Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

scAFGCC: An augmentation-free graph contrastive clustering method for scRNA-seq data analysis☆

Shengwen Tian[a,b,*], Yu Wang[c], Yutian Wang[c], Cunmei Ji[c], Jiancheng Ni[d]

[a] School of Computer Science and Technology, Beijing Institute of Technology, 100081, Beijing, China
[b] Zhongguancun Academy, 100094, Beijing, China
[c] School of Cyber Science and Engineering, Qufu Normal University, 273165, Shandong, China
[d] Network Information Center, Qufu Normal University, 273165, Shandong, China

ARTICLE INFO

Communicated by R. Vimieiro

Keywords:
scRNA-seq
Augmentation-free contrastive learning
Graph convolutional network
Clustering
Visualization

ABSTRACT

The emergence of single-cell RNA sequencing (scRNA-seq) has provided researchers with a powerful tool to investigate cell heterogeneity and human diseases at the level of individual cells. Cell clustering is a crucial step in scRNA-seq data analysis to identify marker genes and recognize cell types. However, scRNA-seq data present challenges for clustering tasks due to their high dimensionality, sparsity, and noise. Although some contrastive learning methods have achieved good results in clustering scRNA-seq data, they are highly sensitive to data augmentation schemes. Here, we propose scAFGCC, a novel augmentation-free graph contrastive clustering method that combines graph convolutional network (GCN) and contrastive learning to exploit inter-cell relationships. scAFGCC does not require data augmentations or negative samples to learn graph representations. Instead, we generate positive samples by exploring the local structural information and the global semantics of the target nodes. We integrate feature representation learning with clustering tasks. Additionally, we introduce a reconstruction module that pretrains the model, facilitating faster training and improved performance. Our experiments on 24 simulated and 13 real datasets show that scAFGCC outperforms seven state-of-the-art methods in terms of accuracy and robustness. We also apply scAFGCC to downstream tasks such as cell annotation and marker gene identification.

1. Introduction

Single cell RNA sequencing (scRNA-seq) is a powerful technique that enables the transcription of RNA in individual cells into cDNA and allows for high-throughput sequencing. This technology provides a means to quantify gene expression at the single-cell level [1,2]. Compared to bulk RNA sequencing techniques, it can better reflect the differences and similarities between different cells, providing a powerful tool for studying the biological processes of individual cells [3]. Cell clustering is one of the most important steps in scRNA-seq data analysis [4,5]. ScRNA-seq can identify new tumor cells and reveal the heterogeneity of tumor cells [6]. Additionally, it can infer the trajectory of cell differentiation, shedding light on the developmental processes occurring within cells [7]. It can reveal the subtypes of cells and identify the expression patterns and functional differences between different subtypes, thereby providing a better understanding of the diversity and complexity of cells [8,9]. Due to the technical limitations of the sequencing process and the inherent biological factors of the data, scRNA-seq data contain a large amount of noise of varying degrees [10,11]. In addition, it has high dimensionality and sparsity, which make cell clustering more complex [12]. Therefore, it is urgent to study and develop effective clustering methods with higher accuracy and broader applicability.

Numerous clustering methods developed specifically for analyzing scRNA-seq data have been widely used in research. SC3 [13] is based on k-means clustering, which integrates clustering results from multiple similarity metrics and feature transformation techniques. The graph-based Seurat [14] first transforms the data into a graph and then applies graph clustering algorithms to identify clusters by discovering subgraphs on a k-nearest neighbor (KNN) graph. The multi-view clustering with graph learning (MCGL) method [15] constructs multiple feature spaces from the raw scRNA-seq data and jointly performs graph learning, graph factorization, and clustering in a unified optimization framework. There

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author at: Zhongguancun Academy, 100094, Beijing, China.
Email address: swtian@bit.edu.cn (S. Tian).

https://doi.org/10.1016/j.neucom.2025.131698
Received 26 May 2025; Received in revised form 25 August 2025; Accepted 27 September 2025
Available online 10 October 2025
0925-2312/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

are subspace clustering methods based on low-rank representation, such as SinNLRR [16] and SCCLRR [17]. The similarity matrices learned from these models have greater advantages in distinguishing cell subpopulations, but they are not suitable for highly sparse data. SIMLR [18] uses multi-kernel learning to transform scRNA-seq data into a low-dimensional space represented as a graph structure, and then applies spectral clustering for cell grouping. However, some methods employ straightforward linear dimensionality reduction followed by clustering. This process may lead to the loss of crucial information as it relies on the data structure or overlooks the manifold structure of the original data, potentially impacting the calculation of the similarity matrix. Additionally, due to the large size and complexity of scRNA-seq datasets, overly complex clustering methods can result in excessive computation and time costs.

To address the challenges posed by the increasing scale of scRNA-seq data, deep learning-based methods have been proposed [19]. For example, scGDC [20] simultaneously learns deep representations and the affinity graph of cells by extending autoencoders with a self-representation layer, and leverages generative adversarial learning to enhance subspace discrimination for rare and heterogeneous cell types. The scDCC [21] incorporates domain-specific knowledge as constraint information through a loss function during the clustering process, enhancing the interpretability of the clustering results. The scziDesk [22] combines weighted soft k-means clustering with a denoising autoencoder to enhance the association of similar cells and cluster cell populations within the learned latent space. The scVI [23], which derives a probabilistic representation of scRNA-seq data from a deep generative model. The DCA [24] introduces the ZINB model into the autoencoder and uses the features extracted by a multi-layer neural network for data clustering. However, these methods only focus on the intrinsic feature information of the data and ignore the relationships between cells, leading to unsatisfactory feature learning effects.

To better capture the topological structure among cells, scGNN [25] utilizes graph neural networks to represent and aggregate cell–cell relationships, and employs a truncated mixture of Gaussians model to simulate heterogeneous gene expression patterns. However, scGNN may introduce false noise edges mixed with true edges, potentially resulting in biased clustering results. The contrastive-sc [26] employs contrastive learning to acquire low-dimensional representations of samples. It learns the similarities and differences between samples and uses this information to cluster cells effectively. The scGCC [27] introduces innovative data augmentation techniques and integrates graph attention networks into contrastive learning to extract feature information. The UMGRL framework [28] employs a bi-level optimization strategy with dual weight networks to enhance multi-view graph contrastive learning, enabling adaptive importance weighting at the node, graph, and edge levels. The DT3OR framework [29] addresses distributional shifts through dual test-time training, using self-distillation and contrastive tasks to improve model adaptability in dynamic environments. These recent advances highlight the growing potential of contrastive learning, particularly in graph-based representation and adaptation tasks. However, existing contrastive learning methods are highly sensitive to data augmentation techniques.

Therefore, this paper introduces a novel and effective graph contrastive learning-based clustering method called scAFGCC, which does not require data augmentation or negative samples. Specifically, this model undergoes a two-stage training process. In the pre-training stage, we fine-tune the parameters of the GCN using the reconstruction module. In the subsequent training stage, we further optimize the parameters of the GCN utilizing the augmentation-free graph contrastive learning module and the clustering module, aiming to achieve optimal performance. Our approach eliminates the need for negative sample pairs and relies solely on the local structural information and global semantic information of target nodes to explore their positive sample space. UMAP reduces dimensions by preserving local relationships between data points, thereby better retaining the structural information among samples in single-cell data. It enables faster processing of large-scale single-cell datasets while maintaining good dimensionality reduction performance. By visualizing single-cell data as points in two-dimensional space, UMAP makes the distribution of data more understandable and interpretable [30].

2. Methods

2.1. The framework of scAFGCC

The workflow of scAFGCC is depicted in Fig. 1 and comprises three key components: a reconstruction module, an augmentation-free graph contrastive learning module, and a clustering module. Specifically, the reconstruction module pretrains the model by reconstructing the gene expression matrix and adjacency matrix to speed up training and capture clustering-friendly features. The augmentation-free graph contrastive learning module utilizes both local and global semantic information to identify positive samples, optimizing the model to minimize distances between similar cells and maximize distances between different cell types. Meanwhile, we combine the contrastive learning module and clustering module to fully extract meaningful features. The model training process involves two stages: first, we pretrain the model using the reconstruction module. The loss function during this pretraining stage is shown as follows:

$$L_{pre-train} = \min\left(L_{recon}\right), \tag{1}$$

$$L_{recon} = (1-\alpha)\, L_{recon-adj} + \alpha L_{recon-exp}. \tag{2}$$

Where $L_{recon-adj}$ and $L_{recon-\mathrm{exp}}$ denote the reconstruction loss of adjacency matrix and expression matrix, respectively. The parameter α is a coefficient that balances the two losses (set to 0.2 in our experiment). Second, we utilize the contrastive loss $\left(L_{cl}\right)$ and the Kullback-Leibler (KL) loss $\left(L_{kl}\right)$ to further train the model. The loss function during this training stage is represented as follows:

$$L_{train} = \min\left[(1-\beta)\, L_{cl} + \beta L_{kl}\right]. \tag{3}$$

Where L_{cl} and L_{kl} represent contrastive loss and KL loss, respectively. The parameter β is also a coefficient that balances the two terms (set to 0.2 in our experiment). The detailed expressions of these losses are illustrated in the following sections.

2.2. Data preprocessing

Due to the high noise and sparsity characteristics of scRNA-seq data, a significant number of entries in the matrix where genes are not expressed may cause noise interference and generate invalid information, so data preprocessing steps are necessary. We use the Scanpy [31] tool to accomplish this task. First, cells with low total gene expression and genes with few counts expressed in cells are filtered out. Second, a logarithmic transformation is applied to the filtered expression matrix to mitigate the adverse effects caused by sequencing depth differences. Third, to facilitate effective comparison between cells, the scRNA-seq data matrix is normalized. Finally, highly variable genes are selected and the count values are scaled to zero mean and unit variance.

2.3. Construction of scRNA-seq data graph

To represent the cell relationships in the expression matrix X, we define an undirected graph for the scRNA-seq data, which is essentially a cell connectivity graph represented as an adjacency matrix A in the model.To align with the positive sample selection strategy in the downstream contrastive clustering module, A is constructed using the k-nearest neighbor (KNN) algorithm. InA, each node corresponds to a cell and edges represent relationships between cells. This design offers the dual benefits of computational efficiency and effective capture of similar biological features. Specifically, if cell j is one of the k-nearest nodes to cell i, an edge is assigned between cell j and i (i.e., $A_{ij} = A_{ji} = 1$;

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

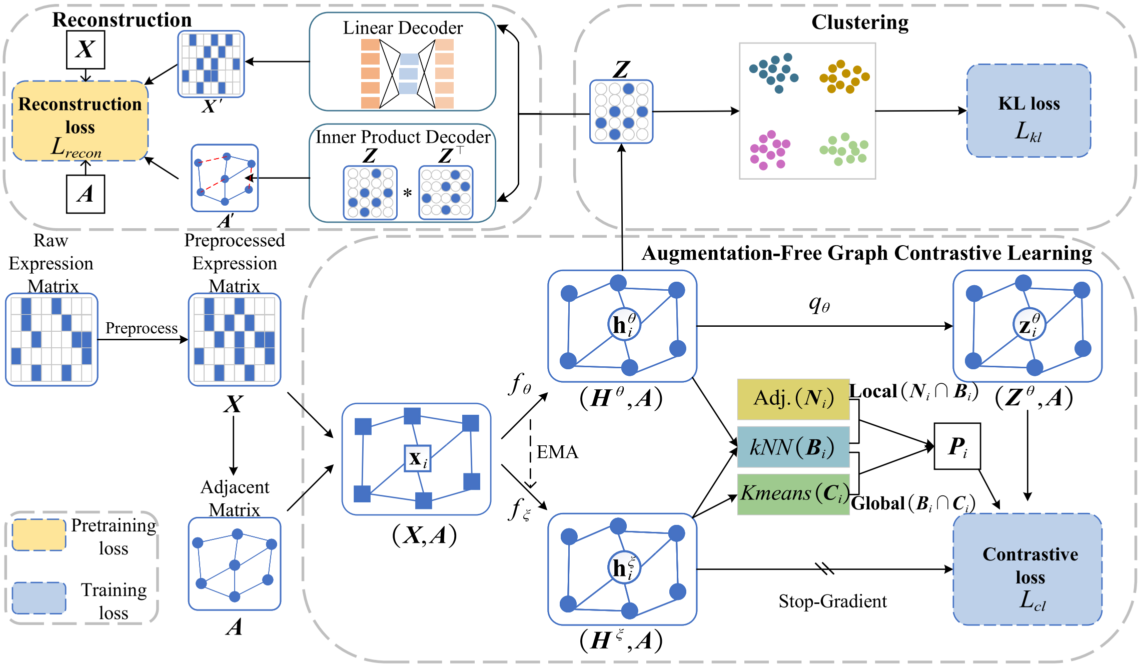

Fig. 1. The workflow of scAFGCC, which consists of three main components: the reconstruction module, the augmentation-free graph contrastive learning module, and the clustering module. First, the reconstruction module pretrains the model by reconstructing the gene expression matrix $\boldsymbol{X}$ and adjacency matrix $\boldsymbol{A}$. Then, the augmentation-free graph contrastive learning module utilizes local structural and global semantic information to identify positive samples. Finally, integrate the contrastive learning module and clustering module together to extract meaningful features by minimizing the contrastive loss and KL loss.

otherwise, $A_{ij} = A_{ji} = 0$). We use cosine similarity to measure the distance between cells, which offers the advantage of maintaining a certain robustness and accuracy in handling high-dimensional sparse data and different distribution situations [32].

2.4. Reconstruction module

To extract as many common features as possible from the training data and reduce the burden of the model on specific learning tasks, we add a reconstruction module in scAFGCC. This module comprises two independent decoders: the Linear Decoder (LD) and the Inner Product Decoder (PD). We fully utilize the cellular feature information and the topological structure between cells to reconstruct the gene expression matrix X and restore the adjacency matrix A. By minimizing the loss between the original data and the reconstructed data, the model is pretrained to capture more clustering-friendly features.

Specifically, LD consists of two linear layers and its output is $X' = \mathrm{LD}(Z)$. The reconstruction loss is defined as:

$$L_{\text{recon-exp}} = \|X - X'\|_2^2 . \tag{4}$$

Z obtains A' through PD, with the formula: $A' = \mathrm{PD}(Z) = \sigma\left(ZZ^{\top}\right)$, where $\sigma(\cdot)$ is an activation function. The reconstruction loss of the adjacency matrix adopts the binary cross-entropy loss function with negative sampling:

$$L_{\text{recon-adj}} = -\frac{1}{E}\left(\sum_i \log\left(\hat{y}_i^{\text{pos}}\right) + \sum_j \log\left(1 - \hat{y}_j^{\text{neg}}\right)\right), \tag{5}$$

where $\hat{y}_i^{\text{pos}}$ and $\hat{y}_j^{\text{neg}}$ are the outputs of the PD, representing positive edges and randomly sampled negative edges. E is the number of edges in the graph.

2.5. Augment-free graph contrastive clustering module

To achieve a reduction in intra-cellular distance among the same cell types and an expansion in inter-cellular distance between different cell types, we introduce the graph contrastive learning module based on the idea of Augment-Free Graph Representation Learning (AFGRL) [33]. AFGRL avoids ignoring structural information due to data augmentation and fully leverages the global semantics and local structural information of scRNA-seq data to better capture pairwise proximity relationships between cells. Unlike AFGRL, we incorporate a reconstruction module and a clustering module. Together, they complement the augmentation-free contrastive objective and are specifically designed to enhance clustering performance for scRNA-seq data. In scAFGCC, we use the GCN to implement two independent encoders: the online encoder $f_\theta(\cdot)$ and the target encoder $f_\xi(\cdot)$. These encoders take the preprocessed expression matrix X and the adjacency matrix A as inputs, generating the online representation $H^\theta = f_\theta(X, A)$ and target representation $H^\xi = f_\xi(X, A)$, respectively. The i-th row elements of H^θ and H^ξ correspond to the distinct node embeddings of the same sample cell x_i, denoted as h_i^θ and h_i^ξ. $f_\xi(\cdot)$ is updated by the exponential moving average (EMA) of $f_\theta(\cdot)$ [34].

For each cell $x_i \in X$, we need to identify the set of real positive nodes P_i that will ultimately be used to calculate the contrastive loss. Nodes similar to x_i are screened and marked as the nearest neighbor node set B_i. The screening process consists of two steps. The first step involves calculating the cosine similarity between cell x_i and all other remaining cells x_j, with the following formula:

$$sim\left(x_i, x_j\right) = \frac{h_i^\theta \cdot h_j^\xi}{\|h_i^\theta\|\|h_j^\xi\|}, \forall x_i \in X. \tag{6}$$

The second step involves using the KNN algorithm to find the top k ($topk$) nodes that are similar to cell x_i, to obtain B_i.

However, relying solely on K-nearest neighbor (KNN) search to establish cell relationships has limitations, as it is unidirectional and does not incorporate label information during computation. This can lead to noise in the nearest neighbor node set B_i, necessitating the removal of false positive samples. We address this issue from two perspectives. From a global perspective, certain nodes may exhibit semantic similarity to x_i and belong to the same cluster, but they may not share an edge in A. To

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

capture the semantic information of the original data, we apply the k-means clustering algorithm to the node embeddings h_i^ξ, resulting in a cell collection named C_i. C_i represents a cluster that includes cells similar to x_i and preserves the global semantic characteristics of the data. Since the k-means is sensitive to the initialization of the cluster centroids, the clustering process is run multiple times, and C_i is the union of the results to ensure robustness. From a local perspective, based on the adjacency matrix A, we can identify and label the nodes that are directly connected to x_i as N_i. The final P_i can be calculated as follows:

$$P_i = \left(B_i \cap C_i\right) \cup \left(B_i \cap N_i\right). \tag{7}$$

The size of the positive sample set P_i is adaptive to both the dataset scale and the diversity of cell types, while its quality is guaranteed through the integration of KNN-based local similarity and K-means-based global semantic consistency.

scAFGCC enhances the representation of the target node x_i and generates better node embeddings. The contrastive loss is defined by maximizing the cosine distance between the target node x_i and P_i. This approach is utilized to quantify the similarity between them, and the contrastive loss is given by the following expression:

$$L_{cl} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{x_j\in P_i}\frac{z_i^\theta f_\xi\left(x_j, A\right)^T}{\left\|z_i^\theta\right\|\left\|f_\xi\left(x_j, A\right)\right\|}, \tag{8}$$

where $z_i^\theta = q_\theta\left(f_\theta\left(x_i, A\right)\right), \forall x_i \in X$. $q_\theta(\cdot)$ represents the predictor network. N is the number of cell nodes.

2.6. Clustering module

Drawing on DEC [35], it can be concluded that clustering performance can be significantly improved by learning the mapping from the data space to a low-dimensional feature space and iteratively optimizing the clustering objective. First, soft assignments between clustering centers μ_j and embedding nodes z_i are calculated using Student's t-distribution:

$$q_{ij} = \frac{\left(1+\left\|z_i-\mu_j\right\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1+\left\|z_i-\mu_{j'}\right\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}, \tag{9}$$

where z_i represents the latent representation of the cell node x_i, and α denotes the degrees of freedom of the t-distribution (set to 1). Based on q_{ij}, we square it and normalize it using soft cluster frequencies to obtain the auxiliary distribution p_{ij}, which is expressed by the following formula:

$$p_{ij} = \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_{j'} q_{ij'}^2/\sum_i q_{ij'}}. \tag{10}$$

Finally, we define the clustering loss as the KL divergence between the soft assignments q_i and the auxiliary distribution p_i, as shown in the following formula:

$$L_{kl} = KL(P\|Q) = \sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}}. \tag{11}$$

3. ScRNA-seq datasets

To thoroughly evaluate the performance of our model, we collected 24 simulated datasets and 13 real-world datasets. For simulated datasets, we generated 12 balanced datasets and 12 imbalanced datasets using the R package “Splatter” [36]. The balance of the datasets is primarily determined based on whether the cell clusters have the same size. For the balanced datasets, they contain 4, 8, 12, or 16 cell clusters, with each cluster consisting of 250 cells. Each cell has 2500 genes, and the data sparsity ranges from 28 % to 47 %. As for the imbalanced datasets, each dataset includes 3000 cells and 2500 genes. They have the same number of cell clusters as the balanced datasets, and the data sparsity ranges from 29 % to 46 %. Further detailed information regarding balanced and imbalanced datasets is recorded in Supplementary Tables S1 and S2.

Furthermore, we collect 13 real-world scRNA-seq datasets from publicly accessible platforms. These datasets are diverse in terms of tissue types, including human brain, human pancreas, mouse embryo, and mouse diaphragm, as well as in terms of biological systems, such as different cell types and varying numbers of cell clusters. The datasets, which come from different sequencing platforms, provide a broad range of cell numbers ranging from 268 to 3605 to evaluate our model's performance. Specifically, the 13 datasets consist of Klein [37], Deng [38], Diaphragm [39], Camp1 [40], Darmanis [41], Muraro [42], Pollen [43], Zeisel [44], Human1, Human2, Human3 [45], Tosches [46] and Shekhar [47] each with detailed information shown in Table 1 and Supplementary Table S3.

4. Evaluation metrics

The evaluation metrics for clustering can be categorized into external evaluation metrics and internal evaluation metrics. External evaluation metrics assess the quality of clustering results using known data labels or class information. In this paper, we employ the external evaluation metrics: Adjusted Rand Index (ARI) [48] and Normalized Mutual Information (NMI) [49] to assess model performance.

The ARI can evaluate the similarity of two assignments by comparing all the sample pairs while ignoring permutations. It evaluates the effectiveness of clustering by calculating the number of the sample pairs that are distributed in the same or different class clusters in the predicted label set $P = \left\{P_1, P_2, \ldots, P_k\right\}$ and real label set $T = \left\{T_1, T_2, \ldots, T_k\right\}$. It is defined specifically as:

$$ARI(P,T) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right] - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}, \tag{12}$$

where n_{ij} represents the number of cells that are present in both P_i and T_j. a_i and b_j represent the respective cell counts in P_i and T_j. ARI ranges from -1 to 1. The larger the value, the more consistent the clustering results are with the real situation.

The NMI is commonly used in clustering, which measures the similarity between the two clustering results, ignoring the permutations. The NMI ranges from 0 to 1. The higher the NMI, the more accurate the

Table 1
Detailed information of the 13 real scRNA-seq datasets.

Dataset	Cells	Genes	Cluster number	Species	Accession ID
Klein	2712	24,175	4	Mouse	GSE65525
Deng	268	22,431	6	Mouse	GSE45719
Diaphragm	1858	22,966	6	Mouse	GSM4505405
Camp1	777	19,020	7	Human	GSE81252
Darmanis	420	22,083	8	Human	GSE67835
Muraro	2126	19,127	10	Human	GSE85241
Pollen	301	23,730	11	Human	SRP041736
Zeisel	3005	19,958	12	Mouse	GSE60361
Human1	1937	20,125	14	Human	GSM2230757
Human2	1724	20,125	14	Human	GSM2230758
Human3	3605	20,125	14	Human	GSM2230759
Tosches	18,664	23,500	15	Reptilian	unknown
Shekhar	26,830	13,166	18	Mouse	unknown

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

Table 2
Comparison of ARI values for ten clustering methods across 13 real scRNA-seq datasets.

Dataset	K-means	scziDesk	scVI	Seurat	SIMLR	SOUP	contrastive-sc	scNAME	scBGEDA	scAFGCC
Human1	0.436 ± 0.08	0.724 ± 0.13	0.442 ± 0.02	0.500 ± 0.00	0.293 ± 0.02	0.268 ± 0.01	0.684 ± 0.09	0.776 ± 0.10	0.820 ± 0.03	**0.961** ± **0.01**
Human2	0.368 ± 0.12	0.764 ± 0.13	0.565 ± 0.02	0.442 ± 0.00	0.366 ± 0.02	0.597 ± 0.14	0.775 ± 0.09	0.807 ± 0.07	0.881 ± 0.07	**0.929** ± **0.02**
Human3	0.650 ± 0.21	0.822 ± 0.08	0.703 ± 0.02	0.514 ± 0.00	0.506 ± 0.03	0.213 ± 0.02	0.611 ± 0.02	0.745 ± 0.12	0.750 ± 0.04	**0.961** ± **0.00**
Diaphragm	0.148 ± 0.02	0.361 ± 0.01	0.312 ± 0.02	0.316 ± 0.00	0.402 ± 0.02	0.431 ± 0.00	0.451 ± 0.02	**0.791** ± **0.08**	0.610 ± 0.01	0.771 ± 0.08
Camp1	0.669 ± 0.12	0.755 ± 0.00	0.662 ± 0.03	0.599 ± 0.00	0.614 ± 0.03	0.546 ± 0.00	0.725 ± 0.03	0.743 ± 0.12	0.607 ± 0.03	**0.788** ± **0.00**
Darmanis	0.580 ± 0.10	0.691 ± 0.05	0.718 ± 0.07	0.628 ± 0.00	0.632 ± 0.06	0.343 ± 0.08	0.704 ± 0.02	0.765 ± 0.11	0.717 ± 0.07	**0.838** ± **0.01**
Deng	0.596 ± 0.27	0.867 ± 0.02	0.335 ± 0.04	0.325 ± 0.00	0.610 ± 0.08	0.824 ± 0.01	0.845 ± 0.02	0.770 ± 0.10	**0.876** ± **0.00**	0.834 ± 0.01
Klein	0.705 ± 0.04	0.815 ± 0.02	0.541 ± 0.05	0.778 ± 0.00	0.510 ± 0.08	0.471 ± 0.00	0.754 ± 0.01	0.747 ± 0.12	0.843 ± 0.04	**0.962** ± **0.00**
Muraro	0.738 ± 0.09	0.789 ± 0.07	0.485 ± 0.02	0.446 ± 0.00	0.321 ± 0.03	0.536 ± 0.08	0.815 ± 0.01	0.772 ± 0.10	0.849 ± 0.09	**0.904** ± **0.00**
Pollen	0.806 ± 0.06	0.881 ± 0.04	0.880 ± 0.04	0.767 ± 0.00	0.806 ± 0.02	0.430 ± 0.02	0.898 ± 0.01	0.797 ± 0.08	0.868 ± 0.03	**0.918** ± **0.02**
Zeisel	0.484 ± 0.01	0.608 ± 0.04	0.364 ± 0.03	0.327 ± 0.00	0.613 ± 0.05	0.509 ± 0.01	0.540 ± 0.06	0.820 ± 0.02	0.623 ± 0.01	**0.820** ± **0.01**
Tosches	0.693 ± 0.13	0.617 ± 0.07	0.566 ± 0.02	0.444 ± 0.00	0.575 ± 0.04	0.480 ± 0.01	0.392 ± 0.04	0.584 ± 0.06	0.712 ± 0.02	**0.891** ± **0.01**
Shekhar	0.507 ± 0.07	0.351 ± 0.06	0.467 ± 0.03	0.714 ± 0.00	0.823 ± 0.05	0.752 ± 0.03	0.463 ± 0.03	0.712 ± 0.09	0.705 ± 0.05	**0.979** ± **0.02**

partition. The formula is shown as follows:

$$NMI\left(P,T\right)=\frac{2\times I\left(P;T\right)}{H\left(P\right)+H\left(T\right)}, \tag{13}$$

where $I\left(P;T\right)$ represents the mutual information, P and T stand for the predicted and true labels of cells, respectively, and $H\left(\cdot\right)$ denotes the cross entropy.

5. Results

To assess the performance of scAFGCC, we select nine state-of-the-art methods for comparison, including traditional PCA-based dimensionality reduction with K-means and Seurat, semisoft clustering method SOUP, multi-kernel similarity learning method SIMLR, as well as three other deep learning-based methods. The detailed information regarding the comparison methods is presented in Supplementary Table S4.

scAFGCC is implemented in Python 3 using the PyTorch framework. Both the online encoder and the target encoder consist of a single layer of GCN. The input dimension is determined by the number of genes, while the output dimension is set to 512, which serves as the dimensionality of the embedding vectors. The predictor network consists of two linear layers with node sizes of (1024, 512). scAFGCC is initially pretrained for 50 epochs and then undergoes a formal training phase of 100 epochs using the AdamW optimizer. The initial learning rate is set to 0.001. In constructing the scRNA-seq graph, k represents the number of neighbors, initially set to 10. $topk$ specifies the selection of the top k nodes with the highest similarity, set to 3. We employ the HDBSCAN [50] algorithm for cell clustering.

5.1. Clustering analysis in simulated datasets

We conduct cluster analysis on 12 balanced and 12 imbalanced datasets using six state-of-the-art methods. The clustering performance is evaluated with ARI and NMI metrics. For each dataset, each method is independently run 5 times, and the average is taken as the final result. The experimental results are presented as box plots in Supplementary Figure S1.

From the figure, it can be seen that our model outperforms the others, whether on balanced or imbalanced datasets. On balanced datasets, the average ARI of our model is 0.04 higher than that of contrastive-sc and 0.17 higher than that of scziDesk. The average NMI of our model is 0.07 higher than that of contrastive-sc and 0.17 higher than that of scziDesk. On imbalanced datasets, the average ARI and NMI of our model are 0.14 and 0.11 higher than those of contrastive-sc, respectively. We also observe that the clustering performance of all methods declines from balanced to imbalanced datasets, which is reasonable. Imbalanced datasets refer to a situation where there is a significant difference in the number of samples among different classes. Usually, the minority class contains far fewer samples compared to the majority class, potentially resulting in clustering algorithms performing poorly in recognizing and separating the minority classes. However, our model still maintains high clustering performance. Real-world data is often imbalanced, and our model exhibits high accuracy and robustness to it.

5.2. Clustering analysis in real datasets

In this experiment, we apply the scAFGCC model to 13 real datasets that are annotated with ground truth labels. We then compare its clustering performance with nine state-of-the-art models to assess its effectiveness. We use ARI and NMI as evaluation metrics. To ensure the stability and reliability of the results, we run all methods 10 times using the default parameters. We then calculate the average value for each method and measure the standard deviation as the error bar to assess the robustness of the models. Next, we perform parameter optimization on contrastive-sc and scziDesk to update the above results.

Table 2 and Supplementary Table S5 present the quantitative values of two distinct metrics for ten different methods across a total of 13 real scRNA-seq datasets. We can observe that scAFGCC outperforms the other nine clustering methods on almost all datasets. Specifically, except for Deng and Diaphragm, scAFGCC achieves the highest ARI and NMI among all the compared methods, with ARI and NMI scores both exceeding 0.77 and 0.70, respectively. Despite not achieving the highest performance in Deng, scAFGCC still ranks fourth in terms of ARI values. This could be because the Deng dataset has a small number of samples, causing the graph to be dense and the model to overfit. In particular, for Human1, Human2, and Human3, scAFGCC demonstrates significantly higher ARI values compared to contrastive-sc, with differences of 0.28, 0.15, and 0.35, respectively. For Darmanis and Pollen, the ARI values of scAFGCC are 0.15 and 0.04 higher than those of scziDesk, respectively. In comparison to contrastive-sc, scAFGCC exhibits a remarkable improvement with an increase of 0.21 in ARI and 0.19 in NMI values for the Klein dataset. The lower error bars in scAFGCC indicate its higher robustness compared to other methods. Furthermore, on large-scale datasets our method consistently surpasses all competing approaches. In particular, it achieves notable improvements in ARI over scBGEDA, the second-best method: 0.18 on the Tosches dataset and 0.27 on the Shekhar dataset. These results underscore the superior robustness and scalability of our approach in addressing large-scale scRNA-seq data. To assess the performance and competitiveness of scAFGCC more accurately, we compare this model with the most recent methods in the field, such as scNAME [51] and scBGEDA [52]. The two methods are executed with 10 different random seeds on 13 real datasets each, and we calculate the mean and standard deviation of ARI and NMI for each dataset. The corresponding experimental outcomes are documented in Supplementary Table S6, revealing that, scAFGCC demonstrates superior performance compared to the two recent methods.

To provide a more intuitive demonstration of scAFGCC's clustering performance, we select two real datasets (Human1 and Klein) and employ UMAP for dimensionality reduction and visualization. To promote uniformity and impartiality in our evaluations across diverse

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر ترجمانو

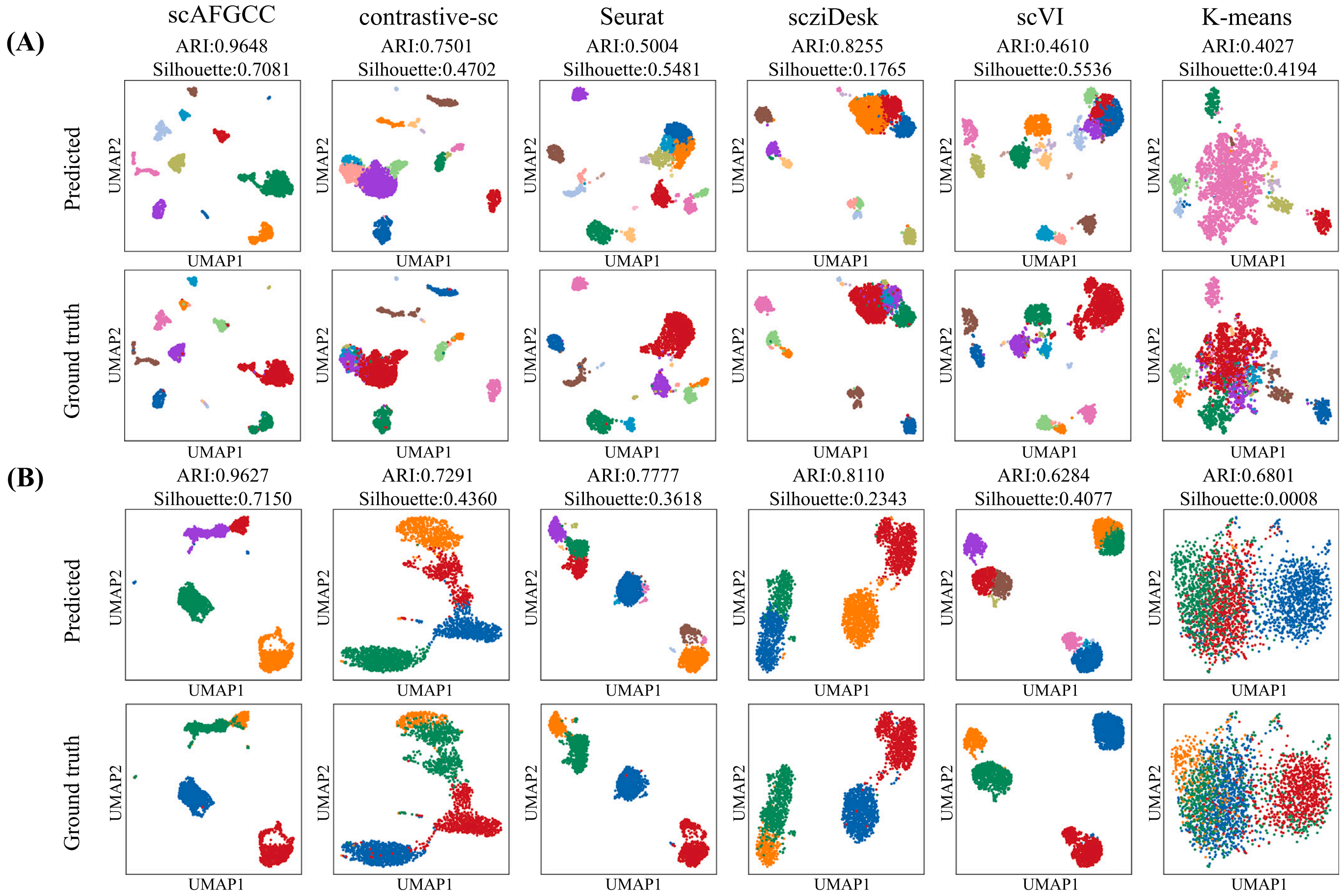

Fig. 2. UMAP visualization of scAFGCC and other clustering methods on two datasets. (A) Human1 dataset. (B) Klein dataset.

datasets and methods, we utilize the default UMAP parameters provided by the Scanpy toolkit consistently throughout all experiments. We choose five models for comparison and utilize ARI and silhouette coefficient [53] to measure their clustering performance. The ground truth labels are also utilized to evaluate and validate the clustering performance. As shown in Fig. 2 (the UMAP plots for the other datasets are displayed in Supplementary Figure S2), scAFGCC demonstrates clear and distinct partitioning of various cell clusters on the both datasets. Furthermore, scAFGCC attains the highest ARI values and silhouette coefficients on the both datasets, surpassing 0.96 and 0.70, respectively. In contrast, other methods tend to mix different cell subtypes together. The clustering results obtained using scAFGCC are visually compelling, clearly demonstrating effective separation.

5.3. Ablation study

To assess the performance gains introduced by the reconstruction module and clustering module, we conduct an ablation experiment on scAFGCC. We define three variant models: scAFGCC-R, which excludes the reconstruction module; scAFGCC-C, which excludes the clustering module; and scAFGCC-R-C, which excludes both the reconstruction and clustering modules. We perform experiments on seven datasets with different species and varying numbers of clusters (Human1, Human2, Diaphragm, Deng, Muraro, and Pollen). The evaluation metrics are ARI and NMI. To obtain reliable results, we execute each variant model 10 times and derive the average outcomes. The corresponding experimental results are presented in Fig. 3. We can observe that scAFGCC outperforms the three variant models. This indicates that both the reconstruction module and clustering module indeed contribute to performance improvements. For example, in the Deng dataset, scAFGCC achieves an ARI improvement of 0.10, 0.06, and 0.06 over scAFGCC-R, scAFGCC-C, and scAFGCC-R-C, respectively.

5.4. Hyperparameter analysis

The k and $topk$ are two hyperparameters in our model. The k represents the number of neighbors for each cell. The $topk$ represents the number of similar cell nodes in the set B_i. We utilize ARI and NMI as evaluation metrics. The reported experimental results are the average values obtained from running the experiments independently 5 times. For $topk$, we select a range from 2 to 10. Fig. 4(A) displays the ARI values for all datasets with different $topk$ values. The clustering performance of scAFGCC remains relatively stable as the $topk$ value increases, which suggests that scAFGCC is not highly sensitive to $topk$. $topk = 6$ achieves the best performance across the majority of the datasets. We conduct scAFGCC on all datasets with $k = \{5, 10, 15, 20, 25, 30\}$. The results are recorded in Fig. 4(B). It is apparent that scAFGCC demonstrates low sensitivity to variations in the k parameter across most datasets. However, it is worth noting that in datasets such as Human1, Diaphragm, Darmanis, and Deng, the performance of scAFGCC exhibits notable changes as the k value increases. On most datasets, $k = 5$ enables scAFGCC to achieve optimal performance. Supplementary Figure S3 shows two line plots depicting the NMI values.

For the above four datasets that are sensitive to variations in the k, we perform joint optimization of both k and $topk$ values. As shown in Fig. 4(C), this further confirms the sensitivity to both $topk$ and k parameters. From the figure, we can observe that in the case of Human1 and Deng datasets, lower values of k result in higher model performance, while the opposite is true for Diaphragm and Darmanis datasets.

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

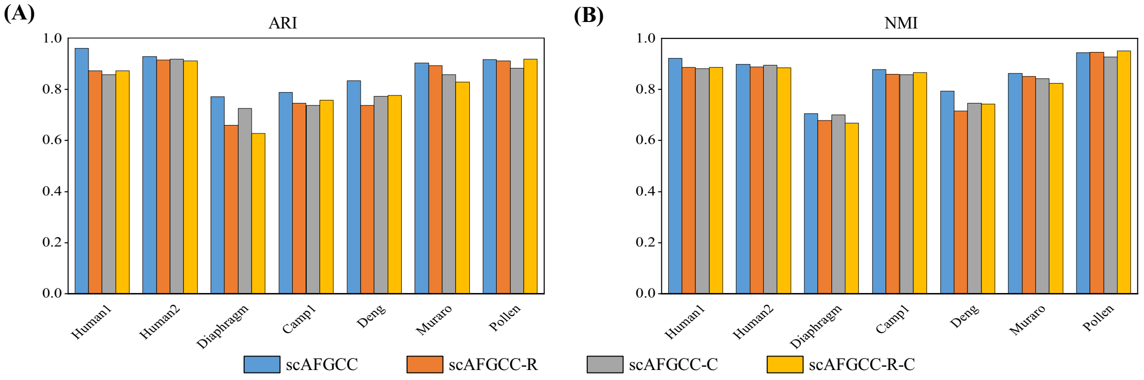

Fig. 3. Performance comparison of scAFGCC and its variant models (scAFGCC-R, scAFGCC-C, scAFGCC-R-C) on seven datasets. (A) ARI scores. (B) NMI scores.

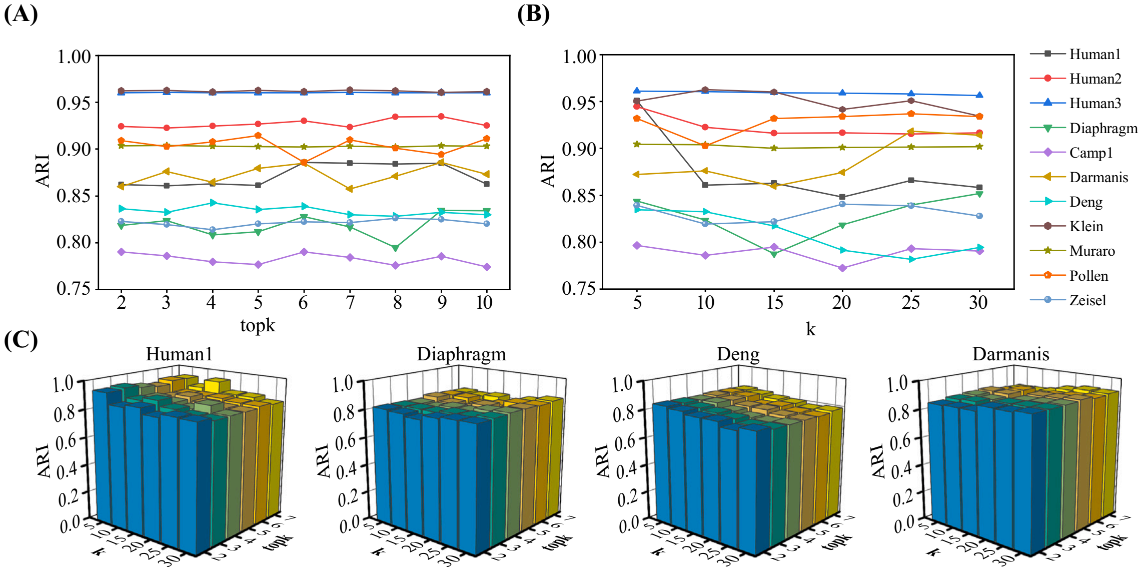

Fig. 4. Analysis of scAFGCC hyperparameters (*topk* and *k*). (A) ARI values for different datasets with varying topk values. (B) ARI values for different datasets with varying *k* values. (C) Joint optimization of *k* and *topk* for sensitive datasets.

5.5. Robustness analysis

In scRNA-seq data analysis, dropout events are a significant challenge. Next, we evaluate the robustness of scAFGCC in handling dropout events. During the data preprocessing stage, we simulate dropout events by setting a specific proportion of non-zero values to zero. We select six datasets with different species and varying cluster numbers, including Human1, Human2, Darmanis, Deng, Klein and Muraro. For fair comparison, we run scAFGCC 5 times using default parameters and utilize the average values of ARI and NMI to evaluate the clustering performance. For each dataset, we respectively set dropout rates of 0, 20, 40, and 60 % to test the robustness of the model. High dropout events result in information loss in scRNA-seq data, posing greater challenges for dimensionality reduction and clustering algorithms. The relevant experimental results are shown in Fig. 5. It is evident that in the majority of datasets, as the dropout rate increases, there is a marginal decrease in the performance of the model. This provides further evidence that scAFGCC exhibits high stability and robustness in handling dropout events.

Furthermore, we conduct a down-sampling experiment on the same six datasets. To ensure fairness, scAFGCC is executed five times, yielding average values for ARI and NMI. We downsample the datasets, extracting 100, 80, 60, and 40 % of the cells, respectively. The experimental results are presented in Supplementary Figure S4. We can conclude that scAFGCC demonstrates stability across different down-sampling datasets, further highlighting its robustness.

5.6. Biological analysis

Biological analysis plays a crucial role in scRNA-seq data analysis. To further investigate the biological significance of the clustering results obtained from scAFGCC, we perform downstream tasks, including cell annotation and marker gene identification for the Darmanis dataset.

In this part, we perform cell annotation. Specifically, based on the predicted labels from scAFGCC and ground truth labels, we separately utilize the Wilcoxon Rank Sum test [54] to identify the top 50 differentially expressed genes (DEGs) for each predicted cluster and ground truth cluster. By comparing the number of overlapping genes between each predicted cluster and ground truth cluster, we calculate the overlapping ratio as a measure of similarity between the predicted clusters and ground truth clusters. To validate the superiority of scAFGCC, we also compare it with four other models. The visual results are presented in Fig. 6(A). We can observe that scAFGCC can assign each cell type to a unique predicted cluster. However, for methods such as contrastive-sc, scziDesk, and K-means, they are unable to assign OPC (Oligodendrocyte Precursor Cell) to a single predicted cluster. For scVI, microglia cannot be unambiguously assigned to a single predicted cluster, and no

Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

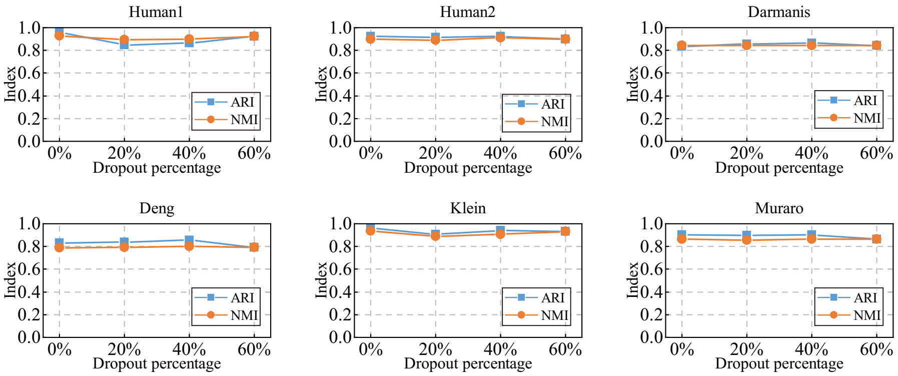

Fig. 5. Clustering performance (ARI and NMI) of scAFGCC on six datasets with varying dropout rates (0 %, 20 %, 40 %, and 60 %).

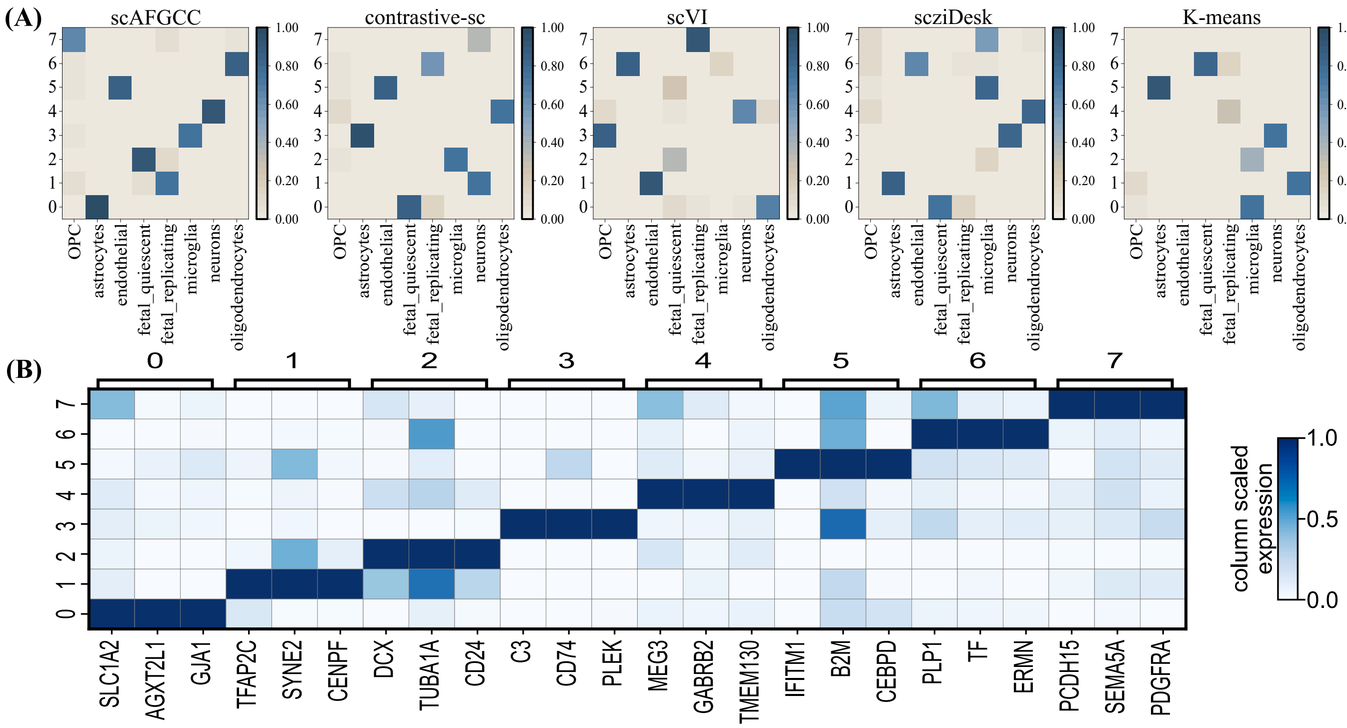

Fig. 6. Result analysis of cell annotation and marker gene identification in the Darmanis dataset. (A) Matrix plots illustrating the overlap between the top 50 DEGs in clusters identified by five methods and the gold standard cell types. (B) The matrix diagram of the first three marker genes for each cluster predicted by scAFGCC.

specific cell type is discovered to correspond to cluster 5. K-means repeatedly assigns microglia to clusters 0 and 2, while clusters 4 and 7 cannot be accurately annotated by K-means. Overall, scAFGCC is able to fully annotate all predicted clusters. Specifically, the predicted clusters (0–7) can be annotated as astrocytes, fetal-replicating, fetal-quiescent, microglia, neurons, endothelial, oligodendrocytes, and OPC.

Furthermore, as a part of the biological analysis, marker gene identification is conducted. We also utilize the Wilcoxon Rank Sum test to identify marker genes for each cluster predicted by scAFGCC. As shown in Fig. 6(B), they respectively present the matrix plot and heat map of the top three marker genes for each cluster predicted by scAFGCC (The top 10 marker genes for different cell clusters are shown in Supplementary Figure S5). We can identify clusters with enriched expression of specific marker genes, which helps in characterizing and distinguishing different cell types within the dataset. To further validate the presence of marker genes in the annotated cell types, we compare them with the marker genes documented in the Cell Marker database [55] and the Darmanis dataset [41]. The results are recorded in Supplementary Table S7. We discover that the majority of marker genes align with the records in the database. However, there are a few specific genes that are not documented, suggesting the possibility of novel candidates for marker genes. For instance, B2M and CEBPD show considerably higher expression levels in endothelial cells compared to others, indicating their potential as marker genes for endothelial cells.

To explore the biological relevance of the identified marker genes, we conduct Gene Ontology (GO) and KEGG pathway analysis. Initially, we select the top 40 marker genes sorted by p-value within each cluster, resulting in a total of 300 genes after converting gene symbols to ensemble IDs. We elucidate the biological functions of these genes. The distribution of genes enriched in GO terms is presented in Fig. 7, while the top 25 terms of the related GO sorted by p-values, are documented in Supplementary Figure S6. In the biological process category, GO terms such as 'nervous system development' (GO:0007399) and 'multicellular organism development' (GO:0007275) exhibit high enrichment levels, both strongly associated with the development of human brain cells [41]. In the cellular component category, the most enriched and

iranpaper Downloaded from https://iranpaper.ir https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

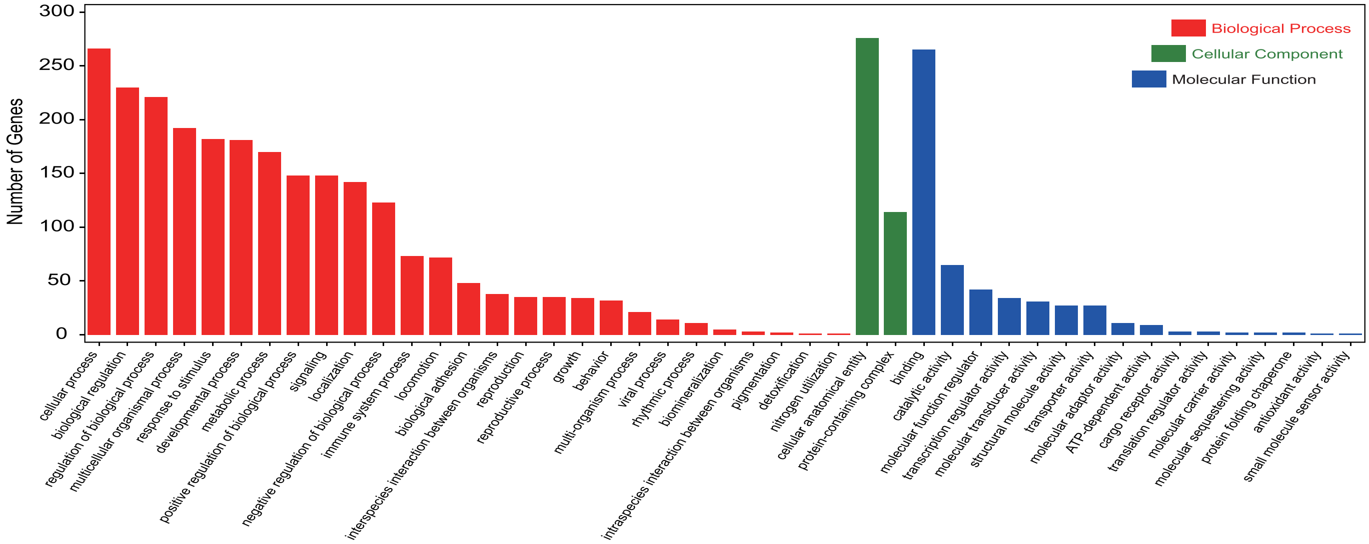

Fig. 7. Genomic interpretability of scAFGCC in the Darmanis dataset: Exploring the gene distribution across three categories of GO enrichments.

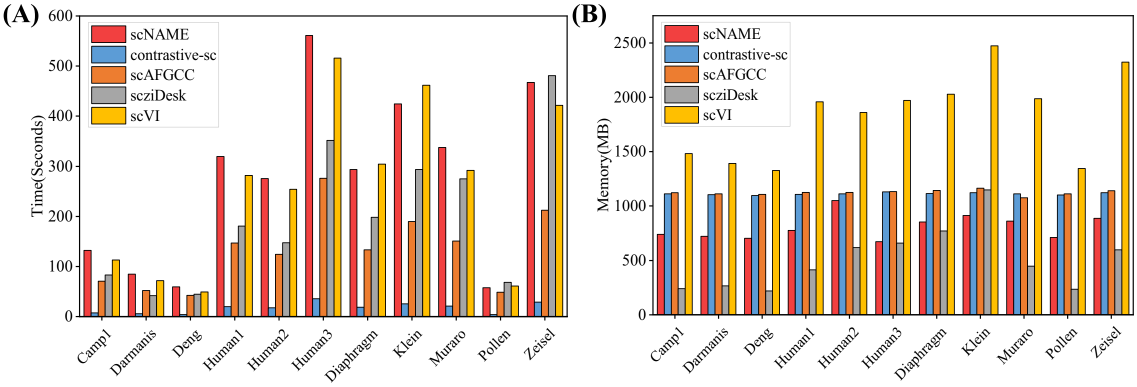

Fig. 8. Efficiency analysis of scAFGCC. (A) Runtime analysis. (B) Memory usage analysis.

notable GO term is ‘cell junction’ (GO:0030054). Cell junctions play a crucial role in the formation and regulation of brain vasculature [56]. In the molecular function category, the GO term ‘protein binding’ (GO:0005515) stands out with a small p-value and a substantial gene distribution, crucial for the formation of adult neural stem cells [57]. Moreover, the top 15 terms of KEGG pathways ranked by p-value are shown in Supplementary Figure S7(A). The top pathway is ‘MicroRNAs in cancer’. Although this pathway is related to cancer, microRNAs also play a crucial role in the nervous system, participating in the development and regulation of neuronal function. The pathway with the highest percentage of genes is ‘PI3K-Akt signaling pathway’, which is also significant in the nervous system, participating in processes such as neuronal growth, survival, and synapse formation [58]. Meanwhile, detailed information about this pathway is visualized in Supplementary Figure S7(B). Overall, these results indicate that scAFGCC can uncover meaningful biological information.

5.7. Efficiency analysis

To validate the efficiency of scAFGCC, we employ the R package “Splatter” to simulate six datasets of varying sizes. These datasets include 1k, 2k, 4k, 6k, 8k, and 10k cells, each consisting of 2000 genes. As the number of cells varies, we measure the runtime and memory usage of scAFGCC. The results are presented in Supplementary Figure S8. It is evident that the runtime and memory usage of scAFGCC do not exhibit quadratic or exponential growth with the number of cells, but instead demonstrate a linear increase.

Furthermore, we perform runtime and memory analysis on four deep learning methods, including scAFGCC, scNAME, contrastive-sc, scVI, and scziDesk, using 11 real datasets. In Fig. 8, we can observe that contrastive-sc is the fastest method among the five methods, possibly due to its fewer trainable parameters. However, it is worth mentioning that our model achieves superior performance compared to contrastive-sc. Our model outperforms scziDesk, scVI and scNAME in terms of runtime. Despite the need to aggregate neighbor information using GCN in our model, we still achieve higher performance with less runtime. In relation to memory usage, scziDesk consumes the least amount of memory among the five methods. Both our model and contrastive-sc exhibit similar memory usage, which is significantly lower compared to scVI. In summary, our scAFGCC method achieves superior clustering performance while maintaining comparable time and memory requirements.

6. Conclusion

High-throughput scRNA-seq provides valuable insights into cellular heterogeneity, rare cell identification, in-depth characterization of cellular states, and the dynamics of biological processes at the single-cell level. In the field of scRNA-seq data analysis, one of the primary and crucial tasks is cell identification by accurately clustering cells into

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

distinct subpopulations based on their molecular profiles. Many clustering methods based on deep learning and contrastive learning have been proposed in recent years. However, these methods often do not fully explore the complex relationships between cells, and some contrastive learning-based approaches can be sensitive to patterns in data augmentation.

Therefore, we propose a novel augmentation-free graph contrastive learning method called scAFGCC for scRNA-seq data analysis. This approach tackles the limitations of existing methods by explicitly capturing the inherent cellular relationships without the need for data augmentation techniques. scRNA-seq data exhibit high-dimensional and sparse characteristics, coupled with dropout events. In our model, we are able to capture the denoised latent representation of scRNA-seq data, which eliminates dropout events and improves the model's performance. Additionally, the biological analysis further emphasizes that scAFGCC yields invaluable insights and information that can greatly contribute to downstream tasks in scRNA-seq data analysis. In summary, the experimental results indicate that our method achieves superior performance in noisy and complex datasets compared to recent advancements in the field, and provides better computational efficiency.

While scAFGCC primarily focuses on capturing cell–cell relationships in its analysis, it may not fully consider the interactions and relationships between individual genes within the cells. Therefore, future research could explore integrating gene–gene relationships into the scAFGCC framework to improve model performance. The code is available at https://github.com/tswstart/scAFGCC.

CRediT authorship contribution statement

Shengwen Tian: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Yu Wang:** Writing – original draft. **Yutian Wang:** Validation, Supervision. **Cunmei Ji:** Writing – review & editing. **Jiancheng Ni:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.neucom.2025.131698.

Data availability

The data that has been used is confidential.

References

[1] E.Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A.R. Bialas, N. Kamitaki, E.M. Martersteck, at al., Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, Cell 161 (5) (2015) 1202–1214.

[2] G.X. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, at al., Massively parallel digital transcriptional profiling of single cells, Nat. Commun. 8 (1) (2017) 14049.

[3] G. Chen, B. Ning, T. Shi, Single-cell RNA-seq technologies and related computational data analysis, Front. Genet. (2019) 317.

[4] E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science, Nat. Rev. Genet. 14 (9) (2013) 618–630.

[5] A.A. Kolodziejczyk, J.K. Kim, V. Svensson, J.C. Marioni, S.A. Teichmann, The technology and biology of single-cell RNA sequencing, Mol. Cell. 58 (4) (2015) 610–620.

[6] A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V. Nahed, W.T. Curry, R.L. Martuza, at al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science 344 (6190) (2014) 1396–1401.

[7] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N.J. Lennon, K.J. Livak, T.S. Mikkelsen, J.L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nat. Biotechnol. 32 (4) (2014) 381–386.

[8] J. Kim, C. Park, K.H. Kim, E.H. Kim, H. Kim, J.K. Woo, J.K. Seong, K.T. Nam, Y.C. Lee, S.Y. Cho, Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity, NPJ Precis. Oncol. 6 (1) (2022) 9.

[9] W. Stephenson, L.T. Donlin, A. Butler, C. Rozo, B. Bracken, A. Rashidfarrokhi, S.M. Goodman, L.B. Ivashkiv, V.P. Bykerk, D.E. Orange, at al., Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation, Nat. Commun. 9 (1) (2018) 791.

[10] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, Mol. Syst. Biol. 15 (6) (2019) e8746.

[11] O. Stegle, S.A. Teichmann, J.C. Marioni, Computational and analytical challenges in single-cell transcriptomics, Nat. Rev. Genet. 16 (3) (2015) 133–145.

[12] V.Y. Kiselev, T.S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data, Nat. Rev. Genet. 20 (5) (2019) 273–282.

[13] V.Y. Kiselev, K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, K.N. Natarajan, W. Reik, M. Barahona, A.R. Green, at al., Sc3: Consensus clustering of single-cell RNA-seq data, Nat. Methods 14 (5) (2017) 483–486.

[14] R. Satija, J.A. Farrell, D. Gennert, A.F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nat. Biotechnol. 33 (5) (2015) 495–502.

[15] W. Wu, W. Zhang, W. Hou, X. Ma, Multi-view clustering with graph learning for scRNA-seq data, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (6) (2023) 3535–3546.

[16] R. Zheng, M. Li, Z. Liang, F.-X. Wu, Y. Pan, J. Wang, Sinnlrr: a robust subspace clustering method for cell type detection by non-negative and low-rank representation, Bioinformatics 35 (19) (2019) 3642–3650.

[17] W. Zhang, Y. Li, X. Zou, Scclrr: a robust computational method for accurate clustering single cell RNA-seq data, IEEE J. Biomed. Health Inform. 25 (1) (2020) 247–256.

[18] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning, Nat. Methods 14 (4) (2017) 414–416.

[19] S. Bao, K. Li, C. Yan, Z. Zhang, J. Qu, M. Zhou, Deep learning-based advances and applications for single-cell RNA-sequencing data analysis, Brief. Bioinform. 23 (1) (2022) bbab473.

[20] H. Wang, X. Ma, Learning deep features and topological structure of cells for clustering of scRNA-sequencing data, Brief. Bioinform. 23 (3) (2022).

[21] T. Tian, J. Zhang, X. Lin, Z. Wei, H. Hakonarson, Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data, Nat. Commun. 12 (1) (2021) 1873.

[22] L. Chen, W. Wang, Y. Zhai, M. Deng, Deep soft k-means clustering with self-training for single-cell RNA sequence data, NAR Genom. Bioinform. 2 (2) (2020) lqaa039.

[23] R. Lopez, J. Regier, M.B. Cole, M.I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics, Nat. Methods 15 (12) (2018) 1053–1058.

[24] G. Eraslan, L.M. Simon, M. Mircea, N.S. Mueller, F.J. Theis, Single-cell RNA-seq denoising using a deep count autoencoder, Nat. Commun. 10 (1) (2019) 390.

[25] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, C. Wang, H. Fu, Q. Ma, D. Xu, Scgnn is a novel graph neural network framework for single-cell RNA-seq analyses., Nat. Commun. 12 (1) (2021) 1882.

[26] M. Ciortan, M. Defrance, Contrastive self-supervised clustering of scRNA-seq data, BMC Bioinform. 22 (1) (2021) 280.

[27] S.-W. Tian, J.-C. Ni, Y.-T. Wang, C.-H. Zheng, C.-M. Ji, Scgcc: Graph contrastive clustering with neighborhood augmentations for scRNA-seq data analysis, IEEE J. Biomed. Health Inform. 27 (12) (2023) 6133–6143.

[28] Y. Mo, H.T. Shen, X. Zhu, Unsupervised multi-view graph representation learning with dual weight-net, Inf. Fusion. 114 (2025) 102669.

[29] X. Yang, Y. Wang, J. Chen, W. Fan, X. Zhao, E. Zhu, X. Liu, D. Lian, Dual test-time training for out-of-distribution recommender system, IEEE Trans. Knowl. Data Eng. 37 (6) (2025) 3312–3326.

[30] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (1) (2019) 38–44.

[31] F.A. Wolf, P. Angerer, F.J. Theis, Scanpy: large-scale single-cell gene expression data analysis, Genome Biol. 19 (2018) 1–5.

[32] H. Wen, G. Ding, C. Liu, J. Wang, Matrix factorization meets cosine similarity: addressing sparsity problem in collaborative filtering recommender system, in: Web Technologies and Applications: 16th Asia-Pacific Web Conference, APWEB 2014, Changsha, China, September 5-7, 2014. Proceedings 16, Springer, 2014, pp. 306–317.

[33] N. Lee, J. Lee, C. Park, Augmentation-free self-supervised learning on graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 7372–7380.

[34] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.

[35] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: International Conference on Machine Learning, PMLR, 2016, pp. 478–487.

[36] L. Zappia, B. Phipson, A. Oshlack, Splatter: simulation of single-cell RNA sequencing data, Genome Biol. 18 (1) (2017) 174.

[37] A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell 161 (5) (2015) 1187–1201.

Downloaded from https://iranpaper.ir

https://www.tarjomano.com
ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

[38] Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells, Science 343 (6167) (2014) 193–196.

[39] N. Schaum, J. Karkanias, N.F. Neff, A.P. May, S.R. Quake, T. Wyss-Coray, S. Darmanis, J. Batson, O. Botvinnik, M.B. Chen, at al., Single-cell transcriptomics of 20 mouse organs creates a tabula muris: the Tabula Muris Consortium, Nature 562 (7727) (2018) 367.

[40] J.G. Camp, F. Badsha, M. Florio, S. Kanton, T. Gerber, M. Wilsch-Bräuninger, E. Lewitus, A. Sykes, W. Hevers, M. Lancaster, at al., Human cerebral organoids recapitulate gene expression programs of fetal neocortex development, Proc. Natl. Acad. Sci. 112 (51) (2015) 15672–15677.

[41] S. Darmanis, S.A. Sloan, Y. Zhang, M. Enge, C. Caneda, L.M. Shuer, M.G. Hayden Gephart, B.A. Barres, S.R. Quake, A survey of human brain transcriptome diversity at the single cell level, Proc. Natl. Acad. Sci. 112 (23) (2015) 7285–7290.

[42] M.J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. Van Gurp, M.A. Engelse, F. Carlotti, E.J. De Koning, at al., A single-cell transcriptome atlas of the human pancreas, Cell Syst. 3 (4) (2016) 385–394.

[43] A.A. Pollen, T.J. Nowakowski, J. Shuga, X. Wang, A.A. Leyrat, J.H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, at al., Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex, Nat. Biotechnol. 32 (10) (2014) 1053–1058.

[44] A. Zeisel, A.B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, at al., Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, Science 347 (6226) (2015) 1138–1142.

[45] M. Baron, A. Veres, S.L. Wolock, A.L. Faust, R. Gaujoux, A. Vetere, J.H. Ryu, B.K. Wagner, S.S. Shen-Orr, A.M. Klein, at al., A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure, Cell Syst. 3 (4) (2016) 346–360.

[46] M.A. Tosches, T.M. Yamawaki, R.K. Naumann, A.A. Jacobi, G. Tushev, G. Laurent, Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles, Science 360 (6391) (2018) 881–888.

[47] K. Shekhar, S.W. Lapan, I.E. Whitney, N.M. Tran, E.Z. Macosko, M. Kowalczyk, X. Adiconis, J.Z. Levin, J. Nemesh, M. Goldman, at al., Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics, Cell 166 (5) (2016) 1308–1323.

[48] J.M. Santos, M. Embrechts, On the use of the adjusted rand index as a metric for evaluating supervised classification, in: International Conference on Artificial Neural Networks, Springer, 2009, pp. 175–184.

[49] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary?, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 1073–1080.

[50] L. McInnes, J. Healy, S. Astels, Hdbscan: hierarchical density based clustering., J. Open Source Softw. 2 (11) (2017) 205.

[51] H. Wan, L. Chen, M. Deng, Scname: Neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data, Bioinformatics 38 (6) (2022) 1575–1583.

[52] Y. Wang, Z. Yu, S. Li, C. Bian, Y. Liang, K.-C. Wong, X. Li, Scbgeda: deep single-cell clustering analysis via a dual denoising autoencoder with bipartite graph ensemble clustering, Bioinformatics 39 (2) (2023) btad075.

[53] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[54] R.F. Woolson, Wilcoxon signed-rank test, Wiley Encycl. Clin. Trials (2007) 1–3.

[55] X. Zhang, Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, L. Xu, G. Liao, M. Yan, at al., Cellmarker: a manually curated resource of cell markers in human and mouse, Nucl. Acids Res. 47 (D1) (2019) D721–D728.

[56] E.S. Lippmann, S.M. Azarin, J.E. Kay, R.A. Nessler, H.K. Wilson, A. Al-Ahmad, S.P. Palecek, E.V. Shusta, Derivation of blood-brain barrier endothelial cells from human pluripotent stem cells, Nat. Biotechnol. 30 (8) (2012) 783–791.

[57] Y.-G. Han, N. Spassky, M. Romaguera-Ros, J.-M. Garcia-Verdugo, A. Aguilar, S. Schneider-Maunoury, A. Alvarez-Buylla, Hedgehog signaling and primary cilia are required for the formation of adult neural stem cells, Nat. Neurosci. 11 (3) (2008) 277–284.

[58] A. Brunet, S.R. Datta, M.E. Greenberg, Transcription-dependent and-independent control of neuronal survival by the PI3K–Akt signaling pathway, Curr. Opin. Neurobiol. 11 (3) (2001) 297–305.

Author biography

Shengwen Tian received the M.S. degree from the School of Cyber Science and Engineering, Qufu Normal University, China, in 2024. He is currently pursuing the Ph.D. degree at the School of Computer Science and Technology, Beijing Institute of Technology, China. His research interests include geometric deep learning, graph representation learning, and bioinformatics.

Yu Wang received the BS degree from the School of Software, Qufu Normal University, China, in 2021, and the MS degree from the School of Cyber Science and Engineering, Qufu Normal University, China, in 2024. Her research interests include graph neural networks and bioinformatics.

Yutian Wang received the BS degree in computer science education from Qufu Normal University, Jining, China, in 1999, and the MS degree in computer application technology from the Beijing University of Technology, China, in 2005. He is a lecturer with the School of Software, Qufu Normal University, China. His research interests include data mining and bioinformatics.

Cunmei Ji received the BS degree from the School of Mechanical Design & Manufacturing and Automation, Nanjing Tech University, Nanjing, China, in 2003, and the MS degree from the School of Computer and Science, Zhejiang University, in 2010. From 2003 to 2006 he worked as an embedded software engineer with the Nanjing Research Institute of Simulation Technology. In 2010, he joined Huawei Technologies Co, Ltd as a software engineer. From 2011 to 2015 he worked as a senior software engineer in Cisco Systems, Inc. He is currently a lecturer with the School of Software, Qufu Normal University, China. His research interests include deep learning and bioinformatics.

Jiancheng Ni received the BS degree in mathematics education, in 1995, the MS degree in mathematics and computer science from Qufu Normal University, in 2002, and the PhD degree in network security from Sichuan University, China, in 2008. He is currently a professor with the Network Information Center, Qufu Normal University, China. His research interests include machine learning, bioinformatics, distributed computing, and network security.