Single-Cell Deep Clustering Method Assisted by Exogenous Gene Information: A Novel Approach to Identifying Cell Types

Dayu Hu, Ke Liang, Hao Yu and Xinwang Liu†, *Senior Member, IEEE*

***Abstract*—In recent years, the field of single-cell data analysis has seen a marked advancement in the development of clustering methods. Despite advancements, most of these algorithms still concentrate on analyzing the provided single-cell matrix data. However, in medical applications, single-cell data often involves a wealth of exogenous information, including gene networks. Overlooking this aspect could lead to information loss and clustering results devoid of significant clinical relevance. An innovative single-cell deep clustering method, incorporating exogenous gene information, has been proposed to overcome this limitation. This model leverages exogenous gene network information to facilitate the clustering process, generating discriminative representations. Specifically, we have developed an attention-enhanced graph autoencoder, which is designed to efficiently capture the topological features between cells. Concurrently, we conducted a random walk on an exogenous Protein-Protein Interaction (PPI) network, thereby acquiring the gene's topological features. Ultimately, during the clustering process, we integrated both sets of information and reconstructed the features of both cells and genes to generate a discriminative representation. Extensive experiments have validated the effectiveness of our proposed method. This research offers enhanced insights into the characteristics and distribution of cells, thereby laying the groundwork for early diagnosis and treatment of diseases.**

***Index Terms*—Exogenous gene information, Clustering, Protein-protein interaction, Node2vec, Deep learning.**

I. Introduction

SINGLE-CELL transcriptome sequencing technology represents a significant advancement in the field of genomics. It elucidates the intricate biological processes at the cellular level and serves as a potent tool for studying the origins and microenvironments of tumors [1]–[4]. Unsupervised clustering represents a pivotal step in this process. By analyzing the gene expression data of individual cells, it precisely differentiates between various cell types and states. This approach provides valuable insights into understanding complex biological systems, such as cancer, neurodegenerative diseases, and developmental processes. However, owing to the complexity of biological systems, devising a clustering algorithm that is both accurate and highly clinically relevant remains a formidable challenge.

In recent years, a significant increase in the development of clustering algorithms tailored for single-cell RNA sequencing (scRNA-seq) data has been observed [5], [6]. Early approaches depended on probabilistic models that estimated high-dimensional cell data through computing the probability of gene expression. For instance, CIDR [7] introduced an interpolation method to handle dropout events, whereas SC3 [8] employed hierarchical k-means clustering to facilitate consensus clustering, presuming Euclidean relationships between cells. However, these methods operate under the assumption that biological data are linear and clear, an assumption that may not always be valid in the real world.

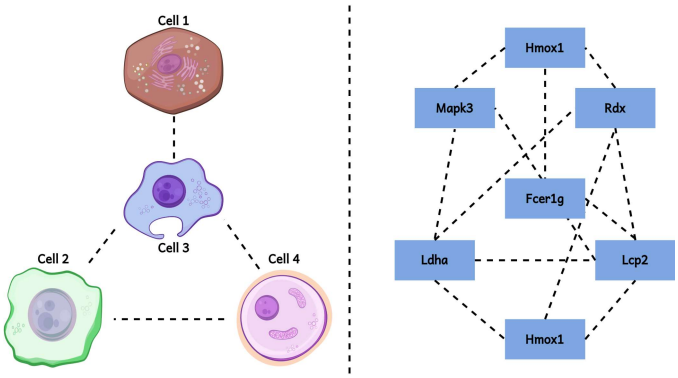


Fig. 1. Cells and genes both exhibit associative relationships. The left illustrates the connections between cells, while the right depicts the associations among genes.

To effectively extract features from scRNA-seq data and circumvent assumptions about data distribution, some researchers have suggested neural networks as a promising approach for mining information from scRNA-seq data [9], [10]. Neural networks, widely used as black box models, can adapt to nearly all data distributions when the parameters are appropriately configured. Numerous single-cell deep clustering algorithms have been proposed to obtain effective representations, A detailed introduction to these methods will be provided in the Related Work section (Section 2A). However, these models often treat cells as isolated entities, overlooking the associations between them.

To integrate cellular interaction relationships into the clustering process, researchers have proposed graph-based approaches for deriving cell embeddings. This approach necessitates constructing a cell graph based on intercellular similarities. These constructed graphs, in conjunction with the original feature matrix, are subsequently inputted into a graph neural network for training. A detailed introduction to graph clustering algorithms will be provided in Section 2B.

Dayu Hu, Ke Liang, Hao Yu, Xinwang Liu are with the School of Computer, National University of Defense Technology, Changsha, China, 410073. Email: hzauhdy@gmail.com, liangke200694@126.com, csyuhao@gmail.com, xinwangliu@nudt.edu.cn.

† Corresponding author.

Although these graph-based deep clustering algorithms have progressed in capturing the topological features of cells, their focus remains primarily on analyzing the provided single-cell matrix data. However, clustering algorithms oriented towards medical applications should integrate external information for a more holistic analysis, as overlooking this aspect could result in clustering outcomes that diverge from clinical conclusions.

Single-cell data inherently contain exogenous information. Unlike other datasets, the features in scRNA datasets are meaningful as they represent genes. Biologists and medical scientists have extensively explored gene relationships. Despite this extensive research, most current clustering methods still overlook these gene relationships, focusing solely on cell connections and neglecting gene associations. However, in reality, genes within each cell participate in complex interrelations due to interactions, regulatory mechanisms, and shared functions and pathways in biological processes. In essence, as is shown in Figure 1, gene topological features are present, yet this aspect remains largely unexplored in single-cell clustering research. By extracting and integrating the topological features of gene interconnectivity into the clustering framework, significant optimization of clustering embeddings and enhancement of clustering outcomes can be achieved. Furthermore, this type of embedding could lead to a more accurate representation of biological characteristics, thereby enhancing the alignment between identified clusters and the actual underlying biological systems.

In light of these considerations, we have developed an exogenous gene information-assisted single-cell deep clustering method (scEGA) that simultaneously focuses on the interaction relationships between cells and genes.To accomplish this, we utilized a graph attention autoencoder (GAT), which captures the topological structure between cells and ensures effective information transmission among them. Additionally, we conducted random walks on the exogenous protein-protein interaction (PPI) network corresponding to the gene set, to obtain embeddings that represent the gene's topological features. During the clustering process, we integrated these two elements and reconstructed the features of both cells and genes, thereby acquiring a discriminative cell representation. Experiments on eight real scRNA datasets demonstrate that our scEGA method is stable and outperforms nine other baseline methods in performance. Our contributions can be summarized as follows:

- The scEGA model simultaneously focuses on the cell features and exogenous gene features, fusing and aligning them during the clustering process to generate a more discriminative representation.
- The scEGA model employs a dual-supervised module to facilitate the optimization of the bottleneck layer, effectively utilizing its own information and requiring no labels.
- The scEGA model is robust and demonstrates superior performance compared to the other nine baseline methods.

II. Related work

A. Single-cell Deep Clustering

Recently, deep learning methods have been widely applied to analyze scRNA-seq data due to their formidable learning capabilities. Li et al. proposed DESC, which iteratively learns the gene expression pattern of each cluster, assigns cells to their respective clusters and continuously mitigates batch effects [11]. Tian et al. propose scDeepCluster [12], a method rooted in the Zero-inflated Negative Binomial (ZINB) model, utilizing a bottleneck layer for deep k-means clustering to enhance clustering outcomes. Tian et al. developed a deep embedding clustering approach for single-cell data (scDCC), integrating the ZINB model with clustering loss and constraint loss [13]. However, these deep neural networks struggle to preserve the topological structure of scRNA-seq data, because they neglect the associations between cells during analysis.

B. Single-cell Deep Graph Clustering

The advent of deep graph autoencoders has addressed the aforementioned concerns, namely, that previous models treated cells as isolated individuals. These graph autoencoders efficiently learn cluster-friendly, low-dimensional representations by incorporating graph topology information of cell-to-cell interactions. Satija et al. proposed Seurat [14], which employs Louvain community detection to construct a cell graph, subsequently analyzed through spectral clustering using Phenograph. Wang et al. proposed scGNN [15], which utilizes a graph neural network to capture and integrate relationships between cells, complemented by a Gaussian model to represent the pattern of heterogeneous gene expression. Yu et al. introduced scTAG [16], a specialized deep graph embedding clustering algorithm tailored for single-cell data, which concurrently optimizes clustering loss, ZINB loss, and cell graph reconstruction loss. Furthermore, Chen proposed scGAC [17], which introduces attention mechanisms based on the cell-to-cell graph, thus ensuring effective information transmission between cells. Meanwhile, our previous model, scDFC [18], combines structural data from cell-to-cell graphs with attribute information from cellular expression patterns, thereby facilitating a comprehensive analysis of scRNA data.

III. Methods

A. Preliminary

Single-cell data refer to genetic expression information obtained through single-cell sequencing technology, which is presented in matrix form. In this work, we provide a simple mathematical description of this data, represented as a numerical matrix denoted by $\mathbf{X} \in \mathbb{R}^{N \times D}$, where D denotes the dimension of genes, and N represents the number of cells.

B. The Framework of scEGA

Figure 2 depicts the comprehensive workflow of the scEGA model, which consists of two main modules. The first module is the dual-matrix alignment module. This module processes the cell dataset and gene set in parallel to construct two

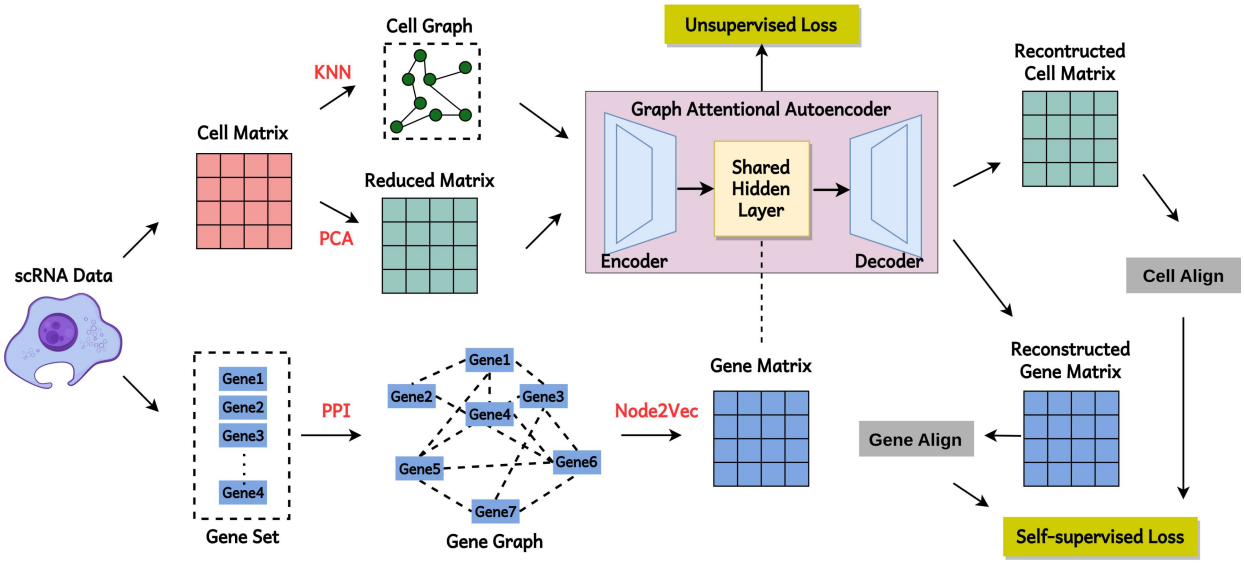


Fig. 2. An illustration of the scEGA model framework. This framework primarily consists of two modules: a dual matrix alignment module, which allows gene representations to participate in the optimization process of deep clustering, thereby fully utilizing the exogenous information of single-cell datasets. The second is a dual-supervised optimization module, which effectively optimizes embeddings through self-supervised and unsupervised loss.

separate matrices: the reconstructed cell matrix and the reconstructed gene matrix. First, to construct the cell-to-cell graph, a k-nearest neighbors (KNN) approach based on similarity measures is employed, and then a specific adjacency matrix $\mathbf{A}$ is obtained. This adjacency matrix, along with the reduced cell matrix, is then fed into a graph attentional autoencoder, as shown in Figure 2. It outputs the reconstructed cell matrix. On the other hand, we also construct the gene-to-gene graph, using exogenous information. Specifically, the gene set is input into the online STRING website to generate a PPI network corresponding to the gene set. This network is then fed as the input of the node2vec algorithm to perform random walks and obtain the final reconstructed gene matrix.

The second module of the model is the dual-supervised optimization module. This module benefits from two supervised mechanisms and relies on no external labels. Initially, the reconstructed matrices of the cell and the gene are obtained. Then, a self-supervised mechanism is used to constrain the stability of these two matrices during the embedding optimization process. Additionally, the shared bottleneck layer is optimized with an unsupervised mechanism to ensure that the embedding exhibits exceptional clustering performance.

C. Dual-matrix Alignment Module

This section provides a detailed description of the dual-matrix fusion module, including the definition of formulas for dual-matrix graphs and the specific computational procedure.

1) Cell Matrix: To accurately learn the cell-to-cell graph information, we designed a graph-based autoencoder (GAT) enhanced by an attention mechanism to fully capture the cell signaling patterns and cell-to-cell relationships. The original feature matrix $\mathbf{X}$, after undergoing principal component analysis (PCA) reduction, yields the dimensionally reduced matrix. The reduced matrix $\hat{\mathbf{X}}$ is encoded to produce the cell embedding, obtained with the following equation:

$$\mathbf{H}_c = \sigma\left(\mathbf{W}_e^c \mathbf{A}\hat{\mathbf{X}}\right), \tag{1}$$

the encoding weight parameter matrix $\mathbf{W}_e^c$ in the Graph Attention Network (GAT) consists of learnable parameters that map the input features to the bottleneck layer. During training, each element of $\mathbf{W}_e^c$ is subject to adjustment. The nonlinear activation function σ facilitates the neural network's ability to learn complex patterns and features.

Subsequently, the embedding $\mathbf{H}_c$, derived from the GAT, is integrated with the embedding produced by the gene graph, which is denoted as $\mathbf{H}_g$. This fusion process is executed as follows:

$$\mathbf{H}_{fusion} = \sigma\left([\mathbf{H}_c||\mathbf{H}_g]\right), \tag{2}$$

the shared embedding of the two graphs is represented as $\mathbf{H}_{fusion}$, and the concatenation operation is represented using $||$. The nonlinear activation function σ remains the same as previously described. Subsequently, the decoding module reconstructs the shared embedding, expressible as:

$$\mathbf{X}_r = \mathbf{W}_d^c \mathbf{H}_{fusion}, \tag{3}$$

where $\mathbf{W}_d^c$ is the learnable decoding matrix of GAT.

2) Gene Matrix: This section offers a comprehensive guide for constructing a gene-to-gene graph from a single-cell dataset. Initially, we processed each single-cell dataset by utilizing Scanpy to identify highly variable genes. We retained the top 2000 genes as the final gene set. Subsequently, the gene set was uploaded to the online platform STRING[1] to generate a PPI network. The network was then saved as an adjacency table to facilitate the subsequent random walk. To generate node embeddings for the PPI network, we utilized the biased approach node2vec for random walk, involving

[1]https://www.string-db.org/

two neighborhood strategies: breadth-first search (BFS) and depth-first search (DFS). BFS focuses on traversing nodes of the same order, while DFS emphasizes traversing higher-order nodes. Figure 3 depicts the detailed procedure of the random walks. By employing these two strategies, node2vec ensures more effective walks, leading to enhanced node embeddings.

Formally, consider $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ as a PPI network, with $\mathbf{V}$ representing the set of nodes, each corresponding to a protein, and $\mathbf{E}$ indicating the interactions between these proteins. We utilize node2vec to generate an embedding vector for each node. Specifically, the sequence $\mathbf{S}_v$ is treated as a corpus, employing the skip-gram model to capture the features. The objective of the skip-gram model is to maximize the probability of observing node v within a specific context. Consequently, for node v, the maximization of the following likelihood function is pursued:

$$\frac{1}{|\mathbf{S}_v|} \sum_{u \in \mathbf{S}_v} \sum_{j \in \mathcal{N}_u} \log \mathbf{P}(v_j | v_u), \tag{4}$$

in this context, $\mathcal{N}_u$ represents the set of neighboring nodes of node u, while $\mathbf{P}(v_j|v_u)$ defines the conditional probability of node j given node u. The calculation of this probability employs the softmax function, defined as follows:

$$\mathbf{P}(v_j|v_u) = \frac{\exp(v_j^\top v_u)}{\sum k \in V \exp(v_k^\top v_u)}, \tag{5}$$

here, v_u and v_j represent the embedding vectors of nodes u and j, respectively. It is assumed that the PPI network comprises n nodes. The gene embedding $\mathbf{Z}_g$ is derived as follows:

$$\mathbf{Z}_g = [v_1, v_2, \ldots, v_n]^\top, \tag{6}$$

the gene embedding $\mathbf{Z}_g$ from the PPI network are inputted into a neural network for joint training and subsequently reconstructed via a decoder. The training process is characterized as follows:

$$\mathbf{H}_g = \sigma\left(\mathbf{W}_e^g \mathbf{Z}_g\right), \tag{7}$$

where $\mathbf{W}_e^g$ is the learnable encoding matrix of gene-to-gene graph, and σ is the nonlinear activation function. In a similar manner, the reconstructed gene matrix $\mathbf{Z}_g^r$ is computed utilizing the equation below:

$$\mathbf{Z}_g^r = \mathbf{W}_d^g \mathbf{H}_{fusion} \tag{8}$$

where $\mathbf{W}_d^g$ denotes the learnable decoding matrix for the gene-to-gene graph.

D. Dual-Supervised Optimization Module

1) Self-supervised Optimization: The aim of self-supervised optimization is to utilize the intrinsic features inherent in the data. In the scEGA framework, the self-supervised loss was employed to ensure the stability of both the cell and gene matrices during the embedding optimization process. This was specifically achieved by aligning the original input data with its reconstructed counterpart. For the cell matrix, the objective was to obtain a reconstructed version that closely resembles the original matrix. This was accomplished by employing cosine loss, a widely-used

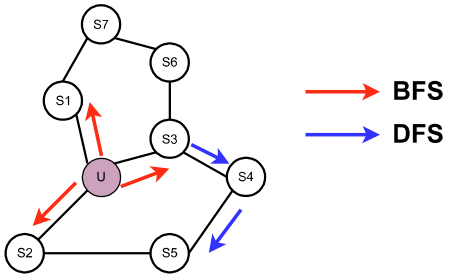


Fig. 3. In BFS and DFS traversals, the node pointed to by the red line is considered a low-order neighbor of the source node, while the node pointed to by the blue line is considered a higher-order neighbor.

similarity metric, to regulate similarity during the clustering process, as demonstrated below:

$$\mathcal{L}_{\text{cell}} = \frac{\hat{\mathbf{X}} \cdot \mathbf{X}_r}{\|\hat{\mathbf{X}}\| \cdot \|\mathbf{X}_r\|}. \tag{9}$$

Regarding the gene matrix, the aim was to maintain the reconstructed gene data unaltered. To accomplish this, Mean Absolute Error (MAE) loss was utilized to guarantee the functional integrity of the genes throughout the clustering optimization process. The particular alignment process unfolds as follows:

$$\mathcal{L}_{\text{gene}} = |\mathbf{Z}_g - \mathbf{Z}_g^r|, \tag{10}$$

the final self-supervised loss is combined as below:

$$\mathcal{L}_{\text{ssl}} = 2\lambda \mathcal{L}_{\text{cell}} + 2(1-\lambda)\mathcal{L}_{\text{gene}}, \tag{11}$$

where λ represents a tunable hyperparameter.

2) Unsupervised Optimization: In this study, the student's t-distribution was utilized to optimize the bottleneck layer. The matrix $\mathbf{Q}$ encapsulates the cluster assignments for each cell, as illustrated below:

$$q_{ij} = \frac{\left(1 + \|z_i - u_j\|^2\right)^{-1}}{\sum_j \left(1 + \|z_i - u_j\|^2\right)^{-1}}, \tag{12}$$

where z represents the embedding of a cell, and u signifies the center of clustering. Subsequently, an auxiliary target distribution $\mathbf{P}$ was constructed based on the clustering distribution $\mathbf{Q}$.

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j \left(q_{ij}^2 / \sum_i q_{ij}\right)}, \tag{13}$$

The objective of optimizing $\mathbf{Q}$ is to closely approximate it to $\mathbf{P}$. To this end, the Kullback-Leibler (KL) divergence was employed as a constraint, termed the unsupervised loss, denoted as follows:

$$\mathcal{L}_{\text{ul}} = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{14}$$

In summary, the dual-supervised optimization module utilizes a two-step learning approach, comprising self-supervised optimization grounded in the data's inherent characteristics,

TABLE I
DETAILS OF THE EIGHT REAL DATASETS.

Datasets	Genes	Clusters	References
Biase	21489	3	[19]
Darmanis	9337	8	[20]
Enge	25929	9	[21]
Bjorklund	26087	4	[22]
Sun	995	6	[23]
Marques	15291	14	[24]
Zeisel	18825	9	[25]
Fink	20932	7	[26]

and unsupervised optimization anchored in the student's t-distribution. Self-supervised optimization is realized via dual-matrix alignment, preserving the consistency of cell matrix and gene matrix throughout clustering, thereby facilitating the learning of a significant compressed representation. Unsupervised optimization seeks to minimize the discrepancy between the actual clustering assignment and the auxiliary target clustering distribution, resulting in embeddings characterized by superior clustering performance. By integrating these two optimization approaches, the total loss of our model $\mathcal{L}_{\mathrm{f}}$ is formulated as follows:

$$\mathcal{L}_{\mathrm{f}} = \mathcal{L}_{\mathrm{ssl}} + \mathcal{L}_{\mathrm{ul}}. \tag{15}$$

IV. EXPERIMENTS

In this section, extensive experiments were conducted to evaluate our model. The subsequent sections will cover Experimental Settings, Clustering Performance, Ablation Study, and Parameter Fine-Tuning.

A. Experimental Settings

1) Datasets: This research presents eight real-world datasets from prevalent species, including humans and mice. To assess the efficacy of various clustering algorithms, each dataset is supplemented with definitive labels. The following provides a succinct introduction to each dataset:

- **Biase** [19], a representative small scRNA dataset, primarily comprises embryonic cells derived from mice, collected during their developmental stages.
- **Darmanis** [20] includes human brain cells, known for their complex composition, leading to its division into multiple clusters.
- **Enge** [21] is comprised of pancreatic cells from humans.
- **Bjroklund** [22] encompasses lymphoid cells from humans, crucial to the immune system.
- **Sun** [23] offers three single-cell datasets, with this study focusing on the first one, containing exclusively mouse lung cells.
- **Marques** [24] includes data from mice, aimed at investigating the developmental origin of oligodendrocyte precursor cells.
- **Zeisel** [25] contains data from mice, sourced from the somatosensory cortex and hippocampus CA1 regions.
- **Fink** [26], sourced from the human adult ureter, may provide insights into metabolic processes.

The initial scRNA data exhibit significant variability in scale and high noise levels, which could potentially lead to erroneous conclusions in subsequent analyses. To mitigate these issues, quality control was conducted on the cellular data prior to clustering. Specifically, cells with expression values within a reasonable range were retained, and outliers with extreme expression values were eliminated. This was achieved by establishing upper and lower thresholds at 75% plus three times the quartile deviation, and 25% minus the quartile deviation, respectively. Following quality control, the data were standardized by scaling to a consistent range. Subsequently, a log2 transformation was applied to the data. To avoid negative infinite values and ensure positive expression values, a pseudo count of 1 was incorporated during the transformation process.

2) Compared Methods: This section offers a concise overview of the baseline methods employed in these experiments.

- **CIDR** [7] utilizes a probabilistic model for evaluating dropout events in cellular data, categorized as a traditional clustering method in this study.
- **SC3** [8] implements a consensus clustering approach using k-means clustering and Euclidean distance, identified as a traditional clustering method in this study.
- **scDeepCluster** [12] introduces a deep autoencoder using the ZINB loss, classified as a deep clustering method in this research.
- **DESC** [11] employs an autoencoder network for cell embedding and batch effect elimination, distinguished as a deep clustering method in this research.
- **scGNN** [15] combines three iterative multimodal autoencoders based on graph neural networks, recognized as a deep graph clustering method in this study.
- **Seurat** [14] features a built-in Phenograph clustering method for constructing cell graphs via community detection, identified as a graph clustering method in this research.
- **scAE** represents a simple deep clustering model constructed for comparative analysis, classified as a deep clustering method in this study.
- **scGAC** [17] introduces an attention mechanism in graph neural networks for efficient cellular graph construction, distinguished as a graph deep clustering method in this research.
- **scDFC** [18] merges cell attribute information with structural inter-cell information for clustering, recognized as a deep fusion clustering method in this study.

3) Implementation Details: The performance of the proposed algorithm was evaluated on an Ubuntu server featuring an Intel Core i7-6800K CPU, 64GB of DDR4 memory, and an NVIDIA TITAN Xp graphics card. The system utilized Ubuntu 22.04.2 LTS, and the algorithm was implemented in Python 3.6, using TensorFlow deep learning framework

TABLE II
ARI SCORES OF SCEGA AND BASELINE METHODS ACROSS EIGHT DATASETS, WITH THE TOP THREE RESULTS HIGHLIGHTED IN BOLD. '-' INDICATES THE ERROR OF THE METHOD ITSELF.

Datasets	Traditional Methods		Deep Clustering Methods			Deep Graph Clustering Methods				
	CIDR	SC3	scDeepCluster	DESC	scAE	scGNN	Seurat	scGAC	scDFC	**scEGA**
Biase	**1.000**	0.948	0.948	0.960	**1.000**	0.330	0.850	**1.000**	**1.000**	**1.000**
Darmanis	0.337	0.470	0.522	**0.536**	0.100	-	0.353	0.508	**0.533**	**0.549**
Enge	0.223	**0.531**	0.218	0.305	0.052	-	0.206	0.261	**0.368**	**0.483**
Bjorklund	0.457	0.721	0.310	0.412	-	0.438	0.056	**0.785**	**0.842**	**0.724**
Sun	0.268	**0.879**	0.783	0.603	0.276	0.465	0.182	**0.784**	**0.784**	**0.834**
Marques	0.100	**0.363**	**0.390**	0.269	-	-	0.172	0.283	0.260	**0.399**
Zeisel	0.167	0.420	**0.736**	0.279	-	-	0.119	**0.657**	0.317	**0.605**
Fink	0.225	0.146	**0.359**	0.212	0.179	-	0.063	0.328	**0.485**	**0.561**

version 1.12.0. The parameters of the random walk on the gene graph in the scEGA model were set to the default parameters of the node2vec algorithm. Encode layer sizes were set to (512, 256, 64), with the bottleneck layer established at 64. The model underwent pre-training for 200 epochs, followed by a training phase lasting 5000 epochs. Learning rates were set at 0.0002 for pre-training and 0.0005 for the training phase.

4) Evaluation: This study utilizes three evaluation metrics, detailed as follows:

- **Adjusted Rand Index (ARI)** [27] is a widely utilized metric for measuring the consistency between clustering results and true labels, necessitating labeled data. The formulation of this index is as follows:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}. \tag{16}$$

- **Normalized Mutual Information (NMI)** [28] is another commonly used metric to assess the similarity between clustering results and true labels, requiring labeled data. The formulation of this index is as follows:

$$\text{NMI} = \frac{2MI(U,V)}{H(U)+H(V)}. \tag{17}$$

- **Silhouette Coefficient (SC)** [29] serves as an internal evaluation metric for assessing clustering quality in an unsupervised manner, eliminating the need for ground truth labels. SC ranges from -1 to 1, where a value of 1 signifies high compactness and distinct separation between clusters, -1 signifies low compactness and poor separation, and 0 indicates overlapping clusters. Contrary to the aforementioned metrics, SC offers a more holistic evaluation by considering both within-cluster and between-cluster distances. In this study, SC is utilized to control the early stopping criterion in our Python implementation.

$$\text{SC} = \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{18}$$

B. Clustering Performance

We carried out a comprehensive series of experiments to assess the effectiveness of established benchmark clustering methods, encompassing our proposed scEGA model and nine other baseline methods. The findings unequivocally indicate that scEGA consistently achieved superior performance in the ARI, as detailed in Table II. scEGA consistently ranked within the top three in all comparative analyses and secured the best in half of the datasets (four out of eight).

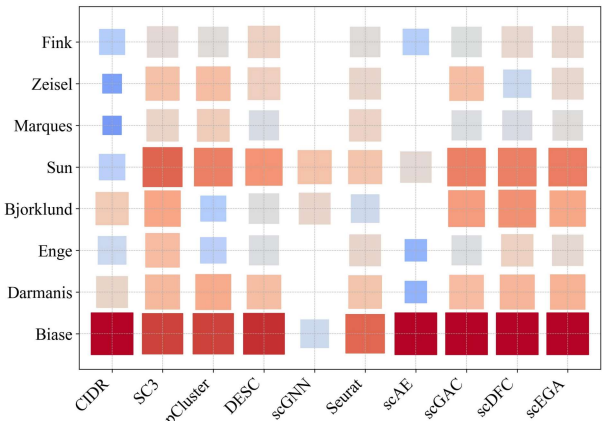


Fig. 4. The heatmap presents the clustering performance of scEGA as evaluated by NMI. The size and color intensity of each square in the heatmap correspond to the NMI values: larger and darker squares indicate higher NMI values.

To visually illustrate these findings, we constructed a heatmap based on NMI for depicting clustering performance, as shown in Figure 4. Each square's size in the heatmap is indicative of the NMI values' magnitude, with a gradation in color from lighter to darker shades to represent ascending NMI values. It is noteworthy that models such as CIDR, scDeepCluster, and scAE displayed comparatively lower clustering performance in NMI terms. Conversely, scEGA not only illustrated its dominance in ARI but also its significant stability in NMI. Given the complexity of the biological environment, which contributes to the intricate distribution of single-cell data, identifying a universally applicable clustering method poses a considerable challenge. Nonetheless, scEGA displayed high performance in nearly all the evaluated tasks.

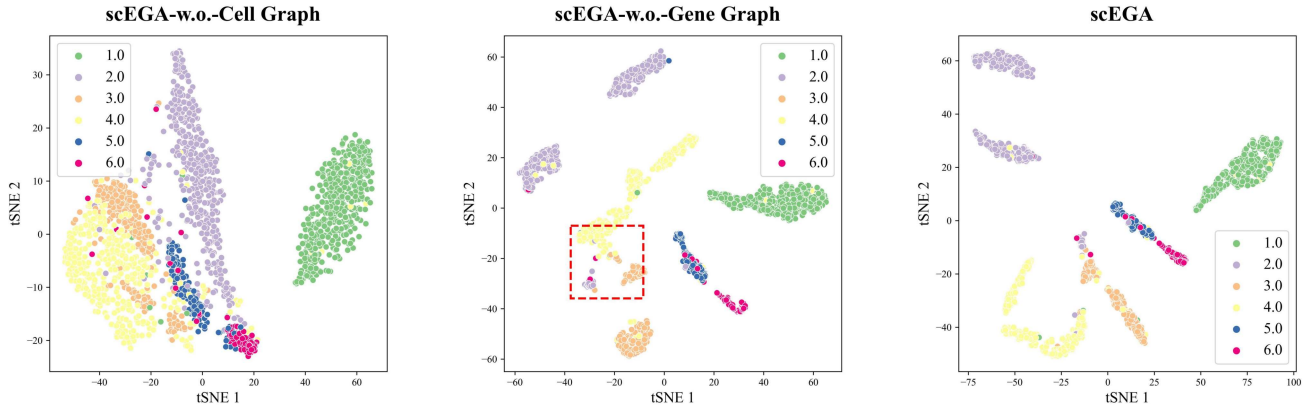


Fig. 5. The visualization illustrates different embeddings on the Sun dataset, resulting from the absence of certain modules in scEMC: scEGA 'w.o.' (without) the cell graph, scEGA 'w.o.' the gene graph, and scEGA inclusive of all modules.

TABLE III
THE ABLATION STUDY UTILIZING ARI VALUES: WE SUCCESSIVELY REMOVED THE CELL GRAPH AND GENE GRAPH FROM THE MODEL TO OBSERVE THE RESPECTIVE IMPACTS ON PERFORMANCE.

Dataset	scEGA-w.o.-Cell Graph	scEGA-w.o.-Gene Graph	scEGA
Biase	0.983	1.000	**1.000**
Darmanis	0.443	0.500	**0.549**
Enge	0.077	0.189	**0.483**
Bjorklund	-	0.513	**0.724**
Sun1	0.282	0.736	**0.834**
Marques	0.235	-	**0.399**
Zeisel	0.246	0.293	**0.605**
Fink	0.110	-	**0.561**

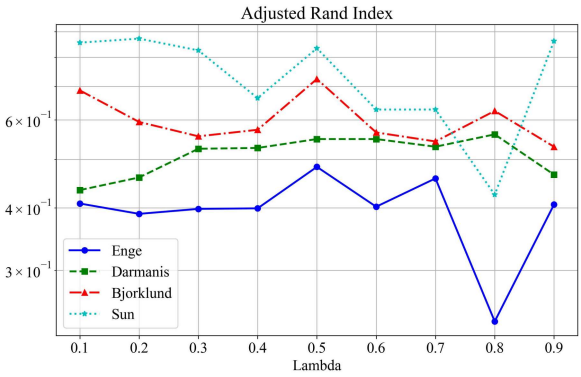


Fig. 6. Investigating optimal values for hyperparameter λ

C. Ablation Study

1) Quantitative Analysis: The dual-matrix alignment module, central to our scEGA model, hinges on the integration of the cell graph and gene graph. To investigate these modules' impact on clustering outcomes, we performed ablation experiments. Specifically, we developed two scEGA variants, one without the cell graph module and the other without the gene graph module. We evaluated their clustering performance in comparison to the complete scEGA model. As indicated in Table III, scEGA demonstrated superior clustering performance. Removing the cell graph module led to a significant decrease in performance, underscoring the cell-to-cell network's importance and its pivotal role in clustering. While the removal of the gene map module did not cause as substantial a decline in clustering performance, the noticeable decrease still emphasized the exogenous gene features' vital role in clustering. The integration of both graphs yielded the best performance, suggesting that our dual-matrix alignment module effectively and reliably enhances clustering.

2) Visualization Analysis: The quality of cell embeddings directly influences clustering performance. In this section, we present a series of visualization analyses on the embeddings of various scEGA variants. Specifically, we performed experiments by separately removing the cell graph module and the gene graph module from the model. We then saved the bottleneck layer and visualized it using t-SNE on the Sun dataset. Our findings indicate a significant decline in the quality of embeddings in the absence of these modules. As illustrated in Figure 5 (left), the omission of the cell graph led to poor clustering, thereby hindering the ability to differentiate between clusters. Highlighted in the red box in Figure 5 (mid), removing the gene graph caused a subset of cells to be incorrectly assigned to clusters. These findings underscore the critical role of both the cell graph and gene graph modules in scEGA in achieving high-quality cluster embeddings.

D. Parameter Sensitivity Analysis.

Our work underscores the considerable influence of dual-matrix alignment on clustering performance and introduces a weight coefficient in equation 11, represented by λ, to regulate the balance between cell loss and gene loss. This section presents an in-depth analysis of the hyperparameter λ and its impact on clustering outcomes. To ascertain the optimal weight partitioning, we carried out experiments across four distinct datasets: Enge, Darmanis, Bjorklund, and Sun. We tested various values of λ in these experiments. A smaller λ value signifies a lesser weight attributed to cell loss, whereas a larger λ value suggests a reduced weight for gene loss. Figure 6 displays the clustering performance of scEGA under varying

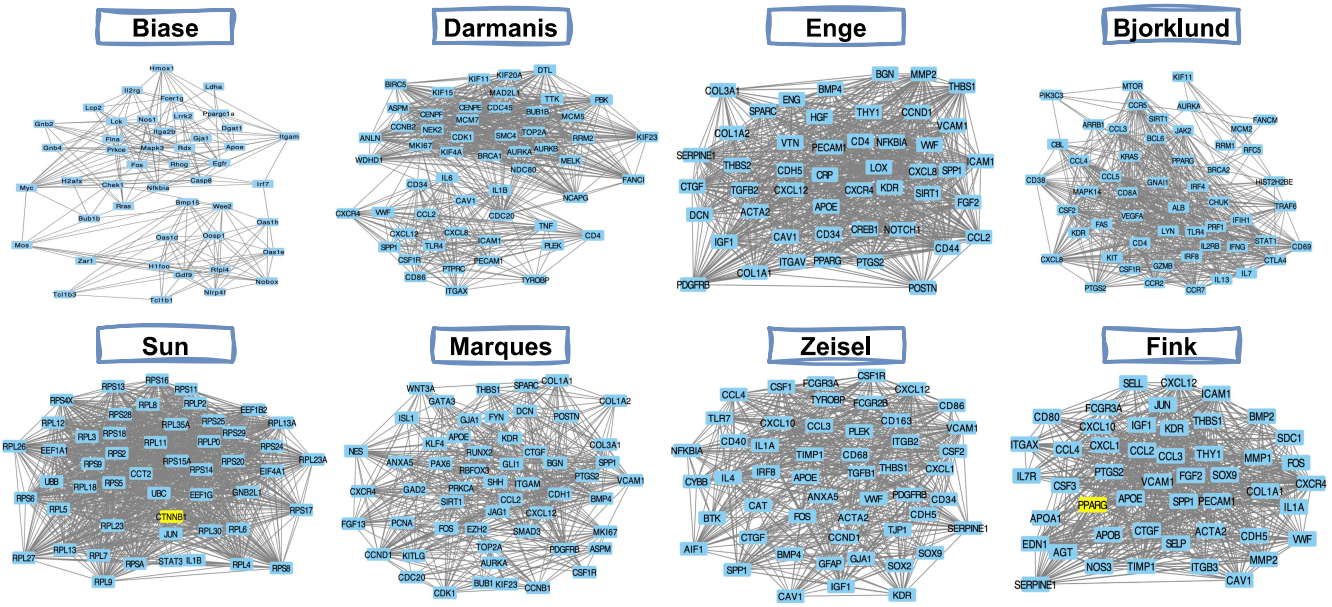


Fig. 7. The visualization of gene graphs pertaining to the eight benchmark datasets utilized in this study.

λ values. The findings indicate that a balanced dual-matrix weight distribution is key to enhancing clustering embeddings, with a λ value of 0.5 yielding consistently superior performance across all datasets.

E. Exogenous Gene Visualization Analysis

The core of this study lies in integrating exogenous gene association information into the clustering process, thereby enhancing the quality of cell embeddings. In this section, to clearly demonstrate the gene-to-gene network, we visualized the gene graph. Specifically, we entered gene sets corresponding to single-cell data into the renowned protein interaction network database, STRING, to acquire the gene adjacency tables. These PPI networks depict the associative relationships among genes. Subsequently, these tables were visualized using Cytoscape software, with the results displayed in Figure 7. We present eight gene graphs derived from a variety of scRNA datasets. This illustration substantiates the gene graph's existence, offering clear biological insights into gene features. Such insights are vital for investigating gene functionality, regulation, and disease mechanisms. Thus, maintaining the stability of gene features during the clustering process is of significant scientific relevance.

V. Conclusion

In conclusion, we have developed scEGA, an effective exogenous gene assisted clustering model for single-cell data, employing a dual-matrix alignment module to constrain cell and gene features. The dual-supervised optimization module enhances cluster embeddings while ensuring the stability of both cell and gene matrices during the optimization process. Our experimental findings show that the cell and gene graphs significantly improve embedding optimization, with our model surpassing other existing methods.

In the future, we aim to investigate innovative approaches to random walks on the gene graph for more effective representations. Additionally, we plan to examine more similarity measures of cells to construct a more precise cell graph [30]–[34]. Collaborative training presents another exciting research direction, as we believe that integrating cell and gene data can mutually enhance clustering effectiveness [35], [36].

References

[1] Dongqing Sun, Jin Wang, Ya Han, Xin Dong, Jun Ge, Rongbin Zheng, Xiaoying Shi, Binbin Wang, Ziyi Li, Pengfei Ren, et al. Tisch: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic acids research*, 49(D1):D1420–D1430, 2021.

[2] Lihong Peng, Feixiang Wang, Zhao Wang, Jingwei Tan, Li Huang, Xiongfei Tian, Guangyi Liu, and Liqian Zhou. Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Briefings in Bioinformatics*, 23(4):bbac234, 2022.

[3] Yunbin Zhang, Jingjing Song, Zhongwei Zhao, Mengxuan Yang, Ming Chen, Chenglong Liu, Jiansong Ji, and Di Zhu. Single-cell transcriptome analysis reveals tumor immune microenvironment heterogenicity and granulocytes enrichment in colorectal cancer liver metastases. *Cancer letters*, 470:84–94, 2020.

[4] Sheng Hu Qian, Meng-Wei Shi, Dan-Yang Wang, Justin M Fear, Lu Chen, Yi-Xuan Tu, Hong-Shan Liu, Yuan Zhang, Shuai-Jie Zhang, Shan-Shan Yu, et al. Integrating massive rna-seq data to elucidate transcriptome dynamics in drosophila melanogaster. *Briefings in Bioinformatics*, page bbad177, 2023.

[5] Qiao Liu, Wanwen Zeng, Wei Zhang, Sicheng Wang, Hongyang Chen, Rui Jiang, Mu Zhou, and Shaoting Zhang. Deep generative modeling and clustering of single cell hi-c data. *Briefings in Bioinformatics*, 24(1):bbac494, 2023.

[6] Jiaqian Yan, Ming Ma, and Zhenhua Yu. bmvae: a variational autoencoder method for clustering single-cell mutation data. *Bioinformatics*, 39(1):btac790, 2023.

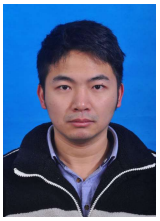
[7] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):1–11, 2017.

[8] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[9] Guo Mao, Zhengbin Pang, Ke Zuo, and Jie Liu. Gene regulatory network inference using convolutional neural networks from scrna-seq data. *Journal of Computational Biology*, 30(5):619–631, 2023.
[10] Xingyan Liu, Qunlun Shen, and Shihua Zhang. Cross-species cell-type assignment from single-cell rna-seq data by a heterogeneous graph neural network. *Genome Research*, 33(1):96–111, 2023.
[11] Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis. *Nature communications*, 11(1):2338, 2020.
[12] Tian Tian, Ji Wan, Qi Song, and Zhi Wei. Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191–198, 2019.
[13] Tian Tian, Jie Zhang, Xiang Lin, Zhi Wei, and Hakon Hakonarson. Model-based deep embedding for constrained clustering analysis of single cell rna-seq data. *Nature communications*, 12(1):1873, 2021.
[14] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.
[15] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1882, 2021.
[16] Zhuohan Yu, Yifu Lu, Yunhe Wang, Fan Tang, Ka-Chun Wong, and Xiangtao Li. Zinb-based graph embedding autoencoder for single-cell rna-seq interpretations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4671–4679, 2022.
[17] Yi Cheng and Xiuli Ma. scgac: a graph attentional architecture for clustering single-cell rna-seq data. *Bioinformatics*, 38(8):2187–2193, 2022.
[18] Dayu Hu, Ke Liang, Sihang Zhou, Wenxuan Tu, Meng Liu, and Xinwang Liu. scdfc: A deep fusion clustering method for single-cell rna-seq data. *Briefings in Bioinformatics*, page bbad216, 2023.
[19] Fernando H Biase, Xiaoyi Cao, and Sheng Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Genome research*, 24(11):1787–1796, 2014.
[20] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.
[21] Martin Enge, H Efsun Arda, Marco Mignardi, John Beausang, Rita Bottino, Seung K Kim, and Stephen R Quake. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell*, 171(2):321–330, 2017.
[22] Åsa K Björklund, Marianne Forkel, Simone Picelli, Viktoria Konya, Jakob Theorell, Danielle Friberg, Rickard Sandberg, and Jenny Mjösberg. The heterogeneity of human cd127+ innate lymphoid cells revealed by single-cell rna sequencing. *Nature immunology*, 17(4):451–460, 2016.
[23] Zhe Sun, Li Chen, Hongyi Xin, Yale Jiang, Qianhui Huang, Anthony R Cillo, Tracy Tabib, Jay K Kolls, Tullia C Bruno, Robert Lafyatis, et al. A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nature communications*, 10(1):1649, 2019.
[24] Sueli Marques, David van Bruggen, Darya Pavlovna Vanichkina, Elisa Mariagrazia Floriddia, Hermany Munguba, Leif Väremo, Stefania Giacomello, Ana Mendanha Falcao, Mandy Meijer, Åsa Kristina Björklund, et al. Transcriptional convergence of oligodendrocyte lineage progenitors during development. *Developmental cell*, 46(4):504–517, 2018.
[25] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.
[26] Emily E Fink, Surbhi Sona, Uyen Tran, Pierre-Emmanuel Desprez, Matthew Bradley, Hong Qiu, Mohamed Eltemamy, Alvin Wee, Madison Wolkov, Marlo Nicolas, et al. Single-cell and spatial mapping identify cell types and signaling networks in the human ureter. *Developmental Cell*, 57(15):1899–1916, 2022.
[27] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.
[28] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.
[29] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
[30] Hong Wang, Xiaoyan Lu, Hewei Zheng, Wencan Wang, Guosi Zhang, Siyu Wang, Peng Lin, Youyuan Zhuang, Chong Chen, Qi Chen, et al. Rnasmc: A integrated tool for comparing rna secondary structure and evaluating allosteric effects. *Computational and Structural Biotechnology Journal*, 2023.
[31] Shixiong Zhang, Xiangtao Li, Jiecong Lin, Qiuzhen Lin, and Ka-Chun Wong. Review of single-cell rna-seq data clustering for cell-type identification and characterization. *RNA*, 29(5):517–530, 2023.
[32] Michelle Ying Ya Lee, Klaus H Kaestner, and Mingyao Li. Benchmarking algorithms for joint integration of unpaired and paired single-cell rna-seq and atac-seq data. *bioRxiv*, pages 2023–02, 2023.
[33] Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326, 2019.
[34] Mei Li, Ya-Wen Zhang, Ze-Chang Zhang, Yu Xiang, Ming-Hui Liu, Ya-Hui Zhou, Jian-Fang Zuo, Han-Qing Zhang, Ying Chen, and Yuan-Ming Zhang. A compressed variance component mixed model for detecting qtns and qtn-by-environment and qtn-by-qtn interactions in genome-wide association studies. *Molecular Plant*, 15(4):630–650, 2022.
[35] Tien-Phat Nguyen, Trong-Thang Pham, Tri Nguyen, Hieu Le, Dung Nguyen, Hau Lam, Phong Nguyen, Jennifer Fowler, Minh-Triet Tran, and Ngan Le. Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1981–1990, 2023.
[36] Jean Ogier du Terrail, Armand Leopold, Clément Joly, Constance Béguier, Mathieu Andreux, Charles Maussion, Benoît Schmauch, Eric W Tramel, Etienne Bendjebbar, Mikhail Zaslavskiy, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nature Medicine*, pages 1–12, 2023.

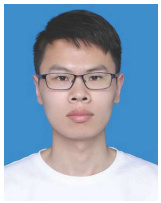
ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (no. 2020AAA0107100), and the National Natural Science Foundation of China (no. 62325604, 62276271).



Dayu Hu is currently pursuing a Ph.D. degree at the National University of Defense Technology (NUDT). Before joining NUDT, he got his BSc degree at Northeastern University (NEU). His current research interests include graph learning and bioinformatics. He has published several papers and served as PC member/ Reviewer in highly regarded journals and conferences such as ACM MM, AAAI, TNNLS, TKDE, TCBB, etc.

Ke Liang is currently pursuing a Ph.D. degree at the National University of Defense Technology (NUDT). Before joining NUDT, he got his BSc degree at Beihang University (BUAA) and received his MSc degree from the Pennsylvania State University (PSU). His current research interests include knowledge graphs, graph learning, and healthcare AI. He has published several papers in highly regarded journals and conferences such as SIGIR, AAAI, ICML, ACM MM, IEEE TNNLS, IEEE TKDE, etc.



Hao Yu is presently pursuing a Ph.D. degree at the National University of Defense Technology (NUDT). He earned a B.Eng degree in computer science from Inner Mongolia University, Hohhot, China, in 2019, and, subsequently, in 2022, obtained an MA.Sc in cyberspace science and technology from Beijing Institute of Technology, Beijing, China. His current research focuses on AI security and Federated Learning. He has authored several papers in top-level journals and conferences, such as IEEE TIFS, TDSC, and ACM MM, and served as a Reviewer for highly regarded journals, such as IEEE TIFS, IEEE TKDE, ACM TOIS, etc.

Xinwang Liu received his PhD degree from National University of Defense Technology (NUDT), China. He is now Professor of School of Computer, NUDT. His current research interests include kernel learning and unsupervised feature learning. Dr. Liu has published 60+ peer-reviewed papers, including those in highly regarded journals and conferences such as IEEE T-PAMI, IEEE T-KDE, IEEE T-IP, IEEE T-NNLS, IEEE T-MM, IEEE T-IFS, ICML, NeurIPS, ICCV, CVPR, AAAI, IJCAI, etc. He serves as the associated editor of TNNLS, TCYB and Information Fusion Journal. More information can be found at https://xinwangliu.github.io/.