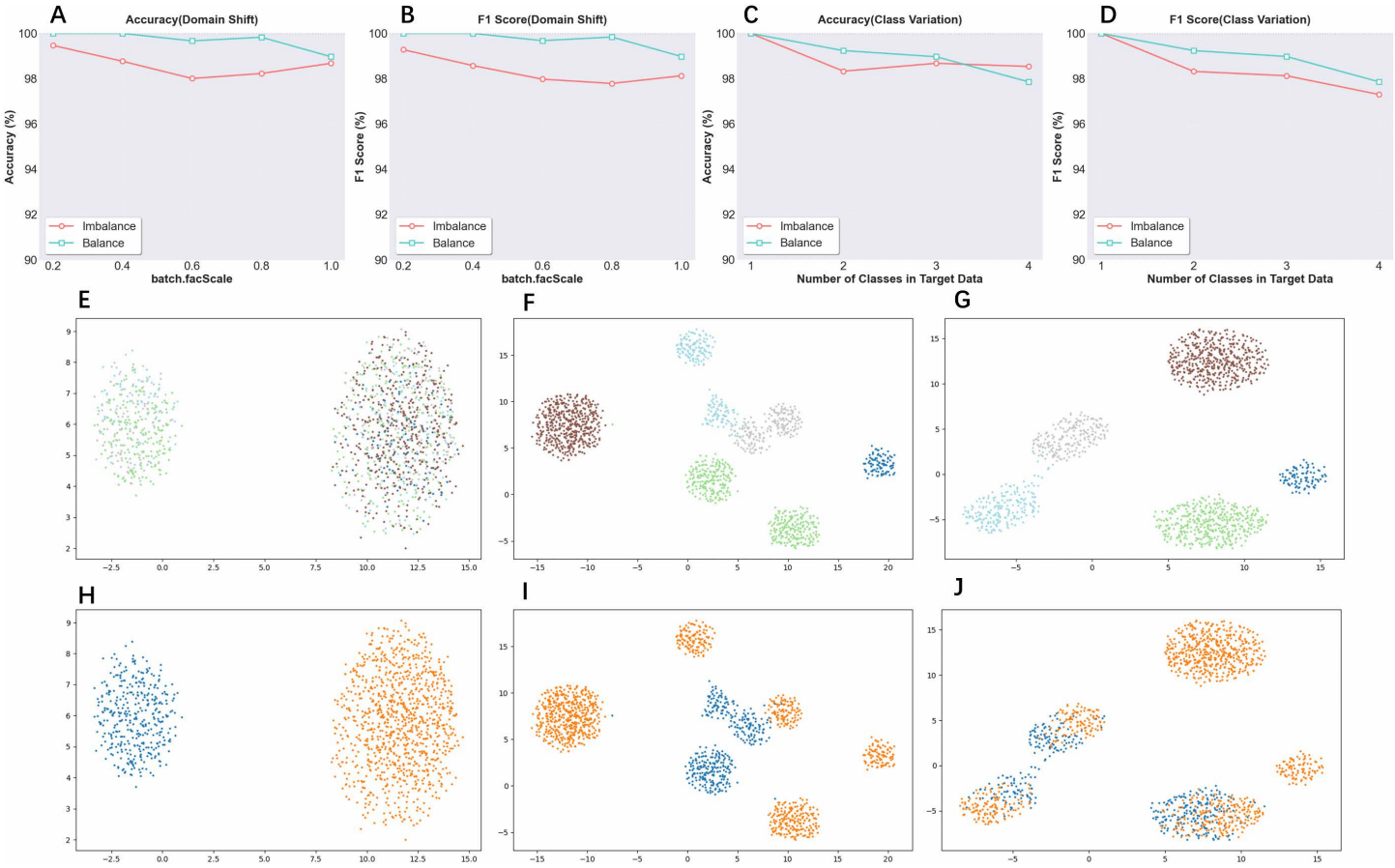

Fig 4. Simulation experiments using splatter. (A) Accuracy on imbalanced and balanced simulated datasets domain shift experiments **(B)** Macro f1-score on imbalanced and balanced simulated datasets domain shift experiments **(C)** Accuracy on imbalanced and balanced simulated datasets class variation experiments **(D)** Macro f1-score on imbalanced and balanced simulated datasets class variation experiments **(E)** Uncorrected simulated dataset of cell types colored (F) cell type representations before source classes and target clusters merging (G) cell type representations after source classes and target clusters merging **(H)** Uncorrected simulated dataset of batch colored (I) batch representations before source classes and target clusters merging (J) batch representations after source classes and target clusters merging.

https://doi.org/10.1371/journal.pcbi.1014223.g004

outperforming SingleCellNet's 91.92 (accuracy) and 63.82 (macro-f1). The same holds for the Stromal dataset, where snRNA-seq serves as the source data and scRNA-seq as the target data. ScNucAdapt achieves an accuracy score of 90.38 and 72.42 on macro-f1 score, outperforming ScAdapt's 89.98 (accuracy) and 71.61 (macro-f1), SingleCellNet's 86.61 (accuracy) and 66.96 (macro-f1), and ScMap's 87.02 (accuracy) and 63.72 (macro-f1). Interestingly, we found that domain adaptation methods often outperformed scRNA-seq classifiers. Visualization results using UMAP for the bladder tissue with three subsets are shown in Supplementary S3 Fig. While most scRNA-seq and snRNA-seq populations merged effectively, we observed a limitation in aligning stromal populations, where ScNucAdapt failed to distinguish between vein endothelial cells and general endothelial cells. This is attributable to the extremely limited training set for vein endothelial cells, which contained only four cells. This makes it a challenging scenario for any alignment method. Despite this, ScNucAdapt maintained the best overall performance among all compared methods in terms of both Accuracy and macro F1-score.

On the unpaired datasets between scRNA-seq and snRNA-seq, which is under a partial setting, ScNucAdapt achieves an accuracy of 87.23 and a macro-f1 of 81.5, outperforming other methods, including ScAdapt. The

Table 3. Classification accuracy of target datasets on bladder and kidney cell types.

Datasets	ScMap	SingleCellNet	ScAdapt	ScNucAdapt
GSE267964-Immune (Sc)→GSE267964-Immune (Sn)	75.06	81.02	90.24	**91.05**
GSE267964-Stromal (Sc)→GSE267964-Stromal (Sn)	79.58	91.92	96.95	**97.80**
GSE267964-Stromal (Sn)→GSE267964-Stromal (Sc)	87.02	86.61	89.98	**90.38**
GSE140989 (Sc)→GSE121862 (Sn)	86.04	70.58	84.01	**87.23**

https://doi.org/10.1371/journal.pcbi.1014223.t003

visualization results of UMAP are shown in Fig 5, where most scRNA-seq and snRNA-seq populations are well-merged across batches while maintaining clear separation by cell type. These results indicate that not only can ScNucAdapt handle distributional differences between scRNA-seq and snRNA-seq, but also under partial settings of unpaired scRNA-seq and snRNA-seq.

All the experimental results show that ScNucAdapt is a robust method in cross-domain annotation between scRNA-seq and snRNA-seq in bladder and kidney cell types.

ScNucAdapt supports reliable cross-annotation of fresh and frozen tumor cells between scRNA-seq and snRNA-seq

This section presents the classification performance of cell types in fresh and frozen tumors. As described in section 2.7, we chose two types of tumors, metastatic breast cancer and chronic lymphocytic leukemia.

The experimental results presented in Table 4 and Supplementary S2 Table show that, under a partial adaptation setting for metastatic breast cancer, where scRNA-seq serves as the source data and snRNA-seq as the target data, ScNucAdapt performed better than ScAdapt's 94.17 (accuracy) and 74.89 (macro-f1), achieving an accuracy of 95.39 and a macro-f1 score of 76.16. Both methods outperform other comparison methods.

Moreover, the cross-domain annotation from scRNA-seq to snRNA-seq on chronic lymphocytic leukemia, ScNucAdapt, achieved an accuracy of 98.39 and a macro-f1 of 94.47, which performed better than existing scRNA-seq cell type classifiers and ScAdapt. These results demonstrate that ScNucAdapt is effective under partial adaptation settings and can reliably handle cross-domain annotation tasks.

Fig 6 presents the UMAP visualization results, illustrating that cells from scRNA-seq and snRNA-seq are well-mixed after integration, yet remain distinctly separated according to their cell type identities.

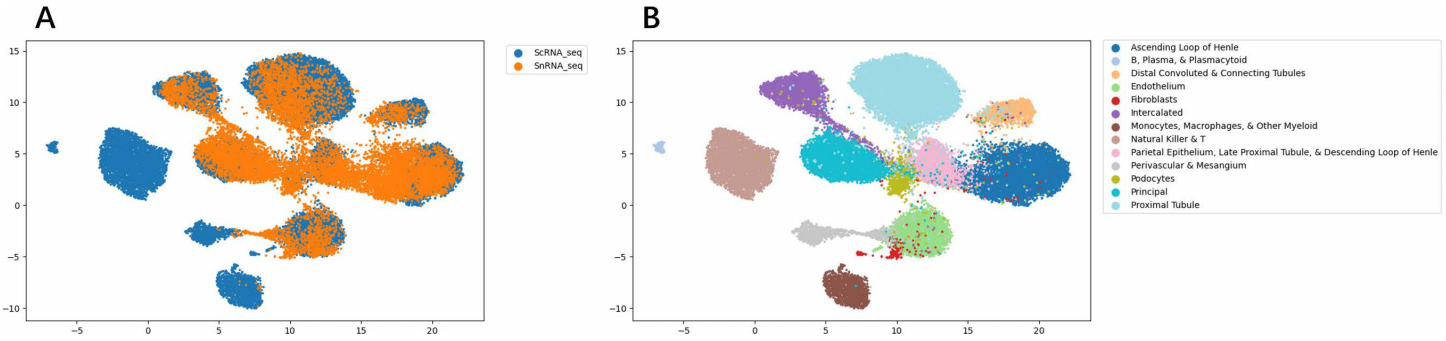

Fig 5. Visualization result of scRNA-seq and snRNA-seq representations. Using UMAP on kidney tissue (A) visualization result on scRNA-seq to snRNA batch representations (B) visualization result on scRNA-seq to snRNA cell type representations.

https://doi.org/10.1371/journal.pcbi.1014223.g005

Table 4. Classification accuracy of target datasets on frozen and fresh tumor cell types.

Datasets	ScMap	SingleCellNet	ScAdapt	ScNucAdapt
GSE140819-CLL (Sc)→ GSE140819-CLL (Sn)	97.64	93.07	96.99	**98.39**
GSE140819-MBC (Sc)→GSE140819-MBC (Sn)	84.88	64.82	94.17	**95.39**

https://doi.org/10.1371/journal.pcbi.1014223.t004

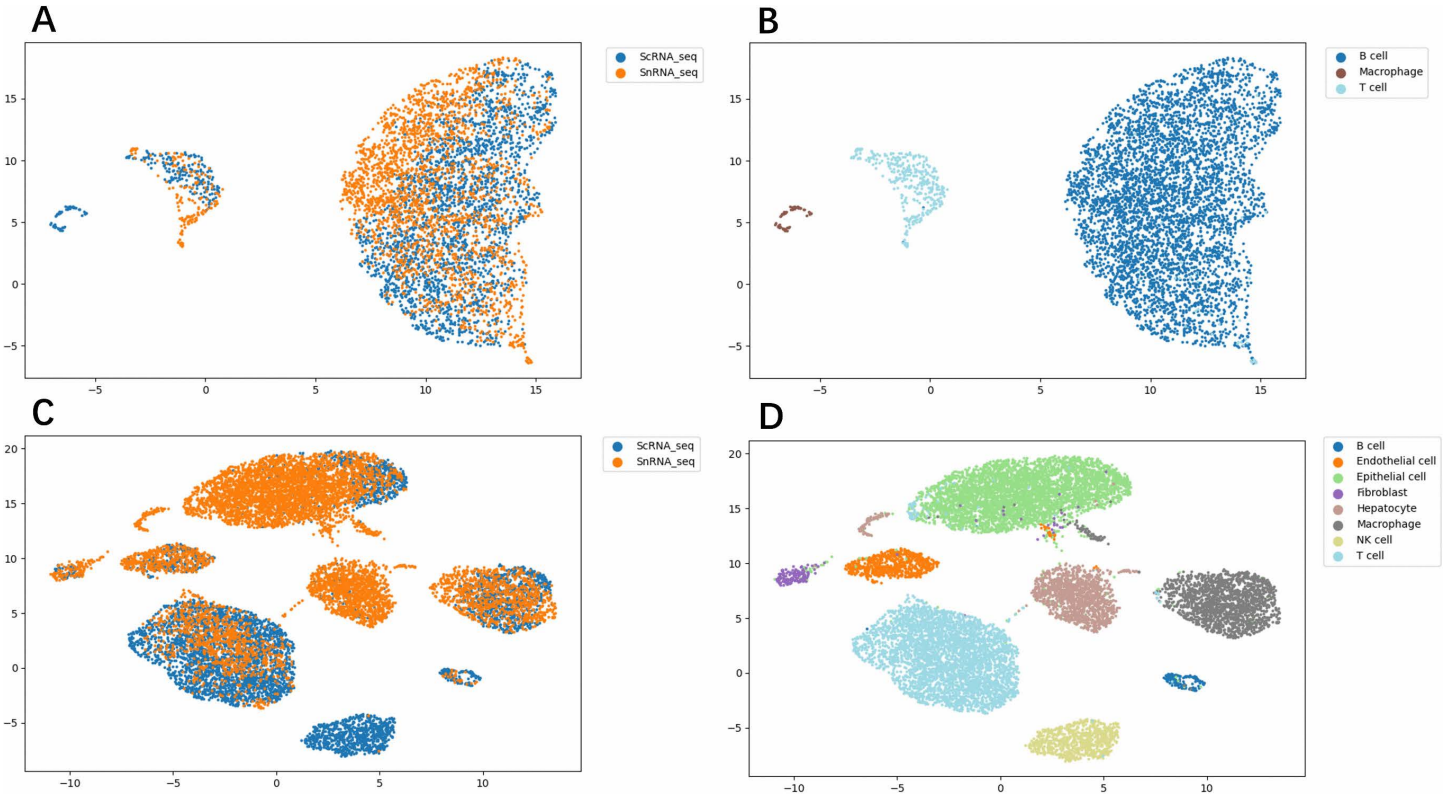

Fig 6. Visualization result of scRNA-seq and snRNA-seq representations. Using UMAP on frozen and fresh tumor tissue (A) visualization result on scRNA-seq to snRNA batch representations using CLL (B) visualization result on scRNA-seq to snRNA cell type representations using CLL (C) visualization result on scRNA-seq to snRNA batch representations using MBC (D) visualization result on scRNA-seq to snRNA cell type representations using MBC.

https://doi.org/10.1371/journal.pcbi.1014223.g006

ScNucAdapt enables cross-annotation of mouse cortical cell types across scRNA-seq and snRNA-seq

We have also included the cross-domain cell type annotation on mouse cortical cell types. The experimental settings are identical to those described in the previous sections.

The experimental results shown in Table 5 and supplementary S2 Table revealed that the cross-annotation from scRNA-seq to snRNA-seq, ScNucAdapt, achieves an accuracy score of 99.78, which performed better than SingleCellNet, ScMap, and scAdapt. On the cross-annotation from snRNA-seq to scRNA-seq, ScNucAdapt, ScAdapt, and SingleCellNet achieved an accuracy of 100.00, while ScMap achieved 99.13 (accuracy) and 97.99 (macro-f1).

As shown in Supplementary S4 Fig, the UMAP embeddings demonstrate successful integration, with cells from both scRNA-seq and snRNA-seq mixing effectively while maintaining clear separation by cell type.

Table 5. Classification accuracy of target datasets on mouse cortical cell types.

Datasets	ScMap	SingleCellNet	ScAdapt	ScNucAdapt
GSE123454 (Sc)→ GSE123454 (Sn)	98.48	99.56	99.56	**99.78**
GSE123454 (Sn)→GSE123454 (Sc)	99.13	**100.00**	**100.00**	**100.00**

https://doi.org/10.1371/journal.pcbi.1014223.t005

Ablation experiments

In this section, we conducted an ablation study to evaluate the contribution of each major component of the proposed ScNucAdapt framework. Two core modules were examined:

- The use of CS divergence to measure and minimize distributional discrepancies between the source subset and the target clusters, thereby aligning their feature distributions.
- The dynamic cluster selection mechanism, which identifies clusters within the target domain without requiring prior knowledge of their number.

We hypothesize that removing the CS divergence would weaken the model's ability to generalize across domains, while omitting the dynamic cluster selection would impair the model's capacity to adapt effectively to target data due to insufficient structural guidance.

To test these hypotheses, ablation experiments were performed on five cross-domain cell-type annotation tasks, including datasets on bladder cell types, kidney cell types, and tumor cell types. The results shown in Fig 7 reveal that excluding

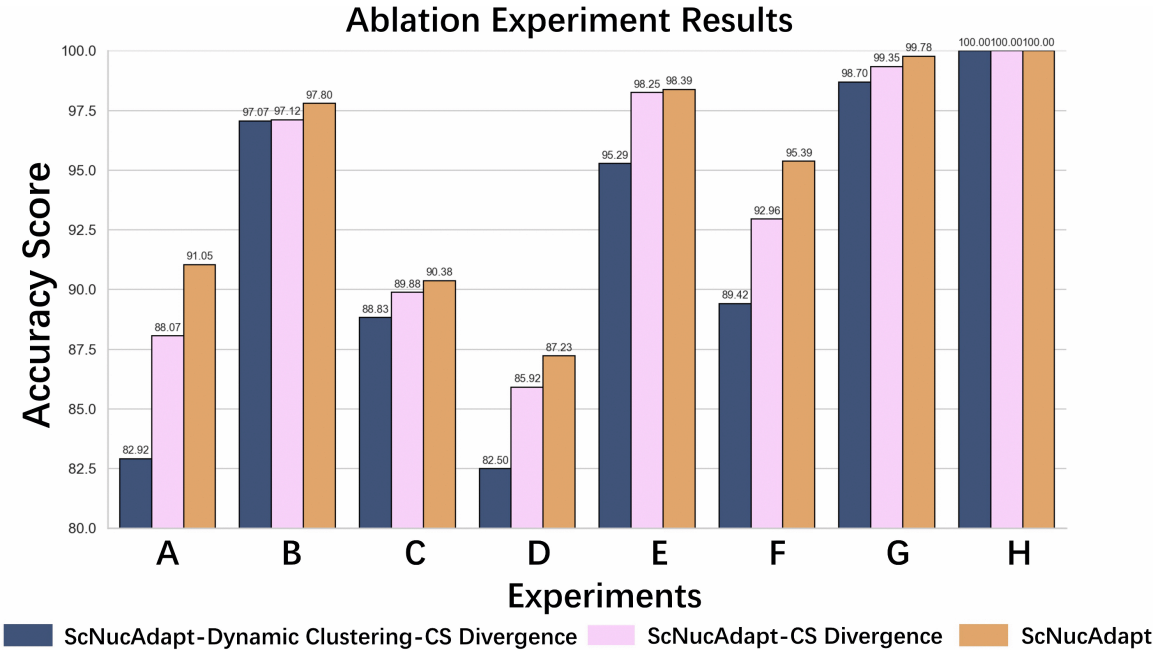

Fig 7. Ablation experiments conducted on 8 cross domain classification. (A) Ablation analysis on GSE267964-Immune (Sc)→GSE267964-Immune (Sn) **(B)** Ablation analysis on GSE267964-Stromal (Sc)→GSE267964-Stromal (Sn) **(C)** GSE267964-Stromal (Sn)→GSE267964-Stromal (Sc) **(D)** Ablation analysis on GSE140989 (Sc)→GSE121862 (Sn) **(E)** Ablation analysis on GSE140819-CLL (Sc)→ GSE140819-CLL (Sn) **(F)** Ablation analysis on Sensitivity analysis on GSE140819-MBC (Sc)→ GSE140819-MBC (Sn) **(G)** Ablation analysis on GSE123454 (Sc)→ GSE123454 (Sn) **(H)** Ablation analysis on GSE123454 (Sn)→ GSE123454 (Sc).

https://doi.org/10.1371/journal.pcbi.1014223.g007

either component substantially decreases generalization performance, indicating that both CS-based distributional alignment and dynamic cluster selection are critical for robust cross-domain annotation.

In particular, removing the dynamic cluster selection module, while retaining the CS divergence, still led to a noticeable decline in accuracy, underscoring the complementary role of adaptive clustering in enhancing ScNucAdapt's ability to generalize between scRNA-seq and snRNA-seq domains. These findings highlight the importance of both proposed components in mitigating modality-specific distributional differences and achieving stable cross-domain cell-type classification.

Sensitive analysis

In the previous section, we introduced the hyperparameter *C*, which is the initial cluster for the Gaussian mixture model. However, the prior knowledge of the real number of clusters in the target dataset is unknown. Our proposed method is capable of dynamically selecting the appropriate number of clusters after being given the hyperparameters. Therefore, we conducted a sensitivity analysis on the hyperparameter *C* to see whether there is extreme fluctuation in the performance when different hyperparameters *C* are given.

A total of six cross-domain cell type annotation experiments are included in the analysis. The experimental results shown in Fig 8 show that on most occasions, ScNucAdapt is insensitive to the hyperparameters. However, we noticed that there's a small fluctuation when conducting sensitive analysis on the immune bladder cell type from scRNA-seq to snRNA-seq, but the performance wasn't significantly degraded. The results suggest that ScNucAdapt can automatically adapt to diverse datasets without the need for extensive hyperparameter tuning, thereby improving its applicability in real-world cross-domain annotation between scRNA-seq and snRNA-seq.

Moreover, we conducted sensitivity analysis on the trade-off hyperparameter λ following the six cross-domain cell type annotation experiments. The experimental results shown in Supplementary S1 Fig indicate that ScNucAdapt is not

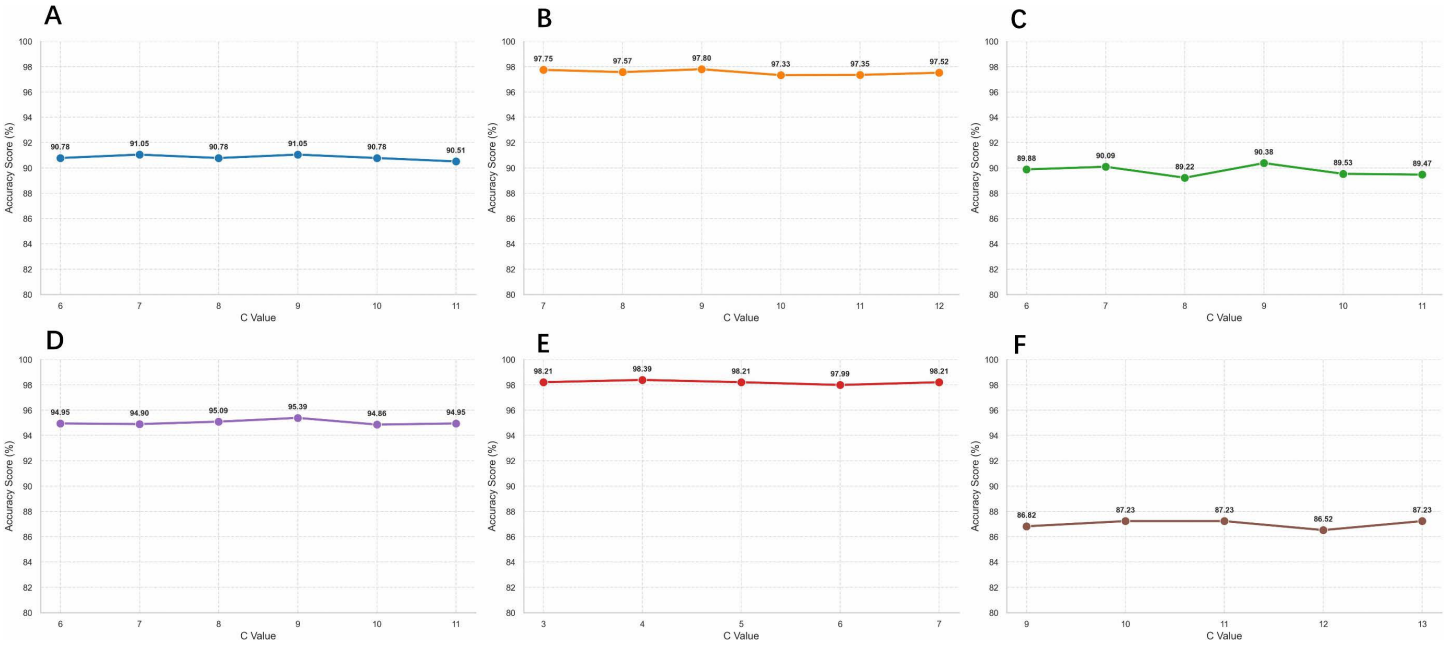

Fig 8. Hyperparameter Sensitive analysis on *C*. (A) Sensitivity analysis on GSE267964-Immune (Sc)→GSE267964-Immune (Sn) **(B)** Sensitivity analysis on GSE267964-Stromal (Sc)→GSE267964-Stromal (Sn) **(C)** Sensitivity analysis on GSE267964-Stromal (Sn)→GSE267964-Stromal (Sc) **(D)** Sensitivity analysis on GSE140819-MBC (Sc)→ GSE140819-MBC (Sn) **(E)** Sensitivity analysis on GSE140819-CLL (Sc)→ GSE140819-CLL (Sn) **(F)** Sensitive analysis GSE140989 (Sc)→GSE121862 (Sn).

https://doi.org/10.1371/journal.pcbi.1014223.g008

sensitive to the changes of the trade-off hyperparameter across a wide range of values. The model maintained stable performance for λ, demonstrating that the proposed method achieves consistent domain adaptation without requiring extensive hyperparameter tuning.

Runtime and peak GPU memory scalability tests

To assess the runtime and peak GPU memory of ScNucAdapt, we performed scaling experiments on simulated single-cell RNA-seq datasets generated with Splatter. We simulated four datasets with approximately 2,000, 5,000, 10,000, and 20,000 cells (10,000 genes, batch.facScale 0.8), which, after removing two cell types from the target batch, resulted in final sizes of 1,613, 4,013, 7,985, and 16,011 cells, respectively. The results shown in Supplementary S2 Fig show that interestingly, memory consumption scaled linearly from 0.065GB to 0.413GB across these dataset sizes, with peak memory remaining stable during encoder updates due to minibatch-based backpropagation. Runtime per epoch increased from 45.7 seconds to 1,436.4 seconds, with the computational bottleneck being the GMM clustering and split-merge operations performed on the full dataset each epoch. Therefore, we could rerun the clustering and matching every n epochs to reduce runtime.

Discussion

In this paper, our study fills the research gap on the cross-domain annotation between scRNA-seq and snRNA-seq, and also addresses the distributional and cell composition differences between the two types of datasets. The experimental results show the robustness of ScNucAdapt on the cross-domain annotation between scRNA-seq and snRNA-seq, and further ablation experiments have proved the effectiveness of the proposed components in ScNucAdapt. Moreover, insensitive to the hyperparameters that need manual controls on the initial clusters.

While ScNucAdapt is proposed for the cross-domain annotation between scRNA-seq and snRNA-seq and shows robustness, several problems and questions could be addressed in future work.

First, label noises [35] that existed in the source datasets could degrade the performance by introducing unreliable supervision signals. These mislabeled samples may hinder domain alignment and reduce the accuracy of downstream annotation.

Second, an exciting direction for future research lies in novel cell type discovery across the target dataset [36]. Current cross-domain annotation frameworks rely heavily on existing cell-type labels and may overlook previously uncharacterized or rare cell populations that are domain-specific. Therefore, future work must integrate Open-Set Domain Adaptation [37] or universal domain adaptation frameworks [38], as these are specifically designed to handle both shared and private label sets, allowing models to flag target-domain-specific cells as “unknown” for further validation.

A more challenging but realistic direction involves scenarios where the gene sets differ substantially between scRNA-seq and snRNA-seq. This moves the problem into the realm of Heterogeneous Domain Adaptation [39], where the feature spaces themselves are mismatched. Future frameworks capable of projecting heterogeneous gene sets into a common latent space would vastly expand the flexibility and applicability of cross-domain annotation.

Moreover, due to high sparsity and high-dimensional spaces occurring in both scRNA-seq and snRNA-seq, ScNucAdapt tends to overfit, although ScNucAdapt tends to generalize well compared to existing methods on cross-domain annotation between scRNA-seq, there’s still room for improvements on the performance. Therefore, future direction could focus on developing algorithms to prevent overfitting and achieve better generalization on target datasets [40].

Finally, imbalanced cell type distributions within each domain could hinder generalization. Overrepresented cell types may dominate the training process, causing the model to underperform on rare populations. Future methods could focus on addressing these within-domain imbalances to improve robustness and cross-domain performance [41].

Conclusion

In this study, we introduced ScNucAdapt, a novel cross-domain annotation framework designed specifically for transferring cell type labels between paired or unpaired scRNA-seq and snRNA-seq datasets. To the best of our knowledge, this

is the first method to address the unique challenges of cross-annotation across these two sequencing protocols. ScNucAdapt tackles both distributional differences and label space mismatches between source and target domains through three key components: a shared encoder that projects both datasets into a common latent space, a dynamic clustering mechanism that adaptively identifies the unknown number of cell types in the target data via split-merge operations, and a Cauchy-Schwarz divergence-based matching strategy that aligns source classes with target clusters while minimizing negative transfer from non-shared cell types.

Extensive experiments on eight cross-domain annotation tasks spanning bladder, kidney, tumor, and mouse cortical tissues demonstrated that ScNucAdapt consistently outperforms existing methods, including scRNA-seq classifiers and domain adaptation baselines, in both accuracy and macro F1-score. The framework proves effective under both closed-set and partial-set scenarios, maintaining robust performance even when target label spaces are subsets of the source. Ablation studies confirmed the necessity of each proposed component, while sensitivity analyses showed that ScNucAdapt is robust to hyperparameter choices, requiring minimal tuning in practice.

Our scalability analysis on simulated datasets confirmed that ScNucAdapt exhibits linear memory scaling and manageable runtime for datasets up to 16,000 cells, with the primary computational bottleneck being the GMM clustering and split-merge operations performed each epoch. For larger datasets, we recommend reducing the frequency of these operations or exploring approximate clustering variants.

Despite these strengths, several promising directions remain for future work. These include handling label noise in source annotations, extending the framework to discover novel cell types in target domains through open-set or universal domain adaptation, addressing heterogeneous feature spaces where gene sets differ substantially between datasets, mitigating overfitting in high-dimensional sparse spaces, and developing strategies to better handle imbalanced cell type distributions. Addressing these challenges will further enhance the applicability and robustness of cross-domain annotation methods in real-world single-cell and single-nucleus studies.

In summary, ScNucAdapt provides a powerful and flexible solution for integrating and annotating scRNA-seq and snRNA-seq data, enabling more consistent and reliable interpretation of cellular identities across experimental protocols and tissue conditions.

Supporting information

S1 Fig. Sensitive analysis on trade-off hyperparameter λ.
(TIF)

S2 Fig. Computational runtime and Peak GPU memory with different input sizes.
(TIF)

S3 Fig. Visualization result of scRNA-seq and snRNA-seq representations. Using UMAP on bladder tissue (A) visualization result on scRNA-seq to snRNA batch representations using immune subset (B) visualization result on scRNA-seq to snRNA cell type representations using immune subset (C) visualization result on scRNA-seq to snRNA batch representations using stromal subset (D) visualization result on scRNA-seq to snRNA cell type representations using stromal subset (E) visualization result on snRNA-seq to scRNA batch representations using stromal subset (F) visualization result on snRNA-seq to scRNA cell type representations using stromal subset.
(TIF)

S4 Fig. Visualization result of scRNA-seq and snRNA-seq representations. Using UMAP on mouse cortical tissue (A) visualization result on scRNA-seq to snRNA (B) visualization result on scRNA-seq to snRNA (C) visualization result on snRNA-seq to scRNA batch representations (D) visualization result on snRNA-seq to scRNA cell type representations.
(TIF)

S5 Fig. Pseudocode of ScNucAdapt.
(TIF)

S1 Table. Hyperparameter details, including the width of the first hiddenlayer, the latent representations' width, the early stopping, the trade-off hyperparameter, the learning rate, batchsize, the epoch before performing initial Gaussian mixture clustering, and the initial cluster number K.
(DOCX)

S2 Table. Macro-f1 score for each experiment.
(DOCX)

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62131004 to QZ, Grant No.62531002 to YW, Grant No.62306051 to WL, Grant No.62481540175 to WL, Grant No.62276035 XFC), the Taishan Scholars Foundation of Shandong Province (Grant No.tsqn202507225 to WL), and the Natural Science Foundation of Chongqing (Grant No.CSTB2025NSCQ-GPX0857 to WL). The Fundamental Research and the Scientific and the Technological Research Program of Chongqing Municipal Education Commission (Grant No.KJQN202300718 to WL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceptualization: Xiran Chen, Quan Zou, Weikai Li, Yansu Wang.

Data curation: Xiran Chen, Qinyu Cai.

Funding acquisition: Quan Zou, Xiaofeng Chen, Weikai Li, Yansu Wang.

Investigation: Xiran Chen.

Methodology: Xiran Chen.

Project administration: Weikai Li, Yansu Wang.

Software: Xiran Chen.

Supervision: Quan Zou, Xiaofeng Chen, Weikai Li, Yansu Wang.

Validation: Xiran Chen, Quan Zou, Qinyu Cai.

Visualization: Xiran Chen.

Writing – original draft: Xiran Chen, Weikai Li, Yansu Wang.

Writing – review & editing: Quan Zou, Qinyu Cai, Xiaofeng Chen, Weikai Li, Yansu Wang.

References

1. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. Elife. 2017;6:e27041. https://doi.org/10.7554/eLife.27041 PMID: 29206104
2. Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. Comput Struct Biotechnol J. 2021;19:961–9. https://doi.org/10.1016/j.csbj.2021.01.015 PMID: 33613863
3. Li T, Wang Z, Liu Y, He S, Zou Q, Zhang Y. An overview of computational methods in single-cell transcriptomic cell type annotation. Brief Bioinform. 2025;26(3):bbaf207. https://doi.org/10.1093/bib/bbaf207 PMID: 40347979
4. Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species. Cell Syst. 2019;9(2):207–213.e2. https://doi.org/10.1016/j.cels.2019.06.004 PMID: 31377170
5. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15(5):359–62. https://doi.org/10.1038/nmeth.4644 PMID: 29608555

6. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020;38(6):737–46. https://doi.org/10.1038/s41587-020-0465-8 PMID: 32341560
7. Wu H, Kirita Y, Donnelly EL, Humphreys BD. Advantages of single-nucleus over single-Cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. J Am Soc Nephrol. 2019;30(1):23–32. https://doi.org/10.1681/ASN.2018090912 PMID: 30510133
8. Zhang C, Tan G, Zhang Y, Zhong X, Zhao Z, Peng Y, et al. Comprehensive analyses of brain cell communications based on multiple scRNA-seq and snRNA-seq datasets for revealing novel mechanism in neurodegenerative diseases. CNS Neurosci Ther. 2023;29(10):2775–86. https://doi.org/10.1111/cns.14280 PMID: 37269061
9. Heuston EF, Doumatey AP, Naz F, Islam S, Anderson S, Kirby MR, et al. Optimized methods for scRNA-seq and snRNA-seq of skeletal muscle stored in nucleic acid stabilizing preservative. Commun Biol. 2025;8(1):10. https://doi.org/10.1038/s42003-024-07445-2 PMID: 39755918
10. Slyper M, Porter CBM, Ashenberg O, Waldman J, Drokhlyansky E, Wakiro I, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat Med. 2020;26(5):792–802. https://doi.org/10.1038/s41591-020-0844-1 PMID: 32405060
11. Park S, Lee S-H, Han S-E, Kim BK, Hwang B. Paired snRNA-seq and scRNA-seq analysis of MASLD patients to identify early-stage markers for disease progression. Hepatol Commun. 2025;9(11):e0820. https://doi.org/10.1097/HC9.0000000000000820 PMID: 41056488
12. Quatredeniers M, Serafin AS, Benmerah A, Rausell A, Saunier S, Viau A. Meta-analysis of single-cell and single-nucleus transcriptomics reveals kidney cell type consensus signatures. Sci Data. 2023;10(1):361. https://doi.org/10.1038/s41597-023-02209-9 PMID: 37280226
13. Le H, Peng B, Uy J, Carrillo D, Zhang Y, Aevermann BD, et al. Machine learning for cell type classification from single nucleus RNA sequencing data. PLoS One. 2022;17(9):e0275070. https://doi.org/10.1371/journal.pone.0275070 PMID: 36149937
14. Tisch A, Madapoosi S, Blough S, Rosa J, Eddy S, Mariani L, et al. Identification of kidney cell types in scRNA-seq and snRNA-seq data using machine learning algorithms. Heliyon. 2024;10(19):e38567. https://doi.org/10.1016/j.heliyon.2024.e38567 PMID: 39403515
15. Andrews TS, Atif J, Liu JC, Perciani CT, Ma X-Z, Thoeni C, et al. Single-cell, single-nucleus, and spatial RNA sequencing of the human liver identifies cholangiocyte and mesenchymal heterogeneity. Hepatol Commun. 2022;6(4):821–40. https://doi.org/10.1002/hep4.1854 PMID: 34792289
16. Li W, Chen S. Partial domain adaptation without domain alignment. IEEE Trans Pattern Anal Mach Intell. 2023;45(7):8787–97. https://doi.org/10.1109/TPAMI.2022.3228937 PMID: 37015373
17. Zhu Y, Pei Y, Wang A, Xie B, Qian Z. A partial domain adaptation scheme based on weighted adversarial nets with improved CBAM for fault diagnosis of wind turbine gearbox. Eng Appl Artif Intell. 2023;125:106674. https://doi.org/10.1016/j.engappai.2023.106674
18. Liu W, Ni Z, Chen Q, Ni L. Attention-guided partial domain adaptation for automated pneumonia diagnosis from chest x-ray images. IEEE J Biomed Health Inform. 2023;27(12):5848–59. https://doi.org/10.1109/JBHI.2023.3313886 PMID: 37695960
19. Wang P, Qi Y, Pan G. Partial domain adaptation for stable neural decoding in disentangled latent subspaces. IEEE Trans Biomed Eng. 2026;73(1):78–89. https://doi.org/10.1109/TBME.2025.3577222 PMID: 40522806
20. Lin Y, Wu T-Y, Wan S, Yang JYH, Wong WH, Wang YXR. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40(5):703–10. https://doi.org/10.1038/s41587-021-01161-6 PMID: 35058621
21. Yan X, Zheng R, Chen J, Li M. scNCL: transferring labels from scRNA-seq to scATAC-seq data with neighborhood contrastive regularization. Bioinformatics. 2023;39(8):btad505. https://doi.org/10.1093/bioinformatics/btad505 PMID: 37584660
22. Zhao B, Song K, Wei D-Q, Xiong Y, Ding J. scCobra allows contrastive cell embedding learning with domain adaptation for single cell data integration and harmonization. Commun Biol. 2025;8(1):233. https://doi.org/10.1038/s42003-025-07692-x PMID: 39948393
23. Liu Y, Pei W, Chen L, Xia Y, Yan H, Hu X. scCorrect: Cross-modality label transfer from scRNA-seq to scATAC-seq using domain adaptation. Anal Biochem. 2025;702:115847. https://doi.org/10.1016/j.ab.2025.115847 PMID: 40154828
24. Ronen M, Finder SE, Freifeld O. Deepdpm: Deep clustering with an unknown number of clusters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. pp. 9861–70.
25. Huang X, Ma Z, Meng D, Liu Y, Ruan S, Sun Q, et al. PRAGA: prototype-aware graph adaptive aggregation for spatial multi-modal omics analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39. 2025. pp. 326–33.
26. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97–109. https://doi.org/10.1093/biomet/57.1.97
27. Yin W, Yu S, Lin Y, Liu J, Sonke JJ, Gavves S. Domain Adaptation with Cauchy-Schwarz Divergence. In: The 40th Conference on Uncertainty in Artificial Intelligence. 2024. Available from: https://openreview.net/forum?id=62m7yvkGIY
28. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0 PMID: 29409532
29. Santo B, Fink EE, Krylova AE, Lin Y-C, Eltemamy M, Wee A, et al. Exploring the utility of snRNA-seq in profiling human bladder tissue: A comprehensive comparison with scRNA-seq. iScience. 2025;28(1):111628. https://doi.org/10.1016/j.isci.2024.111628
30. Menon R, Otto EA, Hoover P, Eddy S, Mariani L, Godfrey B, et al. Single cell transcriptomics identifies focal segmental glomerulosclerosis remission endothelial biomarker. JCI Insight. 2020;5(6):e133267. https://doi.org/10.1172/jci.insight.133267 PMID: 32107344
31. Lake BB, Chen S, Hoshi M, Plongthongkum N, Salamon D, Knoten A, et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. Nat Commun. 2019;10(1):2832. https://doi.org/10.1038/s41467-019-10861-2 PMID: 31249312

32. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One. 2018;13(12):e0209648. https://doi.org/10.1371/journal.pone.0209648 PMID: 30586455
33. Zhou X, Chai H, Zeng Y, Zhao H, Yang Y. scAdapt: virtual adversarial domain adaptation network for single cell RNA-seq data classification across platforms and species. Brief Bioinform. 2021;22(6):bbab281. https://doi.org/10.1093/bib/bbab281 PMID: 34308480
34. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017;18(1):174. https://doi.org/10.1186/s13059-017-1305-0 PMID: 28899397
35. Chen X, Lin S, Chen X, Li W, Li Y. Timestamp calibration for time-series single cell RNA-seq expression data. J Mol Biol. 2025;437(9):169021. https://doi.org/10.1016/j.jmb.2025.169021 PMID: 40010431
36. Shi Y, Ma Y, Chen X, Gao J. scADCA: An Anomaly Detection-Based scRNA-seq Dataset Cell Type Annotation Method for Identifying Novel Cells. Curr Bioinform. 2025;20(10):904–17. https://doi.org/10.2174/0115748936334071240903064630
37. Busto PP, Gall J. Open Set Domain Adaptation. In: Proceedings of the IEEE international conference on computer vision. 2017. pp. 754–63.
38. You K, Long M, Cao Z, Wang J, Jordan MI. Universal Domain Adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 2715–24.
39. Liu F, Zhang G, Lu J. Heterogeneous domain adaptation: an unsupervised approach. IEEE Trans Neural Netw Learn Syst. 2020;31(12):5588–602. https://doi.org/10.1109/TNNLS.2020.2973293 PMID: 32149697
40. He C, Li X, Xia Y, Tang J, Yang J, Ye Z. Addressing the overfitting in partial domain adaptation with self-training and contrastive learning. IEEE Trans Circuits Syst Video Technol. 2024;34(3):1532–45. https://doi.org/10.1109/tcsvt.2023.3296617
41. Wang Y, Chen Q, Liu Y, Li W, Chen S. TITok: a solution for bi-imbalanced unsupervised domain adaptation. Neural Netw. 2023;164:81–90. https://doi.org/10.1016/j.neunet.2023.04.027 PMID: 37148610