biology

MDPI

Article

scVGATAE: A Variational Graph Attentional Autoencoder Model for Clustering Single-Cell RNA-seq Data

Lijun Liu, Xiaoyang Wu, Jun Yu *, Yuduo Zhang, Kaixing Niu and Anli Yu

School of Science, Dalian Minzu University, Dalian 116600, China; manopt@163.com (L.L.); wuxiaoyangzhuxiaoming@outlook.com (X.W.); zhangyuduo@dlnu.edu.cn (Y.Z.); kx65a4@outlook.com (K.N.); alyu1230@163.com (A.Y.)

* Correspondence: yshshng@163.com

Simple Summary: Due to the rapid development of single-cell RNA sequencing technology, the volume of single-cell RNA sequencing data has grown exponentially. Traditional clustering methods have proven increasingly difficult to cluster this large-scale and highly complex single-cell RNA sequencing data. Although many unsupervised clustering methods based on deep neural networks have been developed to cluster cell subpopulations, these methods are complex in models and poor in scalability. In this paper, we propose a novel clustering method for single-cell RNA sequencing, which successfully combines the advantages of these two clustering models, maintaining high clustering performance while also preserving the stable computational efficiency of traditional clustering methods. Experiments conducted on nine public datasets have demonstrated that our proposed novel clustering method for single-cell RNA sequencing outperforms both classic and state-of-the-art clustering methods.

Abstract: Single-cell RNA sequencing (scRNA-seq) is now a successful technology for identifying cell heterogeneity, revealing new cell subpopulations, and predicting developmental trajectories. A crucial component in scRNA-seq is the precise identification of cell subsets. Although many unsupervised clustering methods have been developed for clustering cell subpopulations, the performance of these methods is prone to be affected by dropout, high dimensionality, and technical noise. Additionally, most existing methods are time-consuming and fail to fully consider the potential correlations between cells. In this paper, we propose a novel unsupervised clustering method called scVGATAE (Single-cell Variational Graph Attention Autoencoder) for scRNA-seq data. This method constructs a reliable cell graph through network denoising, utilizes a novel variational graph autoencoder model integrated with graph attention networks to aggregate neighbor information and learn the distribution of the low-dimensional representations of cells, and adaptively determines the model training iterations for various datasets. Finally, the obtained low-dimensional representations of cells are clustered using kmeans. Experiments on nine public datasets show that scVGATAE outperforms classical and state-of-the-art clustering methods.

Keywords: scRNA-seq; unsupervised clustering; variational graph autoencoder; graph attention networks

check for updates

Citation: Liu, L.; Wu, X.; Yu, J.; Zhang, Y.; Niu, K.; Yu, A. scVGATAE: A Variational Graph Attentional Autoencoder Model for Clustering Single-Cell RNA-seq Data. *Biology* **2024**, *13*, 713. https://doi.org/10.3390/biology13090713

Academic Editor: Edward L. Braun

Received: 2 August 2024
Revised: 6 September 2024
Accepted: 7 September 2024
Published: 11 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Single-cell RNA sequencing (scRNA-seq) technology has revolutionized research progress in the field of biology. It can analyze gene expression patterns at the single-cell level, providing an unprecedented perspective for exploring cell heterogeneity and understanding life processes and disease mechanisms [1]. Through scRNA-seq technology, researchers can obtain transcriptome information of individual cells, and then depict complex cell graphs, providing an important basis for precision medicine and drug development [2,3]. The main steps of scRNA-seq analysis usually include data quality control,

gene expression quantification, feature extraction, dimensionality reduction analysis, and cell clustering. Among them, dimensionality reduction analysis and cell clustering undoubtedly play a crucial role, as they are of decisive significance for accurately inferring cell development trajectories and effectively identifying rare cell clusters [4].

However, due to the inherent high dimensionality, technical noise, common dropout events, and batch effects of scRNA-seq data are inevitable [5]. Traditional clustering methods often face multiple challenges such as pattern recognition difficulties, significant noise interference, severe information loss, and unstable results when processing these data, which greatly limits their application effectiveness in single-cell analysis. Therefore, developing efficient and accurate scRNA-seq data clustering methods has become a research hotspot in the current field of bioinformatics.

Cell clustering analysis primarily relies on dimensionality reduction techniques and traditional clustering methods. Given the high-dimensional nature of scRNA-seq data, the primary objective is to derive a low-dimensional representation. Common dimensionality reduction techniques include principal component analysis (PCA), t-distributed stochastic neighborhood embedding (t-SNE) [6], diffusion maps, and uniform manifold approximation and projection (UMAP) [7]. Subsequently, traditional clustering methods, such as k-means and hierarchical clustering, are applied to the learned low-dimensional representations of cells. These methods cluster cells based on their similarity. While these traditional clustering methods are straightforward, efficient, and highly interpretable, they often fall short when dealing with high-dimensional and complex data.

To enhance clustering performance on scRNA-seq data, many methods integrate various similarity metrics and clustering outcomes. For instance, SC3 [8] employs diverse approaches to calculate multiple cell similarity matrices in parallel, applies k-means to these matrices, and subsequently applies hierarchical clustering to a consensus matrix formed from the results. SAME [9], a hybrid model-based clustering technique, seeks to derive diverse solutions from various clustering methods and generates an enhanced ensemble solution by selecting the most diverse subset. RCSL [10] assesses global and local cell relationships to differentiate cell types, evaluates global similarity using Spearman's rank correlation, and learns neighborhood representations for local similarity. These are then linearly combined to produce block diagonal matrices, leading to the final clustering results. While these methods enhance the robustness of clustering results, reflecting cell similarities and differences more accurately, pairwise similarity only captures superficial relationships among cells and fails to capture the deep complex underlying relationships, leading to clustering results that may be insufficiently accurate or comprehensive.

In contrast, deep neural networks possess robust feature learning capabilities, enabling them to more effectively utilize information from scRNA-seq data to enhance clustering accuracy. Examples include scGMAI [11], DCA [12], scDeepCluster [13], and scVAE [14]. Specifically, scGMAI employs autoencoders to concurrently learn feature representations and cluster assignments, reconstructs gene expression values using stacked autoencoders, selects significant independent features via FastICA, and clusters cells using Gaussian mixture models. The DCA method substitutes the conventional mean squared error (MSE) loss function with a loss function based on the zero-inflated negative binomial (ZINB) model to more accurately characterize scRNA-seq data. scDeepCluster also employs an autoencoder based on the ZINB model to enhance the learning of latent features for subsequent clustering. The scVAE method utilizes a deep variational autoencoder for clustering scRNA-seq data.

Although these clustering methods based on deep neural networks have significantly enhanced the accuracy of clustering results, they exhibit notable shortcomings. Firstly, they typically concentrate on learning the intrinsic features of cells, overlooking the structural relationships among them. Indeed, this structural information can effectively unveil the latent similarities among cells and more precisely capture their actual relationships, leading to more accurate clustering outcomes [15]. To naturally embed both the cell feature expression information and the structural relationships among cells into the learned low-dimensional

representations of cells, the clustering analysis method based on Graph Neural Network (GNN) for scRNA-seq data has emerged as the need has arisen. These methods can not only deeply explore the intrinsic structure of single-cell data, but also reveal the complex interactions between cells. Several methods, including CellVGAE [16] and scVGAE, utilize the GNN structure for cell clustering. CellVGAE uses the structural relationships between cells (such as KNN) and gene expression values as input features of the model, integrates graph attention mechanism into variational graph autoencoder, learns high-quality low dimensional representations of cells, and performs k-means clustering on the obtained low dimensional representations of cells. scVGAE integrates Graph Convolutional Networks (GCNs) into the Variational Autoencoder framework and utilizes the ZINB loss function to effectively address missing events in scRNA-seq data. Finally, the interpolated cell expression matrix is clustered. However, since VGAE merely employs the inner product of low-dimensional cell representations as a decoder to reconstruct the original graph and relies on reconstruction loss along with KL divergence regularization as loss functions to optimize model parameters, it fails to accurately capture the complexity and diversity of cellular structural relationships. It exhibits sensitivity to noise and outliers, and the inflexible reconstruction approach severely limits the learning of low-dimensional cell representations. There is a need to employ more sophisticated decoder structures to more effectively capture the critical information and structural features among cells. Recent methods, such as scGNN2.0 [17] and scGAC [18], have employed graph attention autoencoders for the clustering analysis of scRNA-seq data, achieving favorable clustering results. scGNN2.0 leverages GAT and a multimodal autoencoder to formulate and aggregate relationships between cells. scGAC employs graph autoencoders with attention mechanisms to learn cluster-friendly feature representations and utilizes self-optimizing clustering to enhance clustering performance. Nonetheless, these graph autoencoder models based entirely on GAT entail substantial computational costs and memory footprint, and they lack sufficient convenience and scalability for massive and high-dimensional scRNA-seq data.

To overcome these challenges, we propose a novel clustering method for scRNA-seq data named 'scVGATAE' (Single-cell Variational Graph Attention Autoencoder). Firstly, a cell graph is constructed using the Pearson correlation coefficient. Due to the characteristics of scRNA-seq data, scVGATAE employs the diffusion-based denoising strategy of NE to capture the potential similarities among cells from the high-order structure of the cell graph [19]. Based on the cell graph denoised by NE, scVGATAE learns the low-dimensional representations of cells through a variational graph autoencoder integrated with graph attention networks. Specifically, the inner product decoder of VGAE is replaced with two stacked graph attention layers, and a custom loss function is utilized to optimize the model parameters. Subsequently, we adopted an adaptive strategy to select appropriate training times on different datasets and applied k-means clustering analysis. To benchmark the performance of the proposed method, we compared scVGATAE with the six most popular scRNA-seq clustering methods on nine real and effective scRNA-seq datasets. The results showed that scVGATAE not only significantly improved the clustering performance of scRNA-seq data, but also maintained relatively stable running time at various cell orders of magnitude.

2. Materials and Methods

2.1. Architecture of scVGTAE

The initial input of scVGATAE is the raw UMI count matrix of sRNA seq, where each row represents a single cell and each column represents a gene. Each element in the matrix represents the count of Unique Molecular Identifiers (UMIs) for a specific gene in the corresponding cell. However, due to various technical noises and variations that may exist during the experimental process, the UMI counting matrix requires quality control and filtering, as well as pre-processing operations such as matrix normalization and logarithmization, to obtain the preprocessed matrix $X_{n\times m}$. Secondly, scVGATAE calculates the similarity matrix $S_{n\times n}$ between cells based on the Pearson correlation coefficient and

denoises it through network enhancement (NE) to obtain matrix $E_{n\times n}$. Based on matrix $E_{n\times n}$, perform the nearest k-nearest neighbor algorithm and normalize the resulting matrix to obtain cell graph $A_{n\times n}$. The above is the way to obtain cell graph data among the multimodal input data for the variational graph attention autoencoder. The other cell feature expression matrix $Y_{n\times d}$ is obtained by performing PCA dimensionality reduction on the $X_{n\times m}$ matrix. The encoder of the variational graph attention autoencoder (graph convolutional network) takes the cell feature expression matrix $Y_{n\times d}$ and cell graph $A_{n\times n}$ as multimodal inputs, learns the mean μ and variance σ of the low dimensional vector representation of cells, and then uses the reparameterization technique to sample the variable $Z_{n\times b}$ from $N(\mu, \sigma^2)$. Then, the decoder (graph attention network) is used to reconstruct the cell feature expression matrix $Y'_{n\times d}$. The loss function comprises two parts: the mean absolute error between the reconstructed cell feature expression matrix $Y'_{n\times d}$ and the original cell feature expression matrix $Y_{n\times d}$ as well as the divergence between the distribution $N(\mu, \sigma^2)$ of the cell's low-dimensional representation vectors and a normal distribution. Finally, repeat the encoding and decoding operations until the stop condition is reached. The final low dimensional representation vector $Z_{n\times b}$ of cells is obtained through the trained model, and the k-means clustering method is used for cell clustering. The number of cell clusters in the k-means clustering is determined by the Leiden diagram clustering method (based on optimized modular community detection). Firstly, the original cell feature expression matrix $Y_{n\times d}$ is subjected to cell–cell neighborhood graph calculation. Then, the obtained neighborhood graph is subjected to the Leiden clustering algorithm, and finally, the number of cell clusters is obtained. Figure 1 shows the main architecture of scVGATAE.

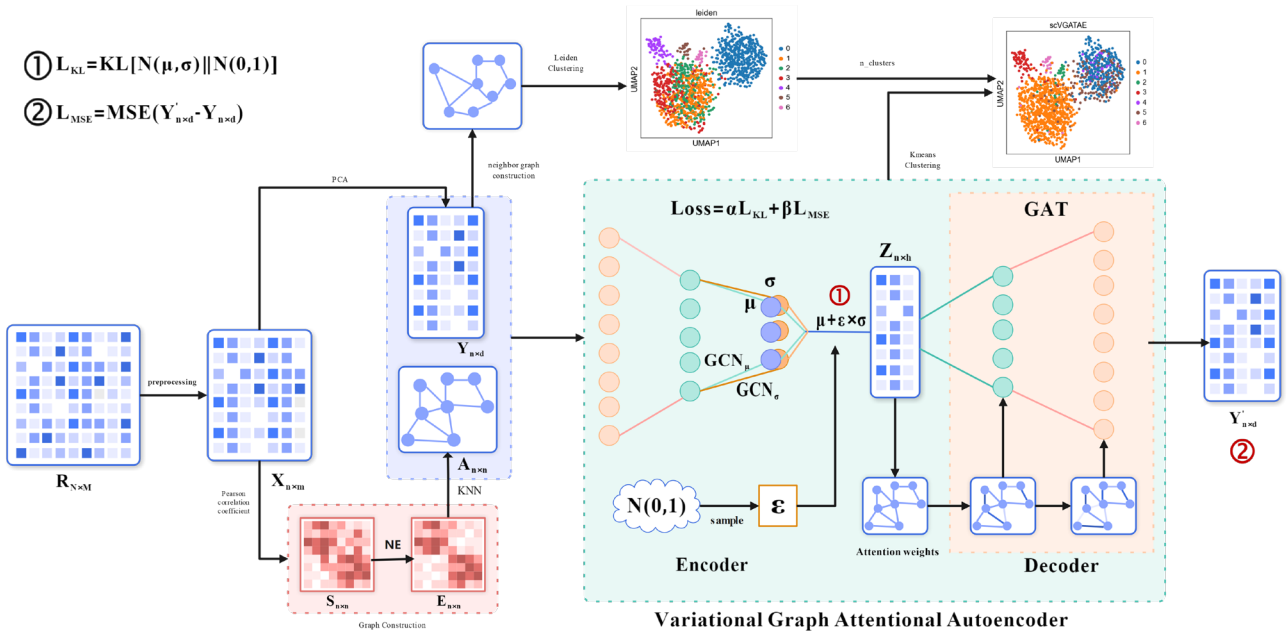


Figure 1. Architecture of scVGATAE. Firstly, the raw gene expression matrix $R_{N\times M}$ undergoes data preprocessing to obtain a preprocessed matrix $X_{n\times m}$. To obtain the two types of input data required for the variational graph attention autoencoder, namely the cell feature expression matrix $Y_{n\times d}$ and the cell graph $A_{n\times n}$, two separate operations are performed on the matrix $X_{n\times m}$. One operation involves applying PCA dimensionality reduction to the matrix $X_{n\times m}$ in order to reduce computational complexity. The other operation utilizes Pearson correlation coefficients to calculate a similarity matrix $S_{n\times n}$ between cells, which is then denoised through Network Enhancement (NE) to obtain the matrix $E_{n\times n}$. Based on the matrix $E_{n\times n}$, the k-nearest neighbors algorithm is executed, and the resulting matrix is normalized to ultimately derive the cell graph $A_{n\times n}$. Next, the cell feature

expression matrix $Y_{n\times d}$ and the cell graph $A_{n\times n}$ are input into the variational graph attention autoencoder (VGATAE) model. The encoder, based on variational inference, learns the mean μ (represented by purple nodes) and variance σ (represented by yellow nodes) of the cells' low-dimensional vector representations. Both μ and σ are learned through a two-layer graph convolutional network (GCN), sharing the first layer but having their own second-layer GCNs, GCN_μ and GCN_σ, respectively. To avoid sampling operations interfering with gradient descent, scVGATAE samples a value ε from the standard normal distribution $N(0,1)$ and then calculates the cells' low-dimensional representation $Z_{n\times h}$ using the formula $Z = \mu + \varepsilon \times \sigma$. Subsequently, a decoder composed of two stacked graph attention networks reconstructs the cell feature expression matrix $Y'_{n\times d}$. This encoding and decoding process is repeated until a stopping condition is met. Finally, the obtained low dimensional representation $Z_{n\times h}$ of cells is subjected to k-means clustering, and the number of cell clusters is calculated using the Leiden algorithm.

2.2. Construction of a Cell Graph Based on NE Denoising

In order to share information between similar cells and learn a more friendly distribution of low dimensional cell feature vectors, the variational graph attention autoencoder needs to construct a graph structure that can more accurately capture the similarity between cells, in order to obtain a highly reliable cell map.

Firstly, the Pearson correlation coefficient is used to construct the initial cell similarity matrix $S_{n\times n}$ to evaluate the degree of similarity in feature expression between different cells. The formula is as follows

$$S_{ij} = \frac{\sum_{i=1,j=1}^{n} (x_i - \overline{x})(x_j - \overline{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{j=1}^{n} (x_j - \overline{x})^2}}, \tag{1}$$

However, scRNA-seq data may contain missing values or outliers, which reduces the reliability of the Pearson correlation coefficients between pairs of cells. This can mislead the sharing of information between similar cells, ultimately leading to suboptimal cell clustering results.

We adopt Network Enhancement (NE) to improve the signal-to-noise ratio of $S_{n\times n}$. NE has designed an effective diffusion denoising strategy that utilizes random walks with a length of no more than 3 and information regularization techniques to optimize the network structure. Among them, a random walk with a length not exceeding 3 means starting from one node and randomly selecting its adjacent nodes for up to three jumps. Through this method, local structural information in the network can be captured, especially the relationships between high weight edges. This local information is crucial for denoising and optimizing network structures. Secondly, information regularization is used to ensure that the diffusion process does not excessively distort the original network structure, while effectively eliminating noise. Finally, NE updates the initial cell similarity matrix $S_{n\times n}$ using a diffusion based method to obtain a cell enhanced similarity matrix $\hat{S}_{n\times n}$.

Specifically, first, based on the initial cell similarity matrix $S_{n\times n}$, calculate the transition matrix P and the local network matrix $\mathcal{T}$, which are, respectively, represented as follows:

$$P_{ij} = \frac{S_{ij}}{\sum_{k\in\mathcal{N}_i} S_{ik}}, j \in \mathcal{N}_i, \tag{2}$$

$$\mathcal{T}_{ij} = \sum_{k=1}^{n} \frac{P_{ik} P_{jk}}{\sum_{v=1}^{n} P_{vk}}, \tag{3}$$

Among them, $\mathcal{N}_i$ represents the set of k neighboring nodes closest to node i; After extensive experimental verification, this paper has $k = 10$.

Subsequently, NE uses the local network matrix $\mathcal{T}$ to update the initial cell similarity matrix $S_{n\times n}$, which is the NE diffusion process. The process is as follows:

$$\hat{S} = \alpha\mathcal{T} \times S \times \mathcal{T} + (1-\alpha)\mathcal{T}, \tag{4}$$

Among them, α is the regularization parameter.

Finally, obtain the denoising similarity matrix $E_{n\times n}$ according to the following formula.

$$E_{ij} = \begin{cases} S_{ij} & if\ \hat{S}_{ij} \geq t \\ 0 & otherwise \end{cases}, \tag{5}$$

where t is a predefined threshold. If $\hat{S}_{ij}$ is too small (less than t), S_{ij} will be considered as a noise edge and E_{ij} will take on a value of zero. Then, we perform the nearest k-nearest neighbor algorithm and normalization on matrix E_{ij} and finally obtain cell graph $A_{n\times n}$.

2.3. Variational Graph Attention Autoencoder

Inspired by the Variational Graph Autoencoder (VGAE) [20] and Graph Attention Network (GAT) [21], we designed a Variational Graph Attention Autoencoder. Firstly, the cell graph $A_{n\times n}$ and cell feature expression matrix $Y_{n\times d}$ are input. An encoder, composed of two graph convolutional network layers, is used to learn the mean μ and variance σ of the low-dimensional vector representations of the cells. Subsequently, the reparameterization trick is employed to sample the low-dimensional vector representations $Z_{n\times b}$ of the cells from $N(\mu, \sigma)$. Finally, a decoder with two stacked graph attention network layers is used to reconstruct the cell feature expression matrix $Y'_{n\times d}$ from the low-dimensional representation $Z_{n\times b}$. The variational graph attention autoencoder learns the network parameters through a custom loss function, embedding the network topology information of the cell graph into the low-dimensional $Z_{n\times b}$ representations of the cells.

2.3.1. Encoder Based on the Graph Convolutional Network

The encoder is a stacked two-layer graph convolutional network

$$q(Z|Y, A) = \prod_{i=1}^{n} q(Z_i|Y, A), \tag{6}$$

where $q(Z_i|Y, A) = N\left(z_i|\mu_i, diag\left(\sigma_i^2\right)\right)$. μ is the mean of the low dimensional vector representation $Z_{n\times b}$ of the cell to be learned, $\mu = GCN_\mu(Y, A)$. σ is the variance of the low dimensional representation $Z_{n\times b}$ of the cell to be learned, and $log\sigma = GCN_\sigma(Y, A)$.

The definition of a two-layer graph convolutional network is as follows:

$$GCN_\mu(Y, A) = \widetilde{A}ReLU\left(\widetilde{A}XW_0\right)W_1, \tag{7}$$

$$GCN_\sigma(Y, A) = \widetilde{A}ReLU\left(\widetilde{A}XW_0\right)W_2, \tag{8}$$

where $\widetilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. GCN_μ and GCN_σ share the first layer of graph convolutional network, but have their own respective second layers of graph convolutional network.

The reparameterization trick helps us sample from the learned distribution of low-dimensional cell representation vectors $N\left(\mu, \sigma^2\right)$ while preserving gradient information, thereby optimizing the model parameters. The formula is as follows:

$$Z = \mu + \varepsilon \times \sigma, \tag{9}$$

where ε is sampled from the standard normal distribution $N(0, 1)$.

2.3.2. Decoder Based on the Graph Attention Network

The Variational Graph Autoencoder uses the inner product of the low-dimensional cell representations $Z_{n\times b}$ as the decoder to reconstruct the cell graph $A_{n\times n}$, as follows:

$$\widetilde{A} = sigmoid\left(ZZ^T\right), \tag{10}$$

In fact, this decoder, which relies solely on the dot product operation of the low-dimensional cell representation matrix Z to reconstruct the cell graph, cannot accurately and flexibly reflect the true relationships between cells, limiting the reconstruction effect. This, in turn, affects the parameter optimization during backpropagation, ultimately restricting the richness and expressiveness of the distribution of the low-dimensional cell representations.

Therefore, we introduce graph attention networks into the variational graph autoencoder, using two stacked graph attention layers as the decoder. The graph attention layers learn the features of cells by aggregating the features of neighboring cells with different weights. Since the weights are dynamically assigned based on the similarity between a cell and its neighboring cells, it can fully capture the complexity and diversity of the data in the low-dimensional cell representations $Z_{n\times b}$ and flexibly embed topological information into the reconstructed cell feature matrix $Y'_{n\times d}$, enhancing the reconstruction effect of the decoder.

Specifically, the computation of the graph attention network layer is divided into the following two steps:

(1) Calculate the attention coefficient

For cell i, calculate the similarity coefficient between its neighbors j and itself one by one, using the formula

$$e_{ij} = LeakyReLU\left(a^T\left[Wh_i \| Wh_j\right]\right), j \in \mathcal{N}_i, \tag{11}$$

where h_i *and* h_j are, respectively, the input features of cell i and cell j, $\mathcal{N}_i$ is the set of neighbors of cell i, W is a learnable linear transformation matrix, $\|$ is a concatenation operation, a is a learnable weight vector, and $LeakyReLU$ is a nonlinear activation function.

Normalize the correlation coefficient using the $softmax$ function to obtain the attention coefficient, with the formula

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k\in\mathcal{N}_i} exp(e_{ik})}, \tag{12}$$

where a_{ij} is the attention coefficient, indicating the importance of cell i to cell j.

(2) Weighted Sum

Based on the calculated attention coefficients, a weighted sum of the features of all neighboring cells is performed, and the formula is

$$h'_i = \sigma\left(\sum_{j\in\mathcal{N}_i} a_{ij} W h_j\right), \tag{13}$$

where σ is a nonlinear activation function. h'_i is a new feature of cell i.

2.3.3. Composition of Loss Function

The Mean Squared Error (MSE) loss function measures the error of the model by calculating the average of the squares of the differences between the reconstructed cell feature matrix $Y'_{n\times d}$ and the original cell feature expression matrix $Y_{n\times d}$. The formula is as follows

$$L_{MSE} = \frac{1}{n}\left(Y - Y'\right)^2, \tag{14}$$

where n is the total number of cells.

Similar to the Variational Graph Autoencoder, we add the KL divergence $\left((\mu, \sigma^2) \| (0,1)\right)$ between each cell's independent normal distribution and the standard normal distribution to the loss function, forcing each normal distribution to approximate the standard normal distribution

$$L_{KL} = \frac{1}{2}\sum_{i=1}^{d}\left(\mu_{(i)}^2 + \sigma_{(i)}^2 - log\sigma_{(i)}^2 - 1\right), \tag{15}$$

Therefore, the loss function of the variational graph attention autoencoder is defined as:

$$L = \alpha L_{MSE} + \beta L_{KL}, \tag{16}$$

where αandβ are coefficients that control the relative weights of two loss functions.

2.3.4. Adaptive Iterative Training

For different datasets, we provide a method to adaptively select the number of epochs for model training. The Silhouette coefficient [22] is calculated based on the low-dimensional cell representations and the predicted labels from k-means clustering to detect a declining trend in the clustering performance of the model. If a significant decrease in performance is detected, the training is stopped.

The Silhouette coefficient of a single cell is defined for each individual cell

$$s = \frac{b - a}{max(a, b)}, \tag{17}$$

where a represents the average distance between a cell and all other cells within the same cluster and b represents the average distance between that cell and all cells in the next closest cluster. The overall average of the Silhouette coefficients across all cells can be used as the Silhouette coefficient for the entire group of cells.

2.4. Datasets

The proposed scVGATAE unsupervised clustering method was evaluated for clustering performance on 9 real and effective scRNA-seq datasets, each containing cells labeled as previously or validated in previous studies. As shown in Table 1, the cell count of these datasets varies significantly, ranging from a few dozen cell samples to thousands of them, covering a wide range of data needs across various scales. Meanwhile, these data are sourced from diverse organizations and databases, fully demonstrating the richness and diversity of data sources. For example, the Adipose [23] dataset comes from the fat of a 36 year old male on the Human Cell Landscape (HCL) database, which contains 1372 cells and 8 cell types. However, the dataset Cerebellum [23] not only has a large scale, containing 7324 cells, but also contains an extremely rich variety of cell types, totaling 19 cell types. In addition, the real cell labels of these datasets are provided by the original database.

Table 1. Details of the 9 evaluated datasets.

Datasets	Databases	Cells	Genes	Subtypes
Biase [24]	GEO	49	21489	3
PBMC6k [25]	GEO	5419	32738	10
Goolam [26]	GEO	124	41480	5
Adipose [23]	HCL	1372	12154	8
Liver [23]	HCL	1811	12631	16
Cerebellum [23]	HCL	7324	19244	19
Heart [23]	HCL	1308	12689	12
Kolodzisjczyk [27]	Array Express	704	10685	3
Baron [28]	GEO	822	14878	13

In order to improve the generalization ability of the scVGATAE model, we adopted different data preprocessing strategies for datasets from different databases and scales. First,

quality control and normalization were performed on the raw Unique Molecular Identifier (UMI) count matrix $R_{M\times N}$ of all scRNA-seq data to obtain the preprocessed matrix $X_{n\times m}$. We preprocessed the raw scRNA-seq data using the Python software package SCANPY (1.10.2) [29]. In order to provide a more reliable and accurate dataset for subsequent analysis, we filtered out low-quality and damaged cells, i.e., cells with less than 200 expressed genes and genes that may have been generated due to sequencing errors or sample contamination, i.e., genes expressed in less than three cells. Then, we normalized and log-transformed the data for the Biase, PBMC6k, Liver, Cerebellum, Heart, Kolodzisjczyk, and Baron datasets, while for the Goolam and Adipose datasets, we only performed log-transformation. After normalizing the dataset, we identified and filtered out highly variable genes to allow our model to focus on the important features that best capture the differences between cells, thereby enhancing the accuracy and interpretability of subsequent analyses. Finally, the filtered highly variable genes underwent scaling with a maximum value of 10.

2.5. Comparison of Clustering Performance

To further evaluate the clustering performance of scVGATAE, we compared it with six single-cell clustering methods. These methods can be divided into three categories, including traditional clustering approaches, deep neural network-based clustering approaches, and GNN (Graph Neural Network)-based clustering approaches.

- Leiden [30] is a community detection algorithm encapsulated in the optimization module of scanpy;
- scVI [31] employs variational inference techniques to efficiently learn model parameters and approximate posterior distributions, enabling high-precision clustering analysis of single-cell RNA sequencing data;
- scDeepCluster [13] combines the Deep Embedding Clustering (DEC) algorithm in the field of image processing with the ZINB model to optimize clustering performance and feature learning;
- scASGC [32] is a cell clustering method based on an adaptive simplified graph convolutional network;
- scGNN2.0 [17] utilizes GAT and multimodal autoencoders to formulate and aggregate intercellular relationships;
- scGAC [18] applies a graph autoencoder that incorporates an attention mechanism to learn cluster-friendly feature representations, and employs self-optimization clustering to improve clustering;
- Among them, the parameters of the six scRNA seq clustering methods to be compared were set according to their tutorials.

To enhance the reliability and reproducibility of the experimental results, we independently performed five clustering analyses for each scRNA-seq clustering method on each dataset and calculated clustering performance evaluation metrics for each analysis. Finally, we took the average of these metrics to evaluate the clustering performance of each clustering method on that dataset.

2.6. Evaluation Metrics

In this paper, two widely used metrics are employed to evaluate the clustering performance of different algorithms: the Adjusted Rand Index (ARI) [33] and Normalized Mutual Information (NMI) [34].

ARI (Adjusted Rand Index) measures the similarity between the predicted cluster labels and the true cluster label assignments, with a value range of $[-1, 1]$: negative values indicate poor labeling, higher values indicate higher clustering accuracy, and 1.0 represents a perfect matching score. The formula for calculating ARI is

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}, \tag{18}$$

where n_{ij} is the number of cells shared between the cluster in the true cluster label assignment and cluster in the predicted cluster label assignment and n represents the total number of cells, $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$.

NMI (Normalized Mutual Information) measures the consistency between the predicted cluster labels and the true cluster label assignments, ignoring arrangement. The formula for calculating NMI is

$$NMI = \frac{2MI(U,V)}{H(U) + H(V)}, \tag{19}$$

where U represents the predicted cluster labels,V represents the true cluster labels, $MI(U,V)$ denotes the mutual information between U and V, and $H(\cdot)$ is the entropy function.

3. Results

3.1. Compared with the Other Six Methods, scVGATAE Significantly Improves the Clustering Performance

The scVGATAE unsupervised scRNA-seq clustering method was compared in clustering performance with six widely popular single-cell clustering methods on nine real and effective scRNA-seq datasets. As shown in the last column of Figure 2, which depicts the average scores of the clustering metrics ARI and NMI for each method across nine datasets, scVGATAE achieves the highest scores for both clustering metrics. Notably, on six datasets, Biase, Goolam, Liver, Cerebellum, Heart, and Baron, scVGATAE attains the highest ARI and NMI scores, significantly outperforming other single-cell clustering methods. Slightly better are the PBMC6k, Adipose, and Kolodzisjczyk datasets. Specifically, on the Goolam dataset, scVGATAE demonstrated remarkable improvements over the suboptimal method scGAC, with a notable 10.1% increase in NMI and a substantial 23.2% enhancement in ARI. Similarly, on the mouse Baron dataset, scVGATAE surpassed scVI by 3.2% in NMI and outperformed scDeepCluster by 8.2% in ARI. In the Cerebellum dataset, compared to scDeepCluster, scVGATAE achieved a 13.6% lift in NMI, while over scGAC, it showed a significant 5% gain in ARI.

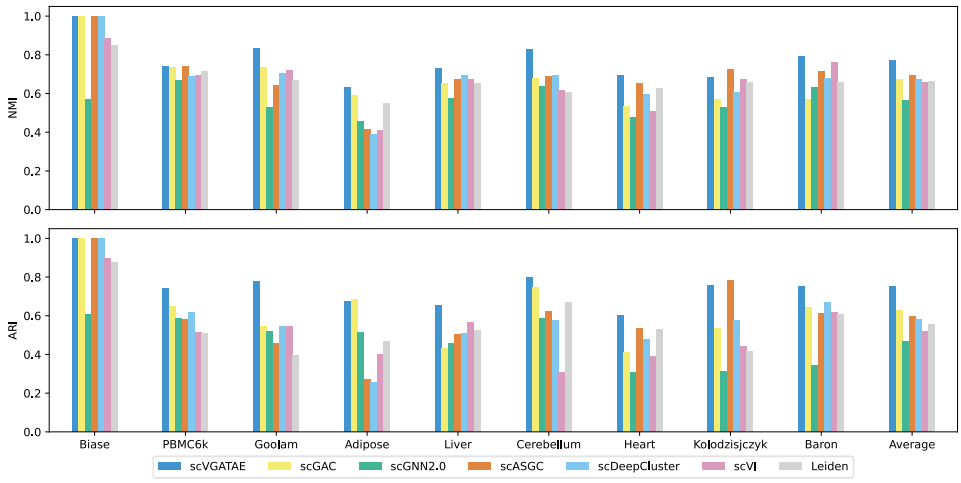


Figure 2. NMI and ARI scores of scVGATAE and six comparison methods in nine datasets, see also Supplementary Tables S1 and S2.

3.2. scVGATAE Learns Cluster-Friendly Embeddings for Cells

To demonstrate that the proposed scVGATAE clustering model not only significantly improves the clustering performance of scRNA-seq data but also improves the low dimensional representations of cells learned through it, which have friendly visual embeddings, firstly, UMAP is used to project the low dimensional representations of cells learned by scVGATAE into a two-dimensional space and visualize the two-dimensional cell representations. As shown in Figure 3, we selected two real datasets with different cell orders

of magnitude and from different tissues, including human adipose cells (Figure 3a) and human cerebellar cells (Figure 3b). Here, scGNN2 represents scGNN2.0.

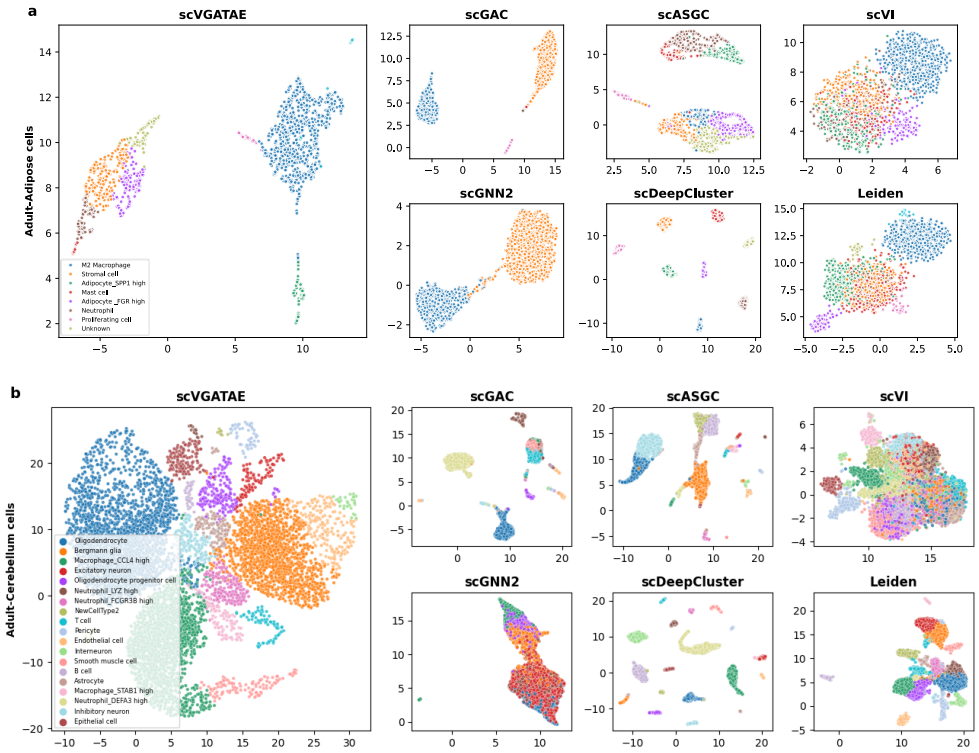


Figure 3. The visualization of the identified clusters of scVGATAE and six competitive methods are drawn. These points represent each sample cells, with different colors indicating different data labels. (**a**) The Adult-Adipose cell. (**b**) The Adult-Cerebellar cells. Where scGNN2 stands for scGNN2.0.

In Figure 3a, we visually demonstrate the segregation of various cell types within a human adipose dataset through color coding. Notably, the scVGATAE method achieved remarkable and distinct separation of eight different cell subtypes, including stromal cells, mast cells, and two types of adipocytes expressing high levels of distinct genes (Adipocyte_FGR high and Adipocyte_SPP1 high). These cell clusters appear as compact and well-demarcated groups in the plot, underscoring scVGATAE's prowess in differentiating subtle variations. In contrast, both Leiden and scVI methods exhibit considerable mixing of cell types in their UMAP embeddings, while scGNN2.0, scGAC, and scASGC, although achieving comparatively good cell embeddings, still display a minor blending of M2 macrophages and stromal cells. This comparison highlights the unique advantage of scVGATAE in single-cell RNA sequencing data analysis. Lastly, although scDeepCluster achieves good separation of distinct cell clusters, it results in overly distanced clusters, potentially obscuring the underlying correlations between them to some extent.

As shown in Figure 3b, for the Cerebellum dataset, which contains over 7000 cells and 19 abundant cell clusters, scVGATAE still achieves a good separation of the dataset with clear boundaries. In comparison, scGNN2.0 and Leiden tend to fuse different cell clusters together. While scGAC, scASGC, and scVI are able to define relatively distinct cell cluster boundaries, they still face the issue of a small number of outliers being mixed in. In summary, compared to the other six methods, these either fail to clearly delineate the boundaries between cell clusters or compromise the intrinsic structural information of the cells' low-dimensional embeddings during the learning process. However, the scVGATAE method demonstrates significant advantages: it not only precisely identifies diverse cell clusters and clearly defines the boundaries between them, but also successfully preserves the rich intrinsic structural information of the cell clusters.

3.3. Ablation Study

We set up four ablation study subjects of scVGATAE: scVGATAE_no_NE (scVGATAE without incorporating NE), scVGATAE_no_attention (a variational graph autoencoder without an attention network), scVGATAE_all_attention (a graph autoencoder solely based on an attention network), and scVGATAE_no_self-adaption-epochs (without self-adaptive model training epochs, using a default of 180 epochs).

As shown in Figure 4, scVGATAE achieves the best clustering performance compared to its four ablation study subjects. The results indicate that the NE (Network Enhancement) denoising module, the variational graph autoencoder model updated based on graph attention networks, and the adaptive adjustment of model training iterations all play significant roles in enhancing the overall performance of the model.

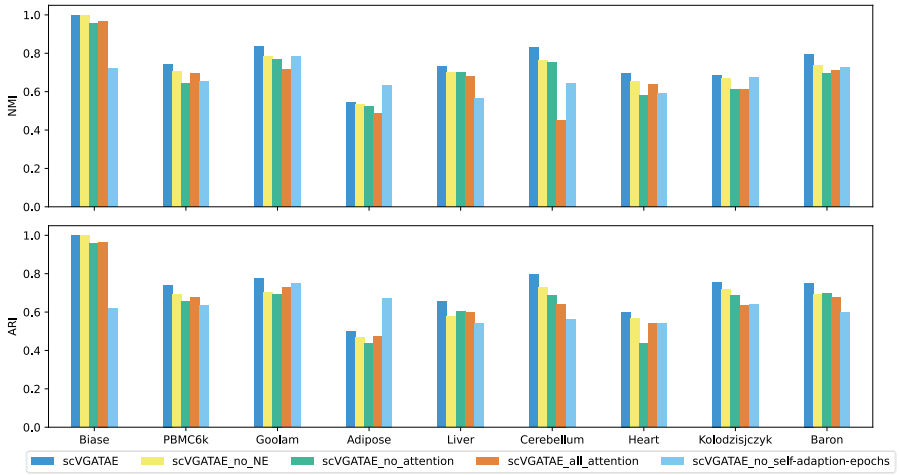


Figure 4. NMI and ARI comparison of scVGATAE and its four different ablation study subjects, see also Supplementary Tables S3 and S4.

3.4. Scalability and Efficiency

To validate the scalability and model execution efficiency of scVGATAE, we compared its runtime with six other scRNA-seq clustering methods. As shown in Figure 5, the runtimes of scGNN2.0, scGAC, and scDeepCluster significantly increase with the increase in the magnitude of cell numbers, whereas scVGATAE maintains a relatively stable runtime across various cell magnitudes. This demonstrates that scVGATAE not only possesses excellent scalability, enabling efficient processing of large-scale scRNA-seq data, but also exhibits a notable advantage in optimizing computational efficiency while maintaining high performance. The data comparison shown in Figure 5 further verifies the efficiency and practicality of scVGATAE among single-cell RNA-seq clustering methods.

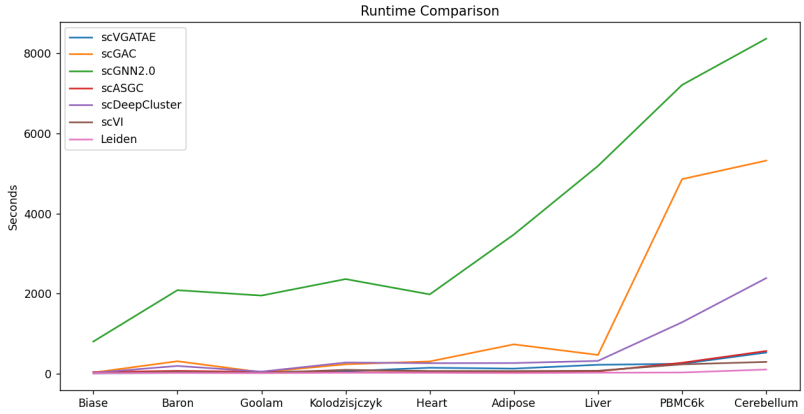


Figure 5. Comparison of the runtime of 7 methods on 9 datasets, see also Supplementary Table S5.

4. Conclusions

With the rapid advancements in single-cell RNA sequencing technology, a vast amount of high-dimensional, sparse, and complex gene expression data has been generated. Efficiently processing these data to extract biologically meaningful information, particularly through unsupervised learning methods for cell clustering, poses a significant challenge in the field of bioinformatics. To address this need, we have developed the scVGATAE model. The success of scVGATAE lies primarily in three aspects. Firstly, it utilizes NE to construct a high-confidence cell graph, effectively eliminating noisy edges in the cell graph. Secondly, by leveraging a variational graph attention autoencoder, it aggregates topological information from the cell network, fully exploiting relationships between cells and significantly enhancing model efficiency. Thirdly, scVGATAE adaptively selects appropriate model training epochs for different datasets, breaking away from the traditional qualitative mindset of manually tuning hyperparameters for neural network training and exhibiting superior generalization capabilities.

Our experimental results show that after removing noisy edges through the NE module, the cellular graph becomes more reliable, and scVGATAE effectively leverages the latent associations between cells to learn a more favorable low-dimensional embedding of cells, significantly boosting the clustering performance.

Despite the fact that scVGATAE outperforms existing methods in terms of efficiency and clustering performance, it still has some shortcomings. The number of cell clusters in scVGATAE is determined using the Leiden method, which may result in weaker generalization capabilities. Moreover, the performance of scVGATAE's ability to adaptively select the appropriate number of model training epochs needs improvement. For instance, in the case of the Adipose dataset, the results obtained when using the adaptive model training epochs were not as good as those achieved through manual parameter tuning (e.g., setting the model training epochs to 180).

In the future, we will focus on enhancing the generalizability of identifying the optimal number of clusters for various datasets and explore other more effective clustering methods.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biology13090713/s1: Table S1: ARI scores of scVGATAE and six comparison methods in 9 datasets; Table S2: NMI scores of scVGATAE and six comparison methods in 9 datasets; Table S3: ARI scores of scVGATAE and its four different ablation study subjects; Table S4: NMI scores of scVGATAE and its four different ablation study subjects; Table S5: Running time of scVGATAE and six comparison methods in 9 datasets.

Author Contributions: Conceptualization L.L. and X.W.; Data curation, K.N.; Formal analysis, A.Y.; Funding acquisition, L.L. and J.Y.; Methodology, L.L. and X.W.; Project administration, L.L. and Y.Z.; Supervision, Y.Z.; Validation X.W. and J.Y.; Visualization, X.W.; Writing—original draft, X.W.; Writing—review and editing, L.L. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: https://github.com/duddubududu/scRNA-seq-data (accessed on 21 August 2024).

Acknowledgments: The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Buettner, F.; Natarajan, K.N.; Casale, F.P.; Proserpio, V.; Scialdone, A.; Theis, F.J.; Teichmann, S.A.; Marioni, J.C.; Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **2015**, *33*, 155–160. [CrossRef] [PubMed]
2. Guo, M.; Du, Y.; Gokey, J.J.; Ray, S.; Bell, S.M.; Adam, M.; Sudha, P.; Perl, A.K.; Deshmukh, H.; Potter, S.S.; et al. Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. *Nat. Commun.* **2019**, *10*, 37. [CrossRef] [PubMed]
3. Hwang, B.; Lee, J.H.; Bang, D. Single-cell RNA sequencing technologies and bioinformaticspipelines. *Exp. Mol. Med.* **2018**, *50*, 1–14. [CrossRef]
4. Sun, S.; Zhu, J.; Ma, Y.; Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.* **2019**, *20*, 269. [CrossRef] [PubMed]
5. Villani, A.C.; Satija, R.; Reynolds, G.; Sarkizova, S.; Shekhar, K.; Fletcher, J.; Griesbeck, M.; Butler, A.; Zheng, S.; Lazo, S.; et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **2017**, *356*, eaah4573. [CrossRef]
6. Hinton, L.v.d.M.a.G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
7. McInnes, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**. [CrossRef]
8. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R.; et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **2017**, *14*, 483–486. [CrossRef]
9. Huh, R.; Yang, Y.; Jiang, Y.; Shen, Y.; Li, Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. *Nucleic Acids Res.* **2020**, *48*, 86–95. [CrossRef]
10. Mei, Q.; Li, G.; Su, Z. Clustering single-cell RNA-seq data by rank constrained similarity learning. *Bioinformatics* **2021**, *37*, 3235–3242. [CrossRef]
11. Yu, B.; Chen, C.; Qi, R.; Zheng, R.; Skillman-Lawrence, P.J.; Wang, X.; Ma, A.; Gu, H. scGMAI: A Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder. *Brief. Bioinform.* **2021**, *22*, bbaa316. [CrossRef]
12. Eraslan, G.; Simon, L.M.; Mircea, M.; Mueller, N.S.; Theis, F.J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **2019**, *10*, 390. [CrossRef] [PubMed]
13. Tian, T.; Wan, J.; Song, Q.; Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* **2019**, *1*, 191–198. [CrossRef]
14. Grønbech, C.H.; Vording, M.F.; Timshel, P.N.; Sønderby, C.K.; Pers, T.H.; Winther, O. scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* **2020**, *36*, 4415–4422. [CrossRef]
15. Gan, Y.; Huang, X.; Zou, G.; Zhou, S.; Guan, J. Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network. *Brief. Bioinform.* **2022**, *23*, bbac018. [CrossRef] [PubMed]
16. Buterez, D.; Bica, I.; Tariq, I.; Andrés-Terré, H.; Liò, P. CellVGAE: An unsupervised scRNA-seq analysis workflow with graph attention networks. *Bioinformatics* **2022**, *38*, 1277–1286. [CrossRef] [PubMed]
17. Gu, H.; Cheng, H.; Ma, A.; Li, Y.; Wang, J.; Xu, D.; Ma, Q. scGNN 2.0: A graph neural network tool for imputation and clustering of single-cell RNA-Seq data. *Bioinformatics* **2022**, *38*, 5322–5325. [CrossRef]
18. Cheng, Y.; Ma, X. scGAC: A graph attentional architecture for clustering single-cell RNA-seq data. *Bioinformatics* **2022**, *38*, 2187–2193. [CrossRef]
19. Wang, B.; Pourshafeie, A.; Zitnik, M.; Zhu, J.; Bustamante, C.D.; Batzoglou, S.; Leskovec, J. Network enhancement as a general method to denoise weighted biological networks. *Nat. Commun.* **2018**, *9*, 3108. [CrossRef]
20. Kipf, T.; Welling, M.J.A. Variational Graph Auto-Encoders. *arXiv* **2016**, arXiv:1611.07308. [CrossRef]
21. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio', P.; Bengio, Y.J.A. Graph Attention Networks. *arXiv* **2017**, arXiv:1710.10903.
22. Peter, J.R. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
23. Han, X.; Zhou, Z.; Fei, L.; Sun, H.; Wang, R.; Chen, Y.; Chen, H.; Wang, J.; Tang, H.; Ge, W.; et al. Construction of a human cell landscape at single-cell level. *Nature* **2020**, *581*, 303–309. [CrossRef]
24. Biase, F.H.; Cao, X.; Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **2014**, *24*, 1787–1796. [CrossRef] [PubMed]
25. Zheng, G.X.; Terry, J.M.; Belgrader, P.; Ryvkin, P.; Bent, Z.W.; Wilson, R.; Ziraldo, S.B.; Wheeler, T.D.; McDermott, G.P.; Zhu, J.; et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **2017**, *8*, 14049. [CrossRef]
26. Mubeen, G.; Antonio, S.; Sarah, J.L.G.; Iain, C.M.; Agnieszka, J.; Anna, H.; Thierry, V.; John, C.M.; Magdalena, Z.-G. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **2016**, *165*, 61–74.
27. Kolodziejczyk, A.A.; Kim, J.K.; Tsang, J.C.; Ilicic, T.; Henriksson, J.; Natarajan, K.N.; Tuck, A.C.; Gao, X.; Bühler, M.; Liu, P.; et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **2015**, *17*, 471–485. [CrossRef] [PubMed]
28. Maayan, B.; Adrian, V.; Samuel, L.W.; Aubrey, L.F.; Renaud, G.; Amedeo, V.; Jennifer Hyoje, R.; Bridget, K.W.; Shai, S.S.-O.; Allon, M.K.; et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **2016**, *3*, 346–360.e4.

29. Aman, R.; Ali, Z.; Butt, H.; Mahas, A.; Aljedaani, F.; Khan, M.Z.; Ding, S.; Mahfouz, M. RNA virus interference via CRISPR/Cas13a system in plants. *Genome Biol.* **2018**, *19*, 1. [CrossRef]
30. Traag, V.A.; Waltman, L.; van Eck, N.J. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **2019**, *9*, 5233. [CrossRef]
31. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef] [PubMed]
32. Shudong, W.; Yu, Z.; Yulin, Z.; Wenhao, W.; Lan, Y.; YunYin, L.; Jionglong, S.; Shanchen, P. scASGC: An adaptive simplified graph convolution model for clustering single-cell RNA-seq data. *Comput. Biol. Med.* **2023**, *163*, 107152.
33. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
34. Strehl, A.; Ghosh, J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2003**, *3*, 583–617.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.