OXFORD UNIVERSITY PRESS | Bioinformatics

scGMM-VGAE: A Gaussian mixture model-based variational graph autoencoder algorithm for clustering single-cell RNA-seq data

Journal:	*Bioinformatics*
Manuscript ID	Draft
Category:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Lin, Eric; University of Toronto - St George Campus, Biostatistics Liu, Boyuan; University of Toronto - St George Campus, Biostatistics Lac, Leann; University of Manitoba, Computer Science Fung, Daryl; University of Manitoba, Computer Science Leung, Carson; University of Manitoba, Computer Science Hu, Pingzhao; Western University Schulich School of Medicine & Dentistry, Biochemistry
Keywords:	Algorithms, Clustering, Machine learning, Single Cell, Statistics, RNA-Seq

scGMM-VGAE: A Gaussian mixture model-based variational graph autoencoder algorithm for clustering single-cell RNA-seq data

Eric Lin[1 *], Boyuan Liu[1 *], Leann Lac[2,3 *], Daryl L.X. Fung[2], Carson K. Leung[2], Pingzhao Hu[1,2,4,5,#]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[2]Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

[3]Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada

[4]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada

[5]Department of Biochemistry, Western University, London, Ontario, Canada

∗ Equal contribution

# To whom correspondence should be addressed.

Dr. Pingzhao Hu, Department of Biochemistry, Western University, Medical Sciences Building Rm. 342, London, Ontario, Canada, N6A 5C1.

Email: phu49@uwo.ca.

Running title: Graph neural network for scRNA-seq data clustering

Abstract

Motivation: Cell type identification using single-cell RNA sequencing (scRNA-seq) is critical for understanding disease mechanisms for disease diagnosis and drug discovery, which involves classifying data into clusters of single cells. Although significant deep learning approaches have been presented and demonstrated to improve clustering performance in high dimensional scRNA-seq data when compared to traditional scRNA-seq clustering methods, cell clustering analysis using an extendable end-to-end framework integrating advanced statistical clustering with deep learning approaches is still understudied.

Results: To address this issue, we proposed a Gaussian mixture model-based variational graph autoencoder on scRNA-seq (scGMM-VGAE) which is a combination of an advanced statistical method and a deep learning model. The unsupervised clustering algorithm scGMM-VGAE clusters cell types by applying the GMM clustering module to latent data encoding by inputting cell-cell graph and a gene feature matrix. We clustered single cells from four publicly available scRNA-seq datasets and three simulated datasets using the proposed method and compared its performance with four baseline models. By successfully incorporating a statistical model into an unsupervised deep learning algorithm to cluster cell types in scRNA-seq data, the scGMM-VGAE algorithm significantly outperformed the selected baseline models in both the real and simulated scRNA-seq datasets using the adjusted Rand index (ARI), normalized mutual information (NMI), and Silhouette coefficient (SC) as model performance measures.

Availability and implementation: All source codes used in this study can be found at https://github.com/ericlin1230/scGMM-VGAE.

Contact: phu49@uwo.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

1. Introduction

Significant advances in single cell RNA-sequencing (scRNA-seq) methods for measuring gene expression at the single-cell level have been widely developed and applied to research in microbiology, biomedical sciences, neuroscience, and oncology (Wei et al., 2022). In contrast to traditional bulk RNA-sequencing, scRNA-seq begins with cell isolation before RNA extraction (Chaudhry et al., 2019), which is useful for a small population of target cells and for comparing the various cell types in a sample. Due to the nature of heterogeneity and high dimensionality, scale, and noise, analyzing scRNA-seq data is more challenging than analyzing RNA sequencing data from bulk cell populations (Wang et al., 2022). Single cell RNA-seq data are high resolution, thus, effectively analyzing and mining either the complex relationships or potential regulation patterns between cells are critical to the success of downstream research and personalized medicine (Su et al., 2021).

Clustering analysis has been received a significant attention from bioinformatics researchers in analyzing gene expression data (Govek et al., 2019), cancer research, and personalized medicine (Wu et al., 2021). Cell clustering is a fundamental step in scRNA-seq data computational analysis, in which clustering is performed to classify a set of cells from unknown cell groups into meaningful cell types based on transcriptome similarity or existing biological knowledge (Krzak et al., 2019; Li et al., 2020; Shiga et al., 2021). The results of cell clustering are important for a variety of downstream analyses, including cell type differential expression, cellular composition estimation, and rare cell type discovery (Chen et al.,2019). Thus, methods for cell clustering analysis have been in high demand for answering research questions using scRNA-seq data. However, due to the gene expression nature of these data, issues such as high noise and a large proportion of missing values (zeroes) due to the fraction of non-sequenced transcripts data pose a challenge to clustering performance accuracy (Haque et al., 2017). Traditional clustering methods have limitations in terms of accuracy, scalability, and interpretability when analyzing thousands of gene features in high-dimensional scRNA-seq data (Wu et al., 2021).

As a result, several state-of-the-art clustering methods in machine learning have been recently developed to cluster scRNA-seq data to improve accuracy. Seurat includes a graph-based clustering analysis tool for scRNA-seq data that uses a shared nearest neighbour (SNN) modularity optimization-based clustering algorithm to identify cell clusters (Butler et al., 2018; Stuart et al., 2019; Hao et al., 2021). Lopez et al. (2018) presented the single-cell variational inference (scVI) algorithm, which groups similar cells by first generating conditional distributions with a deep neural network and then applying them to a hierarchical Bayesian model. Deep count autoencoder (DCA) networks used a negative binomial noise model with or without zero-inflation to capture nonlinear gene-gene dependencies that was used to denoise scRNA-seq datasets while accounting for count distribution, over-dispersion, and sparsity (Eraslan et al., 2019). In graph-structured data, variational graph auto-encoder (VGAE) (Kipf & Welling, 2016), an unsupervised learning framework, is a latent variable model which minimizes the reconstruction error of the data by optimizing the variational lower bound using a single Gaussian distribution. Buterez et al. (2021) introduced CellVGAE which applies variational graph autoencoder (VGAE) to scRNA-seq data.

Despite the fact that these promising deep learning approaches outperform traditional scRNA-seq clustering methods in terms of clustering performance, cell clustering analysis using an extendable end-to-end framework integrating statistical clustering with powerful deep learning approaches is still not extensively studied. Jiang et al. (2017) applied Gaussian mixture model (GMM) to variational deep embedding (VaDE) using images for both input and output. To overcome the limitation in distribution and data structure, Hui et al. (2020) later developed Gaussian mixture model-based variational graph autoencoder (GMM-VGAE) to analyze benchmark graph datasets (Pubmed, Citeseer and Cora). This unsupervised generative clustering framework combines a statistical model GMM and a deep learning model VGAE for clustering tasks where the Gaussian prior distribution is replaced by a mixture of Gaussians prior distribution. In this study, we considered single-cell Gaussian mixture model-based variational graph autoencoder (scGMM-VGAE) in cell clustering on scRNA-Seq data, where GMM module is used to cluster cell types from latent data encoded from VGAE. As benchmark methods, we considered Seurat, DCA+leiden, scVI, and CellVGAE. The clustering analysis results on real and simulated data showed that the proposed algorithm outperforms baseline models in cell clustering performance on seqRNA-seq data.

2. Materials and methods

2.1. Datasets

Real datasets

We considered three gold standard real labelled scRNA-seq datasets in this study: Baron3, Baron4 (Baron et al., 2016), and Darmanis (Darmanis et al., 2015). The truth cell labels are included along the scRNA cells × genes count matrix. The sequenced cells in both Baron3 and Baron4 datasets are from human pancreatic islets. Specifically, Baron3 data were collected from a male donor aged 38 with a BMI of 27.5 and no diabetes status, while Baron4 data was collected from a female donor aged 59 with a BMI of 29.9 and type 2 diabetes mellitus. Baron3 and Baron4 data share the same scRNA experimental protocol, sequenced genes, and numbers of clusters but have different numbers of cells. In our study, these two datasets were named as *Baron3 data* and *Baron4 data*, respectively. Darmanis et al. (2015) generated a human brain scRNA-seq capturing the cellular complexity of fetal human brain in adults at a whole transcriptome level. This dataset is named as *Darmanis data*. The cells are divided into 8 categories, which correspond to 8 major brain cell types (e.g., strocytes, oligodendro-cytes, oligodendrocyte precursor cells (OPCs), neurons, microglia, and vascular cell).

In addition, we considered a dataset with no true labels generated which includes liver cells (hepatoblasts, hepatocytes, and cholangiocytes) of E10.5 to E17.5 mouse embryos (Yang et al., 2017). We denoted this dataset as *Yang data*. The summary of number of cells, genes, and clusters for both the labelled and the unlabeled datasets is described in **Table 1**.

Table 1. Summary of labelled and unlabeled datasets.

Datasets	# cells	# genes	Labelled (Y/N)	# clusters
Baron3 data (Baron et al., 2016)	3650	20125	Y	14
Baron4 data (Baron et al., 2016)	1303	20125	Y	14
Darmanis data (Darmanis et al., 2015)	420	22085	Y	8
Yang data (Yang et al., 2017)	529	40916	N	N/A

The R-package Seurat v3 (Stuart et al., 2019) was used for pre-processing scRNA-seq data. Since scRNA-seq data are usually highly variable (Cui et al., 2021), to stabilize the variation within feature matrix, we applied log-normalization which is commonly used pre-processing procedure (Booeshaghi & Pachter, 2021). In this research, we considered top 1200 genes, which are selected by variance stabilizing transformation (VST) selection method adapted in *FindVariableFeatures* function of Seurat v3 package. VST chooses top variable genes using the relationship between variance and mean (Buterez et al., 2021).

Simulated datasets

To further evaluate the performance of scGMM-VGAE, we considered simulated data. These datasets were generated using a state-of-the-art simulation algorithm SPARSim (Baruzzo et al., 2020). In the simulation, SPARSim first estimates the gene expression level intensities (Z), gene expression level variabilities (Φ), and library size (L) from a given template to simulate a dataset with similar properties. The expression level (X_{ij}) of gene i in cell j is later modelled using a gamma distribution

$$X_{ij}\ Gamma\left(shape = \frac{1}{\Phi_i}, scale = Z_i \bullet \Phi_i\right). \qquad (1)$$

To represent the systematic and sample-dependent bias, SPARSim simulates technical variability (Y_j) for the expression of cell j where Y_j is modelled using multivariate hypergeometric distribution parameterized with $n = L_j$ and $m = X_j$.

In our study, to simulate single-cell datasets, we considered Zheng preset template (Zheng et al., 2017) implemented in SPARSim to generate the simulated datasets. The parameters, i.e., intensities, variabilities, and library size, in the preset are estimated from the peripheral blood mononuclear cells (PBMC) sample, and the original template dataset contains 3388 cells, 15041 genes, and 4 labelled cell types (Zheng et al., 2017). The parameters estimated from the Zheng dataset are preset in the algorithm. To evaluate the scGMM-VGAE clustering stability of different sizes' datasets, we simulated two additional datasets with 1694 cells and 6776 cells, respectively. These datasets were generated using the Zheng preset by halving and doubling the cell numbers of each cell type. As for the simulated genes, we selected the top 1200 genes from the original templates using variance stabilizing transformation to be consistent to the real datasets. We considered 10 replications for each simulation setting and calculated the average mean and standard deviation of the selected evaluation metric to evaluate clustering performance of proposed method. Details of simulated datasets are summarized in Table 2.

Table 2. Summary of simulated datasets using Zheng preset template (Zheng et al., 2017).

	Zheng's template			**Simulation**		
Simulated Datasets	**# cells**	**# genes**	**# clusters**	**# cells**	**# genes**	**# clusters**
1694-cell data	3388	15041	4	1694	1200	4
3388-cell data	3388	15041	4	3388	1200	4
6776-cell data	3388	15041	4	6776	1200	4

2.2. Architecture of scGMM-VGAE

The single-cell GMM variational graph auto encoder (scGMM-VGAE) is an end-to-end deep learning framework utilizing the benefit of GMM in discovering the distribution of scRNA-Seq data to deliver more precise cluster result. Given the filtered and normalized cells × genes matrix as input of scGMM-VGAE, we used R-package SingleCellExperiment 3.14 to build a $\mathcal{K}$-nearest neighbour (KNN)-based cell-cell graph for each data set. The optimal $\mathcal{K}$ to produce the best results for our selected datasets was reported as 5 (Booeshaghi & Pachter, 2021). The scGMM-VGAE framework considers a processed feature matrix $\mathbf{X}$ of dimension $n \times p$ (i.e., cells × genes) and a cell-cell graph $\mathbf{A}$ of dimension $n \times n$ as inputs. The inputs are later passed through a variational auto encoder-decoder network of two-layer GCN integrating with GMM. The end-to-end scGMM-VGAE model can be trained together by optimizing the variational evidence lower bound (ELBO) loss (Kingma & Welling, 2014). The unsupervised learning algorithm scGMM-VGAE aims to partition the cells captured in a graph into K clusters without using cell labels. The algorithm will then return outputs as a reconstructed cell-cell graph $\hat{\mathbf{A}}$ and cell clustering results. However, in this study, we focused on cell clustering task of the algorithm. The overview of scGMM-VGAE in clustering scRNA-seq data is illustrated in **Figure 1**. To define the scGMM-VGAE structure and clustering procedure in scRNA-Seq data, we will first define some basic notations and GMM.

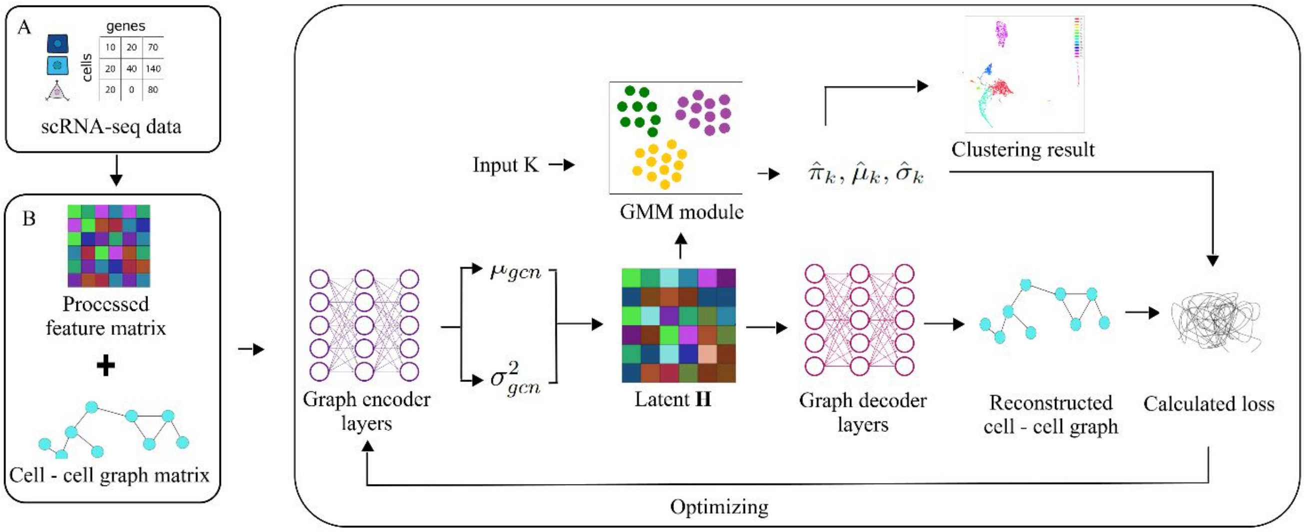

Figure 1: The diagram of scGMM-VGAE. The end-to-end framework is divided into three processes: (A) Data processing. In each dataset, top 1200 genes are selected, the rows represent cells and columns represent genes. (B) Data inputting. The model takes the processed feature matrix **X** and

cell-cell graph matrix $\boldsymbol{A}$ as inputs. (C) scGMM-VGAE module. The processed feature matrix and cell-cell graph are fed into autoencoder GCN to generate latent matrix $\boldsymbol{H}$. Given $\boldsymbol{H}$, GMM initializes clusters with estimates of cluster mean, standard deviation, and weight which are used to produce clustering results. Decoder decodes $\boldsymbol{H}$ into reconstructed cell-cell graph. Parameters from both GMM and VGAE modules are combined to calculate and optimize the designed losses. The algorithm returns clustering results as final output.

Background and notations

Suppose $\mathcal{G} = (\mathbf{V},\mathbf{E},\mathbf{X})$ is an undirected graph representation of the scRNA data where $\mathbf{V} = \{v_1,v_2,...,v_n\}$ is a set of nodes, $\mathbf{E}$ is a set of edges between nodes, and $\mathbf{X} \in \mathrm{R}^{n \times p}$ is an attribute feature matrix of all nodes. In the context of scRNA-Seq data, n and p represent the number of cells and the size of genes, respectively. To better understand the structure of scGMM-VGAE which is also a graph convolutional network, we considered a graph convolutional layer which is a layer-wise convolution to transform the l^{th} hidden layer $\mathbf{H}^{(l)}$ to $\mathbf{H}^{(l+1)}$ (Kipf & Welling, 2017). Given cell-cell graph $\mathbf{A}$ and processed feature matrix $\mathbf{X}$, here $\mathbf{A} = [\alpha_{ij}] \in \mathrm{R}^{n \times n}$ is topological structure of $\mathcal{G}$ where a_{ij} $= 1$ if nodes v_i and v_j are connected as edge e_{ij}, otherwise $\alpha_{ij} = 0$. The spectral convolution function f_ϕ is defined as

$$f_\phi\left(\mathbf{H}^{(l)},\mathbf{A}|\mathbf{W}^{(l)}\right) = \phi\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \qquad (2)$$

where $\mathbf{H}^{(0)} = \mathbf{X}$, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n$, and $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$. Here $\mathbf{W}^{(l)}$ denotes weight parameters, ϕ is an activation function (e.g., ReLU, Sigmoid, tanh), and $\mathbf{I}_n$ is an identity matrix.

Encoder and decoder modules

Autoencoder is used for embedding the input scRNA-seq gene expression data $\mathbf{X} \in \mathrm{R}^{n \times p}$ into low dimensional space, where n and p are the number of cells and the size of genes, respectively. Autoencoder is an unsupervised neural network that consists of the encoder and decoder modules. Kipf & Welling (2016) proposed a simple inference model parameterized by a two-layer graph convolutional network (GCN) which is defined as

$$q(\mathbf{H}|\mathbf{X},\mathbf{A}) = \prod_{i=1}^{n} q(h_i|\mathbf{X},\mathbf{A}) \quad \text{and} \quad q(h_i|\mathbf{X},\mathbf{A}) = \mathrm{N}\left(h_i|\boldsymbol{\mu}_i, diag\left(\boldsymbol{\sigma}_i^2\right)\right), \tag{4}$$

where h_i is the stochastic latent variable, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are vectors of GCN mean and standard deviation, respectively. Now, the **encoder** takes cell-cell graph $\mathbf{A}$ and the processed feature matrix $\mathbf{X}$ and generate the latent variable matrix $\mathbf{H}$.

For simplicity, we define $\boldsymbol{\beta}_{gcn} = \left(\boldsymbol{\mu}_{gcn}, \boldsymbol{\sigma}_{gcn}\right)$ where $\boldsymbol{\mu}_{gcn}$ and $\boldsymbol{\sigma}_{gcn}$ are matrices of $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_{\mathbf{i}}$, respectively. Let $\mathbf{W}^{(0)}$ be the weight of the first layer and $\mathbf{W}_{\boldsymbol{\beta}}^{(1)}$ be the weight of $\boldsymbol{\beta}_{gcn}$, the architecture for the encoder is follows.

$$\mathbf{H}^{(1)} = \phi_1\left(\mathbf{H}^{(0)}, \mathbf{A}|\mathbf{W}^{(0)}\right), \tag{5}$$

$$\mathbf{H}_{\boldsymbol{\beta}_{gcn}}^{(2)} = \phi_2\left(\mathbf{H}^{(1)}, \mathbf{A}|\mathbf{W}_{\boldsymbol{\beta}_{gcn}}^{(1)}\right), \tag{6}$$

where $\mathbf{H}^{(1)}$ is defined as the first layer, $\mathbf{H}_{\boldsymbol{\beta}}^{(2)}$ is the second layer of GCN parameterized by $\boldsymbol{\beta}$. Here we considered ϕ_1 as $Relu(\cdot)$ and ϕ_2 as a linear function. Since $\mathbf{H}_{\boldsymbol{\beta}_{gcn}}^{(2)}$ share the same weight $\mathbf{W}^{(0)}$, we have $\mathbf{H}_{\boldsymbol{\beta}_{gcn}}^{(2)} = \boldsymbol{\mu}_{\mathbf{gcn}}$ for $\boldsymbol{\beta}_{gcn} = \boldsymbol{\mu}_{gcn}$, otherwise $\mathbf{H}_{\boldsymbol{\beta}_{gcn}}^{(2)} = \log \boldsymbol{\sigma}_{gcn}^2$. The **decoder** is conducted to reconstruct the cell-cell graph structure $\mathbf{A}$ using generative model defined as an inner product between latent variables,

$$p(\hat{\mathbf{A}}|\mathbf{H}) = \prod_{i=1}^{n} \prod_{j=1}^{n} p\left(\hat{\mathbf{A}}_{ij}|h_i, h_j\right). \tag{7}$$

Here $p\left(\hat{\mathbf{A}}_{ij}|h_i, h_j\right) = \phi_D\left(h_i^{\top}, h_j\right)$ where ϕ_D is an activation function and $\hat{\mathbf{A}}_{ij}$ is the reconstructed cell-cell graph structure from decoder.

GMM module

Gaussian mixture model (GMM) has been widely used for unsupervised model-based clustering for complex data (Reynolds, 2015; McLachlan et al., 2019). In our study, we considered latent matrix $\mathbf{H} = [\mathbf{x_{id}}]_{\mathbf{n} \times \mathbf{p_e}}$ where $i = 1,2,\ldots,n$ and $\mathrm{d} = 1,2,\ldots,\mathrm{p_e}$. Each row $\mathbf{h}_i$ represents one cell and $\mathrm{p_e}$ is the latent embedding size. The goal of using GMM is to cluster $\mathbf{h}_i$ into K components or clusters. The probability density function (PDF) of K-component GMM is given by

$$f_{gmm}(\mathbf{h};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{h};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{3}$$

where $\mathbf{u}_k$ and $\boldsymbol{\Sigma}_\mathbf{k}$ are mean vectors and covariance matrices of mixture components, respectively. For $k = 1,2,...,K$, $\pi_k > 0$ is the weight of the k^{th} mixture component restricted to $\sum_{k=1}^{K}\pi_k = 1$. The GMM can be parameterized by $\boldsymbol{\theta} = (\pi_1,...,\pi_k,\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_\mathbf{k},\boldsymbol{\Sigma}_1,...,\boldsymbol{\Sigma}_k)$. Here $\boldsymbol{\theta}$ can be estimated using the well-known statistical algorithm expectation-maximization (El Assaad et al., 2016; Garriga et al., 2016; Uykan, 2021; Tsumoto et al., 2022), with K-means (Yu et al., 2018; Malik, 2019; Sinaga & Yang, 2020) as initialization clustering method. In our proposed method, we considered $\boldsymbol{\Sigma}_\mathbf{k}$ as a diagonal covariance.

Clustering optimization

Hui et al. (2020) proposed GMM-VGAE as an attributed graph clustering model. We extended this model to cluster cells in scRNA-Seq data which contains a significant number of biological and technical zeros. In this end-to-end algorithm, the number of cell clusters K will be either pre-determined based on the prior information (e.g., the number of labels in the labeled data) or estimated based on the unlabeled data. Given K, a cell-cell graph $\mathbf{A}$, and a processed feature matrix $\mathbf{X}$, an initial clustering assignment is conducted using GMM to obtain the estimates of cluster parameters $\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k$ which are probability, mean, and standard deviation of cluster k such that $\pi \in \mathrm{R}_+^K$ and $\sum_{k=1}^{K}\pi_k = 1$. These parameters later are merged to obtain the combined designed loss function for the end-to-end framework. The scGMM-VGAE is then trained and optimized by maximizing the evidence lower bound (ELBO) using the stochastic gradient variational Bayes (SGVB) estimator and the reparameterization trick (Kingma & Welling, 2014). The clustering loss function is defined as

$$\mathcal{L}_{ELBO} = \mathrm{E}_{q(h,k|x,\alpha)}\left[\log \frac{p(\alpha,h,k)}{q(h,k|x,\alpha)}\right], \tag{8}$$

where $p(\alpha,h,k) = p(\alpha|h)p(h|k)p(k)$ and $q(h,k|x,\alpha) = q(h|x,\alpha)q(k|x,\alpha)$ is the variational posterior. The $\mathcal{L}_{ELBO}$ can also be divided into reconstruction loss $\mathcal{L}_{REC}$ and Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$. Once the training is completed by maximizing **Equation 8**, the latent representation h can be obtained by $q(h|x,\alpha) = \mathrm{N}(h;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\sigma}}^2 I)$ where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}^2$ are derived by $\mathbf{H}^{(2)}_{\mu_{gcn}}$ and $\mathbf{H}^{(2)}_{\sigma_{gcn}}$ (**Equation 6**). The clustering assignments can be obtained by

$$q(k|x,\alpha) = p(k|h) = \frac{p(k)p(h|k)}{\sum_{k'=1}^{K} p(k')p(h|k')}. \qquad (9)$$

The general approach in scGMM-VGAE clustering is presented in **Algorithm 1**.

Algorithm 1. scGMM-VGAE clustering process

```
input number of clusters $K$, processed feature matrix $\mathbf{X}$, cell-cell graph $\mathbf{A}$,
      and max iterations $T_{max}$.

initialize
    Obtain latent $\mathbf{H}$ from encoder module.
    Given $\mathbf{H}$, obtain $\hat{\pi}_k$, $\hat{\mu}_k$, and $\hat{\sigma}_k$ of cluster $k$ by GMM.
    Sample $h \sim \mathrm{N}(\hat{\mu}_k,\hat{\sigma}_k^2 I)$ and update $\mathbf{H}$.

for $t = 1,2,\ldots,T_{max}$ do
    Compute $\alpha \sim \mathrm{Ber}(\mu_\alpha)$ where $\mu_\alpha = \phi(\mathbf{H}_i^\top,\mathbf{H}_j)$ and $\phi$ is activation function.
    Compute loss $\mathcal{L}_{ELBO}$ using Equation 8.
Update $\mathbf{H}$ and derive probabilities $q(h|x,\alpha)$.
Obtain $q(k|x,\alpha)$ using Equation 9.

output cell cluster labels
```

Hyper-parameters setting

In scGMM-VGAE framework, number of neurons, embedding size, and the optimizer are required parameters of activation functions. We considered a wide range of the number of neurons (from 24 to 64) and embedding size (12 to 32) as the numbers out of the range either produced errors frequently or created unstable results. The options for activation functions are varied such as Sigmoid, Tanh, and ReLU functions. For activation function, we considered Sigmoid since other functions may either produce errors or show extremely sub-optimal results. Among several available optimizers (e.g., Adam, SGD, and RMSProp), only Adam optimizer shows no errors. The final parameters applied on all datasets for scGMM-VGAE algorithm are summarized in **Table 3**. The optimal parameters corresponding to each dataset is reported in **Supplementary information**.

Table 3. Summary of hyper-parameter setting for scGMM-VGAE

# neurons	Embedding size	Activation function	Optimizer
24	12	Sigmoid	Adam
32	16	Sigmoid	Adam
48	24	Sigmoid	Adam
50	25	Sigmoid	Adam
56	28	Sigmoid	Adam
64	32	Sigmoid	Adam

2.3. Benchmark methods

To compare the clustering results with scGMM-VGAE method, we considered four baseline methods. The same pre-processed cell × genes feature matrix is used in all the baseline methods. While scVI requires the feature matrix as the input, CellVGAE requires both the feature matrix and the cell-cell graph.

Single-cell variational inference (scVI): scVI is a deep neural network, which first generates conditional distributions and applies them to a hierarchical Bayesian model (Lopez et al., 2018). While encoder of the network encodes cell information using non-linear transformation, the decoder uses another non-linear transformation to create a posterior estimate of the genes. This allows the similar cells to be grouped.

Cell variational graph autoencoder (CellVGAE): This method uses VGAE with encoders and decoders to learn latent variables (Buterez et al., 2021). Cell-VGAE encoders employ graph attention networks (GAT) which is a neural framework for graph data. The graph is later reconstructed from the latent variables, which is used to create clusters.

Seurat: This clustering method includes a graph-based clustering analysis tool for scRNA-seq data, which applies a shared nearest neighbor (SNN) modularity optimization-based clustering algorithm to identify the cell clusters (Stuart et al., 2019).

DCA+Leiden: Deep count autoencoder (DCA) networks uses a negative binomial noise model with or without zero-inflation to capture the nonlinear gene-gene dependencies (Eraslan et al., 2019). The network first denoises scRNA-seq datasets and then clusters the cells by Leiden clustering.

2.4. Evaluation metrics and visualization for clustering

To evaluate the clustering performance of scGMM-VGAE in comparison to four baseline models, we considered three metrics: adjusted Rand index (ARI), normalized mutual information (NMI), and Silhouette score. However, for unlabeled datasets, only Silhouette metric can be considered since this metric does not require truth labels. Furthermore, to visualize the clustering results, we considered uniform manifold approximation and projection (UMAP).

Adjusted Rand index (ARI): ARI calculates the similarities between true labels and the predicted labels of a given cluster. ARI values range between 0 and 1 where 0 represents a random result and 1 represents a complete agreement between the clusters (Yeung & Ruzzo, 2001). The equation for calculating the ARI value is defined as

$$ARI = \frac{\Sigma_{ij}\binom{N_{ij}}{2} - \left[\Sigma_i\binom{N_i}{2}\Sigma_j\binom{N_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\Sigma_i\binom{N_i}{2} + \Sigma_j\binom{N_j}{2}\right] - \left[\Sigma_i\binom{N_i}{2}\Sigma_j\binom{N_j}{2}\right]/\binom{N}{2}}, \qquad (10)$$

where N_{ij} represents the number of cells with true label i assigned to cluster j. N_i represents the number of cells assigned in cluster i, N_j represents the number of cells with true label j.

Normalized mutual information (NMI): The mutual information (MI) score is used to measure the similarity and exploits the grouping property (Kraskov & Grassberger, 2009). NMI is a normalization of MI in which MI is normalized by generalized mean of the true labels and predicted labels. NMI is defined as

$$NMI(Y,C) = \frac{2MI(Y,C)}{[H(Y) + H(C)]}, \tag{11}$$

where Y is the predicted cell labels, C is the true cell labels, $H(Y)$ and $H(C)$ are the entropy of true and predicted cell labels, respectively. Thus, NMI scales the result from 0 (no mutual information) to 1 (perfect correlation).

Silhouette score: Silhouette was developed as a graphical tool for the validation of clustering results. This metric evaluates the degree of separation between clusters by measuring the tightness and separation between clusters (Zhao et al., 2018). SC is given as

$$SC_i = \frac{(b_i - a_i)}{max(b_i,a_i)}, \tag{12}$$

where a is dissimilarity within a cluster i and b is the dissimilarity between cluster i and its nearest cluster. The score ranges from -1 to 1 where the larger value indicates a higher degree of separation among clusters.

Uniform manifold approximation and projection (UMAP): UMAP is a visualization method for clustering results (McInnes et al., 2018). This method performs dimension reduction on the feature matrix and represents the cells in a two-dimensional setting. Since UMAP uses a framework based on Riemannian geometry, the structure of the original data in a reduced dimension can be preserved. In UMAP visualization, cells visualized in a two-dimensional setting and labels are plotted in different colors to present the clustering results.

3. Results

3.1. Clustering performance in real datasets

As discussed in **Section 2.3**, we applied scGMM-VGAE and other four baseline clustering methods (i.e., Seurat, DCA+Leiden, scVI, and CellVGAE) to three real labelled datasets (i.e., *Baron3 data*, *Baron4 data*, and *Darmanis data*) and one real unlabeled dataset (i.e., *Yang data*). To evaluate the performance of scGMM-VGAE, we compared the results to four baseline clustering methods based on the two true label-based metrics (i.e., ARI and NMI) and the internal metric silhouette score (SC). **Figure 2** illustrates the clustering results of the three real labelled datasets (**Panels A, B, and C**) and one unlabeled dataset (**Panel D**). For scGMM-VGAE, we used all six parameter settings (**Table 3**) for each measurement.

For the labelled data, scGMM-VGAE outperforms the other four baseline methods on three evaluation metrics, although these labelled datasets contain many zero values. The second-best results are from either Cell-VGAE or Seurat, depending on the datasets and metrics. Specifically, the ARI and NMI results of scGMM-VGAE surpass the ARI and NMI of the second ranked method by 0.02 to 0.06 and 0.01 to 0.1, respectively. These findings are evidence to show that scGMM-VGAE produces the best results although a significant number of zeros presented in scRNA-seq data.

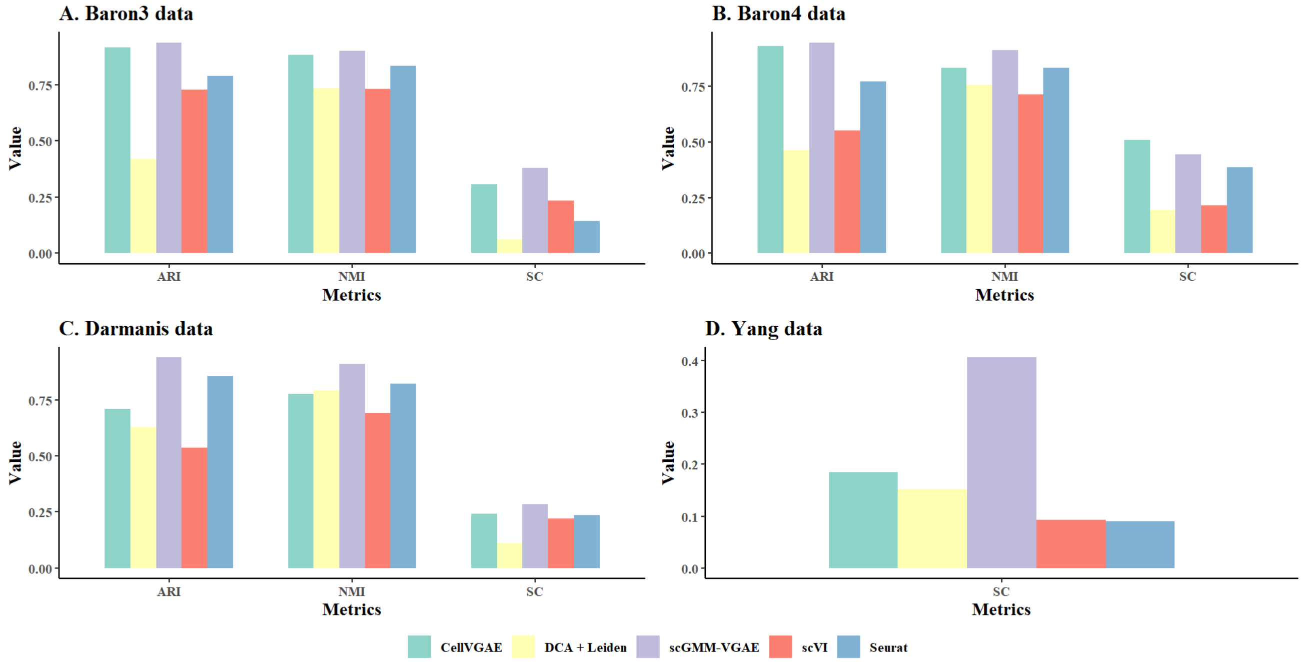

Figure 2: Performance scores of scGMM-VGAE and four baseline methods on three labelled and one unlabelled dataset: (A.) *Baron3 data*; (B.) *Baron4 data*; (C.) *Darmanis data*; (D.) *Yang data*.

On the other hand, to compare the clustering performance of scGMM-VGAE with four baseline clustering methods on unlabeled *Yang data*, only silhouette score (SC) metric was used for evaluation. The clustering results for *Yang data* are illustrated in **Figure 2D**. As shown in the histogram, scGMM-VGAE is the best candidate in cell clustering in compared to other four methods on *Yang data.*

3.2. UMAP visualization

We consider UMAP visualization to demonstrate the clustering results of the proposed scGMM-VGAE and the four selected baseline methods in comparison to the true labels (i.e., *Truth*). The UMAP visualization of *Baron3 data* (**Figure 3**) and each method show that the DCA + Leiden overestimated cluster counts. When compared to the true labels, many clusters are located at the top of the UMAP cluster. Other methods correctly clustered the top portion of the UMAP visualization. As seen in the bottom right clusters, the Seurat and DCA + Leiden methods produce over-clustered results with more clusters than the truth label. Other methods appear to have appropriate comparisons to the truth.

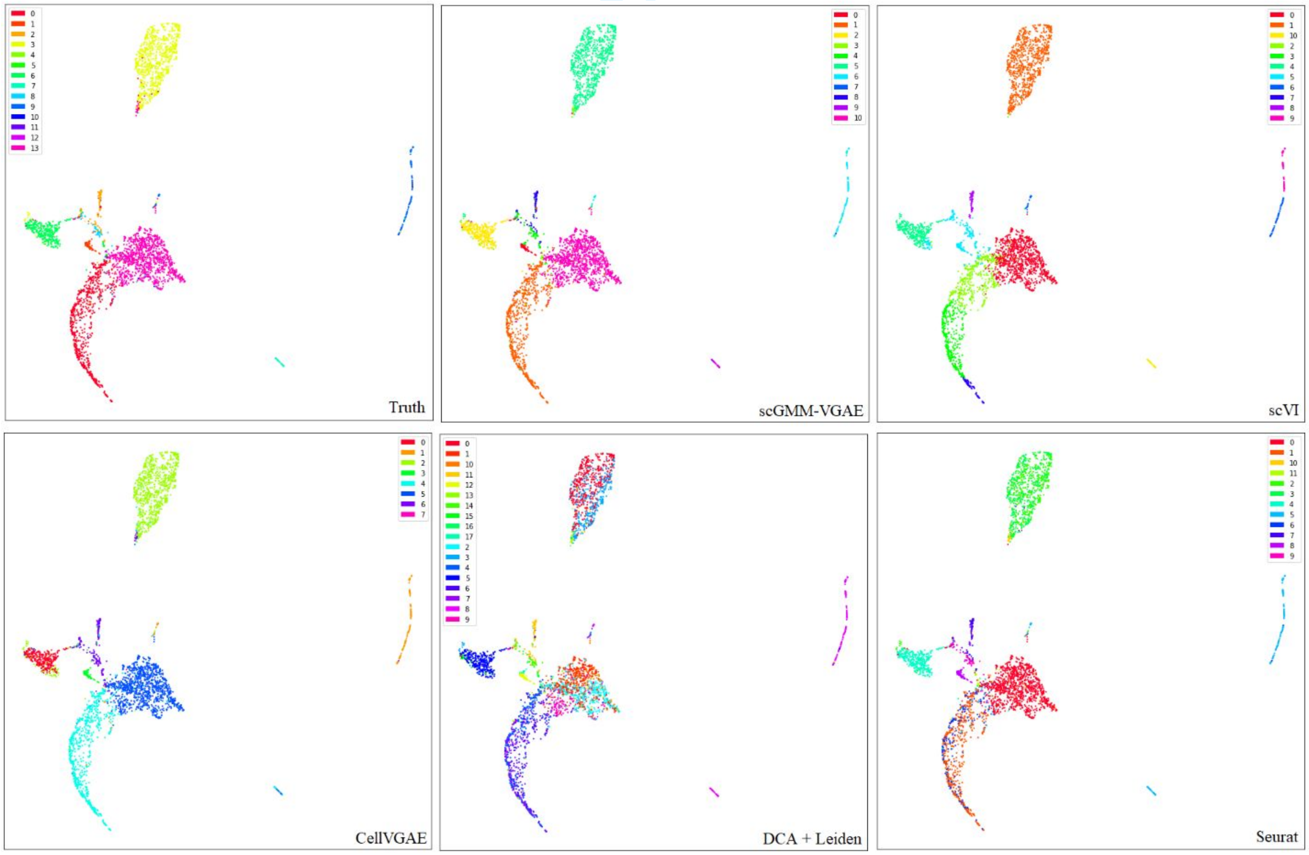

Figure 3: UMAP visualization of clustering result for *Baron3 data*.

The same results can be observed in *Baron4 dataset*. Over-clustering can be observed in DAC + Leiden method and Seurat method, with DCA + Leiden methods having more over-clustering in both the top and the bottom sections. UMAP visualization for clustering results of *Baron4 data* and *Darmanis data* can be found in **Supplementary information**.

3.3. Clustering performance in simulated datasets

To further evaluate performance and stability of scGMM-VGAE in cell clustering, we applied the internal metric SC along with the label-based ARI and NMI metrics on the simulated datasets with three different cell numbers generated by SPARSim using the Zheng preset (**Section 2.1**), as summarized in **Table 2**. For each size (i.e., number of cells) of the simulated data, we generated ten copies of datasets and calculated the average mean and standard deviation of each metric. The histograms illustrating clustering performance of the five methods in three simulated datasets is presented in **Figure 4**.

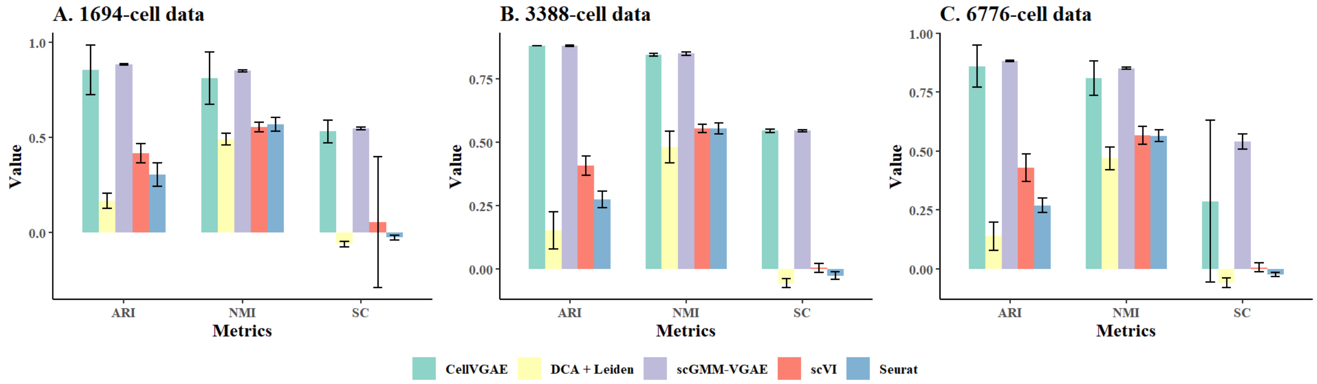

Figure 4: Average performance scores (mean and standard deviation) of scGMM-VGAE and four baseline methods on simulated datasets with (A.) *1694-cell data*, (B.) *3388-cell data*, and (C.) *6776-cell data.* The error bars represent 95% confidence interval of the ten copies simulated dataset for each size.

Similar to the previous results of real labelled and unlabeled datasets, scGMM-VGAE consistently shows the best results among the five methods on all simulated data for all three metrics. Cell-VGAE achieves the second-best results for ARI, NMI, and SC. Specifically, for *1694-cell data*, ARI, NMI, and SC of scGMM-VGAE surpass those of Cell-VGAE by 0.02, 0.04, and 0.02. For

3388-cell data, scGMM-VGAE produces similar results as Cell-VGAE. The difference of each metric score between two methods are within 0.005. As for *6776-cell data*, scGMM-VGAE surpasses Cell-VGAE by 0.02 for ARI, 0.04 for NMI, and 0.25 for SC. To test the clustering stability of scGMM-VGAE among different-size datasets, we group all the simulation results of scGMM-VGAE by evaluation metrics, as shown in **Figure 5**. The results from three different sizes of simulated data are almost the same for all three metrics. Thus, scGMM-VGAE shows great stability.

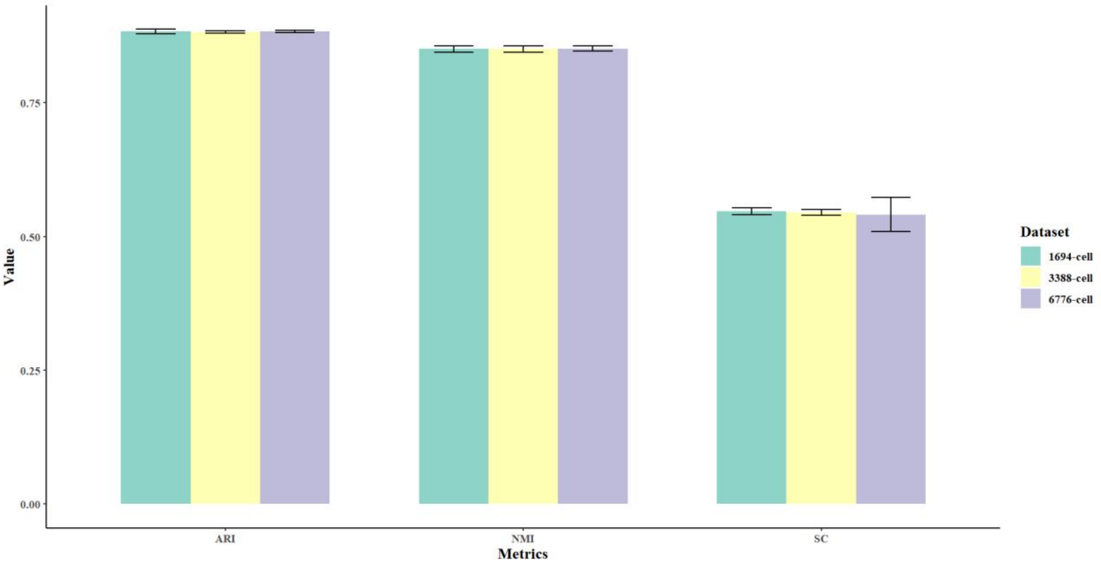

Figure 5: Stability test for scGMM-VGAE on simulated data. The ARI, NMI, and SC values of scGMM-VGAE clustering results for 1694-cell (green), 3388-cell (yellow), and 6776-cell (purple) simulated datasets. The error bars represent 95% confidence interval of the ten copies simulated dataset for each size.

3.4. Time efficiency

When analyzing the largest simulated dataset (i.e., *6776-cell data*), we recorded the time/run for each algorithm, which is summarized in **Table 4**. The scGMM-VGAE achieves 2.75 minutes per run, approximately 2.5 times faster than another VGAE-based algorithm Cell-VGAE (7 min/run). Seurat is the fastest algorithm among five algorithms with less than 10 seconds per run since the algorithm does not require a deep neuron network. However, among the deep learning-based methods, scGMM-VGAE is one of the fastest algorithms.

Table 4. Summary of running speed (min/run) of scGMM-VGAE and four baseline methods on 6776-cell simulated dataset.

Clustering methods	Running speed
scGMM-VGAE	2.75 min/run
ScVI	22 min/run
Cell-VGAE	7 min/run
DCA+Leiden	1.75 min/run
Seurat	≤ 10 sec/run

4. Conclusion and discussion

Based on the clustering results of the labelled data and unlabelled data (**Section 3.1**), and simulated data (**Section 3.3**), we can conclude that scGMM-VGAE shows better performance over other selected baseline methods for all selected datasets. Especially, scGMM-VGAE outperforms another VGAE-based clustering method Cell-VGAE, which achieves second-best results for most datasets. Thus, this study reconfirms that scGMM-VGAE can be effectively used for clustering scRNA data containing a great amount of biological and technical zeroes since the algorithm incorporates a Gaussian mixture model to the encoder data generated using VGAE to further enhance the clustering performance. From the stability and time efficiency tests, we also observe that scGMM-VGAE shows adequate stability and running speed. In conclusion, scGMM-VGAE successfully combines an unsupervised deep learning model (VGAE) with a statistical model (GMM) to enhance the clustering performance of scRNA data.

The algorithm requires a predefined number of clusters for initialization. However, knowing the exact number of cell types in advance is difficult, and experienced determination is not always reliable. Another limitation is that only the top 1200 genes were selected in our study. With different numbers of top genes, different feature matrices and cell-cell graphs will be produced, which might lead to different clustering results. Thus, in future applications, we will consider more parameter settings and the selected number of cells and aim to extend the scGMM-VGAE without requiring predefined number of clusters. We also aim to explore other possibilities for utilizing the scGMM-VGAE model in biological data analysis (e.g., clustering microbiome data).

5. Data availability

The various publicly available scRNA-seq datasets that support the findings of this study are available at Gene Expression Omnibus (GEO) database. The accession numbers of Baron3 and Baron4 data are GSM2230759 and GSM2230760, respectively. Darmanis data was downloaded at https://github.com/ciortanmadalina/single-cell-sota/tree/master/input/brainCIDR. Raw Darmanis data can also be found at GEO (accession no. GSE67835). The accession number for the raw data files of Yang data reported in this study is GSE90047.

6. Funding

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Conflict of interest: none declare.

References

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503-527. https://doi.org/10.1016/j.datak.2007.03.016

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, *3*(4),346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Baruzzo, G., Patuzzi, I., & Di Camillo, B. (2020). SPARSim single cell: A count data simulator for scRNA-seq data. *Bioinformatics, 36*(5),1468-1475. https://doi.org/10.1093/bioinformatics/btz752

Booeshaghi, A. S., & Pachter, L. (2021). Normalization of single-cell RNA-seq counts by log (x + 1) or log(1 + x). *Bioinformatics, 37*(15), 2223-2224. https://doi.org/10.1093/bioinformatics/btab085

Buterez, D., Bica, I., Tariq, I., Andrés-Terré, H., & Liò, P. (2021). CellVGAE: An unsupervised scRNA-seq analysis workflow with graph attention networks. *Bioinformatics, 38*(5),1277-1286. https://doi.org/10.1093/bioinformatics/btab804

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*, 411-420. https://doi.org/10.1038/nbt.4096

Chaudhry, F., Isherwood, J., Bawa, T., Patel, D., Gurdziel, K., Lanfear, D. E., Ruden, D. M., & Levy, P. D. (2019). Single-cell RNA sequencing of the cardiovascular system: New looks for old diseases. *Frontiers in Cardiovascular Medicine, 6*, 173. https://doi.org/10.3389/fcvm.2019.00173

Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in genetics, 10*,317. https://doi.org/10.3389/fgene.2019.00317

Cui, Y., Zhang, S., Liang, Y., Wang, X., Ferraro, T. N., & Chen, Y. (2021). Consensus clustering of single-cell RNA-seq data by enhancing network affinity. *Briefings in Bioinformatics, 22*(6). https://doi.org/10.1093/bib/bbab236.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A., & Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences, 112*(23), 7285-7290. https://doi.org/10.1073/pnas.1507125112.

El Assaad, H., Samé, A., Govaert, G., & Aknin, P. (2016). A variational expectation-maximization algorithm for temporal data clustering. *Computational Statistics & Data Analysis, 103*, 206-228. https://doi.org/10.1016/j.csda.2016.05.007

Eraslan, G., Simon, L., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-018-07931-2.

Garriga, J., Palmer, J. R. B., Oltra, A., & Bartumeus, F. (2016). Expectation-maximization binary clustering for behavioural annotation. *PloS one, 11*(3), e0151984-e0151984. https://doi.org/10.1371/journal.pone.0151984

Govek, K., Yamajala, V., & Camara, P. (2019). Clustering-independent analysis of genomic data using spectral simplicial theory. PLoS *Computational Biology, 15*(11), e1007509. https://doi.org/10.1371/journal.pcbi.1007509.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … & Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell, 184*(13), 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine, 9*(1), 75. https://doi.org/10.1186/s13073-017-0467-4

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science (American Association for the Advancement of Science), 313*(5786), 504-507. https://doi.org/10.1126/science.1127647

Hui, B., Zhu, P., & Hu, Q. (2020). Collaborative graph convolutional networks: Unsupervised learning meets semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 4215-4222. https://doi.org/10.1609/aaai.v34i04.5843.

Jiang, Z., Zheng, Y., Tan, H., Tang, B., & Zhou, H. (2017). Variational deep embedding: An unsupervised and generative approach to clustering. *IJCAI.* https://doi.org/10.48550/arXiv.1611.05148

Kingma, D., & Welling, M. (2014). Auto-encoding variational Bayes. *ICLR.* https://doi.org/10.48550/arXiv.1312.6114

Kipf, T. N., & Welling, M. (2016). Variational graph autoencoders. *NeurIPS Workshop on Bayesian Deep Learning*. https://doi.org/10.48550/arXiv.1611.07308

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *ICLR.* https://doi.org/10.48550/arXiv.1609.02907

Kraskov, A., & Grassberger, P. (2009). MIC: Mutual information based hierarchical clustering. In F. Emmert-Streib, & M. Dehmer (Eds), *Information theory and statistical learning*. (pp. 101-123). Boston, MA.

Krzak, M., Raykov, Y., Boukouvalas, A., Cutillo, L., & Angelini, C. (2019). Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. *Frontiers in Genetics, 10*, 1253-1253. https://doi.org/10.3389/fgene.2019.01253

Li, R., Guan, J., & Zhou, S. (2020). Single-cell RNA-seq data clustering: A survey with performance comparison study. *Journal of Bioinformatics and Computational Biology, 4*(18). https://doi.org/10.1142/S0219720020400053

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods, 15*, 1053-1058. https://doi.org/10.1038/s41592-018-0229-2

Malik, A. (2019). *Applied unsupervised learning with R: uncover hidden relationships and patterns with K-Means clustering, hierarchical clustering, and PCA*. (1st ed.) Birmingham.

McInnes, L., Healy, J., Saul, N., & Grosberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open-source Software, 3*(29), 861. https://doi.org/10.21105/joss.00861

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application, 6*(1), 355-378. https://doi.org/10.1146/annurevstatistics031017-100325.

Rao, J., Zhou, X., Lu, Y., Zhao, H., & Yang, Y. (2021). Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. *iScience, 24*(5), 102393. https://doi.org/10.1016/j.isci.2021.102393

Reynolds, D. (2015). Gaussian mixture models. In S. Z. Li, & A. K. Jain (Eds.), *Encyclopedia of Biometrics*. (pp. 827-832). Springer US, Boston, MA.

Shiga, M., Seno, S., Onizuka, M., & Matsuda, H. (2021). SC-JNMF: Single-cell clustering integrating multiple quantification methods based on joint non-negative matrix factorization. *PeerJ.* https://doi.org/10.7717/peerj.12087.

Sinaga, K. P. & Yang, M. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, *8*, 80716-80727. https://doi.org/10.1109/ACCESS.2020.2988796

Stuart, T., Butler, A., Ho_man, P., Hafemeister, C., Papalexi, E., Mauck, W. M. I., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell, 177*, 1888-1902. https://doi.org/10.1016/j.cell.2019.05.031.

Su, K., Yu, T., & Wu, H. (2021). Accurate feature selection improves single-cell RNA-seq cell clustering. *Briefings in Bioinformatics, 22*(5). https://doi.org/10.1093/bib/bbab034.

Tsumoto, S., Kimura, T., and Hirano, S. (2022). Expectation-maximization (EM) clustering as a preprocessing method for clinical pathway mining. *The review of socionetwork strategies, 16*(1), 25-52. https://doi.org/10.1007/s12626-021-00100-w

Uykan, Z. (2021). Fusion of centroid-based clustering with graph clustering: An expectation maximization-based hybrid clustering. In *IEEE transaction on neural networks and learning systems*. https://doi.org/ 10.1109/TNNLS.2021.3121224

Wang, J., Zou, Q., & Lin, C. (2022). A comparison of deep learning-based pre-processing and clustering approaches for single-cell RNA sequencing data. *Briefings in Bioinformatics, 23*(1). https://doi.org/10.1093/bib/bbab345

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

Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., Wang, T., Larsson, L., Kaczorowski, D., Weisenfeld, N. I., Uytingco, C. R., Chew, J. G., Bent, Z. W., Chan, C-L., …, Swarbrick, A. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics, 53*(9), 1334-1347. https://doi.org/10.1038/s41588-021-00911-1.

Yang, L., Wang, W., Qiu, W., Guo, Z., Bi, E., & Xu, C. (2017). A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology, 66*(5), 1387-1401. https://doi.org/10.1002/hep.29353

Yeung, K. Y. & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics, 17*(9), 763-774. https://doi.org/10.1093/bioinformatics/17.9.763.

Yu, S., Chu, S.,Wang, C., Chan, Y., & Chang, T. (2018). Two improved k-means algorithms. *Applied Soft Computing, 68*, 747-755. https://doi.org/10.1016/j.asoc.2017.08.032

Zhao, S., Sun, J., Shimizu, K., and Kadota, K. (2018). Silhouette scores for arbitrary defined groups in gene expression data and insights into differential expression results. *Biological Procedures Online, 20*(1), 5. https://doi.org/10.1186/s12575-018-0067-8

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., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., …, Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049