JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 32, Number 10, 2025
© Mary Ann Liebert, Inc.
Pp. 935–950
DOI: 10.1177/15578666251359688

Network-Guided Sparse Subspace Clustering on Single-Cell Data

CHENYANG YUAN,[1,*] SHUNZHOU JIANG,[1,*] SONGYUN LI,[2,*] JICONG FAN,[1,3]
and TIANWEI YU[1,3]

ABSTRACT

With the rapid development of single-cell RNA sequencing (scRNA-seq) technology, researchers can now investigate gene expression at the individual cell level. Identifying cell types via unsupervised clustering is a fundamental challenge in analyzing single-cell data. However, due to the high dimensionality of expression profiles, traditional clustering methods often fail to produce satisfactory results. To address this problem, we developed NetworkSSC, a network-guided sparse subspace clustering (SSC) approach. NetworkSSC operates on the same assumption as SSC that cells of the same type have gene expressions lying within the same subspace. In addition, it integrates a regularization term incorporating the gene network's Laplacian matrix, which captures functional associations between genes. Comparative analysis on nine scRNA-seq datasets shows that NetworkSSC outperforms traditional SSC and other unsupervised methods in most cases.

Keywords: cell type identification, gene network, single-cell RNA sequencing, sparse subspace clustering, upsupervised learning.

1. INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a powerful technology that enables measurement of transcriptomic profiles at the individual cell level (Tang et al., 2009). It achieves unprecedented resolution for studying the transcriptome and provides critical insights into the genomic heterogeneity among individual cells (Buettner et al., 2015).

Many computational approaches have been developed for different stages of scRNA-seq data analysis, including data normalization and denoising (Amodio et al., 2019; Eraslan et al., 2019), data correction and batch effect removal (Hie et al., 2019; Polański et al., 2020), dimensionality reduction and feature selection (Deng et al., 2023; Ding et al., 2018), clustering and cell type annotation (Kiselev et al., 2019; Wei et al., 2022), cell–cell communication (Efremova et al., 2020; Jin et al., 2021), RNA velocity (Bergen et al., 2020; La Manno et al., 2018), and multimodal integration (Stark et al., 2020; Zuo and Chen, 2021). Among these stages, a critical component of scRNA-seq analysis is the classification of cell subpopulations through

[1] School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, China.
[2] School of Medicine, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, China.
[3] Shenzhen Research Institute of Big Data, Shenzhen, China.
*These authors have contributed equally.

This work was submitted as an extended abstract to the 20th International Symposium on Bioinformatics Research and Applications (ISBRA 2024: https://bio.csu.edu.cn/ISBRA2024/ISBRA2024_Home.html).

clustering into biologically meaningful groups. This is typically achieved via unsupervised clustering, which groups cells based on transcriptomic similarity and subsequently annotates these groups with corresponding cell types or subtypes. Cell clustering facilitates many downstream analyses, including cellular composition estimation and rare cell type discovery (Chen et al., 2019), making it a fundamental step in scRNA-seq data analysis. However, expression profiles in scRNA-seq data are often noisy and high-dimensional, with the number of features (genes) far exceeding the number of samples (cells). This imbalance poses statistical challenges when comparing cells in such high-dimensional spaces (Andrews and Hemberg, 2018).

Various traditional clustering methods have been directly applied to scRNA-seq analysis, including K-Means (Hartigan and Wong, 1979) and spectral clustering (SC; Ng et al., 2001). In addition, several methods have been specifically developed for cell type identification in scRNA-seq data. These include Seurat 2.0 (Butler et al., 2018), SIMLR (single-cell interpretation via multikernel learning) (Wang et al., 2017), and MPSSC (spectral clustering based on learning similarity matrix) (Park and Zhao, 2018), most of which aim to learn cell–cell similarities. Consensus clustering approaches such as SC3 (Kiselev et al., 2017) enhance clustering accuracy by aggregating results from multiple clustering algorithms. CIDR (Clustering through Imputation and Dimensionality Reduction) (Lin et al., 2017) and scImpute (Li and Li, 2018) improve clustering performance by imputing dropout events in scRNA-seq data. Other studies have also applied Nonnegative Matrix Factorization to clustering scRNA-seq data (Shao and Höfer, 2017). Most methods, however, only consider pairwise cell similarities, making it difficult to capture the complex relationships among cells.

In recent years, deep learning-based methods have also been developed for clustering scRNA-seq data due to their strong representation learning capabilities (Erfanian et al., 2023). Most methods adopt the Deep Embedded Clustering (DEC) framework by performing clustering on the embedded features of an autoencoder, including DESC (Deep Embedding for Single-cell Clustering) (Li et al., 2020), scAnCluster (Chen et al., 2020), scDeepCluster (single-cell model-based deep embedded clustering; Zeng et al., 2022), GOAE (Gene Ontology AutoEncoder) and GONN (Gene Ontology Neural Network) (Peng et al., 2019), scVAE (variational auto-encoders for single-cell gene expression data) (Grønbech et al., 2020), scDCCA (deep contrastive clustering for single-cell RNA-seq data) (Wang et al., 2023), G3DC (Gene-Graph-Guided Selective Deep Clustering; He et al., 2024), and scDFC (single-cell deep fusion clustering) (Hu et al., 2023). While these DEC-based methods have achieved considerable success, their performance heavily depends on the neural network architecture, hyperparameter tuning, and optimization algorithms, all of which require substantial domain expertise. Moreover, they generally lack the interpretability offered by traditional approaches such as K-Means and SC. For scRNA-seq data analysis, there remains a strong need for clustering algorithms that are not only accurate but also convenient, stable, and highly interpretable.

Sparse subspace clustering (SSC, Elhamifar and Vidal, 2013) is a variant of SC that has demonstrated promising performance in clustering high-dimensional data. SSC assumes that high-dimensional data lie in a union of multiple low-dimensional subspaces, with each data point represented as a linear combination of other points within the same subspace. SinNLRR (subspace clustering for cell type detection by non-negative and low-rank representation) (Zheng et al., 2019) supports the application of SSC to cell clustering by leveraging the subspace characteristics of cells' expression profiles, assuming that cells of the same type occupy the same subspace. Building on the same assumption, AdaptiveSSC (Zheng et al., 2020) modifies SSC by employing a data-driven adaptive sparse constraint to construct the affinity matrix for improved cell type identification. However, SSC-based methods face certain limitations. In cell clustering, most genes may be noisy and nondiscriminative, leading to low clustering accuracy. Although some researchers have proposed using principal component analysis (PCA) to reduce the dimensionality of data (Usoskin et al., 2015; žurauskienė and Yau, 2016), SSC and its variants lack an automatic and effective feature selection mechanism for denoising. Moreover, prior knowledge from gene networks, which is highly valuable for capturing functional relationships, is not utilized by SSC or its variants.

In this article, we propose a network-guided SSC method, NetworkSSC. NetworkSSC follows the same subspace assumption as SSC but incorporates the Laplacian matrix of the gene network as a regularization term. This approach leverages the functional relationships among genes encoded in the network structure, making it better suited for clustering gene expression data. NetworkSSC demonstrates improved performance compared to traditional SSC on multiple real datasets.

The remainder of the article is organized as follows: Section 2 introduces the concept of traditional SSC, which serves as the foundation for our method. Section 3 describes the design of NetworkSSC and outlines the procedures for solving the optimization problem. Section 4 applies NetworkSSC to nine real datasets, benchmarks its performance against seven competing methods, and validates its design and robustness via multiple analyses. Finally, Section 5 concludes the article with a discussion of potential improvements and future directions for NetworkSSC.

2. SPARSE SUBSPACE CLUSTERING

High dimensionality often poses significant challenges to data processing and analysis. SSC addresses these challenges by employing subspace representation to simplify clustering in high-dimensional spaces. According to the subspace assumption, high-dimensional data are distributed within a union of multiple low-dimensional subspaces (Parsons et al., 2004). Each sample can be expressed as a linear combination of other samples within the same subspace.

Applying this assumption to scRNA-seq data, the expression profile of a cell can be represented as a linear combination of other cells' profiles. Specifically, let $X \in \mathcal{R}^{p \times n}$ denote the expression matrix, where each row corresponds to a gene, and each column corresponds to a cell. The expression vector for cell i can then be written as:

$$x_i = z_1 x_1 + \cdots + z_{i-1} x_{i-1} + z_{i+1} x_{i+1} + \cdots + z_n x_n, \tag{1}$$

where $z_j \geq 0 \; (j \neq i)$ represents the similarity between cells i and j. If $z_j = 0$, it indicates that cells i and j do not belong to the same type. Extending this relationship to all cells, the matrix representation becomes:

$$X = XZ, \tag{2}$$

where $Z \in \mathcal{R}^{n \times n}$ is the coefficient matrix, and z_{ij} represents the similarity score between cells i and j.

To reveal the subspace structure, sparsity constraints are imposed on Z, encouraging it to have a block diagonal structure. The sparsity of a vector is measured using its l_0-norm, which counts the number of nonzero elements. However, solving optimization problems with the l_0-norm is typically NP-hard (Nondeterministic Polynomial-time hard) in real applications. To address this, the l_1-norm is commonly used as a convex relaxation of the l_0-norm (Chen et al., 2001). With this relaxation, the coefficient matrix Z can be computed by solving the following optimization problem:

$$min_Z \, \|Z\|_1 \quad s.t. \quad X = XZ, diag(Z) = 0. \tag{3}$$

The constraint $diag(Z) = 0$ is imposed to prevent the trivial solution where each data point is only represented by itself, which would result in Z being the identity matrix. After relaxing the optimization problem in Equation (3), it can be reformulated as:

$$min_Z \, \frac{1}{2} \|X - XZ\|_F^2 + \lambda \|Z\|_1 \quad s.t. \quad diag(Z) = 0, \tag{4}$$

where λ is a is a regularization parameter that controls the sparsity of the coefficient matrix Z.

Once the optimal Z is obtained, a symmetric affinity matrix $W \in \mathcal{R}^{n \times n}$ is constructed as:

$$W = \left(|Z| + |Z^{\mathrm{T}}| \right)/2. \tag{5}$$

The matrix W inherits desirable properties from Z. Specifically, it is a sparse matrix whose elements are nonnegative and represent the similarity between each pair of data points. The diagonal entries of W are set to zero since they represent a point's similarity to itself, which is not meaningful for clustering.

Finally, SC is performed on the learned affinity matrix W to obtain the clustering result. In this step, W is interpreted as the adjacency matrix of a newly learned graph. It is used to construct the graph Laplacian, whose eigenvectors reveal the intrinsic structure of the data, enabling effective clustering in a lower-dimensional space.

3. NETWORK-GUIDED SPARSE SUBSPACE CLUSTERING

We present NetworkSSC, an adaptation of SSC tailored for gene expression data. NetworkSSC builds upon the subspace assumption in SSC that cells of the same type reside in a common subspace. In addition, it incorporates gene network information to capture intrinsic relationships within the data. The workflow of NetworkSSC is illustrated in Figure 1. Starting with an input expression matrix X, NetworkSSC constructs the

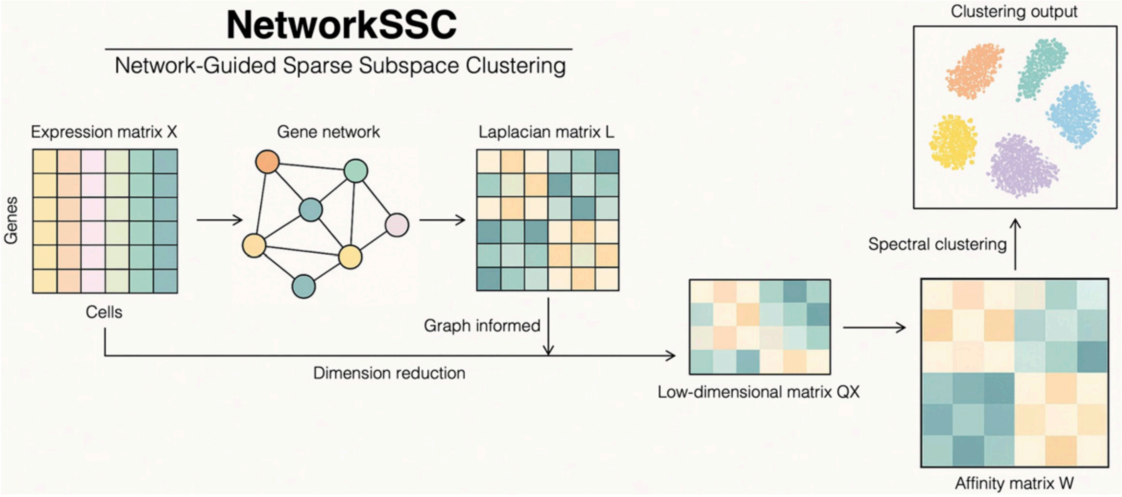

FIG. 1. Workflow of NetworkSSC. NetworkSSC, network-guided sparse subspace clustering.

gene network and the corresponding Laplacian matrix L. Guided by the information encoded in L, NetworkSSC then performs dimensionality reduction on X and computes the affinity matrix W. Finally, SC is applied to W to produce the clustering result. Overall, the inclusion of the Laplacian matrix L directly informs the dimensionality reduction of X, a step that is missing in traditional SSC. This integration enables efficient feature selection, thereby improving the quality of the downstream affinity matrix construction.

3.1. Model design

The gene network encodes information about the functional relationships between genes. As different cell types are often characterized by functional differences, leveraging the gene network provides valuable guidance for the clustering process. To incorporate this information, we introduce a graph regularization term into the original SSC cost function. This term is embedded with the Laplacian matrix of the gene network. According to the properties of the Laplacian matrix, the regularization term increases the weights of coefficients corresponding to intercluster samples while decreasing those associated with intracluster samples (He et al., 2017). This adjustment effectively captures the local geometric structure of the data and exploits intrinsic relationships within it. In addition, we perform dimensionality reduction on the expression matrix X to select important features and enhance computational efficiency. Under this new formulation, the coefficient matrix Z is computed as follows:

$$min_{Q,Z}\, f \triangleq \frac{1}{2}\|QX - QXZ\|_F^2 + \lambda_1\|Z\|_1 + \frac{\lambda_2}{2} tr\left(QLQ^{\mathrm{T}}\right) \quad s.t. \quad Q^{\mathrm{T}}Q = I, diag(Z) = 0, \tag{6}$$

where $Q \in \mathcal{R}^{d\times p}$ is a projection matrix designed to reduce the dimensionality of X from p to d, and $L \in \mathcal{R}^{p\times p}$ represents the Laplacian matrix of the gene network. This framework allows the intrinsic relationships within the expression matrix X to be leveraged iteratively during the optimization of both Q and Z.

The subsequent steps align closely with those in traditional SSC. Once the optimal Z is obtained, the affinity matrix W is constructed accordingly. SC is then applied to W, yielding the final clustering result.

3.2. Optimization

To solve Q and Z in the optimization problem outlined in Equation (6), we employ a technique known as Proximal Alternating Linearized Minimization (Bolte et al., 2014). This approach involves iteratively updating Q and Z in an alternating manner, ensuring efficient and stable optimization.

Consider the optimization steps in the t-th iteration. We first fix Q_{t-1} and solve Z, that is,

$$min_Z\, \mathcal{L}_1 \triangleq \frac{1}{2}\|Q_{t-1}X - Q_{t-1}XZ\|_F^2 + \lambda_1\|Z\|_1 \quad s.t. \quad diag(Z) = 0. \tag{7}$$

Let $\mu_{t-1} \geq 1.01 \cdot \|(Q_{t-1}X)^{\mathrm{T}}Q_{t-1}X\|_2$, then

$$\begin{aligned}\mathcal{L}_1 = \frac{1}{2}\|Q_{t-1}X - Q_{t-1}XZ\|_F^2 + \langle -(Q_{t-1}X)^{\mathrm{T}}(Q_{t-1}X - Q_{t-1}XZ_{t-1}), Z - Z_{t-1}\rangle \\ + \frac{\mu_{t-1}}{2}\|Z - Z_{t-1}\|_F^2 + \lambda_1\|Z\|_1.\end{aligned} \tag{8}$$

Let $G \triangleq -(Q_{t-1}X)^{\mathrm{T}}(Q_{t-1}X - Q_{t-1}XZ_{t-1})$, then

$$\mathcal{L}_1 = \frac{\mu_{t-1}}{2}\left\|Z - Z_{t-1} + \frac{G}{\mu_{t-1}}\right\|_F^2 + \lambda_1\|Z\|_1. \tag{9}$$

Let $C \triangleq Z_{t-1} - G/\mu_{t-1}$, then we update Z_t by:

$$Z_t = max\left(C - \frac{\lambda_1}{\mu_{t-1}}, 0\right) + min\left(C + \frac{\lambda_1}{\mu_{t-1}}, 0\right), \tag{10}$$

$$Z_t = Z_t - diag(Z_t). \tag{11}$$

Then, we fix Z_t and update Q_t by:

$$H \triangleq X - XZ_t, \tag{12}$$

$$M \triangleq HH^{\mathrm{T}} + \lambda_2 L, \tag{13}$$

$$Q_t : \text{last } d \text{ principal components of } M. \tag{14}$$

We also update μ_t by:

$$\mu_t = 1.01 \cdot \|(Q_tX)^{\mathrm{T}}Q_tX\|_2. \tag{15}$$

The update process continues iteratively until one of the following stopping criteria is met:

- The number of iterations exceeds a predefined upper limit, *kmax*;
- The updates of both Z and the cost function f in successive steps fall below a small threshold value, ϵ:

$$\frac{\|Z_t - Z_{t-1}\|_F}{\|Z_{t-1}\|_F} \leq \epsilon, \tag{16}$$

$$f(Q_t, Z_t) - f(Q_{t-1}, Z_{t-1}) \leq \epsilon. \tag{17}$$

A possible value for *kmax* is 1000, and a possible value for ϵ is 0.001.

Finally, we consider the initialization of Q and Z. We initialize Q as:

$$Q_0 : \text{first } d \text{ principal components of } X. \tag{18}$$

The initialization of Z can be found by solving

$$-(Q_0X)^{\mathrm{T}}(Q_0X - Q_0XZ_0) + \lambda_3 = 0, \tag{19}$$

where we set $\lambda_3 = 1$ as default. Thus,

$$Z_0 = ((Q_0X)^{\mathrm{T}}Q_0X + I)^{-1}(Q_0X)^{\mathrm{T}}Q_0X, \tag{20}$$

$$Z_0 = Z_0 - diag(Z_0). \tag{21}$$

The complete procedures of NetworkSSC are summarized in Algorithm 1.

Algorithm 1. Network-Guided Sparse Subspace Clustering (NetworkSSC)

```
**Input:** expression matrix $X$, normalized Laplacian matrix $L$, hyperparameters $d$, $\lambda_1$, $\lambda_2$, maximum number of iterations
        $kmax$, threshold value $\epsilon$, number of clusters $m$
$Q_0 \leftarrow$ first $d$ principal components of $X$;
$Z_0 \leftarrow ((Q_0X)^{\mathrm{T}}Q_0X + I)^{-1}(Q_0X)^{\mathrm{T}}Q_0X$;
$Z_0 \leftarrow Z_0 - diag(Z_0)$;
$\mu_0 \leftarrow 1.01 \cdot \|(Q_0X)^{\mathrm{T}}Q_0X\|_2$;
$k \leftarrow 0$;
**while** $k < kmax$ **do**
    $f_k \leftarrow \frac{1}{2}\|Q_kX - Q_kXZ_k\|_F^2 + \lambda_1\|Z_k\|_1 + \frac{\lambda_2}{2}tr\big(Q_kLQ_k{}^{\mathrm{T}}\big)$;
    $G \leftarrow -(Q_kX)^{\mathrm{T}}(Q_kX - Q_kXZ_k)$;
    $C \leftarrow Z_k - G/\mu_k$;
    $Z_{k+1} \leftarrow max(C - \lambda_1/\mu_k, 0) + min(C + \lambda_1/\mu_k, 0)$;
    $Z_{k+1} \leftarrow Z_{k+1} - diag(Z_{k+1})$;
    $H \leftarrow X - XZ_{k+1}$;
    $M \leftarrow HH^{\mathrm{T}} + \lambda_2L$;
    $Q_{k+1} \leftarrow$ last $d$ principal components of $M$;
    $\mu_{k+1} \leftarrow 1.01 \cdot \|(Q_{k+1}X)^{\mathrm{T}}Q_{k+1}X\|_2$;
    $f_{k+1} \leftarrow \frac{1}{2}\|Q_{k+1}X - Q_{k+1}XZ_{k+1}\|_F^2 + \lambda_1\|Z_{k+1}\|_1 + \frac{\lambda_2}{2}tr\big(Q_{k+1}LQ_{k+1}{}^{\mathrm{T}}\big)$;
    $k \leftarrow k+1$;
    **if** $\|Z_{k+1} - Z_k\|_F/\|Z_k\|_F \leq \epsilon$ and $f_{k+1} - f_k \leq \epsilon$
        **then break**;
**end**;
Return optimal $Q$ and $Z$;
$W \leftarrow (\,|Z| + |Z^{\mathrm{T}}|\,)/2$;
Perform SC on $W$ with prespecified number of clusters $m$;
**Output:** clustering result on $X$, containing $m$ clusters
```

3.3. Hyperparameter tuning

NetworkSSC involves three key hyperparameters that must be defined prior to model training: the reduced data dimension d, the l_1 penalty factor λ_1, and the Laplacian penalty factor λ_2. An empirical rule is to set d to approximately 10%–20% of the original dimension. This reduction strikes a balance between preserving essential data structures and reducing the computational burden. Due to the dimensionality reduction, λ_1 and λ_2 should take relatively large values to ensure effective regularization. In practice, we perform a coarse grid search over a list of empirical values ranging from 10 to 1000 and select the combination that optimizes the clustering performance on each specific dataset. The impact of different hyperparameter settings is further evaluated in Supplementary Data S1 and Supplementary Figure S1.

3.4. Handling ultra-high dimensional data

In practice, directly applying NetworkSSC to real data can be computationally challenging due to the ultra-high dimensionality, which often results in slow execution and excessive memory usage. To address these issues and improve computational efficiency, we propose a shortcut for handling datasets with a relatively large number of genes, typically when $p > 5000$. The core idea is to partition the original dataset into q equally sized blocks based on the p genes, where

$$q = \left\lceil \frac{p}{5000} \right\rceil. \tag{22}$$

The partitioning is performed completely at random. NetworkSSC is then applied separately to each of the q blocks, producing q affinity matrices: $W_1, W_2, \ldots, W_q$. These matrices are subsequently averaged to generate a single affinity matrix W, which effectively integrates information across the entire dataset. Finally, SC is performed on W to produce the final clustering result.

4. RESULTS

4.1. Experimental datasets

We evaluated NetworkSSC using nine scRNA-seq datasets from human organs (Baron et al., 2016; De Micheli et al., 2020; He et al., 2020; Khaliq et al., 2022; La Manno et al., 2016; Muraro et al., 2016; Wu et al., 2021; Zilionis et al., 2019), as summarized in Table 1. These datasets were generated from distinct studies and platforms, encompassing a diverse range of tissue types and disease conditions, which allowed for a comprehensive and unbiased evaluation of our method. Among them, two datasets (Baron and Midbrain) were preprocessed by retaining only the top 2000 genes with the highest expression variation, reflecting scenarios where gene screening is applied. The remaining seven datasets were used in their original scale without additional gene filtering. As a result, these datasets cover a wide range of dimensions, as indicated by the relative number of genes compared with cells, allowing us to comprehensively evaluate NetworkSSC's performance. For example, Zilionis had a gene count more than 15 times the cell count, while Baron had a comparable number of genes and cells. Furthermore, each dataset provided ground truth labels for the cells, enabling an objective evaluation of clustering performance by comparing results with true labels.

For each dataset, we extracted the corresponding gene network from the human gene network available in the HINT (High-quality INTeractomes) database (Das and Yu, 2012). Genes not matching entries in the full network were retained and treated as isolated nodes. This ensured the preservation of all genes in the analysis, irrespective of their presence in the external network. In addition, we normalized the expression data for each gene to have zero mean and unit variance.

4.2. Benchmarking with other clustering methods

We benchmarked NetworkSSC against a broad range of unsupervised clustering algorithms, including the aforementioned traditional SSC and AdaptiveSSC, as well as K-Means, SC, Louvain clustering, scDeepCluster, and G3DC. K-Means and spectral clustering are representative of traditional clustering methods. K-Means aims to minimize within-cluster distances, while spectral clustering operates by clustering the eigenvectors of the graph Laplacian matrix. Louvain clustering detects communities in a k-nearest neighbor graph of cells by optimizing modularity, effectively grouping cells with similar gene expression profiles. It is the default clustering algorithm in SCANPY (Wolf et al., 2018), one of the most widely used Python packages for single-cell data analysis, and thus represents a highly popular and well-established method in the single-cell genomics community. scDeepCluster and G3DC represent state-of-the-art deep learning-based approaches. scDeepCluster learns a robust low-dimensional representation of scRNA-seq data using a ZINB-based denoising autoencoder, explicitly optimized for clustering through Kullback–Leibler divergence in the latent space. G3DC integrates graph regularization based on a prior gene network with feature selection via $l_{2,1}$-norm regularization and includes a reconstruction loss to generate informative embeddings for clustering. Since NetworkSSC is fundamentally an extension and enhancement of the traditional SSC framework, we emphasize and prioritize its comparison with SSC and AdaptiveSSC in this study.

To assess the performance of different clustering methods, we employed two widely used evaluation metrics: normalized mutual information (NMI, Pfitzner et al., 2009) and adjusted rand index (ARI, Hubert and Arabie, 1985). Given the true labels X and the clustering labels Y, these metrics are defined as follows:

TABLE 1. SUMMARY OF BASIC CHARACTERISTICS OF EXPERIMENTAL DATASETS

Dataset	*Platform*	*Source organ*	*Genes*	*Cells*	*Cell types*	*Gene filtering*	*Transformation*
BC_CID4067	Chromium	Breast Cancer	18,176	3752	7	No	No
Muscle_HeOrgan	Chromium	Muscle	16,024	3000	11	No	No
Zilionis	inDrop	Lung	41,861	2702	12	No	No
Muraro	CEL-Seq2	Pancreas	19,127	2126	10	No	Log-transformed
Baron	inDrop	Pancreas	2000	2812	6	Yes	No
Midbrain	STRT-seq	Brain	2000	1695	25	Yes	No
HCA_CC	Chromium	Colorectal Cancer	21,977	2948	5	No	No
Muscle	Chromium	Muscle	12,798	968	8	No	No
Blood_HeOrgan	Chromium	Blood	14,552	1407	9	No	No

CEL-Seq2, cell expression by linear amplification and sequencing 2; Chromium, chromium single single-cell gene expression platform; inDrop, indexed droplet single-cell RNA sequencing; STRT-Seq, single-cell tagged reverse transcription sequencing.

$$NMI(X,Y) = \frac{I(X,Y)}{H(X)+H(Y)}, \tag{23}$$

where $I(X,Y)$ represents the mutual information between X and Y, and $H(X)$ and $H(Y)$ denote the entropy of the true and clustering labels, respectively.

$$ARI(X,Y) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] \Big/ \binom{n}{2}}, \tag{24}$$

where n_{ij} is the number of cells belonging to group i in the true labels and group j in the clustering labels, $n_{i.}$ is the total number of cells in group i in the true labels, $n_{.j}$ is the total number of cells in group j in the clustering labels, and n is the total number of cells. Both NMI and ARI provide a measure of consistency between the clustering results and the true labels. NMI captures the shared information between the true and predicted labels, normalized to account for variations in label distributions. ARI adjusts the Rand Index to correct for chance groupings, making it particularly effective for datasets with imbalanced clusters.

Figure 2 presents the overall performance of the eight methods on the nine experimental datasets. Since most methods exhibit some level of randomness due to the embedded K-Means procedure, we ran each method 20 times and reported the average NMI and ARI. The standard deviation across runs for each method was reduced to negligible levels.

The results demonstrate the superior performance of NetworkSSC compared with traditional SSC and its variant. Specifically, NetworkSSC achieved comparable NMI and ARI scores to SSC on the BC_CID4067 and Muscle_HeOrgan datasets, while delivering significantly higher scores on the remaining datasets. It also

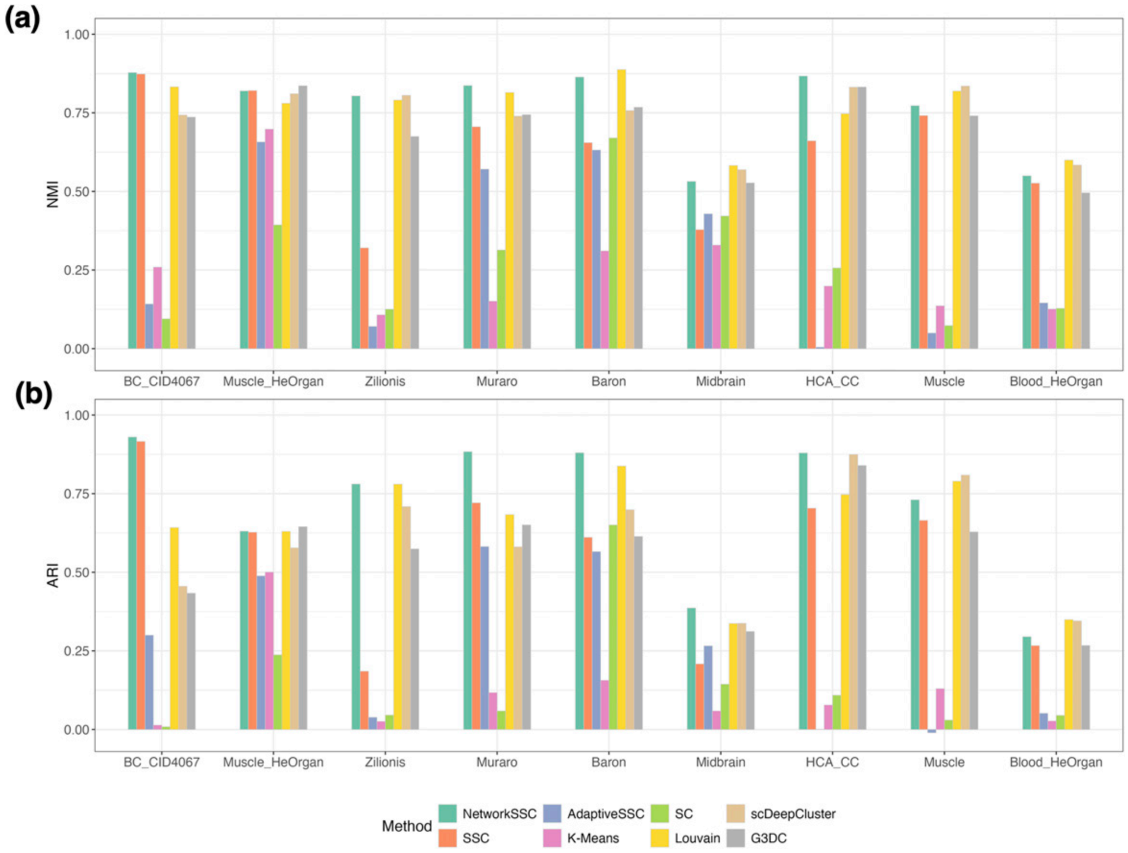

FIG. 2. Bar plots showing the average **(a)** NMI and **(b)** ARI for all compared methods across experimental datasets. x-axis: experimental datasets. y-axis: average NMI or ARI values calculated from 20 independent runs of each method. ARI, adjusted rand index; NMI, normalized mutual information.

consistently outperformed AdaptiveSSC across all nine datasets. This consistent superiority underscores the clear advantage of NetworkSSC as an enhancement of the traditional SSC framework. Moreover, NetworkSSC showed even stronger performance over traditional clustering methods such as K-Means and spectral clustering, achieving significantly higher NMI and ARI scores on all experimental datasets. These results indicate NetworkSSC's ability to more effectively capture complex structures in high-dimensional, noisy scRNA-seq data compared with traditional approaches.

Next, we compared NetworkSSC with widely used and deep learning-based methods. Compared with Louvain clustering, NetworkSSC achieved higher NMI and ARI on six and seven out of the nine datasets, respectively. Slightly lower NMI and ARI scores were observed only on the Muscle and Blood_HeOrgan datasets. This strong performance across most datasets highlights the competitiveness of NetworkSSC relative to widely adopted methods in the field. A similar trend was observed in the comparison with scDeepCluster, where NetworkSSC outperformed scDeepCluster on the majority of datasets, while scDeepCluster held a marginal advantage on the Muscle and Blood_HeOrgan datasets. We attribute this to the gene coverage in these two datasets, which resulted in highly sparse gene networks and reduced the effectiveness of network-guided feature selection and low-dimensional embedding in NetworkSSC. Finally, NetworkSSC outperformed G3DC on eight out of nine datasets, demonstrating its distinct advantage among gene network-based clustering approaches.

Based on the benchmarking results, it is noteworthy that the superior performance of NetworkSSC, particularly over traditional and baseline methods, remained consistent regardless of the source study, platform, tissue type, or dimensionality of the dataset. Furthermore, NetworkSSC offers better user convenience than certain other methods. For example, Louvain clustering does not accept the number of clusters as a direct input; instead, it relies on a resolution parameter to control granularity, which requires manual tuning and lacks straightforward interpretability. G3DC, on the other hand, requires manual adjustment of the neural network architecture based on the input data, rather than offering a fully automated, end-to-end solution. Taken together, NetworkSSC combines strong clustering performance with practical usability, making it a superior alternative to existing benchmarking methods.

Besides clustering accuracy, we also examined the computational time of NetworkSSC across all datasets, as detailed in Supplementary Table S1. Notably, the running time of NetworkSSC scales well with the number of genes in a dataset. For datasets with prescreened genes, NetworkSSC is often faster than traditional SSC, taking less than a minute to group thousands of cells. For datasets with ultra-high dimensionality, NetworkSSC requires more effort for feature selection, which can extend the running time to up to an hour.

4.3. Visualization of clustering results

In addition to the quantitative evaluation using NMI and ARI, we also visualized the clustering results from NetworkSSC and SSC-based approaches. Specifically, for each method, we first performed PCA to reduce the learned affinity matrix W to 50 dimensions and then embedded the PCA-transformed data into a two-dimensional space using t-distributed stochastic neighbor embedding technique (t-SNE, Kobak and Berens, 2019). In these visualizations, cells were projected onto the two-dimensional embeddings and colored according to their ground truth labels. High clustering quality is indicated when cells of the same type are grouped closely together.

As visualized in Figure 3, on the Zilionis dataset, NetworkSSC demonstrates a marked improvement over traditional SSC and AdaptiveSSC in its ability to group cells of the same type together. The clusters produced by NetworkSSC exhibit more distinct and cohesive silhouettes, whereas those generated by traditional SSC and AdaptiveSSC show significant overlap, with different cell types blending together in the t-SNE plots, and no clear clustering patterns emerge. This observation holds true for the Muraro dataset as well. These visualization results further underscore the superior capability of NetworkSSC to facilitate the visualization and interpretation of scRNA-seq data.

4.4. Ablation study and sensitivity analysis

To assess the importance of each model component in NetworkSSC, particularly the Laplacian regularization term, we conducted ablation studies on selected datasets. Specifically, we disabled the sparsity and Laplacian regularization components by setting the hyperparameters λ_1 and λ_2 to zero, respectively, and evaluated the clustering performance of the resulting reduced models. As shown in Figure 4a, on all three test datasets,

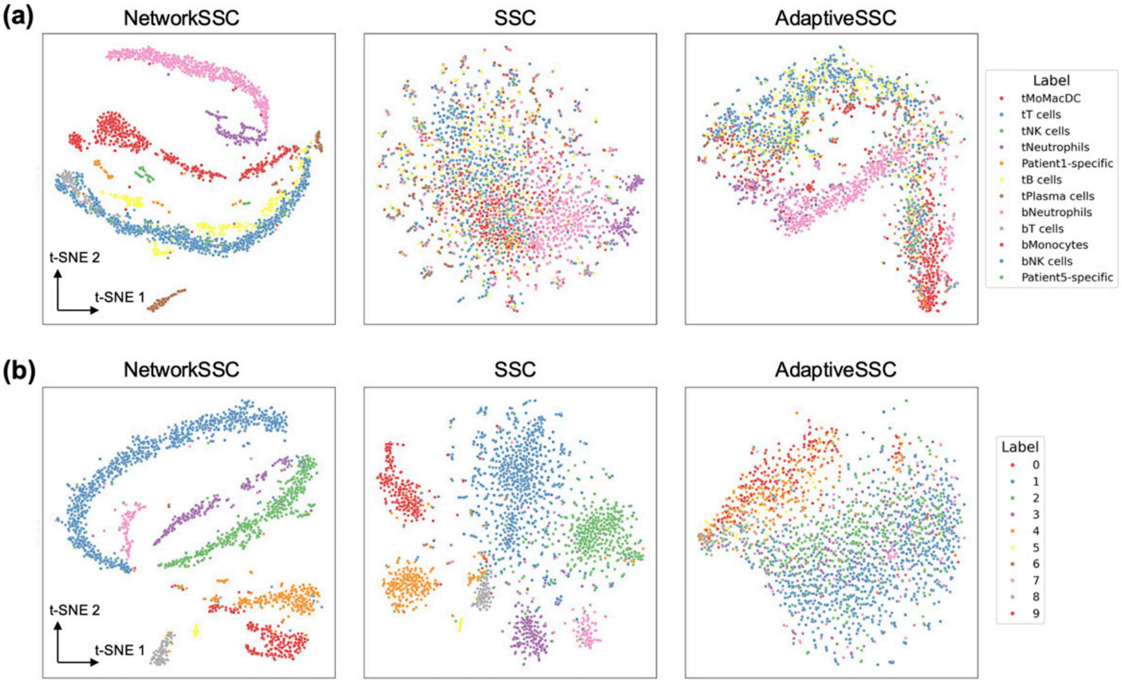

FIG. 3. Two-dimensional t-SNE embeddings of the affinity matrices learned by NetworkSSC, SSC, and AdaptiveSSC on the **(a)** Zilionis and **(b)** Baron datasets. Cells are color-coded based on their ground truth cell type labels. SSC, sparse subspace clustering; t-SNE: t-distributed stochastic neighbor embedding.

BC_CID4067, Zilionis, and Muraro, removing the sparsity regularization led to a slight fluctuation in NMI and ARI. In contrast, removing the Laplacian regularization resulted in substantial drops in both metrics. For example, NMI values dropped to near zero across all datasets, indicating that the clustering results were almost entirely independent of the ground truth. The ARI scores for the BC_CID4067 and Muraro datasets were even negative, suggesting worse-than-chance agreement. These findings collectively highlight the significance of incorporating gene network information. In other words, the Laplacian regularization term in NetworkSSC provides essential structural guidance and plays a critical role in the model's effectiveness.

NetworkSSC relies on the input gene network to capture functional relationships among genes. To evaluate the potential impact of external gene networks on NetworkSSC's performance, we conducted a sensitivity analysis on the Baron and Midbrain datasets by randomly removing edges from the dataset-specific gene networks and assessing the clustering results based on the resulting incomplete networks. As shown in Figure 4b, on both datasets, removing 30% of the network edges slightly reduced the NMI and ARI. However, the reduced NMI and ARI scores remained high and still outperformed those achieved by the original SSC and AdaptiveSSC. Furthermore, removing an additional 20% of the edges (from 30% to 50%) caused only negligible disturbance to the performance. These observations can be attributed to the fact that the originally constructed gene networks are already very sparse, with most elements in the adjacency matrix being zero. Nevertheless, NetworkSSC is able to effectively leverage the sparse network information for downstream dimensionality reduction and cell embedding. This ability to handle sparse gene networks directly contributes to NetworkSSC's robustness against incomplete or erroneous networks.

4.5. Downstream analysis

The graph-informed dimensionality reduction step in NetworkSSC enables automatic and efficient selection of important features, with gene importance scores directly embedded in the learned projection matrix Q. Specifically, we use the l_2-norm of the columns of Q to quantify the importance of each gene, where a larger l_2-norm indicates greater relevance for cell-type identification. Leveraging this property, we performed downstream analyses on the Baron and BC_CID4067 datasets. Since the Baron dataset underwent prescreening and included only the 2000 most variable genes, we extracted the top 200 genes based on our feature importance estimates. The BC_CID4067 dataset contained 18,176 genes, from which we retained the top 1000 important

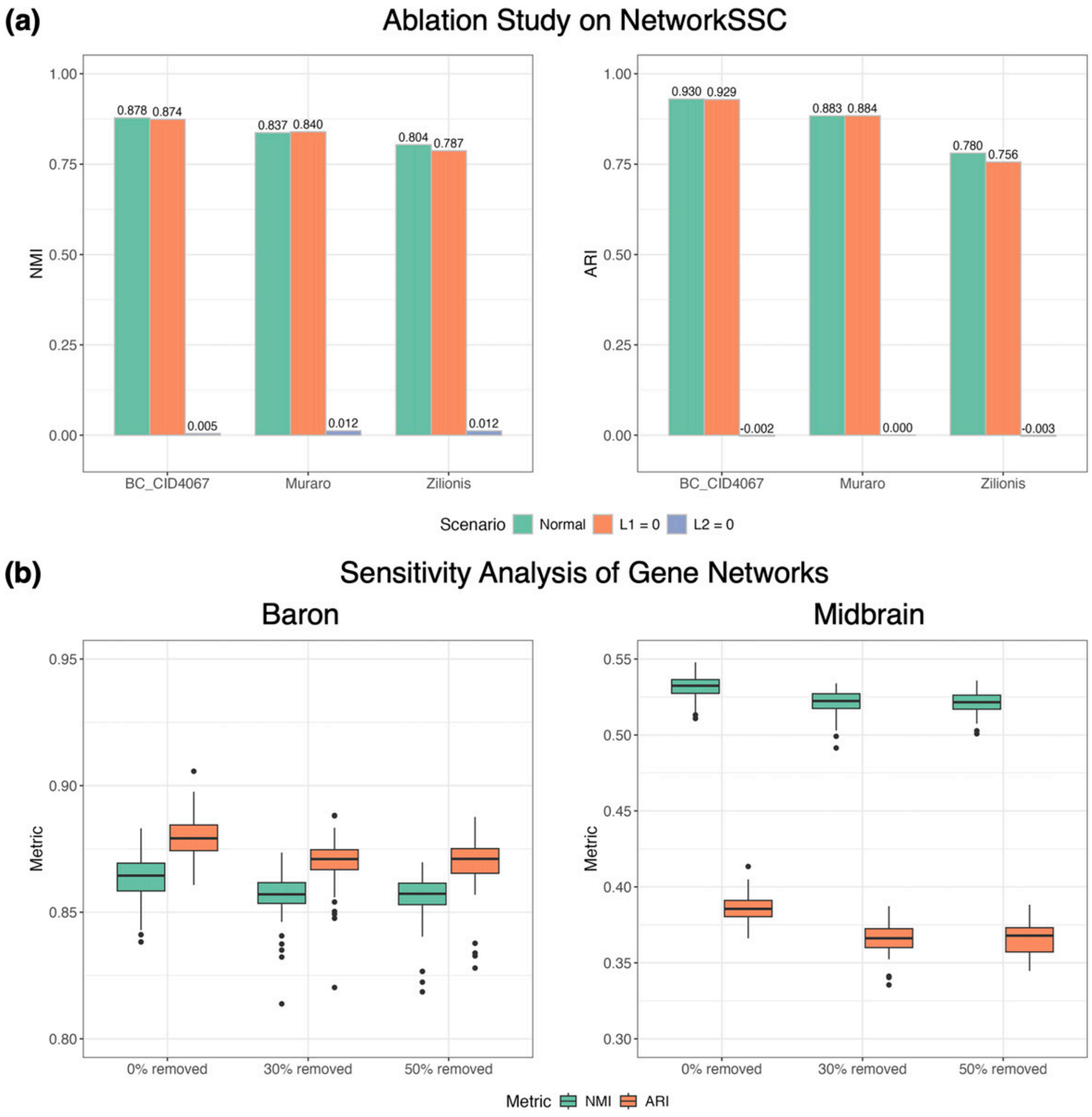

FIG. 4. Ablation study and sensitivity analysis of NetworkSSC. **(a)** Bar plots showing the average NMI and ARI of NetworkSSC under three scenarios: the original model, without sparsity regularization ($\lambda_1 = 0$), and without Laplacian regularization ($\lambda_2 = 0$), on the BC_CID4067, Muraro, and Zilionis datasets. **(b)** Box plots showing the NMI and ARI of NetworkSSC after randomly removing 0%, 30%, and 50% of the edges from the input gene network across 100 independent runs on the Baron and Muraro datasets.

genes. The identified genes and their interconnections are visualized and described in Supplementary Data S2 and Supplementary Figure S2.

We further analyzed the selected genes using gene set enrichment analysis (GSEA) via GOstat (Beissbarth and Speed, 2004). Given the redundancy often present in Gene Ontology (GO) terms, we first selected enriched GO terms based on a significance threshold defined by GOstat, and then removed redundant or overly broad terms. Specifically, we excluded GO terms containing more than 200 genes, as they tended to be nonspecific. Next, we examined the pairwise relationships among the remaining significant GO terms by evaluating their overlapping gene sets. If the shared genes accounted for more than a specified proportion of the larger GO term, we removed the larger term. Otherwise, if the shared genes accounted for more than that proportion of the smaller term, we removed the smaller one.

The top enriched biological processes in the Baron dataset identified by GSEA are visualized in the Chord diagram in Figure 5a. These processes contributed most significantly to the clustering results and were closely

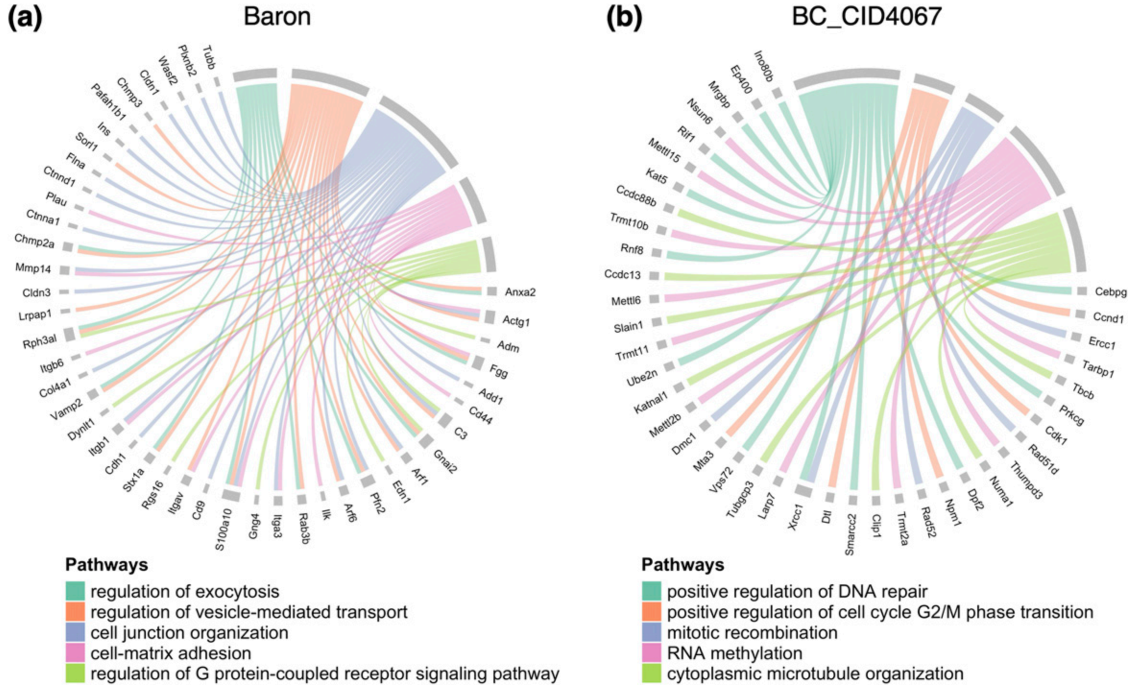

FIG. 5. Chord diagrams illustrating the top important genes identified by NetworkSSC and the top enriched biological processes identified by GSEA in the **(a)** Baron and **(b)** BC_CID4067 datasets. Genes are connected by edges to GO terms in which they are involved. GO, Gene Ontology; GSEA, gene set enrichment analysis.

associated with pancreatic function. For example, 10 of the selected important genes were involved in the process “regulation of exocytosis,” which plays a critical role in insulin and enzyme secretion by pancreatic cells (Lang, 1999). Additionally, 18 important genes were associated with the process “regulation of vesicle-mediated transport,” which is central to secretory granule trafficking in the endocrine pancreas (ARVAN and CASTLE, 1998). Overall, a substantial proportion of the selected genes were enriched in signaling-related processes, such as “regulation of G protein-coupled receptor signaling pathway” and “integrin-mediated signaling pathway,” highlighting their roles in nutrient response and overall regulation of pancreatic function.

Figure 5b illustrates the top enriched biological processes identified by GSEA in the BC_CID4067 dataset. Similar to the Baron dataset, the processes contributing significantly to the clustering results were also functionally relevant. For example, 13 important genes were associated with the process “positive regulation of DNA repair,” defects which are central to breast cancer development (Majidinia and Yousefi, 2017). In addition, five important genes were involved in the process “positive regulation of cell cycle G2/M phase transition,” a pathway commonly dysregulated during breast cancer proliferation (Cappelletti et al., 2000). Together, the results from both datasets suggest that the genes distinguishing cell clusters are functionally meaningful and consistent with the biological characteristics of the cells under study.

5. DISCUSSION

Cell type identification through unsupervised clustering is a fundamental challenge in scRNA-seq data analysis. In this study, we propose NetworkSSC, a method that integrates gene network information into SSC for scRNA-seq data, enabling biologically informed dimensionality reduction and improved clustering performance. We demonstrate that NetworkSSC consistently outperforms traditional SSC and its variants, achieving higher NMI and ARI across a range of scRNA-seq datasets. It also surpasses five additional benchmarking methods, including traditional, widely used, and deep learning-based approaches, on most experimental datasets. The clear clustering patterns observed in t-SNE visualizations further highlight NetworkSSC’s ability to enhance the interpretability of scRNA-seq data. Beyond model performance, we conducted ablation studies and sensitivity analyses to validate the model design and demonstrate its robustness to sparsity in the input gene network. In addition, downstream analyses confirmed the effectiveness of NetworkSSC in identifying important and functionally relevant genes involved in the clustering process. Together, these advancements establish NetworkSSC as a robust and valuable tool for high-resolution cell type discovery.

Despite its advantages, NetworkSSC has certain limitations. Although we examined its robustness to sparsity in the input gene network, the method's intrinsic reliance on external gene networks may introduce variability depending on the quality and coverage of the network. In addition, grid search for hyperparameter tuning can be computationally intensive, particularly for larger datasets. Future directions include extending NetworkSSC to integrate gene networks inferred from multiple modalities and to handle datasets with incomplete or noisy network information. Automating the parameter selection process could also improve efficiency and broaden applicability. Furthermore, integrating multiomics data may further enhance the clustering accuracy of NetworkSSC. Additional validation on more complex tissues and across species will also help strengthen its generalizability and impact.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their constructive and helpful comments on this article.

AUTHORS' CONTRIBUTIONS

This study was conceived and led by T.Y. and J.F. C.Y. and S.J. designed the model and algorithm with input from J.F. and T.Y. and implemented the NetworkSSC software. S.L., C.Y., and S.J. led the data analyses with evaluation from J.F. and T.Y. C.Y. wrote the article with feedback from all the coauthors.

CODE AND DATA AVAILABILITY

An open-source implementation of the NetworkSSC algorithm is available on GitHub: https://github.com/chen-yang-yuan/NetworkSSC. All scRNA-seq data analyzed in this study have been uploaded to Zenodo: https://zenodo.org/records/15603351.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

FUNDING INFORMATION

This work was supported by the National Key R&D Program of China (No. 2022ZD0116004), the National Natural Science Foundation of China (Nos. 62106211 and 62376236), the Guangdong Talent Program (No. 2021CX02Y145), the Guangdong Provincial Key Laboratory of Big Data Computing, and Shenzhen Science and Technology Program JCYJ20240813113536047 and ZDSYS20230626091302006.

SUPPLEMENTARY MATERIAL

Supplementary Data

REFERENCES

Amodio M, van Dijk D, Srinivasan K, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods 2019;16(11):1139–1145; doi: 10.1038/s41592-019-0576-7

Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Aspects Med 2018;59:114–122; doi: 10.1016/j.mam.2017.07.002

Arvan P, Castle D. Sorting and storage during secretory granule biogenesis: Looking backward and looking forward. Biochem J 1998;332(Pt 3):593–610; doi: 10.1042/bj3320593

Baron M, Veres A, Wolock SL, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst 2016;3(4):346–360.e4; doi: 10.1016/j.cels.2016.08.011

Beissbarth T, Speed TP. GOstat: Find statistically overrepresented gene ontologies within a group of genes. Bioinformatics 2004;20(9):1464–1465; doi: 10.1093/bioinformatics/bth088

Bergen V, Lange M, Peidli S, et al. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol 2020;38(12):1408–1414; doi: 10.1038/s41587-020-0591-3

Bolte J, Sabach S, Teboulle M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math Program 2014;146(1–2):459–494; doi: 10.1007/s10107-013-0701-9

Buettner F, Natarajan KN, Casale FP, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-Sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 2015;33(2):155–160; doi: 10.1038/nbt.3102

Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36(5):411–420; doi: 10.1038/nbt.4096

Cappelletti V, Fioravanti L, Miodini P, et al. Genistein blocks breast cancer cells in the G2M phase of the cell cycle. J Cell Biochem 2000;79(4):594–600; doi: 10.1002/1097-4644(20001215)79:4<594::AID-JCB80>3.0.CO;2-4

Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. SIAM Rev 2001;43(1):129–159; doi: 10.1137/S003614450037906X

Chen G, Ning B, Shi T. Single-Cell RNA-Seq technologies and related computational data analysis. Front Genet 2019; 10:317; doi: 10.3389/fgene.2019.00317

Chen L, Zhai Y, He Q, et al. Integrating deep supervised, self-supervised and unsupervised learning for Single-Cell RNA-Seq clustering and annotation. Genes (Basel) 2020;11(7):792; doi: 10.3390/genes11070792

Das J, Yu H. HINT: High-Quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 2012;6(1):92; doi: 10.1186/1752-0509-6-92

De Micheli AJ, Spector JA, Elemento O, et al. A Reference Single-Cell transcriptomic atlas of human skeletal muscle tissue reveals bifurcated muscle stem cell populations. Skelet Muscle 2020;10(1):19; doi: 10.1186/s13395-020-00236-3

Deng T, Chen S, Zhang Y, et al. A cofunctional grouping-based approach for non-redundant feature gene selection in unannotated Single-Cell RNA-Seq analysis. Brief Bioinform 2023;24(2):bbad042; doi: 10.1093/bib/bbad042

Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun 2018;9(1):2002; doi: 10.1038/s41467-018-04368-5

Efremova M, Vento-Tormo M, Teichmann SA, et al. CellPhoneDB: Inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc 2020;15(4):1484–1506; doi: 10.1038/s41596-020-0292-x

Elhamifar E, Vidal R. Sparse subspace clustering: Algorithm, theory, and applications. arXiv 2013; doi: 10.48550/arXiv.1203.1005

Eraslan G, Simon LM, Mircea M, et al. Single-Cell RNA-Seq denoising using a deep count autoencoder. Nat Commun 2019;10(1):390; doi: 10.1038/s41467-018-07931-2

Erfanian N, Heydari AA, Feriz AM, et al. Deep learning applications in single-cell genomics and transcriptomics data analysis. Biomed Pharmacother 2023;165:115077; doi: 10.1016/j.biopha.2023.115077

Grønbech CH, Vording MF, Timshel PN, et al. scVAE: Variational auto-encoders for single-cell gene expression data. Bioinformatics 2020;36(16):4415–4422; doi: 10.1093/bioinformatics/btaa293

Hartigan JA, Wong MA. Algorithm AS 136: A K-Means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics) 1979;28(1):100–108; doi: 10.2307/2346830

He W, Chen JX, Zhang W. Low-Rank representation with graph regularization for subspace clustering. Soft Comput 2017;21(6):1569–1581; doi: 10.1007/s00500-015-1869-0

He S, Fan J, Yu T. G3DC: A gene-graph-guided selective deep clustering method for single cell RNA-Seq data. Big Data Min Anal 2024;7(3):809–827; doi: 10.26599/BDMA.2024.9020011

He S, Wang L-H, Liu Y, et al. Single-Cell transcriptome profiling of an adult human cell atlas of 15 major organs. Genome Biol 2020;21(1):294; doi: 10.1186/s13059-020-02210-0

Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. Nat Biotechnol 2019;37(6):685–691; doi: 10.1038/s41587-019-0113-3

Hu D, Liang K, Zhou S, et al. scDFC: A deep fusion clustering method for single-cell RNA-Seq data. Brief Bioinform 2023;24(4):bbad216; doi: 10.1093/bib/bbad216

Hubert L, Arabie P. Comparing partitions. Journal of Classification 1985;2(1):193–218; doi: 10.1007/BF01908075

Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell-cell communication using cellchat. Nat Commun 2021;12(1):1088; doi: 10.1038/s41467-021-21246-9

Khaliq AM, Erdogan C, Kurt Z, et al. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. Genome Biol 2022;23(1):113; doi: 10.1186/s13059-022-02677-z

Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-Seq data. Nat Rev Genet 2019;20(5):273–282; doi: 10.1038/s41576-018-0088-9

Kiselev VY, Kirschner K, Schaub MT, et al. SC3: Consensus clustering of single-cell RNA-Seq data. Nat Methods 2017;14(5):483–486; doi: 10.1038/nmeth.4236

Kobak D, Berens P. The art of using T-SNE for single-cell transcriptomics. Nat Commun 2019;10(1):5416; doi: 10.1038/s41467-019-13056-x

La Manno G, Gyllborg D, Codeluppi S, et al. Molecular diversity of midbrain development in mouse, human, and stem cells. Cell 2016;167(2):566–580.e19; doi: 10.1016/j.cell.2016.09.027

La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. Nature 2018;560(7719):494–498; doi: 10.1038/s41586-018-0414-6

Lang J. Molecular mechanisms and regulation of insulin exocytosis as a paradigm of endocrine secretion. Eur J Biochem 1999;259(1–2):3–17; doi: 10.1046/j.1432-1327.1999.00043.x

Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell RNA-Seq data. Nat Commun 2018; 9(1):997; doi: 10.1038/s41467-018-03405-7

Li X, Wang K, Lyu Y, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-Seq analysis. Nat Commun 2020;11(1):2338; doi: 10.1038/s41467-020-15851-3

Lin C, Jain S, Kim H, et al. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res 2017;45(17):e156; doi: 10.1093/nar/gkx681

Majidinia M, Yousefi B. DNA repair and damage pathways in breast cancer development and therapy. DNA Repair (Amst) 2017;54:22–29; doi: 10.1016/j.dnarep.2017.03.009

Muraro MJ, Dharmadhikari G, Grün D, et al. A Single-Cell transcriptome Atlas of the human pancreas. Cell Syst 2016; 3(4):385–394.e3; doi: 10.1016/j.cels.2016.09.002

Ng A, Jordan M, Weiss Y. On Spectral Clustering: Analysis and an Algorithm. In: Advances in Neural Information Processing Systems. MIT Press; 2001.

Park S, Zhao H. Spectral clustering based on learning similarity matrix. Bioinformatics 2018;34(12):2069–2076; doi: 10.1093/bioinformatics/bty050

Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: A review. SIGKDD Explor Newsl 2004; 6(1):90–105; doi: 10.1145/1007730.1007731

Peng J, Wang X, Shang X. Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq Data. BMC Bioinformatics 2019;20(Suppl 8):284; doi: 10.1186/s12859-019-2769-6

Pfitzner D, Leibbrandt R, Powers D. Characterization and evaluation of similarity measures for pairs of clusterings. Knowl Inf Syst 2009;19(3):361–394; doi: 10.1007/s10115-008-0150-6

Polański K, Young MD, Miao Z, et al. BBKNN: Fast batch alignment of single cell transcriptomes. Bioinformatics 2020;36(3):964–965; doi: 10.1093/bioinformatics/btz625

Shao C, Höfer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. Bioinformatics 2017;33(2):235–242; doi: 10.1093/bioinformatics/btw607

Stark SG, Ficek J, Locatello F, et al.; Tumor Profiler Consortium. SCIM: Universal Single-Cell matching with unpaired feature sets. Bioinformatics 2020;36(Suppl_2):i919–i927; doi: 10.1093/bioinformatics/btaa843

Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq Whole-Transcriptome analysis of a single cell. Nat Methods 2009; 6(5):377–382; doi: 10.1038/nmeth.1315

Usoskin D, Furlan A, Islam S, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci 2015;18(1):145–153; doi: 10.1038/nn.3881

Wang J, Xia J, Wang H, et al. scDCCA: Deep contrastive clustering for single-cell RNA-Seq data based on autoencoder network. Brief Bioinform 2023;24(1):bbac625; doi: 10.1093/bib/bbac625

Wang B, Zhu J, Pierson E, et al. Visualization and analysis of single-cell RNA-Seq Data by Kernel-Based similarity learning. Nat Methods 2017;14(4):414–416; doi: 10.1038/nmeth.4207

Wei X, Li Z, Ji H, et al. EDClust: An EM–MM hybrid method for cell clustering in multiple-subject single-cell RNA sequencing. Bioinformatics 2022;38(10):2692–2699; doi: 10.1093/bioinformatics/btac168

Wolf FA, Angerer P, Theis FJ. SCANPY: Large-Scale single-cell gene expression data analysis. Genome Biol 2018; 19(1):15; doi: 10.1186/s13059-017-1382-0

Wu SZ, Al-Eryani G, Roden DL, et al. A Single-Cell and spatially resolved atlas of human breast cancers. Nat Genet 2021;53(9):1334–1347; doi: 10.1038/s41588-021-00911-1

Zeng Y, Wei Z, Zhong F, et al. A Parameter-Free deep embedded clustering method for single-cell RNA-Seq data. Brief Bioinform 2022;23(5):bbac172; doi: 10.1093/bib/bbac172

Zheng R, Li M, Liang Z, et al. SinNLRR: A robust subspace clustering method for cell type detection by non-negative and low-rank representation. Bioinformatics 2019;35(19):3642–3650; doi: 10.1093/bioinformatics/btz139

Zheng R, Liang Z, Chen X, et al. An adaptive sparse subspace clustering for cell type identification. Front Genet 2020; 11:407; doi: 10.3389/fgene.2020.00407

Zilionis R, Engblom C, Pfirschke C, et al. Single-Cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. Immunity 2019;50(5):1317–1334.e10; doi: 10.1016/j.immuni.2019.03.009

Zuo C, Chen L. Deep-Joint-Learning analysis model of single cell transcriptome and open chromatin accessibility data. Brief Bioinform 2021;22(4):bbaa287; doi: 10.1093/bib/bbaa287

Žurauskienė J, Yau C. pcaReduce: Hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics 2016;17(1):140; doi: 10.1186/s12859-016-0984-y

Address correspondence to:
Jicong Fan
School of Data Science
The Chinese University of Hong Kong
Shenzhen (CUHK-Shenzhen)
Shenzhen
Guangdong 518172
China

E-mail: fanjicong@cuhk.edu.cn

Tianwei Yu
School of Data Science
The Chinese University of Hong Kong
Shenzhen (CUHK-Shenzhen)
Shenzhen
Guangdong 518172
China

E-mail: yutianwei@cuhk.edu.cn