IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 March 2024, accepted 22 April 2024, date of publication 26 April 2024, date of current version 3 May 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3394038*

RESEARCH ARTICLE

# Robust scRNA-seq Cell Types Identification by Self-Guided Deep Clustering Network

**YISONG WANG AND MINGZHI WANG**
College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China
Corresponding author: Mingzhi Wang (wmz@nefu.edu.cn)

**ABSTRACT** The emergence of single-cell RNA sequencing (scRNA-seq) has brought to light the critical need for scrutinizing transcriptomes at the individual cellular level with unparalleled precision. A pivotal aspect of scRNA-seq data analysis involves cell identification, commonly accomplished through diverse clustering methodologies. However, scRNA-seq datasets frequently encounter missing values due to technical limitations, posing a significant challenge that can compromise the accuracy of clustering outcomes. In response, we present a novel approach that seamlessly integrates missing value estimation with scRNA-seq clustering. Our method harnesses the power of an imputation autoencoder network to predict missing values, coupled with the deployment of a deep clustering network for efficient cell categorization. To mitigate the risk of deep clustering networks converging towards suboptimal local minima, we have devised a self-guided learning strategy. This approach exploits shared parameters between the imputation and clustering networks, fostering a symbiotic relationship that enhances overall performance. Through rigorous empirical evaluations, we substantiate the effectiveness of our methodology, demonstrating its comparability to, or surpassing, several established single-cell clustering techniques. Furthermore, our analysis of cellular trajectories underscores the proficiency of the proposed method in accurately deducing cellular trajectories by leveraging the clustering results to discern biologically meaningful cell types.

**INDEX TERMS** Single cell clustering, data imputation, subspace clustering.

## I. INTRODUCTION

The advancement of single-cell RNA sequencing (scRNA-seq) has revolutionized transcriptomic studies, enabling large-scale, high-throughput analysis with unprecedented single-cell accuracy. This technique has become crucial in biological research for dissecting gene expression at the individual cell level, thus revealing key insights into both known and novel cellular processes. A critical component of scRNA-seq analysis is cell identification, a process vital for distinguishing and characterizing cells in diverse populations. This is efficiently achieved through clustering methods, which group cells by similar gene expression patterns, facilitating the identification of distinct subpopulations without the need for pre-established cell type definitions.

For this purpose, several single-cell clustering methods have been developed to enhance the analysis of scRNA-seq data. For instance, SNN-Cliq [2] leverages the shared nearest neighbors (SNN) concept to define similarities between cells and maps them using a quasi-clique based clustering algorithm. Pcareduce [3] focuses on understanding the link between dimensionally reduced data through PCA and the subsequent identification of cell clusters. Another notable method, SC3 [4], adopts a multi-dimensional approach to clustering by integrating various distance metrics. It begins by calculating distances between cells using a range of metrics to form a comprehensive distance matrix, followed by employing k-means clustering to group cells according to this matrix. The advancement of deep learning has further enriched scRNA-seq data analysis, leading to the development of innovative single-cell clustering techniques. DISCERN [5], for example, uses reference datasets to accurately reconstruct missing gene expression values with a deep generative model, demonstrating enhanced clustering performance. ScDSSC [6] integrates noise reduction and dimensionality reduction within a deep sparse subspace

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

clustering framework. This method learns feature representations and performs clustering simultaneously by explicitly modeling scRNA-seq data generation. ScGPCL [7] is a graph-based prototype contrast learning method that thoroughly exploits cell-cell relationships. It applies graph neural networks to a bipartite cell-gene graph, significantly improving clustering accuracy and cell type identification [8], [9], [10], [11].

However, scRNA-seq data often contains missing values due to technical limitations, including inefficient RNA capture during sequencing, PCR amplification bias, and variable sequencing depth. Missing data poses a critical challenge for clustering analysis of scRNA-seq data. Since missing values do not reflect true underlying expression levels, they can obstruct accurate identification of cell types when applying clustering algorithms. Therefore, developing computational approaches to address the issue of missing data is crucial for robust analysis and biological interpretation of scRNA-seq datasets. Several computational approaches have been developed to address the issue of missing data in scRNA-seq, with each method based on distinct principles and models [12], [13], [14], [15], [16], [17], [18], [19], [20]. Notable methods include MAGIC [12], which constructs a Markov transition matrix to facilitate data smoothing based on cell-to-cell interactions. ScImpute [13] employs a LASSO regression model to iteratively impute values for individual cells. SAVER [14] utilizes a Bayesian approach with different prior probability functions, while DrImpute [15] aggregates values imputed from multiple cluster priors or distance matrices. More recent methods like scFEA [19] incorporate probabilistic models accounting for flux balance constraints and graph neural networks for optimization. The DCA method [20] trains an autoencoder to model gene expression distributions using a zero-inflated negative binomial prior.

In contrast to many existing methods that treat scRNA-seq data imputation and single-cell clustering as separate entities [21], our research focuses on their integration, combining the imputation of missing values with clustering in scRNA-seq data analysis [22]. Our proposed solution introduces an imputation autoencoder network specifically designed to effectively address missing values in single-cell data. This model, enhanced with a variant reconstruction loss, adeptly recovers missing values by capitalizing on shared information across cells. Following the imputation phase, we utilize a deep subspace clustering network to facilitate precise cell identification, based on the imputed dataset. Aware of the inherent challenges in deep subspace clustering, such as the tendency to converge on suboptimal local solutions, we have implemented a self-guided learning strategy. This strategy involves the computation of a guiding graph to steer the training process, leading to more refined clustering results. A key innovation of our approach is the parameter sharing between the imputation and clustering networks. This synergy promotes mutual reinforcement, where improvements in imputation directly enhance clustering performance and vice versa. This collaborative learning framework not only refines each task individually but also demonstrates the potential for significantly improved overall performance in both imputation and clustering of scRNA-seq data.

The main contributions in this paper are summarized as follows:

1) We introduce a novel imputation autoencoder network, specifically designed to address the missing values in scRNA-seq data. This network, equipped with a variant reconstruction loss, effectively recovers missing values, leveraging common information across individual cell samples.
2) To overcome the limitations of deep subspace clustering networks, such as their susceptibility to suboptimal local solutions, our method includes a self-guided learning strategy. This strategy enhances the training process and leads to more accurate clustering outcomes.
3) Our work innovatively integrates these two analytical processes. This integration ensures a more cohesive and efficient analysis of scRNA-seq data.

The subsequent sections of this paper are organized as follows: In Section II, we review closely related work in conjunction with our proposed method. Section III outlines the details of the proposed method. In Section IV, we present experimental results to demonstrate the effectiveness of our approach. Finally, we conclude and outline future directions for our work.

## II. RELATED WORK

The proposed model is designed to seamlessly integrate two pivotal tasks in scRNA-seq data analysis: missing data imputation and single-cell type clustering. In this section, we explore these two critical aspects related to our proposed model.

### A. SINGLE CELL DATA IMPUTATION

A prevalent issue in RNA-Seq data is the sparsity of expression matrices, often characterized by a large number of zeros. Many of these zero or near-zero values are artificially introduced by technical deficiencies, such as insufficient mRNA molecules, low capture rates, sequencing depth, or other technical factors, collectively known as drop-out. Several methods have been proposed to address the challenges posed by excessive zeros in scRNA-seq data. For example, ScImpute [13] Learns the dropout probability for each gene in each cell and extrapolates dropout values by considering information from other similar cells selected based on genes unlikely to be affected by dropout events. DrImpute [15] Inferences dropouts by averaging the expression values of similar cells defined by clustering. DCA [20] Utilizes a deep count autoencoder network to de-noise and enhance scRNA-seq datasets by learning the count distribution, overdispersion, and sparsity of the data.

While these approaches offer effective solutions, our proposed model takes a unique stance by integrating missing data imputation seamlessly with single-cell type clustering.

### B. CELL TYPES CLUSTERING

The primary objective of deep clustering is to group samples into distinct clusters without relying on explicit ground truth labels. While clustering are commonly employed in various tasks. Accurate identification and categorization of distinct cell types are imperative for unraveling the biological insights embedded in scRNA-seq data. For this purpose, various single cell clustering methods have been proposed [23], [24], [25], [26], [27], [28], [29], [30]. For example, to address the challenges posed by high-dimensional scRNA-seq data, early approaches focused on dimensionality reduction before applying classical clustering methods. For instance, SC3 [4] facilitates the quantitative characterization of cell types in single-cell RNA-seq by leveraging transcriptome features. It is a user-friendly unsupervised clustering tool that amalgamates multiple clustering solutions consistently, yielding high-precision and robust clustering results. SIMLR [1] measures cell similarities using multiple cores and applies k-means for clustering. Zheng et al. [31] proposed SinNLRR by adding low-rank and non-negative structure to the similar matrix based on Spectral clustering. Wang et al. [1] proposed SIMLR algorithm based on Spectral clustering, learned inter-cell distance measurement from gene expression data through multi-core learning and constructed similarity matrix, which not only improved the clustering effect but also effectively adapted to multiple downstream steps. As single-cell experiments continue to scale up in cell numbers but face issues with data quality, there is a growing need for imputation methods that are not only fast but highly scalable [32], [33], [34], [35], [36], [37].

The success of deep clustering methods has led to the development of recent deep models that have shown promising results. For instance, DESC [38] is an unsupervised deep embedding algorithm designed to cluster data from single-cell RNA sequencing. It iteratively optimizes the clustering objective function and effectively eliminates batch effects. Comprehensive evaluations have shown that DESC achieves a proper balance between cluster accuracy and stability, possesses a small memory footprint, and can eliminate batch effects without relying on batch information. scTAG [39] learns cell-cell topological representation and identifies cell clusters using a depth map convolutional network. Notably, SCVIS [40] employs a generative approach to learn cell representation through t-SNE regularization, preserving local structure.

## III. MATERIALS AND METHOD

In this section, we introduce the proposed approach that integrates missing value estimation with scRNAseq clustering. The proposed approach consists of two integral modules: the missing data imputation module and the deep subspace clustering module. Figure. 1 provides a comprehensive visual representation of the proposed method, offering an overview of the interaction and workflow between these two modules. This illustration serves to elucidate the key components and the seamless integration of the imputation and clustering processes within the proposed framework. In subsequent sections, we explain the proposed method in detail including the data preparation, each module and training strategy.

### A. IMPUTATION MODULE

Consider a dataset comprising single-cell information denoted as $X \in R^{n\times d}$. Notably, this dataset may exhibit missing entries $X_{i,j} = NaN$. where the absence of data occurs between the $i$-th and $j$-th samples. To identify these missing entries, a missing index matrix, is defined to highlight the specific locations where data is absent:

$$S_{i,j} = \begin{cases} 1 & If\ i-th\ and\ j-th\ is\ missing \\ 0 & Otherwise \end{cases} \tag{1}$$

To address the challenge of imputing missing values, we employ an imputation autoencoder framework, comprising an encoder and a decoder. In this autoencoder architecture, the encoder initially transforms the input data matrix into a latent space representation. Subsequently, the decoder reconstructs the latent space back to the original input space, formulated mathematically as:

$$\hat{X} = \mathrm{D(E(X))} \tag{2}$$

where $\hat{X}$ represents the reconstructed matrix, and $D(\cdot)$ and $E(\cdot)$ denote the functions of the encoder and decoder, respectively. The objective is to utilize this framework for data completion by employing an imputation autoencoder loss. The imputation autoencoder loss represents a specialized autoencoder variant, notable for processing input data where a segment is intentionally ''masked'' or hidden. This model's decoder is tasked with reconstructing the full dataset, necessitating the filling or recovery of the masked section. Excelling in managing datasets with missing values, these models use the available (unmasked) data to refine the prediction and imputation of the absent values. The integration of a mask index matrix, identifying the missing data portions, is crucial. It works in tandem with the tailored loss function, which is designed to assess the model's accuracy in predicting and reconstructing these missing segments. Therefore, in conjunction with the mask index matrix, the associated loss function can be defined as:

$$\mathcal{L}_{\mathrm{maskAE}} = \left\| \mathrm{S} \odot X - \mathrm{S} \odot \hat{\mathrm{X}} \right\|_{\mathrm{F}}^{2} \tag{3}$$

The reconstructed $\hat{\mathrm{X}}$ matrix encompasses both imputed data and reconstructed data derived from $X$. Subsequently, the completed scRNA-seq data can be obtained through the following expression:

$$\mathrm{Y} = (1 - \mathrm{S}) \odot \mathrm{X} + \mathrm{S} \odot \hat{\mathrm{X}} \tag{4}$$

In this equation, Y represents the completed scRNA-seq data, where the element-wise product with the complement of the mask matrix $(1 - \mathrm{S})$ is applied to the original data matrix $X$, and the element-wise product with the mask matrix S is applied to the reconstructed matrix $\hat{\mathrm{X}}$. This formulation

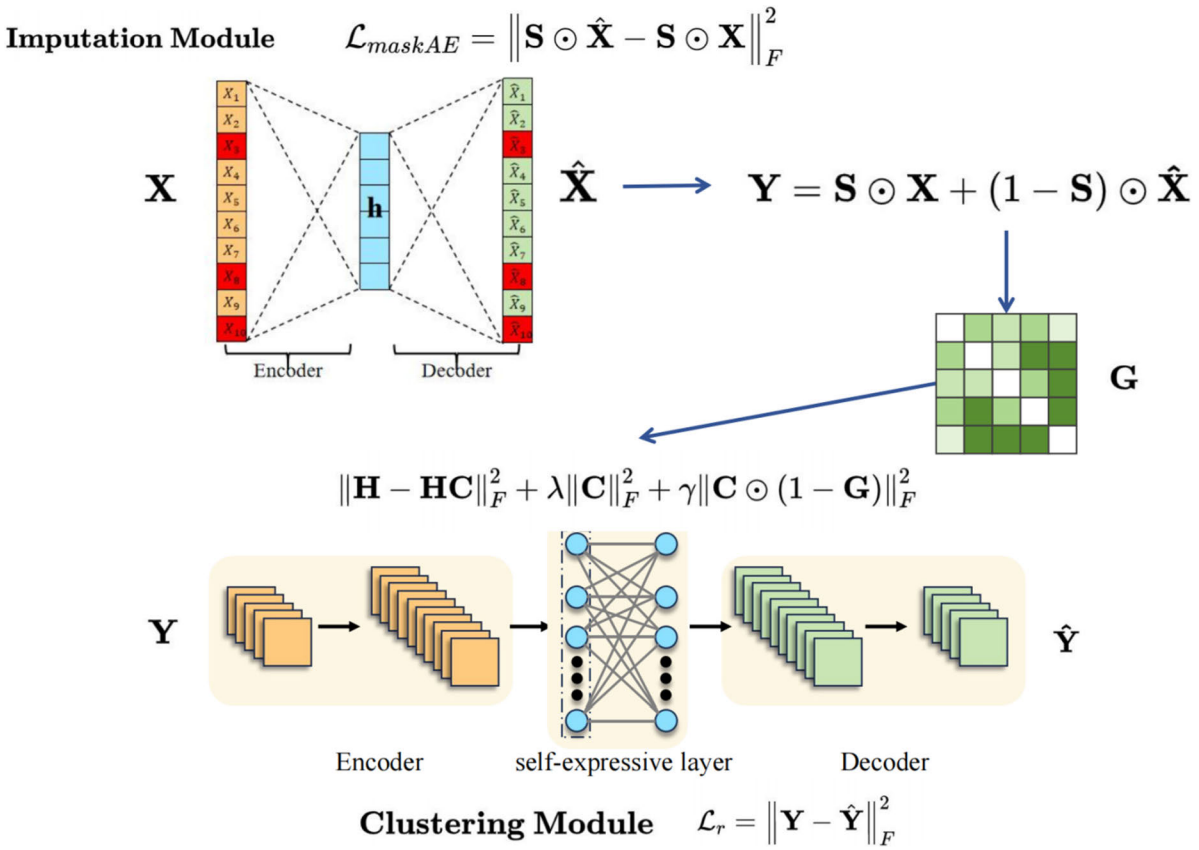


**FIGURE 1.** Illustrates the schematic overview of the proposed methodology. Comprising two core modules, namely the missing data imputation module and the deep subspace clustering module, this model encapsulates the key components of our approach.

effectively integrates the observed and imputed values to generate a comprehensive and enhanced scRNA-seq dataset.

### B. DEEP CLUSTERING MODULE

Once the completed scRNA-seq data $Y$ is obtained, the objective is to identify the associated cell types through a deep clustering approach. The chosen method is the Deep Clustering Network, a deep neural network architecture tailored for unsupervised subspace clustering. The network architecture is constructed around a deep autoencoder, which non-linearly maps the input data to an underlying space. Notably, a distinctive feature of this architecture is the introduction of a novel self-expression layer situated between the encoder and decoder. This self-expression layer emulates the ''self-expression'' properties found effective in traditional subspace clustering. Particularly, the expression layer facilitates a straightforward yet effective method to learn the pairwise affinity between all data points through a standard propagation process. The neural network-based approach, being inherently nonlinear, is adept at clustering data points with intricate nonlinear structures. The incorporation of self-representation layers in this module enhances their capability to handle nonlinearly separable data, making them particularly suited for discerning complex relationships in scRNA-seq datasets. In detail, Similar to the autoencoder, the formulation of encoder and decoder can be defined as:

$$\mathrm{H} = \mathrm{E}(\mathrm{Y}) \tag{5}$$

$$\hat{\mathrm{Y}} = \mathrm{D}(\mathrm{H}) \tag{6}$$

where, $H$ represents the latent space obtained through the encoder, and $\hat{Y}$ represents the reconstructed scRNA-seq data obtained through the decoder. The related reconstruction loss of the clustering network is defined as follows:

$$\mathcal{L}_{\mathrm{dscRec}} = \left\| \hat{\mathrm{Y}} - \mathrm{Y} \right\|_{\mathrm{F}}^2 \tag{7}$$

Moreover, a fundamental component of this clustering module is the self-expressive layer, dedicated to learning the self-expression matrix for subsequent clustering tasks:

$$\mathcal{L}_{\mathrm{dscSelf}} = \|\mathrm{H} - \mathrm{HC}\|_{\mathrm{F}}^2 + \lambda \left\| \mathrm{C} \right\|_{\mathrm{F}}^2 \tag{8}$$

In this equation, $C$ represents the self-expression matrix, and $\lambda$ is a regularization parameter. This loss function encapsulates the self-expression properties critical for discerning underlying subspaces and facilitating effective unsupervised clustering. As previously discussed, deep subspace clustering networks are prone to being trapped in suboptimal local solutions. To address this challenge, we introduce a self-guided learning strategy aimed at calculating a graph that guides the

training process, thereby leading to more accurate clustering outcomes. In particular, we incorporate a weight graph regularization to guide the self-expression learning process. The weight matrix, representing an affinity matrix, can be estimated from the completed $Y$. Consequently, the corresponding loss function for self-expression learning is defined as:

$$\mathcal{L}_{\text{dscSelf}} = \left\| \mathrm{H} - \mathrm{HC} \right\|_{\mathrm{F}}^2 + \lambda \left\| \mathrm{C} \right\|_{\mathrm{F}}^2 + \gamma \left\| (1 - \mathrm{G}) \odot \mathrm{C} \right\|_{\mathrm{F}}^2 \quad (9)$$

where $\gamma$ is a weighting factor for the introduced graph regularization term. This term incorporates graph-guided information through the weight matrix $G$, which is estimated from the original imputed data $Y$ as:

$$\mathrm{G} = \frac{\langle \mathrm{Y}, \mathrm{Y}^{\mathrm{T}} \rangle}{\|\mathrm{Y}\|, \|\mathrm{Y}\|} \quad (10)$$

The additional regularization term $\|(1 - G) \odot C\|_F^2$ helps guide the self-expression learning process using the similarity graph structure. This aids in mitigating suboptimal local solutions and ultimately contributes to improved clustering performance.

### C. MUTUAL TRAINING STRATEY

Our proposed Mutual Training Strategy addresses the limitations inherent in using an independent imputation autoencoder and deep clustering network. While the regularization term $\|(1 - G) \odot C\|_F^2$ in self-expression learning helps avoid suboptimal solutions, the deep clustering network may initially produce unsatisfactory results. Similarly, the imputation autoencoder, though equipped with an imputation loss function, operates separately from the clustering process. To overcome these challenges, we introduce a joint optimization approach for both the imputation and clustering components. This is achieved by enabling the imputation autoencoder and the deep clustering modules to share parameters during end-to-end training. Through this method, the imputed data generated by the autoencoder is not static but is iteratively updated and fed into the clustering module across each training epoch. Consequently, the outputs from the deep clustering process can reciprocally enhance the quality of imputation.

This synergistic framework allows for a dynamic interaction between clustering and imputation. The performance of the clustering module acts as a self-supervising signal for the imputation process, while the improved imputed data, in turn, refines the effectiveness of the clustering. Such a coupled approach ensures that enhancements in one component directly benefit the other, leading to a more integrated and efficient analysis process.

## IV. EXPERIMENT

In this section, we conduct an evaluation of the proposed method using several benchmark datasets, comparing its performance against various representative clustering methods and two baseline methods. Additionally, a comprehensive analysis of the model is presented.

**TABLE 1. The summary of the used datasets.**

| Datasets | Cells | Genes | Subtypes |
|---|---|---|---|
| Human Brain | 466 | 22085 | 8 |
| Human Embryonic | 1018 | 19189 | 7 |
| Mouse Bladder | 2746 | 20670 | 16 |
| Embryonic stem | 758 | 19189 | 11 |

### A. DATA PREPARATION

Four datasets are employed to assess the efficacy of our method, with details provided below:

1) Human Brain: This scRNA-seq data is obtained from epileptic patients undergoing temporal lobectomy for medically refractory seizures [38]. The dataset initially includes 466 samples, each characterized by the expression levels of 22,085 genes, encompassing 8 distinct cell types.
2) Human Embryonic: This dataset comprises 1,018 single cells, each characterized by the expression of 19,189 transcripts. The data naturally forms 7 distinct clusters [41].
3) Mouse Bladder: This dataset consists of 2,746 cells originating from 16 different types of mouse bladder cells [42]. Each sample is characterized by the expression levels of 20,670 genes.
4) Embryonic stem: This dataset represents scRNA-seq data from the differentiation process of embryonic stem cells to definitive endoderm cells [41]. The dataset encompasses a total of 758 cells, with each cell characterized by the expression levels of 19,189 genes.

The summary of the above data can be found in Table 1.

### B. COMPETING METHODS

To benchmark the proposed method, we compare it against two baseline methods and several representative clustering approaches:

1) AE+kmeans: This baseline method initially employs an autoencoder to acquire discriminative latent representations, followed by the application of k-means clustering on the learned latent representation.
2) AE+SC: Similar to the AE+k-means approach, this method utilizes spectral clustering on the acquired latent representation.
3) scDSC [43]: scDSC comprises a Zero-Inflated Negative Binomial (ZINB) model, an autoencoder, and a Graph Neural Network (GNN) module.
4) GraphSC [44]: GraphSC utilizes a graph autoencoder for clustering scRNA-seq data.
5) scNAME [45]: scNAME involves a mask estimation task for gene correlation mining and a neighborhood contrast learning framework for developing cell intrinsic structure. The learned patterns through mask

estimation aid in revealing uncorrupted data structures and denoising the original single-cell data.

6) ScFseCluster [46]:scFse clustering is a scRNA-seq clustering analysis basd on the feature selection module supported by the Quantum Squirrel Search algorithm (FSQSSA).
7) JLONMFSC [47]:JLONMFSC is a clustering model that jointly learns nonnegative matrix factorization and subspace clustering, using graph regularization matrix factorization to learn local features. Global features are learned through low-rank representation subspace clustering. Finally, the joint learning of local features and global features is carried out to improve the clustering effect.

### C. EVALUATION METRIC

In our evaluation, we employ standard metrics to quantitatively assess the performance of our clustering method: Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Clustering Accuracy (ACC). These metrics provide a comprehensive evaluation of both the accuracy and the similarity in our cluster recognition results. In order to enhance the robustness of our algorithmic validation, we repeated all comparative experiments a total of 10 times. Subsequently, we reported the outcomes for accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI), along with their respective variances.

NMI is a measure of the similarity between two sets of clusterings. It is derived from the mutual information and the entropy of the cluster labels and is normalized to a scale of 0 to 1. Higher NMI values indicate a greater degree of similarity between the compared clusterings.

ARI metric is an adjusted version of the Rand Index, accounting for random clustering effects. It contrasts the actual Rand Index with an expected index that would result from random clustering, operating within a range of 0 to 1. A higher ARI suggests that the clustering outcome is more closely aligned with the actual data distribution. This metric is particularly relevant in complex analysis scenarios, such as single-cell gene expression data analysis.

ACC measures the accuracy of the clustering results by determining the proportion of correctly classified samples. A high ACC value reflects a clustering outcome that closely mirrors the real distribution of the data, indicating a more effective clustering performance.

These metrics collectively enable a robust evaluation of our method, ensuring its reliability and effectiveness in clustering tasks.

### D. RESULTS AND ANALYSIS

Overall, our study implements the proposed method using Python with PyTorch and rigorously evaluates it on four standard datasets. The comparative analysis, detailed in Tables 2, Tables 3 and Tables 4, utilizes three key evaluation metrics, NMI, ARI and ACC, across these benchmark datasets. The findings consistently demonstrate the superior performance of our proposed method compared to others, Particular emphasis is placed on its effectiveness in cRhallenging situations. Key observations from our analysis include:

1) Our results affirm that deep clustering methods surpass deep feature learning combined with shallow clustering approaches (such as AE+kmeans, AE+spectral). This advantage is attributed to the end-to-end architecture of deep clustering, which enhances representation learning capabilities. These capabilities enable deep clustering methods to uncover meaningful structures and achieve more accurate clustering outcomes, thereby outperforming their shallow counterparts.
2) Our method consistently outperforms competing methods, with the best results achieved overall and ScFseCluster obtaining the second-best outcome. Notably, significant improvements are observed across various datasets, highlighting the superior efficacy of our proposed approach. Particularly noteworthy is the substantial enhancement observed on the Embryonic stem datasets, with improvements of 10.1% on ARI and 7.34% on NMI compared to scDSC, alongside a 11.9% increase in ACC. On the Human Brain dataset, our method demonstrates notable improvements of 20.1% on ARI, 20.3% on NMI, and 22.8% on ACC compared to the GraphSC method. Similarly, on the Mouse dataset, our method outperforms scNAME by 32.9% on ARI, 20.2% on NMI, and 22.1% on ACC. These substantial performance gains underscore the robustness and effectiveness of our proposed method, particularly in challenging scenarios such as the analysis of the Human Brain dataset, highlighting its capability to deliver more accurate and reliable clustering results.
3) Comparing our method's imputation approach with other techniques, it's evident that our method consistently achieves superior clustering performance. This underlines the effectiveness of our imputation module in handling missing data accurately. The method's capacity to enhance subsequent clustering outcomes through effective data imputation demonstrates its robustness and superiority over other imputation approaches. This finding suggests that our proposed imputation module is instrumental in improving the overall method's performance, leading to more precise and dependable clustering results, especially in scenarios with missing data.

Overall, the proposed method stands out as a highly effective and versatile tool in the field of deep clustering. It demonstrates exceptional capability in managing complex and varied datasets, consistently delivering reliable and precise clustering outcomes. This effectiveness underscores the method's potential for broad applicability, particularly in scenarios that demand robust data analysis capabilities. Its proficiency is especially noteworthy in situations involving intricate data structures and datasets with missing information, where traditional methods might struggle.

**TABLE 2.** Performances of all competitive methods in terms of NMI(Std).

| Method | ours | scDSC | graph-sc | scNAME | AE-kmeans | AE+SC | ScFseCluster | JLONMFSC |
|---|---|---|---|---|---|---|---|---|
| Human Brain | 92.35(1.724) | 83.19(3.182) | 73.62(4.135) | 76.46(3.972) | 82.55(2.956) | 87.68(2.739) | 91.27(2.058) | 89.83(2.346) |
| Human Embryonic | 91.48(1.967) | 89.93(3.078) | 78.36(4.059) | 74.24(4.032) | 87.12(3.135) | 88.96(3.048) | 90.35(2.193) | 91.07(2.482) |
| Mouse Bladder | 76.90(1.872) | 73.05(3.153) | 58.21(3.919) | 61.38(3.961) | 63.46(2.829) | 73.89(2.717) | 76.13(2.368) | 74.52(2.293) |
| Embryonic stem | 80.54(2.019) | 74.63(2.937) | 71.29(4.375) | 76.52(3.745) | 73.15(3.268) | 77.24(3.083) | 78.49(2.136) | 76.96(2.204) |

**TABLE 3.** Performances of all competitive methods in terms of ARI(Std).

| Method | ours | scDSC | graph-sc | scNAME | AE-kmeans | AE+SC | ScFseCluster | JLONMFSC |
|---|---|---|---|---|---|---|---|---|
| Human Brain | 90.43(1.859) | 81.79(3.124) | 72.14(3.927) | 74.58(3.894) | 82.37(2.862) | 81.54(2.908) | 88.29(1.945) | 83.68(2.414) |
| Human Embryonic | 80.84(2.135) | 75.46(3.229) | 73.68(4.182) | 48.35(4.156) | 74.29(3.148) | 75.09(3.072) | 76.81(2.264) | 77.97(2.659) |
| Mouse Bladder | 69.81(2.063) | 58.37(3.089) | 42.55(3.846) | 46.83(3.725) | 56.12(2.853) | 63.24(2.743) | 67.08(2.437) | 64.65(2.075) |
| Embryonic stem | 70.36(2.184) | 63.28(3.168) | 50.87(4.105) | 64.03(3.978) | 62.54(3.039) | 65.62(3.125) | 67.59(2.175) | 66.43(2.322) |

**TABLE 4.** Performances of all competitive methods in terms of ACC(Std).

| Method | ours | scDSC | graph-sc | scNAME | AE-kmeans | AE+SC | ScFseCluster | JLONMFSC |
|---|---|---|---|---|---|---|---|---|
| Human Brain | 95.25(1.767) | 82.76(3.234) | 73.52(4.009) | 77.92(3.820) | 82.51(2.813) | 89.21(2.865) | 92.52(1.989) | 89.35(2.369) |
| Human Embryonic | 81.97(2.007) | 77.32(3.107) | 57.46(3.992) | 43.06(4.174) | 71.18(3.078) | 73.53(3.115) | 77.52(2.104) | 78.67(2.511) |
| Mouse Bladder | 64.06(1.901) | 53.51(3.004) | 48.20(3.798) | 49.91(3.937) | 52.25(2.760) | 57.21(2.873) | 60.91(2.465) | 58.14(2.126) |
| Embryonic stem | 65.67(2.163) | 57.88(2.992) | 43.67(4.293) | 59.18(3.895) | 51.10(3.155) | 60.96(2.984) | 64.90(2.298) | 61.02(2.294) |

### *E. ABLATION STUDY*

To assess the effectiveness of the data imputation and self-guided regularization in our proposed method, we compare it with two variants:

- *Variant 1: Without data imputation, we directly apply the proposed self-guided deep subspace clustering on raw data.*
- *Variant 2: Remove the self-guided learning regularization.*

Performance comparison of different variants of the proposed method on Human Embryonic and Mouse Bladder datasets in terms of NMI is illustrated in Figure. 2.

The decline in performance upon excluding self-guided learning suggests that this component plays a vital role in the deep clustering module. It implies that self-guided learning contributes substantially to the refinement and enhancement of the clustering process, aiding in the accurate grouping of data points.Furthermore, disregarding the issue of missing values in the samples and omitting data interpolation also leads to a deterioration in clustering performance. This outcome emphasizes the importance of addressing missing data in clustering tasks. Experimental results demonstrate that the full model outperforms its two variants, indicating the effectiveness of the proposed approach.

### *F. VISUALIZATION ANALYSIS*

To gain a more intuitive biological interpretation of the clustering results, we employed visual graphs on a

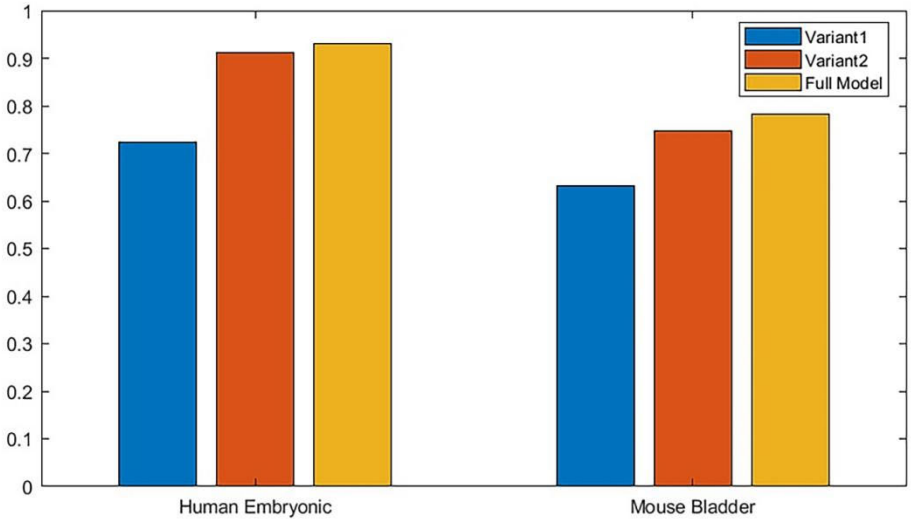


**FIGURE 2.** Performance comparsion of different variants of the proposed method in terms of NMI.

two-dimensional plane using four real datasets. Specifically, we compared our approach to several competitive approaches, namely SCDSC, Graphsc, SCNAME, AUTOENCODER, ScFseCluster and JLONMFSC. The low-dimensional latent features were extracted and visualized using t-distributed stochastic neighbor embedding (t-SNE) for each method.

In the dataset presented in the Figure. 3, we observe that the proposed method effectively distinguishes different cell types, irrespective of cluster sizes. In contrast, SCDSC, Graphsc, SCNAME and AUTOENCODER struggle to clearly delineate clusters. These visualization plots offer visual evidence that aligns with our numerical results, confirming that our method excels in detecting various cell types, concentrating cells within a cluster, and effectively separating different cell types.

### G. CONVERGENCE ANALYSIS

The assessment of convergence behavior is a crucial aspect of evaluating the stability and effectiveness of a model during its training phase. To demonstrate the convergence of our method, we closely monitored the loss value throughout the training process. For our model's architecture, we opted for a deep convolutional autoencoder structure. The size of the encoding layer was set to (512, 256, 256), while the embedding layer was configured to (256, 128, 32), complementing the structure of the decoding layer. We commenced by pre-training the basic autoencoder using all available data to establish a well-prepared initial model. The training parameters were meticulously chosen: we set the learning rate to 0.001, the number of epochs to 200, and the batch size to 256. The Adam optimizer was employed to adjust the learning rate dynamically. It is important to note that each layer in the autoencoder module during formal training mirrored the size of the pre-trained autoencoder, maintaining consistency in the model's structure. The initial learning rate was maintained at 0.001 with a batch size of 256, and the model underwent 200 epochs of training using the Adam optimizer. The convergence of the model is depicted in Figure. 4, where a consistent decrease in the loss value is observed with each training iteration. This trend is indicative of the model's robust convergence properties.

Additionally, the relationship between clustering evaluation metrics and training convergence is further elucidated in the results presented in Figure. 5, It's important to note that both Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are critical metrics for evaluating clustering performance, and interestingly, they tend to exhibit parallel trends in their evaluations. Moreover, a significant observation is made regarding the convergence of the method. As the loss value decreases with each training iteration and eventually stabilizes, a similar trend of stabilization is noticed in the values of NMI and ARI. This correlation is a strong indicator of the method's robust convergence properties. The stability in these metric scores, aligning with the stabilization of the loss values, underscores the reliability and effectiveness of the training process. This consistency between the decreasing and stabilizing loss values and the corresponding stability in NMI and ARI scores demonstrates not only the method's good convergence but also its capability to produce reliable and accurate clustering results. It affirms the method's potential for practical and effective applications in clustering tasks, where stable and consistent performance is crucial.

### H. CASE STUDY

Single-cell clustering is a fundamental preprocessing step in single-cell RNA sequencing (scRNA-seq) analysis, setting the stage for subsequent in-depth investigations. Beyond clustering, cell trajectory analysis stands as another pivotal task in scRNA-seq studies. This analysis utilizes time-series gene expression data from individual cells, enabling researchers to deduce the trajectories that explain the dynamics of cell

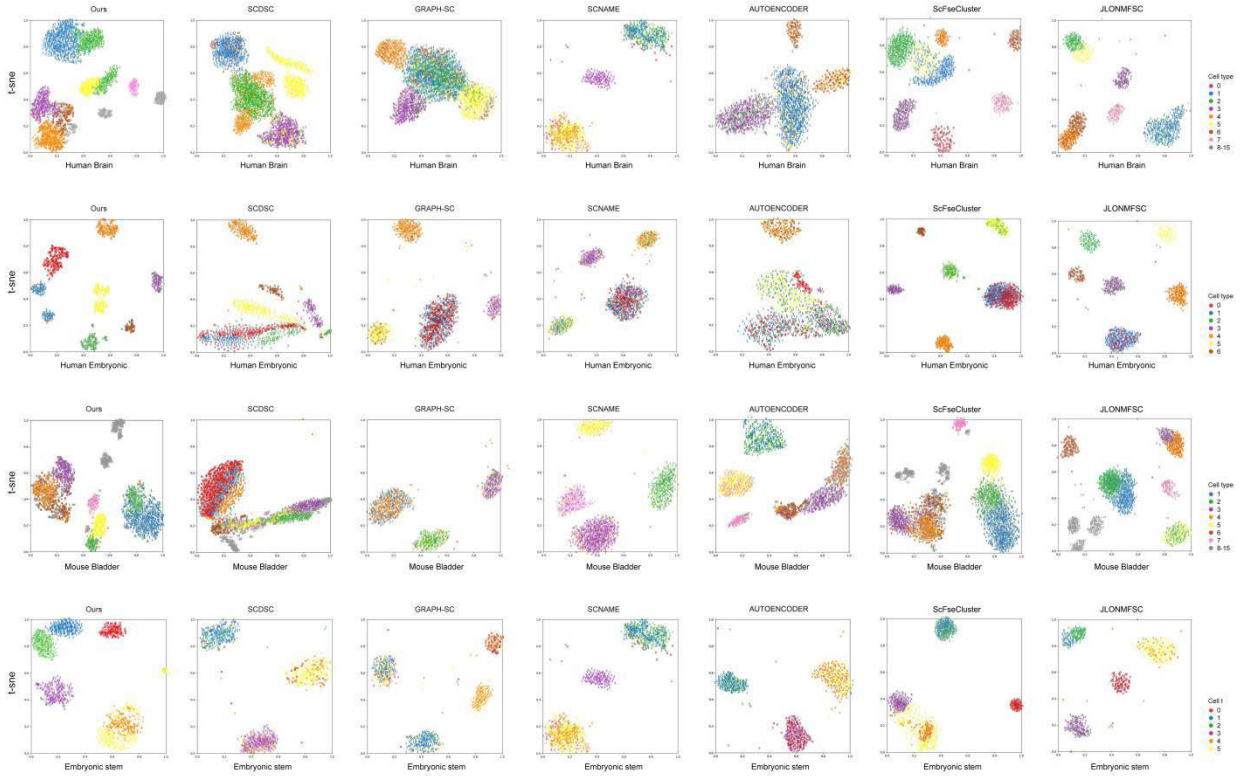


**FIGURE 3.** Depicts the t-SNE visualization showcasing the learned features obtained from various clustering methods alongside the autoencoder. Each distinct color represents a different cluster, facilitating the comparison of clustering outcomes.

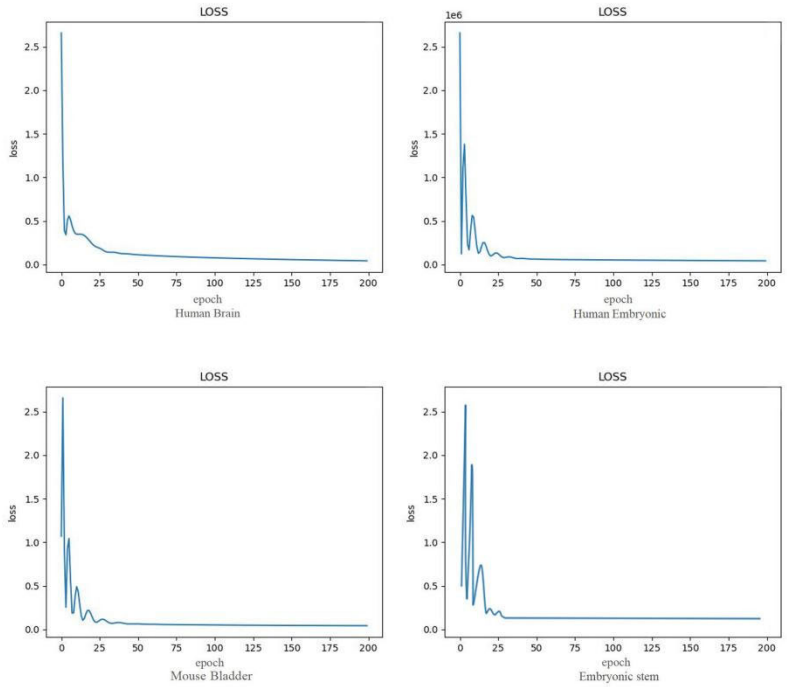


**FIGURE 4.** Displays the convergence analysis, illustrating the loss value trajectory throughout the learning process across four representative datasets. This analysis provides insights into the optimization progress of the method.

state transformations over time. In our study, we conduct cell trajectory analysis experiments using embryonic stem cell data collected at six distinct time points. This experiment is designed to evaluate the efficacy of our method

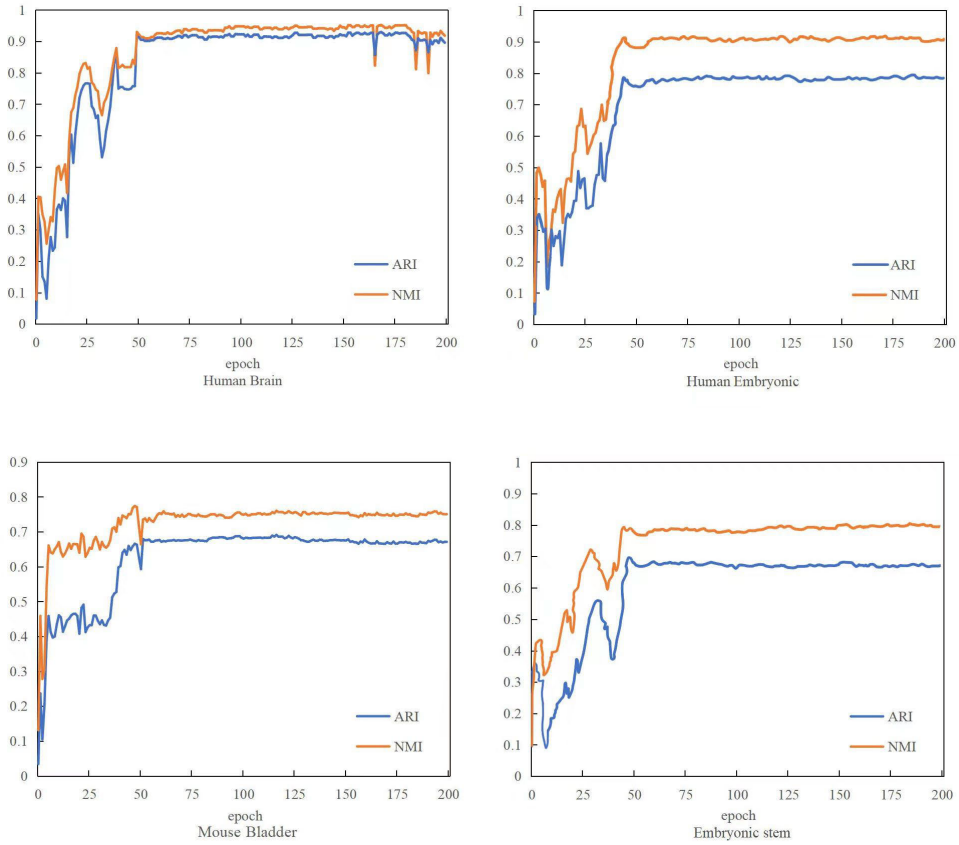


**FIGURE 5.** Analysis of convergence patterns and clustering performance.

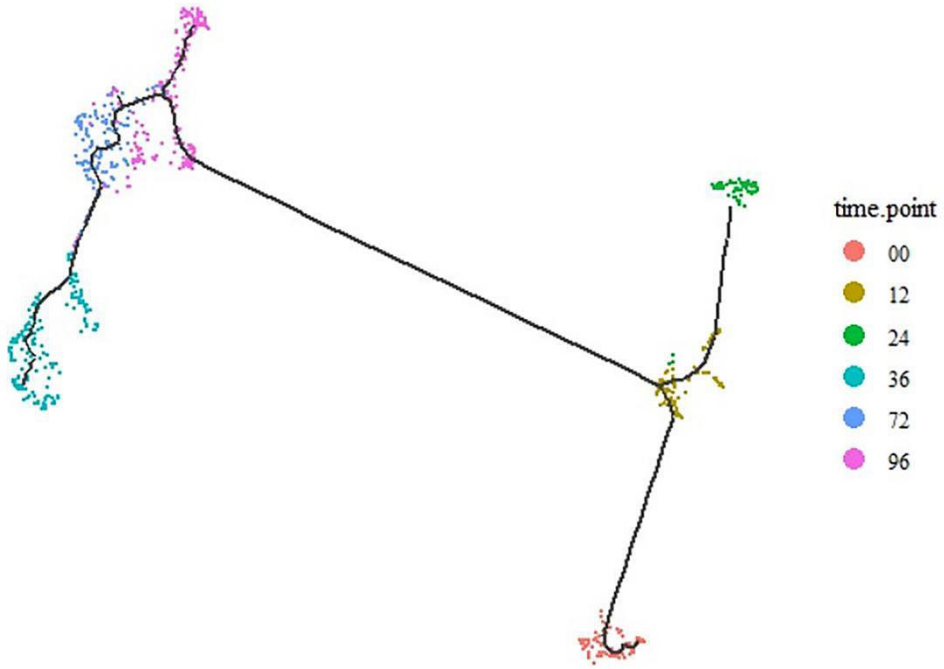


**FIGURE 6.** Illustrates the cell trajectory analysis conducted using the proposed method, with different colors representing various time points (0, 12, 24, 36, 48, 72, 96 hours). This visualization enables the examination of temporal changes in cell types over the specified time intervals.

in accurately inferring cellular trajectories, with a particular focus on understanding the developmental pathways of cells.

The results, illustrated in Figure 6, provide a clear demonstration of the capability of our proposed method. They show

that our approach not only accurately infers cell trajectories but also offers valuable insights into the complex processes of cellular development and differentiation. This outcome highlights the method's utility in unraveling the intricate mechanisms driving cell state changes, thereby contributing significantly to the field of developmental biology and related research areas.

## V. CONCLUSION

In this work, we implemented an imputation autoencoder network to estimate missing values, followed by the utilization of a deep clustering network for precise cell identification. To mitigate the susceptibility of deep clustering networks to local suboptimal solutions, we introduced a novel self-guided learning approach, which integrates parameter sharing between the imputation and clustering networks, mutually enhancing both processes. Experimental results indicate that the clustering performance of our method surpasses that of the other competing clustering methods. Notably, our utilization of the imputation autoencoder network not only enhances the accuracy of model clustering scRNA-seq data but also offers a novel approach for generating scRNA-seq data and advancing biological research. However, a notable limitation of our approach is its operational efficiency, particularly as the runtime escalates with the increasing number of cells in scRNA-seq datasets. Therefore, optimizing the clustering efficiency for large-scale single-cell data is crucial for practical application. Future research directions include leveraging anchor graph technology to expedite the learning process and reduce runtime. Furthermore, given the emergence of diverse data types in areas such as gene expression, nucleotide sequencing, and protein abundance, we are committed to broadening our research scope. We aim to apply our deep clustering framework to multi-omics studies, employing more efficient deep learning methods to extract information from scRNA-seq data and enhance clustering accuracy. This effort is geared towards improving performance and expanding the applicability of our proposed method.

## DATA AVAILABILITY STATEMENT

The source code and dataset of Sgdsc have been up loaded to https://github.com/wangyisong66/experiment.git.

## REFERENCES

[1] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning," *Nature Methods*, vol. 14, no. 4, pp. 414–416, Apr. 2017.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

[3] J. Žurauskienė and C. Yau, "PcaReduce: Hierarchical clustering of single cell transcriptional profiles," *BMC Bioinf.*, vol. 17, no. 1, p. 140, Mar. 2016.

[4] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hemberg, "SC3: Consensus clustering of single-cell RNA-seq data," *Nature Methods*, vol. 14, no. 5, pp. 483–486, May 2017.

[5] F. Hausmann, C. Ergen, R. Khatri, M. Marouf, S. Hänzelmann, N. Gagliani, S. Huber, P. Machart, and S. Bonn, "DISCERN: Deep single-cell expression reconstruction for improved cell clustering and cell subtype and state detection," *Genome Biol.*, vol. 24, no. 1, p. 212, Sep. 2023.

[6] H. Wang, J. Zhao, C. Zheng, and Y. Su, "ScDSSC: Deep sparse subspace clustering for scRNA-seq data," *PLOS Comput. Biol.*, vol. 18, no. 12, Dec. 2022, Art. no. e1010772.

[7] J. Lee, S. Kim, D. Hyun, N. Lee, Y. Kim, and C. Park, "Deep single-cell RNA-seq data clustering with graph prototypical contrastive learning," *Bioinformatics*, vol. 39, no. 6, Jun. 2023, Art. no. btad342.

[8] T. S. Andrews and M. Hemberg. *Modelling Dropouts Allows for Unbiased Identification of Marker Genes in scRNASeq Experiments*. bioRxiv. Accessed: Feb. 1, 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/065094

[9] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: Simulation of single-cell RNA sequencing data," *Genome Biol.*, vol. 18, no. 1, p. 174, Sep. 2017.

[10] X. Zhu, T. Ching, X. Pan, S. M. Weissman, and L. Garmire, "Detecting heterogeneity in single-cell RNA-seq data by non-negative matrix factorization," *PeerJ*, vol. 5, p. e2888, Jan. 2017.

[11] O. Poirion, X. Zhu, T. Ching, and L. X. Garmire, "Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage," *Nature Commun.*, vol. 9, no. 1, p. 4892, Nov. 2018.

[12] D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, and B. Bierie, "Recovering gene interactions from single-cell data using data diffusion," *Cell*, vol. 174, no. 3, pp. 716–729, Jul. 2018.

[13] W. V. Li and J. J. Li, "An accurate and robust imputation method scImpute for single-cell RNA-seq data," *Nature Commun.*, vol. 9, no. 1, p. 997, Mar. 2018.

[14] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang, "SAVER: Gene expression recovery for single-cell RNA sequencing," *Nature Methods*, vol. 15, no. 7, pp. 539–542, Jul. 2018.

[15] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry, "DrImpute: Imputing dropout events in single cell RNA sequencing data," *BMC Bioinf.*, vol. 19, no. 1, p. 220, Jun. 2018.

[16] M. Chen and X. Zhou, "VIPER: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies," *Genome Biol.*, vol. 19, no. 1, p. 196, Nov. 2018.

[17] P. van Galen, V. Hovestadt, M. H. Wadsworth II, T. K. Hughes, G. K. Griffin, S. Battaglia, J. A. Verga, J. Stephansky, T. J. Pastika, J. L. Story, and G. S. Pinkus, "Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity," *Cell*, vol. 176, no. 6, pp. 1265–1281, Mar. 2019.

[18] Z. Huang, W. J. Lu, C. Hong, and J. Ding, "Cheetah: Lean and fast secure two-party deep neural network inference," in *Proc. 31st USENIX Secur. Symp.*, Boston, MA, USA, Aug. 2022, pp. 809–826.

[19] N. Alghamdi, W. Chang, P. Dang, X. Lu, C. Wan, S. Gampala, Z. Huang, J. Wang, Q. Ma, Y. Zang, M. Fishel, S. Cao, and C. Zhang, "A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data," *Genome Res.*, vol. 31, no. 10, pp. 1867–1884, Oct. 2021.

[20] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Commun.*, vol. 10, no. 1, p. 390, Jan. 2019.

[21] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell RNA-sequencing experiments," *Genome Biol.*, vol. 17, no. 1, p. 63, Apr. 2016.

[22] D. Sengupta, N. A. Rayan, M. Lim, B. Lim, and S. Prabhakar. *Fast, Scalable and Accurate Differential Expression Analysis for Single Cells*. bioRxiv. Accessed: Feb. 1, 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/049734

[23] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, Mar. 2014.

[24] J. Zhu, J. Zhang, L. Wang, H. Huang, Z. Zhang, K. Song, and X. Zhang, "Progressively helical multi-omics data fusion GCN and its application in lung adenocarcinoma," *IEEE Access*, vol. 11, pp. 73568–73582, 2023.

[25] Z.-J. Cao and G. Gao, "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding," *Nature Biotechnol.*, vol. 40, no. 10, pp. 1458–1466, Oct. 2022.

[26] D. Bo, X. Wang, C. Shi, and H. Shen, "Beyond low-frequency information in graph convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 5, pp. 3950–3957.
[27] C. Che, Y. Fu, W. Shi, Z. Zhu, and D. Wang, "Dual feature fusion tracking with combined cross-correlation and transformer," *IEEE Access*, vol. 11, pp. 144966–144977, 2023.
[28] H. V. Ribeiro, D. D. Lopes, A. A. B. Pessa, A. F. Martins, B. R. da Cunha, S. Gonçalves, E. K. Lenzi, Q. S. Hanley, and M. Perc, "Deep learning criminal networks," *Chaos, Solitons Fractals*, vol. 172, Jul. 2023, Art. no. 113579.
[29] Z. Gao, W. Dang, X. Wang, X. Hong, L. Hou, K. Ma, and M. Perc, "Complex networks and deep learning for EEG signal analysis," *Cognit. Neurodyn.*, vol. 15, no. 3, pp. 369–388, Jun. 2021.
[30] P. Ji, J. Ye, Y. Mu, W. Lin, Y. Tian, C. Hens, M. Perc, Y. Tang, J. Sun, and J. Kurths, "Signal propagation in complex networks," *Phys. Rep.*, vol. 1017, pp. 1–96, May 2023.
[31] R. Zheng, M. Li, Z. Liang, F.-X. Wu, Y. Pan, and J. Wang, "SinNLRR: A robust subspace clustering method for cell type detection by non-negative and low-rank representation," *Bioinformatics*, vol. 35, no. 19, pp. 3642–3650, Oct. 2019.
[32] P. Lin, M. Troup, and J. W. K. Ho, "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data," *Genome Biol.*, vol. 18, no. 1, p. 59, Mar. 2017.
[33] J. Ronen and A. Akalin, "netSmooth: Network-smoothing based imputation for single cell RNA-seq," *F1000Research*, vol. 7, p. 8, Jan. 2018.
[34] L. Zhang and S. Zhang, "Comparison of computational methods for imputing single-cell RNA-sequencing data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 376–389, Mar. 2020.
[35] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, pp. 740–742, Jul. 2014.
[36] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Rev. Genet.*, vol. 16, no. 3, pp. 133–145, Mar. 2015.
[37] D. Grün, L. Kester, and A. van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nature Methods*, vol. 11, no. 6, pp. 637–640, Jun. 2014.
[38] X. Li, K. Wang, Y. Lyu, H. Pan, J. Zhang, D. Stambolian, K. Susztak, M. P. Reilly, G. Hu, and M. Li, "Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis," *Nature Commun.*, vol. 11, no. 1, p. 2338, May 2020.
[39] Z. Yu, Y. Lu, Y. Wang, F. Tang, K. C. Wong, and X. Li, "ZINB-based graph embedding autoencoder for single-cell RNA-seq interpretations," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 4, pp. 4671–4679.
[40] Y. Wu and K. Zhang, "Tools for the analysis of high-dimensional single-cell RNA sequencing data," *Nature Rev. Nephrol.*, vol. 16, no. 7, pp. 408–421, Jul. 2020.
[41] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson, "Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm," *Genome Biol.*, vol. 17, no. 1, p. 173, Aug. 2016.
[42] X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, and D. Huang, "Mapping the mouse cell atlas by microwell-seq," *Cell*, vol. 172, no. 5, pp. 1091–1107, Feb. 2018.
[43] Y. Gan, X. Huang, G. Zou, S. Zhou, and J. Guan, "Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network," *Briefings Bioinf.*, vol. 23, no. 2, Mar. 2022, Art. no. bbac018.
[44] M. Ciortan and M. Defrance, "GNN-based embedding for clustering scRNA-seq data," *Bioinformatics*, vol. 38, no. 4, pp. 1037–1044, Jan. 2022.
[45] H. Wan, L. Chen, and M. Deng, "ScNAME: Neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data," *Bioinformatics*, vol. 38, no. 6, pp. 1575–1583, Mar. 2022.
[46] Z. Wang, X. Xie, S. Liu, and Z. Ji, "ScFseCluster: A feature selection-enhanced clustering for single-cell RNA-seq data," *Life Sci. Alliance*, vol. 6, no. 12, Dec. 2023, Art. no. e202302103.
[47] W. Lan, M. Liu, J. Chen, J. Ye, R. Zheng, X. Zhu, and W. Peng, "JLONMFSC: Clustering scRNA-seq data based on joint learning of non-negative matrix factorization and subspace clustering," *Methods*, vol. 222, pp. 1–9, Feb. 2024.

**YISONG WANG** was born in Suzhou, Anhui, China, in 1998. He is currently pursuing the degree with the School of Computer and Control Engineering, Northeast Forestry University. His research interests include bioinformatics, deep learning, machine learning, and artificial intelligence.

**MINGZHI WANG** was born in Tengzhou, Shandong, China, in 1997. He received the B.S. degree from the College of Information Science and Engineering, Linyi University, in 2019. His research interests include artificial intelligence, deep learning, and medical image processing.