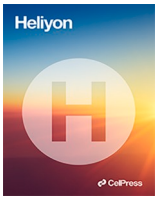
Heliyon 10 (2024) e40732

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

scVAG: Unified single-cell clustering via variational-autoencoder integration with Graph Attention Autoencoder

Seyedpouria Laghaee [a], Morteza Eskandarian [b], Mohammadamin Fereidoon [a], Somayyeh Koohi [a,*]

[a] Department of Computer Engineering, Sharif University of Technology, Tehran, Tehran, 1458889694, Iran
[b] Department of Computer Science, University of Tehran, Tehran, Tehran, 1417614411, Iran

ARTICLE INFO

Keywords:
Single-cell RNA sequencing
Variational-autoencoder
Graph attention autoencoder
Clustering
Dimensionality reduction

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) enables high-resolution transcriptional profiling of cell heterogeneity. However, analyzing this noisy, high-dimensional matrix remains challenging. We present scVAG, an integrated deep learning framework combining Variational-Autoencoder (VAE) and Graph Attention Autoencoder (GATE) for enhanced single-cell clustering. Building upon scGAC, our approach replaces its restrictive linear principal component analysis (PCA) with nonlinear dimensionality reduction better suited for scRNA-seq data. Specifically, we integrate VAE and GATE to enable more flexible latent space encoding. Extensive experiments on 20 datasets demonstrate scVAG's superior performance over previous state-of-the-art methods including scGAC, SCEA, SC3, Seurat, scGNN, scASGC, DESC, NIC, scLDS2, DRJCC, sLMIC, and jSRC. On average, scVAG improves clustering accuracy by 5 percent in ARI and 4 percent in NMI parameters. Visualizations highlight scVAG's capacity to recover interpretable biological structures. Our VAE-GATE pipeline extracts intricate expression patterns into compact representations that precisely delineate cell subpopulations consistent with ground truth labels. Overall, scVAG establishes a robust architecture for elucidating cell taxonomies from noisy transcriptomic inputs.

1. Introduction

Cells are the fundamental units governing organism development through intricate gene interaction networks, giving rise to diverse, specialized cell types. While cellular heterogeneity enables population adaptation, it can also lead to disease if abnormal changes occur. Developing targeted treatments requires comprehensively decoding cell functionality across contexts [1]. By using RNA sequencing technology, it becomes possible to profile expression patterns of regulatory genes that govern cell behaviour. While traditional sequencing (Bulk RNA sequencing) measures average gene expression and ignores cell-to-cell variability, single-cell RNA sequencing changes cellular biology. Single-cell profiling has revolutionized our understanding of tissue formation, disease progression, and clinical response by unveiling novel cell types, trajectories, and communication networks. By uniquely barcoding mRNA from thousands of cells and capturing sequences, expression levels of more than ten thousand genes are quantified simultaneously across a cell population [2]. This provides unprecedented resolution into cell heterogeneity that is obscured in traditional bulk sequencing approaches [3]. However, realizing the full potential of scRNA-seq remains challenging due to the high dimensionality and

* Corresponding author.
E-mail addresses: poria.laghayee@gmail.com (S. Laghaee), morteza.eskandarian@gmail.com (M. Eskandarian), mohammadamin.fereidoon@gmail.com (M. Fereidoon), koohi@sharif.edu (S. Koohi).

https://doi.org/10.1016/j.heliyon.2024.e40732
Received 27 April 2024; Received in revised form 29 October 2024; Accepted 25 November 2024
Available online 27 November 2024

technical issues inherent to the data. A major characteristic is the abundance of zero counts, stemming from either biological or technical factors. Biological zeros result from genes not expressed in a cell, while technical zeros occur when low abundance transcripts are lost during library processing and amplification. Discriminating between these zero types poses analytical difficulties [4]. Thus, appropriate computational approaches are critical to handle the sparsity and high dimensionality of scRNA-seq data [5].

Clustering analysis is crucial for understanding intricate cellular patterns. Moreover, clustering often forms the foundation for downstream differential expression testing, trajectory reconstruction, and cell annotation. It is important to note that, traditional clustering methods face limitations in handling noisy, high-dimensional data. They make assumptions about data distributions and require manual tuning of parameters like cluster numbers, risking suboptimal, biased results. In contrast, deep learning and graph-based methods provide compelling solutions by learning representations directly from gene expression patterns without restrictive assumptions. Complex neural networks and graph structures can intrinsically model nonlinear relationships in high-dimensional data, sensitively revealing rare subpopulations and transitional states indistinguishable to prior techniques [6].

To comprehensively evaluate our model, we compared it against many state-of-the-art single-cell gene expression clustering methods. Several state-of-the-art single-cell gene expression clustering methods have been introduced in recent years. Seurat [7] is widely used for integrating various data types and enabling both clustering and single-cell data inference. DESC [8], an autoencoder-based model, maps scRNA-seq data to a lower-dimensional space for clustering. SC3 [9] integrates multiple techniques and utilizes consensus clustering for robust results. The scASGC [10] model employs adaptive graph convolution with an attention mechanism for cell representation. scGNN [11] uses Graph Neural Networks for embedding, imputation, and clustering but is limited by ineffective graph denoising. scGAC [12] and SCEA [13] leverage Graph Attention mechanisms and use Network Enhancement [14] to denoise graphs. However, scGAC relies on Principal Component Analysis (PCA) for initial dimensionality reduction, which may not be optimal for complex high-dimensional data. SCEA uses a Multi-Layer Perceptron (MLP)-based encoder but lacks mechanisms to regulate learning effectively. Our model was also evaluated against several MATLAB-based methods, including sLMIC [15], NIC [16], jSRC [17], and DRjCC [18]. sLMIC integrates multi-omics data through self-representation learning, extracting shared and specific features. NIC clusters single-cell data using adaptive graph learning for integrated analysis. jSRC combines dimension reduction and clustering to enhance interpretability and scalability. DRjCC jointly learns dimension reduction and clustering, integrating projected matrix decomposition with non-negative matrix factorization. Additionally, scLDS2 [19] uses adversarial learning to effectively identify rare cell types. These comparisons demonstrate the robustness and versatility of our approach, tested against diverse baseline models with different clustering strategies.

In our study, we present scVAG, a novel approach integrating Variational-Autoencoder (VAE) [20] and Graph Attention Autoencoder (GATE) [21] to enhance single-cell analysis. Building upon the scGAC framework, scVAG incorporates diverse Autoencoders and replaces PCA with a nonlinear dimensionality reduction structure. This improves downstream clustering accuracy.

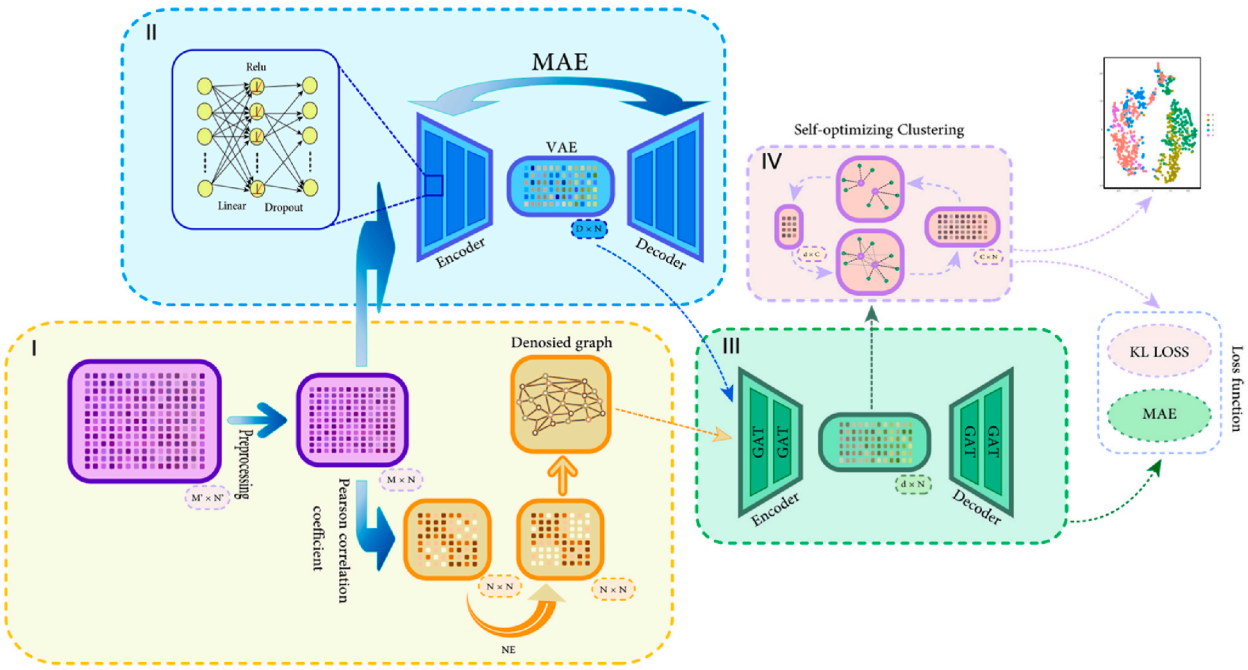


Fig. 1. Overview of the scVAG methodology. The workflow is divided into four main stages: I. Data preprocessing, graph construction, and graph denoising using Network Enhancement (NE); II. Variational Autoencoder (VAE) for dimensionality reduction of preprocessed data; III. Graph Attention Autoencoder (GATE) combining VAE results and denoised graph data to generate new embeddings using attention mechanisms; IV. Self-optimized clustering for final cell type identification. This pipeline integrates single-cell gene expression data processing, graph-based analysis, and deep learning techniques to achieve accurate and interpretable cell clustering results.

By leveraging the strengths of VAE in capturing latent representations of single-cell data and GATE in modeling complex relationships within graph-structured data, scVAG aims to capitalize on the advantages of both structures. The enhancements are attributed to the integration of VAE and GATE components, enabling more precise representation learning and capturing intricate relationships in single-cell gene expression patterns. The performance of scVAG was assessed on 19 diverse real and simulated datasets, using three widely-adopted clustering evaluation metrics. ScVAG outperformed state-of-the-art methods, including graph attention-based approaches like SCEA and scGAC.

2. Results

2.1. Overview of scVAG workflow

A robust analysis of single-cell RNA sequencing data necessitates a well-defined pipeline including data preprocessing for cleanliness and normalization of data, dimensionality reduction for creating a meaningful low-dimensional representation, and clustering for identifying boundaries between cell subpopulations. This integrated approach ensures effective visualization and discovery of cellular heterogeneity in the high-dimensional gene expression profiles [22]. Building upon this need, our method, scVAG, implements a comprehensive four-step process to ensure robust and accurate clustering outcomes in single-cell gene expression analysis. As shown in Fig. 1, scVAG's methodology comprises four stages: I) data preprocessing and graph denoising; II) VAE-based dimension reduction; III) embedding generation via Graph Attention Autoencoder; and IV) self-optimized clustering for cell identification. This multi-step approach enables scVAG to effectively reduce high-dimensional single-cell data into a meaningful low-dimensional representation for revealing distinct cell characteristics, delineating boundaries, and accurately clustering cell subpopulations.

Data preprocessing is a critical first step for reliable single-cell clustering. Careful data preprocessing, including filtering low-quality cells and genes, can reduce technical noise in scRNA-seq measurements. As demonstrated by Scanpy [23], tailored preprocessing enhances the quality of gene expression matrix for subsequent clustering. The cell-cell association graph is then constructed, revealing topological connections between similar cells and easing boundary identification. Subsequently, denoising methods are applied to refine the graph, enhancing its reliability and suitability for further analysis.

Building upon the denoised graph, scVAG employ an integrated pipeline that combines a Variational-Autoencoder (VAE) with a Graph Attention Autoencoder (GATE) for nonlinear dimensionality reduction. By using these two different structures in dimension reduction part, this approach leverages both global data distributions and local graph structures to effectively reveal patterns among cell subpopulations. The decision not to use VAE or GATE independently stems from practical considerations. GATE's performance tends to decrease with a large number of features, and it requires substantial time to run on extensive gene expression profiles. This is primarily due to the calculation of attention weights, which scales with the number of input features. On the other hand, using VAE alone would miss out on graph information, which is essential for capturing intercellular relationships. Therefore, scVAG integrates VAE and GATE components to take advantage of their complementary strengths.

We performed ablation experiments to assess the individual contributions of VAE and GATE components. According to Fig. 2, the best result was consistently achieved when the model used a combination of VAE and GATE in the pipeline across various datasets. This combination not only improves clustering accuracy but also enhances the model's scalability to larger datasets.

In addition, we observed significant time efficiency improvements in our development process. For instance, using the Chung

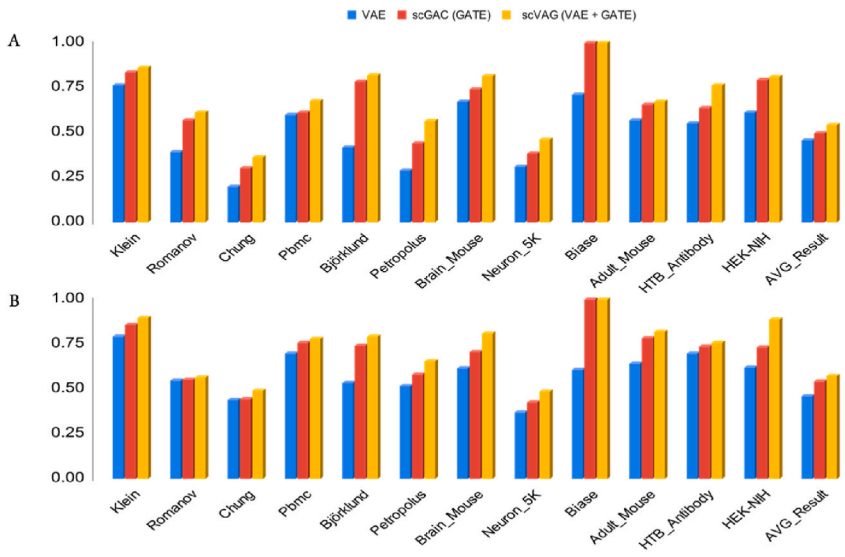


Fig. 2. Results of the ablation experiment, comparing model performance using VAE, GATE, and their combination (ARI (A) and NMI (B)). Compared to using VAE or GATE alone, the combined approach exhibits superior performance.

dataset (approximately 500 cells) on a Google Colab GPU, the full scVAG model completed its run in about 15 min, whereas GATE alone on the full feature set required approximately 319 min. This example illustrates the potential computational benefits of our combined approach. These observations underscore the synergistic benefits of combining VAE and GATE in the scVAG pipeline, demonstrating improvements in both performance and practical applicability.

Finally, a self-optimizing clustering method is used for more concrete clustering. A standard algorithm such as K-means may produce suboptimal results due to random center initialization. To address this, scVAG integrates a self-optimizing mechanism into K-means. In the methods section, each component of the pipeline will be discussed in greater detail.

2.2. Evaluation metrics for clustering

In the assessment of our proposed method, we utilized three widely recognized metrics: Adjusted Rand Index (ARI) [24], Normalized Mutual Information (NMI) [25], and Cosine Similarity (Cs) [26]. These quantify the similarity between predicted and true cluster labels, providing a comprehensive evaluation. As shown in Eq. (1), the Adjusted Rand Index is calculated using the following formula.

$$ARI = \frac{\sum ij \binom{n(i.\ j)}{2} - \frac{\left[\sum_i \binom{a(i)}{2} \cdot \sum_j \binom{b(j)}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum i \binom{a(i)}{2} \cdot \sum j \binom{b(j)}{2}\right] - \frac{\left[\sum_i \binom{a(i)}{2} \cdot \sum_j \binom{b(j)}{2}\right]}{\binom{n}{2}}} \tag{1}$$

In the Adjusted Rand Index formula, $n(i.\ j)$ represents the number of cells assigned to cluster i based on the model's predicted labels and to cluster j based on the true labels. $a(i)$ denotes the total count of cells predicted to be in cluster i. Similarly, $a(j)$ signifies the total count of cells actually belonging to cluster j based on true labels. Finally, n corresponds to the total number of clusters. The Adjusted Rand Index accounts for chance in cluster assignments, making it suitable for random labeling scenarios. The next metric, Normalized Mutual Information is calculated using the formula shown in Eq. (2).

$$NMI(X.\ Y) = \frac{MI(X.\ Y)}{H(X) + H(Y)} \tag{2}$$

where, $MI(X.\ Y)$ represents the mutual information between the assigned clustering X and true labels Y, $H(X)$ denotes the information entropy for prediction X, and $H(Y)$ is the information entropy of the true labels. Normalized Mutual Information offers a normalized perspective of mutual information values between assigned and true labels, facilitating interpretation. Both ARI and NMI offer valuable insights into the clustering accuracy, particularly in scenarios where direct label matching may be impractical due to the nature of clustering algorithms. In addition to the ARI and NMI, we utilized Cosine Similarity as another evaluation parameter for assessing the efficacy of our proposed clustering method. Cosine Similarity measures the cosine of the angle between two vectors and is commonly used to quantify the similarity between two clusters. The Cosine Similarity between two vectors X and Y is calculated as follows:

Table 1
Summary of real single-cell datasets used to evaluate scVAG.

Dataset Name	Type	# of clusters	# of cells	# of genes	sequencing platform	Date of released	Access link
Klein [27]	Homo sapiens	4	2717	24021	Illumina HiSeq	2015	GSE65525
Romanov [28]	Mus musculus	7	2863	18496	Illumina HiSeq	2017	GSE74672
Chung [1]	Homo sapiens	5	515	27420	Illumina HiSeq	2017	GSE75688
PBMC [29]	Homo sapiens	8	4220	16412	10X	2017	10x Genomics
Björklund [30]	Homo sapiens	4	647	26,087	Smart-Seq2	2016	GSE70580
Petropolus [31]	Homo sapiens	5	1518	21627	Illumina HiSeq	2016	E-MTAB-3929
Brain_Mouse [32]	Mus musculus	5	501	19465	spaceranger-2.0.1	2023	10x Genomics
Neuron_5K [33]	Homo sapiens	11	5483	32286	cellranger-6.0.0	2020	10x Genomics
Biase [34]	Mus musculus	3	49	21489	Illumina HiSeq	2014	GSE57249
Human_TBNK_Antibody [35]	Homo sapiens	5	892	36601	cellranger-6.0.0	2021	10x Genomics
Adult_Mouse_Heart_5k [36]	Mus musculus	10	3220	19925	cellranger-7.0.0	2022	10x Genomics
HEK293T & NIH3T3(HEK-NIH) [37]	Homo sapiens & Mus musculus	3	5923	72302	cellranger-8.0.0	2024	10x Genomics

$$\mathrm{CS}(X.Y) = \frac{X \cdot Y}{\| X\| \times \| Y \|} \tag{3}$$

where $X \cdot Y$ represents the dot product of vectors X and Y, and $\| X \|$ and $\| Y \|$ denote the Euclidean norms of vectors X and Y, respectively. In the context of clustering evaluation, Cosine Similarity offers insights into the alignment between the cluster assignments generated by our method and the true labels of the data. By quantifying the similarity of cluster assignments to the ground truth, Cosine Similarity provides an additional perspective on the clustering performance beyond what ARI and NMI offer.

2.3. Overview of datasets

We used 20 challenging real and simulated single-cell datasets from diverse biological contexts in order to provide a comprehensive evaluation of scVAG. The real single-cell datasets used in the project range from human to mouse samples, as summarized in Table 1. As listed in this table, the gene counts range from 16,412 to 36,601 and the cell numbers range from 49 to 5483. Our dataset package covers data with varying cell counts - from small datasets like Biase with 49 cells to large datasets like PBMC with 5483 cells.

In addition, we generated 8 synthetic datasets with splatter [38] that covered a variety of situations, including varying numbers of clusters and imbalanced cluster distributions, to evaluate the model's performance under varying data characteristics.

Table 2 provides detailed information of these datasets. Finally, it is worth noting that the datasets used in this project were obtained from public data repositories and original publications. The cell type labels for each dataset were derived from the metadata provided with the downloaded data files.

2.4. Data analysis

In this section, we delve into a comprehensive analysis of scVAG's performance, emphasizing its clustering accuracy compared to twelve other baseline methods (SCEA, scGAC, scGNN, Seurat, DESC, SC3, scASGC, DRJCC, NIC, scLDS2, jSRC, and sLMIC). The primary metrics under consideration are the ARI and NMI, offering a well-rounded evaluation across diverse single-cell datasets. ARI and NMI comparisons between scVAG and other baselines are shown in Fig. 3. As illustrated, scVAG consistently outperforms the selected baselines across both ARI and NMI metrics. According to the average subplot of ARI and NMI, scVAG is capable of capturing meaningful subpopulations of cells in various biological conditions. To evaluate scVAG's robustness, we present two additional visualizations in Fig. 4. The first is a diagram depicting clustering accuracy vs. dataset size, with color brightness indicating performance for both ARI and NMI metrics. The second uses a heatmap showing accuracy values within each cell, with color brightness again representing clustering accuracy. Both visualizations demonstrate scVAG's robust performance.

Furthermore, to evaluate the noise-handling capability and effectiveness of our proposed model under challenging conditions, we conducted experiments using 8 synthetically generated datasets exhibiting varying degrees of class number, cluster imbalance, and noise exposure. Fig. 5 shows how our model consistently achieved better results on simulated datasets.

Fig. 6 utilizes t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize scVAG's clustering findings on seven datasets: Chung, Human_TBNK_Antibody, HEK-NIH, Klein, PBMC, Petroplus, and Romanov.

The visualizations demonstrate scVAG's interpretability in accurately identifying meaningful biological structures. In the Chung data, scVAG distinctly separates B cells and T cells, superior to other methods. For the Human_TBNK_Antibody data, scVAG's clusters strongly correspond to the true labels, surpassing even current state-of-the-art baselines. These interpretations highlight scVAG's ability to uncover insightful single-cell structures by distinguishing underlying subpopulations in both datasets more accurately than other approaches.

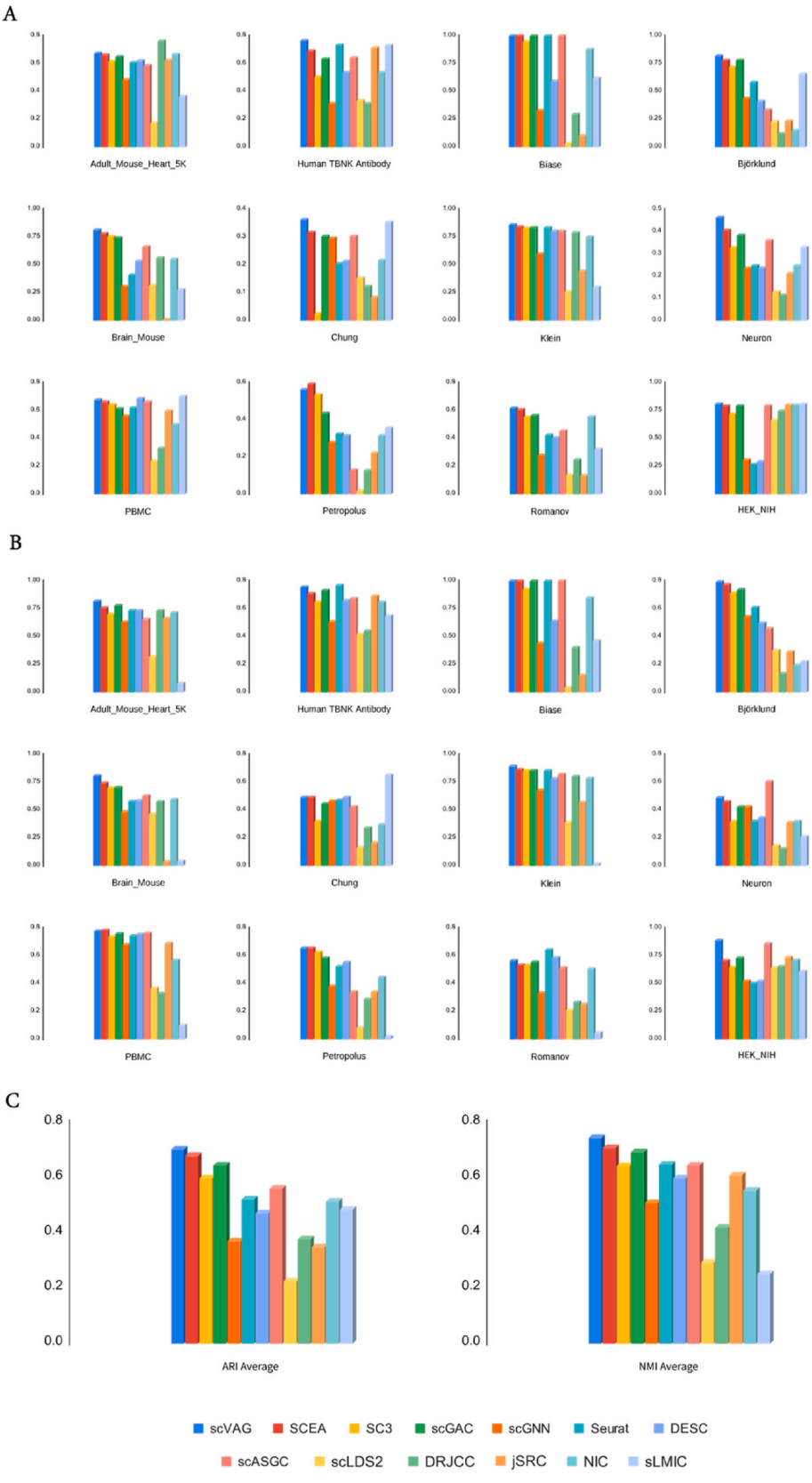
2.5. Statistical analysis

To demonstrate the statistical significance of scVAG's performance improvements over state-of-the-art methods, we conducted rigorous statistical analyses. For each dataset and each evaluation metric (ARI and NMI), we ran each method 5 times to account for potential variability in performance. We then employed two-tailed paired t-tests to compare the mean performance of scVAG against each baseline method across all datasets.

The null hypothesis for each test stated that there was no significant difference in the mean performance (measured by ARI and

Table 2
Summary of synthesis single-cell datasets used to evaluate scVAG.

Datasets #	distributions	# of cells	# of genes	# of clusters
Data1	Unbalanced	1000	2000	4
Data2	Balanced	1000	2000	4
Data3	Balanced	1040	2000	8
Data4	Unbalanced	1000	2000	8
Data5	Unbalanced	2000	2000	8
Data6	Balanced	2080	2000	8
Data7	Unbalanced	2240	2000	16
Data8	Balanced	2240	2000	16



(caption on next page)

Fig. 3. Comparison of clustering performance for scVAG versus baseline methods on 12 single-cell datasets. (A) Adjusted Rand Index (ARI) scores across all datasets. (B) Normalized Mutual Information (NMI) scores across all datasets. (C) Average ARI and NMI scores across datasets. ScVAG consistently achieves higher ARI and NMI scores compared to other methods, demonstrating its superior clustering performance in single-cell analysis.

NMI) between scVAG and the compared method. By using multiple runs, we ensure a more robust statistical comparison that accounts for the inherent variability in the clustering algorithms.

Fig. 7 presents 24 subplots illustrating the variations in ARI and NMI values for scVAG, SCEA, SC3, scGAC, scGNN, Seurat, DESC, and scASGC across all real datasets. The p-values resulting from these t-tests, based on the 5 runs for each method-dataset combination, are indicated on each plot.

We used the following significance levels.

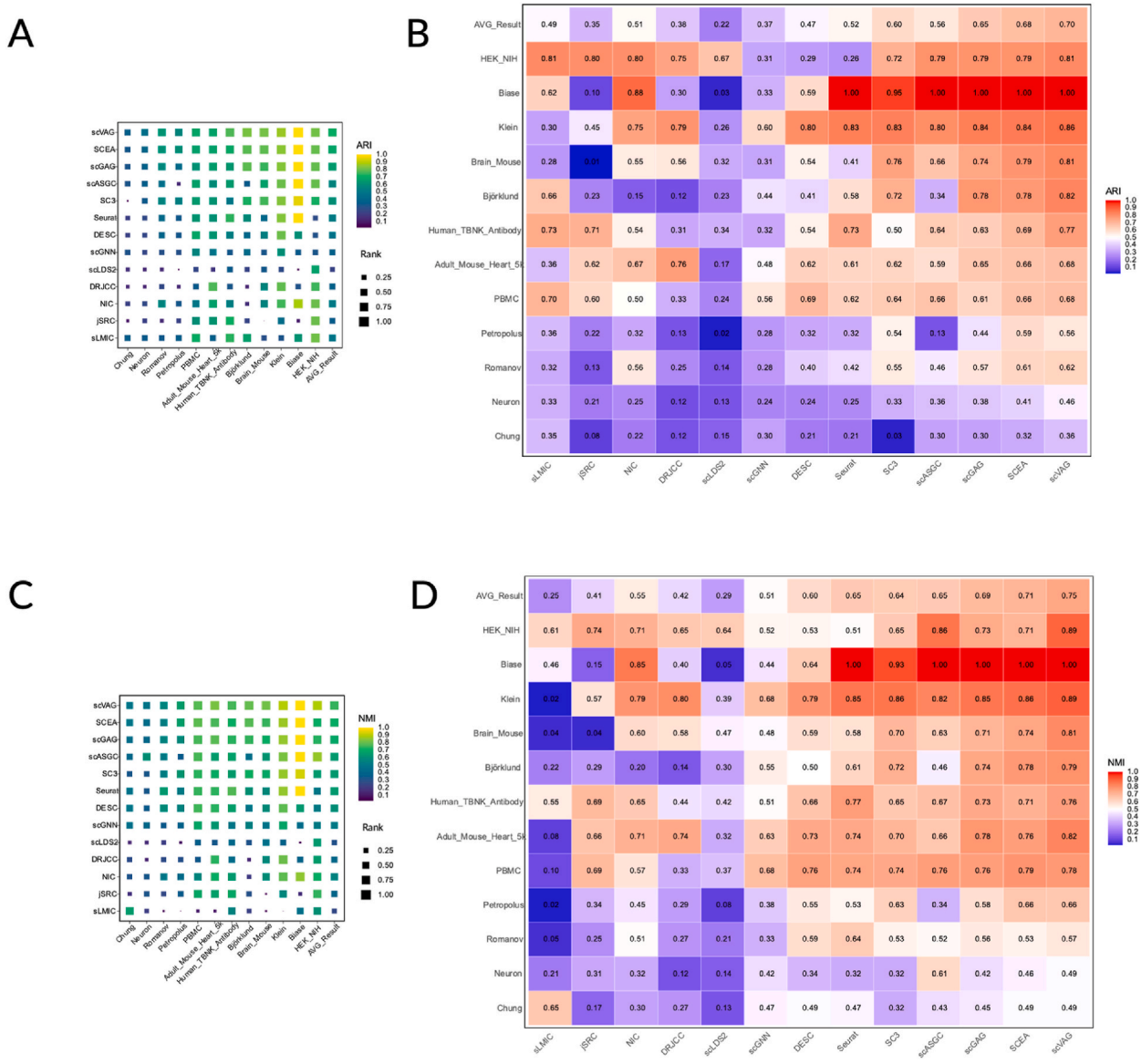


Fig. 4. Comprehensive Robustness Evaluation of scVAG Across Two Metrics. Top row (Panels A and B) presents the Adjusted Rand Index (ARI) evaluation, with dot plots (A) and heatmap (B) demonstrating scVAG's robust performance across datasets compared to baselines. The bottom row (Panels C and D) illustrates the Normalized Mutual Information (NMI) assessment, mirroring the structure of the top row with corresponding dot plots (C) and heatmap (D). Consistent bright colors across all panels highlight the reliable clustering capabilities of scVAG, surpassing other methods on average.

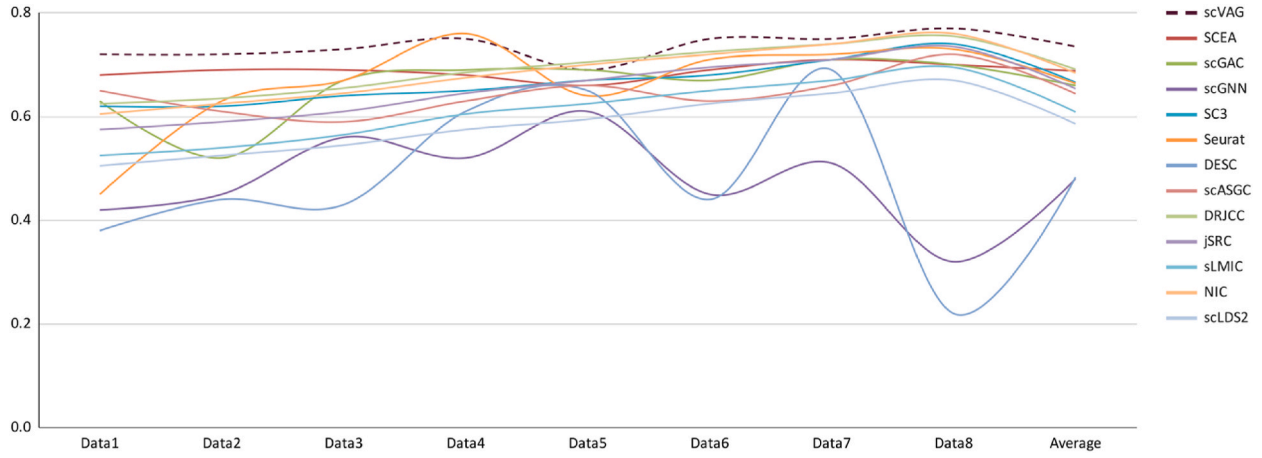


Fig. 5. Comparison of Cosine Similarity scores for scVAG across 8 simulated datasets. Overall, scVAG demonstrates effective performance in measuring cell similarity across diverse simulated datasets.

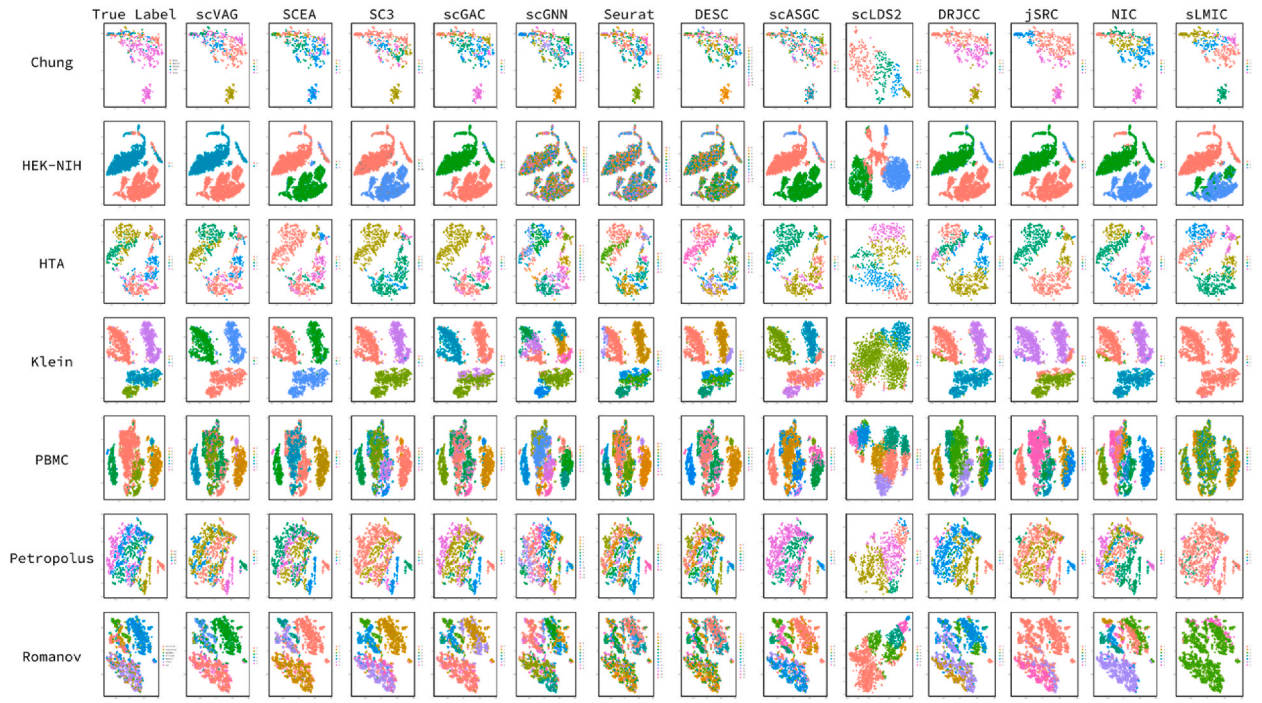


Fig. 6. Visualizations comparing scVAG against baseline models on seven diverse single-cell datasets. The datasets are presented in rows from top to bottom: Chung, HEK-NIH, HTA(Human_TBNK_Antibody), Klein, PBMC, Petropolus, and Romanov. Each row presents results from 14 different methods alongside True labels for direct comparison of clustering performance. Across all datasets, scVAG's clusters consistently show strong agreement with the true cell type labels, accurately capturing intricate patterns in the single-cell data. This comprehensive comparison demonstrates scVAG's advanced ability for accurate and interpretable clustering across a wide range of biological datasets, outperforming current state-of-the-art methods in various single-cell contexts.

- * indicates $p < 0.05$ (significant)
- ** indicates $p < 0.01$ (highly significant)
- *** indicates $p < 0.001$ (very highly significant)
- **** indicates $p < 0.0001$ (extremely highly significant)
- ns indicates $p \geq 0.05$ (not significant)

For the p-value test, a smaller value indicates stronger evidence against the null hypothesis, suggesting a more significant difference

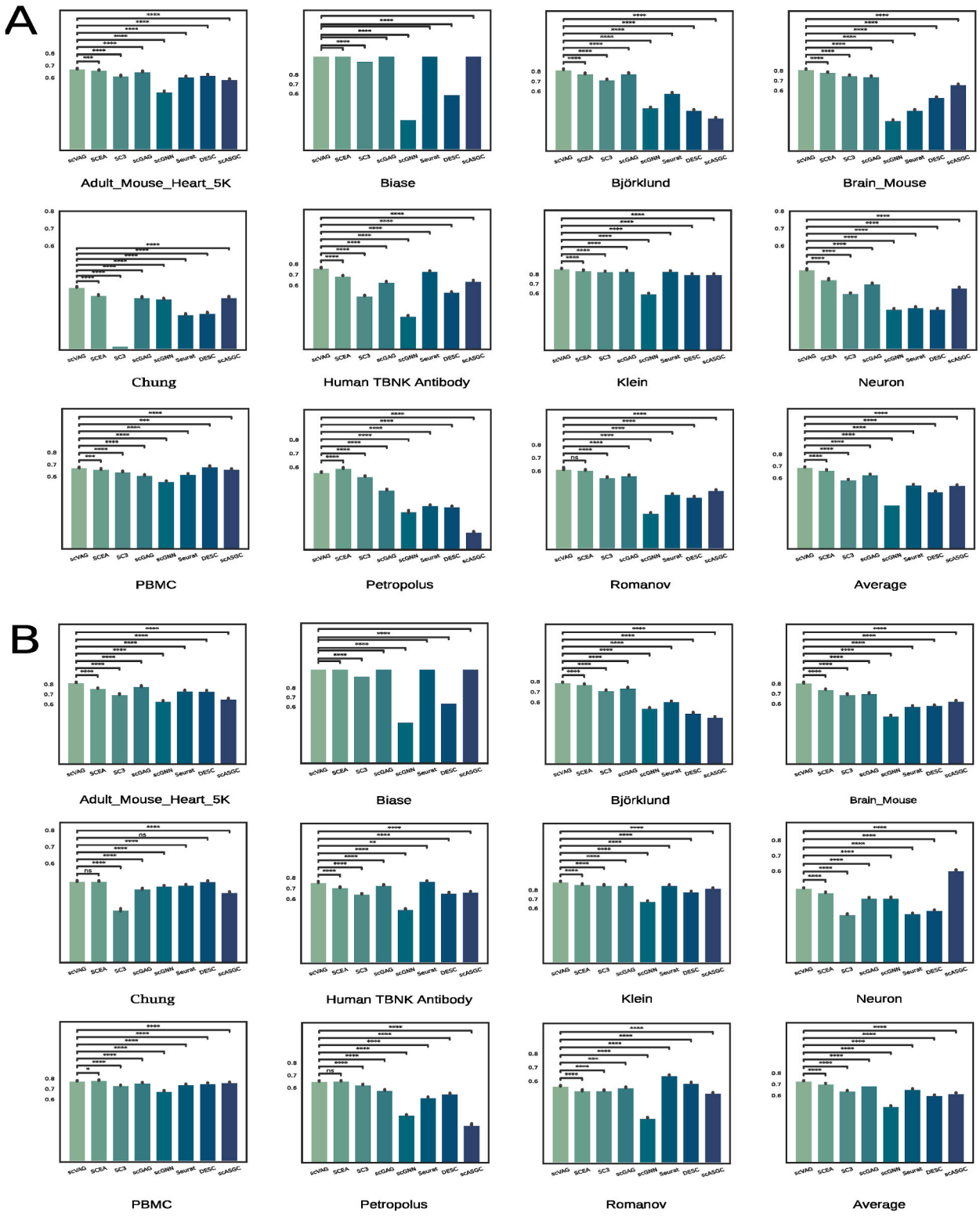


Fig. 7. Statistical analysis of scVAG performance compared to state-of-the-art methods (SCEA, SC3, scGAC, scGNN, Seurat, DESC, and scASGC) across real datasets. (A) includes Adjusted Rand Index (ARI) subplots and (B) Normalized Mutual Information (NMI). Stars indicate the significance levels of the p-values obtained from t-tests comparing scVAG's performance against other methods, with the null hypothesis stating no average improvement by scVAG.

between the methods. As an example, in the subplot for ARI on the Adult_Mouse_Heart_5k dataset (first row, first column), statistical tests indicate that scVAG outperformed all other methods (SCEA, SC3, scGAC, scGNN, DESC, Seurat, and scASGC) with $p < 0.0001$, denoted by ****. This suggests an extremely highly significant improvement across the multiple runs. Across all the test cases, we passed the statistical test in over 90 % of the situations.

2.6. Analysis of Marker genes

To demonstrate the practicality of our model, we used the clustering results of the scVAG to conduct differential expression gene analyses on breast cancer data (GSE75688) [39]. We utilized the DESeq2 package to perform differential expression analysis. DESeq2 applies a negative binomial distribution to model the read counts and employs statistical techniques to identify genes with significant expression changes between the tumor and healthy conditions. The results were ordered by adjusted p-value, and genes with an adjusted p-value less than 0.05 and a positive log2 fold change were selected as differentially expressed. As listed in Table 3, CALML5, CXADR, ID4, TSPAN8, CLDN4, NDST4, KLK5, KRT81 are among the differentially expressed genes with adjusted p-values less than 1e-26.

There is evidence that all of these differentially expressed genes contribute to the development and progression of breast cancer. CALML5, a member of the calmodulin-like protein family, has been shown to have significant activity in primary breast cancers [40]. Loss of CXADR expression has been associated with epithelial-mesenchymal plasticity, a key process in breast cancer [41]. ID4, an inhibitor of differentiation proteins, is linked to a stem-like phenotype and poor prognosis in basal-type breast cancer [42].

Tspan8 is expressed in breast cancer and regulates E-cadherin/catenin signaling and metastasis accompanied by increased circulating extracellular vesicles [43]. CLDN4, a tight junction protein, enhances cell proliferation, migration, tumor growth, and metabolic reprogramming in breast cancer cells [44]. NDST4, a gene encoding an enzyme involved in heparan sulfate biosynthesis, is frequently mutated in breast cancer [45]. KLK5 and KRT81 have been identified as potential serological biomarkers and indicators of aggressive pathological features and poor outcomes in breast cancer [46,47]. The significant enrichment of these genes among the DEGs highlights their importance in the molecular mechanisms driving breast cancer progression and metastasis, further validating the significance of our model.

3. Discussion

In this paper, we introduce scVAG, a novel deep learning framework for single-cell RNA-seq clustering analysis that integrates a Variational-Autoencoder and Graph Attention Autoencoder. Our approach outperforms the current graph clustering technique (scGAC) by replacing restrictive linear assumptions with flexible nonlinear dimensionality reduction specifically adapted to single-cell transcriptional profiles. Extensive comparative evaluations across 20 datasets spanning diverse biological contexts demonstrate scVAG's superior performance. Additionally, scVAG surpasses graph neural network techniques such as SCEA, scGAC, scASGC, and scGNN on all real datasets tested. This demonstrates its superiority as the top graph-based method for single-cell clustering. On average, scVAG achieves an Adjusted Rand Index of 0.693 and a Normalized Mutual Information of 0.733, outperforming previous methods including SC3, Seurat, scGAC, SCEA, scASGC, scGNN, and DESC. Based on statistical tests (t-tests), our model significantly outperformed other baseline models in more than 90 % of cases. Additionally, we evaluated scVAG in terms of noise and imbalanced clustering distributions using 8 challenging synthesis datasets. scVAG demonstrated higher performance than baseline methods across adjusted Rand index, cosine similarity, and normalized mutual information metrics, suggesting potential advantages for single-cell clustering tasks. Visualizations of the learned cellular embeddings further demonstrate scVAG's capacity to recover interpretable biological structures. As shown in t-SNE projections, scVAG embeddings clearly distinguish cell subtypes in a manner consistent with ground truth labels, accurately recovering known cell types.

In summary, the proposed scVAG architecture establishes a robust framework for elucidating cell taxonomies from noisy, high-dimensional transcriptomic data. ScVAG enables sensitive and accurate clustering at a resolution of single-cell types by combining global data distribution learning with local graph-based feature extraction. This method facilitates biological discovery across applications like development, disease modeling, and monitoring therapeutic responses.

Future directions include extending scVAG to integrate multi-omic single cell datasets for further enhanced cell annotation. Rich information beyond gene expression such as chromatin accessibility, DNA methylation, and protein levels will enable more

Table 3
Differentially Expressed Genes Identified by scVAG Clustering Results in Breast Cancer Data (GSE75688), Selected Using DESeq2 Analysis with Adjusted p-Values < 1e-26 and Positive Log2 Fold Change.

ENS Id	Gene Symbol	log2FoldChange	p-value	adjusted p-value
ENSG00000178372	CALML5	7.762030765	4.76E-52	1.19E-48
ENSG00000154639	CXADR	5.062255781	2.29E-41	3.82E-38
ENSG00000138653	NDST4	9.469921971	1.33E-33	1.43E-30
ENSG00000167754	KLK5	11.20203572	1.77E-32	1.78E-29
ENSG00000205426	KRT81	7.751981186	1.63E-31	1.54E-28
ENSG00000172201	ID4	5.972873064	3.80E-31	3.17E-28
ENSG00000127324	TSPAN8	7.31810505	1.57E-29	1.12E-26
ENSG00000189143	CLDN4	2.973165401	9.91E-29	6.20E-26

comprehensive cell state characterization. We will also explore predictive modeling scenarios leveraging scVAG's powerful representations, including imputation of missing cell modalities and projection of differentiation trajectories.

4. Limitations of the study

For the model to run, three components are required. Initially, there is a preprocessing section, followed by the VAE section, and finally the GATE section. Due to this implementation, the model takes longer to complete than some other techniques, particularly traditional methods like Seurat and SC3. Our experiments show that scVAG requires comparable time and memory to other deep learning-based methods (e.g., scGAC, SCEA), but significantly more than faster traditional approaches. For instance, on the HEK-NIH dataset, scVAG took 4 h to run compared to 0.7 h for Seurat, while using 25 GB of memory. Moreover, if the client does not have a GPU, the VAE and the GATE may take even longer to run. This increased computational demand is a trade-off for the improved clustering accuracy and ability to capture complex cellular relationships demonstrated in our evaluations.

5. Material and methods

5.1. Data preprocessing

In order to achieve robust clustering, we implement a targeted pre-processing approach based on data-driven quality control. The expression matrix $d(M'*N')$ consisting of M′ cells and N′ genes is converted to $d(M*N)$, where $M < M'$ and $N < N'$. Cells are evaluated based on detected gene counts. Extremely low counts indicate low mRNA content or poor library preparation. Extremely high mRNA levels are not useful for clustering, as they probably represent housekeeping genes and are not useful for analysis. Furthermore, genes with very low or very high expression across all cells are uninformative and not useful for clustering, as they lack the variability to distinguish cell populations.

Quality thresholds are determined by the 1st and 3rd quartiles of cell and gene detection distributions. Cells and genes outside the interquartile range are filtered out. The resulting refined expression matrix is used as input for the scVAG pipeline, enabling robust graph-based deep learning analysis. In addition, it accelerates the learning process by reducing the size of the input matrix.

5.2. Graph construction and denoising

Constructing a graph representing intercellular connections is vital for cell clustering based on topological relationships. Using Pearson correlation coefficients [48] between gene expression profiles, scVAG first computes a cell-to-cell similarity matrix. According to Eq. (4), the scVAG calculates the correlation between each pair of cells. x and y represent the gene expression profiles of individual cells.

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \tag{4}$$

A gene expression profile is represented as a vector of expression values, where each element represents the level of expression of a specific gene in the cell. After calculation, the Pearson correlation coefficient (r) ranges from −1 to 1.

- r = 1 indicates a perfect positive linear relationship.
- r = −1 indicates a perfect negative linear relationship.
- r = 0 indicates no linear relationship between the variables.

Consequentially, the model defines a threshold based on r to assign edges between nodes (cells). Since gene expression data contain noise, the initial similarity graph can contain misleading edges that distort true cluster assignments. To accurately reflect topological relationships, biological and technical noise must be removed. To address this issue, we used a state-of-the-art network denoising technique known as Network Enhancement (NE). To refine the original network, NE model utilizes multiple local similarity networks. Each local similarity network updates the original network through multiplying by its current state. In contrast to noise, which typically manifests as local inconsistencies, true biological signals are always consistent. As a result, this iterative updating process aims to attenuate noise while enhancing true biological signals. As shown in Eq. (5), mathematically, NE utilizes the following process. Specifically, we perform NE to $S(n.\ n)$ to obtain a re-weighting similarity matrix $E(n.\ n)$ and calculate the denoised similarity matrix $E'(n.\ n)$ as follows:

$$E'(i.\ j) = \begin{cases} S(i.\ j) & if \quad E(i.\ j) >= t \\ 0 & Otherwise \end{cases} \tag{5}$$

where, t is a predefined threshold. If $E(i.\ j)$ is too small (less than t), $S(i.\ j)$ is considered noise and $E'(i.\ j)$ is set to zero. Our model selects the K most similar cells as neighbors for each cell. This process establishes a reliable structure for graph attentional learning and robust cluster resolution.

5.3. Dimensionality reduction with VAE

This paper introduces a Variational-Autoencoder (VAE) framework as the first step in our comprehensive dimensionality reduction approach. The meticulously crafted VAE architecture is designed to unravel the detailed relationships that exist within high-dimensional gene expression data. The encoder section of the VAE comprises multiple dense layers with Rectified Linear Unit (ReLU) activation functions, facilitating the extraction of complex features crucial for subsequent clustering. According to Eq. (6), the encoder function $q_\phi(z|d)$ maps the input gene expression data d to the latent space z through a series of affine transformations and nonlinear activations.

$$h_i = \mathrm{ReLU}(W_i h_{i-1} + b_i)\ i = 1 \ldots L \tag{6}$$

where $h_0 = d_i W_i$ are the weight matrices, b_i are the bias vectors, and L denotes the number of layers. The final layer outputs the mean (μ) and log-variance $(\log \sigma^2)$ of the latent distribution. The decoder module inversely transforms the latent variables z back to the original data space $\widetilde{d}$, attempting to reconstruct the input data. This process is represented as the decoder function $p_\theta(d|z)$, which is a conditional distribution parameterized by θ. The decoder architecture involves affine transformations followed by ReLU activation functions according to Eq. (7):

$$\widetilde{d} = \mathrm{ReLU}(W' z + b') \tag{7}$$

where $\widetilde{d}$ represents the reconstructed data, W' are the weight matrices of the decoder, and b' are the bias vectors. This symmetry achieves an effective balance between feature extraction and error minimization. To prevent overfitting and improve generalization, dropout layers are incorporated during both the encoding and decoding stages. Dropout randomly drops input and hidden units, preventing the model from overly relying on specific features [49]. Additionally, weight decay, a form of L2 regularization, is applied in the Adam optimizer to penalize large weights, discouraging noise fitting [50].

This model undergoes a rigorous training process that consists of 200 epochs with a batch size of 64 samples per iteration. In order to optimize the network parameters, we use the Adam optimizer, a popular choice due to its adaptive learning rates. The loss function used for training is the Mean Absolute Error (MAE) [51], which is particularly well suited to regression tasks like the assessment of reconstruction fidelity. According to Eq. (8), MAE can be mathematically defined as:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{8}$$

where y_i represents the true value and $\widehat{y}_i$ represents the predicted value, and n is the total number of samples. During backpropagation, we use gradient clipping to enhance training stability and prevent issues such as exploding gradients. Gradient clipping constrains gradients to a predefined range during backpropagation. Additionally, we perform periodic evaluations throughout training to ensure that the model is learning the underlying distribution effectively. In summary, the VAE leverages a symmetrical architecture between the encoding and decoding sections to transform single-cell expression into a lower-dimensional latent representation. Dual regularization via dropout and weight decay allows meaningful pattern extraction without overfitting.

5.4. Final dimensionality reduction with GATE

In the final stage of our dimensionality reduction approach, we used the Graph Attention Autoencoder (GATE). GATE demands both the VAE output matrix and the pruned cell graph as inputs. As shown in Fig. 1, the GATE architecture consists of two overlapping graph attention layers [52] and a symmetric decoder. This design enables the capture of intricate cellular relationships within a condensed feature space, crucial for understanding complex biological systems. Eq. (9) describes the graph attention layer in more detail:

$$\mathbf{h}_i' = \sigma\left(\sum_{j \in N_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j \right) \tag{9}$$

where $\mathbf{h}_i'$ is the new feature of cell i. N_i is the set of neighbors of cell i. $\mathbf{h}_j$ is the input feature of cell j. W is a learnable transformation matrix, and σ is a non-linear activation function. α_{ij} is the attention coefficient, representing the importance of cell j to cell i. Based on these attention coefficients, a cell's attributes are shaped by the attributes of its neighboring cells. In Eq. (10), GATE explicitly integrates similarity information into the attention coefficient, which is calculated by transforming the distance between two cells by a Gaussian kernel.

$$e_{ij} = \exp\left(- \left| \boldsymbol{a}_1^T \boldsymbol{W} \boldsymbol{h}_i - \boldsymbol{a}_2^T \boldsymbol{W} \boldsymbol{h}_j \right|^2 \right) \tag{10}$$

where $\mathbf{a}_1$ and $\mathbf{a}_2$ are the learnable weight vectors for cell i and its neighbor j, respectively. As a general rule, the attention coefficient will be normalized by a Softmax function in order to be comparable across different cells, as shown in Eq. (11):

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \tag{11}$$

In addition, Softmax normalization is employed to scale the attention coefficients, enhancing interpretability and comparability. Then the model employs multi-head attention [53] to stabilize the learning process. According to Eq. (12), X independent attention modules are used to jointly learn the representation.

$$\mathbf{h}_i' = \oplus_{x=1}^{X} \sigma\left(\sum_{j \in N_i} {}^{x}_{i}{}_{ij} \mathbf{W}^x \mathbf{h}_j\right) \tag{12}$$

where $\oplus$ is an aggregation operation. The concatenation and averaging function are used as aggregation operation for the first three layers and the last layer of the Autoencoder, respectively. Multi-head attention strategy combines outputs from different attention layers to promote robust learning. This approach allows the model to interpret cellular connectivity data from diverse perspectives.

To constrain the learning process, the model minimizes the Mean Absolute Error (MAE) between the reconstructed and input matrices, ensuring that low-dimensional representations accurately capture interaction patterns.

5.5. Self-optimizing clustering

After training the GATE to generate the cell hidden representation matrix, scVAG leverages this representation for clustering. A simple clustering result can obtain from algorithm like K-means. scVAG employs a self-optimized clustering module to enhance the coherence and interpretability of identified cell groups. The process starts by calculating the similarities between each cell and cluster centroids (Initialized by K-means) using the t-student distribution. After per-cell normalization, a clustering membership matrix $q_{c\times n}$ is created to represent the association of each cell with different clusters. This process showed in Eq. (13):

$$q_{ij} = \frac{\left(1 + \| \mathbf{z}_i - \mathbf{u}_j \|^2\right)^{-1}}{\sum_{k=1}^{c} \left(1 + \| \mathbf{z}_i - \mathbf{u}_k \|^2\right)^{-1}} \tag{13}$$

where $\mathbf{z}_i$ is the embedding of cell i. $\mathbf{u}_j$ is the embedding of cluster center j, and c is the number of cell types. Subsequently, an optimized membership matrix $p_{c\times n}$ is constructed based on $q_{c\times n}$, which is defined as:

$$p_{ij} = \frac{q_{ij}^2 \Big/ \sum_{i=1}^{n} q_{ij}}{\sum_{k=1}^{c} \left(q_{ik}^2 \Big/ \sum_{i=1}^{n} q_{ik}\right)} \tag{14}$$

In this process, the previous matrix is redistributed in order to achieve more deterministic cluster points. The process can be described as follow:

- Normalization of each row (sum equal to 1)
- Normalization of each column (sum equals 1)

As part of our training process, we calculate a silhouette score based on latent representations to monitor the clustering performance for early stopping. This process can be summarized as Eq. (15):

$$s_i = \frac{b_i - a_i}{\max(a_i,\ b_i)} \tag{15}$$

where a_i is the mean distance between cell i and cells in the same cluster, and b_i is the mean distance between cell i and cells in different clusters. A clustering result's silhouette score is the average silhouette score across all cells. As the silhouette score converges, the latent representation of cells and cluster centers will be iteratively updated.

For the optimization part, as showed in Eq. (16), we minimize the Kullback-Leibler Divergence (KLD) [54] between the current and previous membership matrices to monitor the learning process and reinforce clustering patterns.

$$L_c = \sum_{i=1}^{c} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{16}$$

As showed in Eq. (17), the final loss function of scVAG is the sum of MAE for the graph section and KLD for the self-optimized clustering.

$$L = L_r + \gamma L_c \tag{17}$$

where L_r is the reconstruction loss, L_C is the clustering loss and γ is a hyperparameter that balances two losses. This total loss balances

the Autoencoder reconstruction loss (Lr) and clustering loss (Lc) using a weighting hyperparameter (γ). This consolidated framework enhances both the cellular representation and the cluster accuracy. By integrating learned cellular attributes and dynamically optimizing cluster centers, scVAG achieves more reliable and accurate cluster assignments. This enhances stability and accuracy beyond traditional K-means for single-cell clustering.

CRediT authorship contribution statement

Seyedpouria Laghaee: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Morteza Eskandarian:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis. **Mohammadamin Fereidoon:** Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Somayyeh Koohi:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Data availability

All real datasets and the source codes for this study are available at GitHub: https://github.com/pourialaghayee/scVAG.git.

Funding

Authors do not use any source of funding.

Declaration of competing interest

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

Acknowledgements

We acknowledge that portions of this code are derived from the scGAC project and express our gratitude to the scGAC authors for their contributions. Additionally, we would like to thank Mrs. Sara Keshavarz Zadeh and Dr. Mahmood Kalemati (Members of the "EtaDac" lab at Sharif University of Technology) for their valuable guidance and consultation during the preparation of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e40732.

References

[1] W. Chung, H.H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, et al., Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer, Nat. Commun. 8 (2017) 15081, https://doi.org/10.1038/ncomms15081.
[2] M.A. Skinnider, J.W. Squair, L.J. Foster, Evaluating measures of association for single-cell transcriptomics, Nat. Methods 16 (2019) 381–386, https://doi.org/10.1038/s41592-019-0372-4.
[3] X. Ran, L. Tong, W. Chenghao, L. Qi, P. Bo, Z. Jiaying, Heliyon Single-cell data analysis of malignant epithelial cell heterogeneity in lung adenocarcinoma for patient classification and prognosis prediction 9 (2023) e20164, https://doi.org/10.1016/j.heliyon.2023.e20164.
[4] Y. Shi, J. Wan, X. Zhang, Y. Yin, CL-Impute: a contrastive learning-based imputation for dropout single-cell RNA-seq data, Comput. Biol. Med. 164 (2023) 107263, https://doi.org/10.1016/j.compbiomed.2023.107263.
[5] Z. He, Q. Zhou, J. Du, Y. Huang, B. Wu, Z. Xu, et al., Heliyon Integrated single-cell and bulk RNA sequencing reveals CREM is involved in the pathogenesis of ulcerative colitis, Heliyon 10 (2024) e27805, https://doi.org/10.1016/j.heliyon.2024.e27805.
[6] Z. Luo, C. Xu, Z. Zhang, W. Jin, A topology-preserving dimensionality reduction method for single-cell RNA-seq data using graph autoencoder, Sci. Rep. 11 (2021) 1–8, https://doi.org/10.1038/s41598-021-99003-7.
[7] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. Mauck, S. Zheng, A. Butler, et al., Integrated analysis of multimodal single-cell data, Cell 184 (2021) 3573–3587.e29, https://doi.org/10.1016/j.cell.2021.04.048.
[8] X. Li, K. Wang, Y. Lyu, H. Pan, J. Zhang, D. Stambolian, et al., Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis, Nat. Commun. 11 (2020) 2338, https://doi.org/10.1038/s41467-020-15851-3.
[9] V.Y. Kiselev, K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, et al., SC3: consensus clustering of single-cell RNA-seq data, Nat. Methods 14 (2017) 483–486, https://doi.org/10.1038/nmeth.4236.
[10] S. Wang, Y. Zhang, Y. Zhang, W. Wu, L. Ye, Y. Li, et al., scASGC: an adaptive simplified graph convolution model for clustering single-cell RNA-seq data, Comput. Biol. Med. 163 (2023) 107152, https://doi.org/10.1016/j.compbiomed.2023.107152.
[11] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, et al., scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses, Nat. Commun. 12 (2021) 1882, https://doi.org/10.1038/s41467-021-22197-x.

[12] Y. Cheng, X. Ma, scGAC: a graph attentional architecture for clustering single-cell RNA-seq data, Bioinformatics 38 (2022) 2187–2193, https://doi.org/10.1093/bioinformatics/btac099.

[13] S.A.R. Abadi, S.P. Laghaee, S. Koohi, An optimized graph-based structure for single-cell RNA-seq cell-type classification based on non-linear dimension reduction, BMC Genom. 24 (2023) 227, https://doi.org/10.1186/s12864-023-09344-y.

[14] B. Wang, A. Pourshafeie, M. Zitnik, J. Zhu, C.D. Bustamante, S. Batzoglou, et al., Network enhancement as a general method to denoise weighted biological networks, Nat. Commun. 9 (2018) 3108, https://doi.org/10.1038/s41467-018-05469-x.

[15] H. Wang, Z. Liu, X. Ma, Learning consistency and specificity of cells from single-cell multi-omic data, IEEE J Biomed Health Inform 28 (2024) 3134–3145, https://doi.org/10.1109/JBHI.2024.3370868.

[16] W. Wu, W. Zhang, X. Ma, Network-based integrative analysis of single-cell transcriptomic and epigenomic data for cell types, Brief Bioinform 23 (2022), https://doi.org/10.1093/bib/bbab546.

[17] W. Wu, Z. Liu, X. Ma, jSRC: a flexible and accurate joint learning algorithm for clustering of single-cell RNA-sequencing data, Brief Bioinform 22 (2021), https://doi.org/10.1093/bib/bbaa433.

[18] W. Wu, X. Ma, Joint learning dimension reduction and clustering of single-cell RNA-sequencing data, Bioinformatics 36 (2020) 3825–3832, https://doi.org/10.1093/bioinformatics/btaa231.

[19] H. Wang, X. Ma, Learning discriminative and structural samples for rare cell types with deep generative model, Brief Bioinform 23 (2022), https://doi.org/10.1093/bib/bbac317.

[20] D.P. Kingma, M. Welling, An introduction to variational autoencoders, Foundations and Trends® in Machine Learning 12 (2019) 307–392, https://doi.org/10.1561/2200000056.

[21] A. Salehi, H. Davulcu, Graph attention auto-encoders, Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI 2020– Novem (2020) 989–996, https://doi.org/10.1109/ICTAI50040.2020.00154.

[22] R. Qi, A. Ma, Q. Ma, Q. Zou, Clustering and classification methods for single-cell RNA-sequencing data, Brief Bioinform 21 (2020) 1196–1208, https://doi.org/10.1093/bib/bbz062.

[23] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. 19 (2018) 15, https://doi.org/10.1186/s13059-017-1382-0.

[24] J.M. Santos, M. Embrechts, On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification, 2009, pp. 175–184, https://doi.org/10.1007/978-3-642-04277-5_18.

[25] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, J. Mach. Learn. Res. 11 (2010).

[26] A.R. Lahitani, A.E. Permanasari, N.A. Setiawan, Cosine similarity to determine similarity measure: study case in online essay assessment, in: 2016 4th International Conference on Cyber and IT Service Management, IEEE, 2016, pp. 1–6, https://doi.org/10.1109/CITSM.2016.7577578.

[27] A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, et al., Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell 161 (2015) 1187–1201, https://doi.org/10.1016/j.cell.2015.04.044.

[28] R.A. Romanov, A. Zeisel, J. Bakker, F. Girach, A. Hellysaz, R. Tomer, et al., Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes, Nat. Neurosci. 20 (2017) 176–188, https://doi.org/10.1038/nn.4462.

[29] 10xgenomics.com. 4k PBMCs. 10XgenomicsCom 2017. https://support.10xgenomics.com/single-cell-gene-e. (Accessed 10 March 2023). https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k?.

[30] Å.K. Björklund, M. Forkel, S. Picelli, V. Konya, J. Theorell, D. Friberg, et al., The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing, Nat. Immunol. 17 (2016) 451–460, https://doi.org/10.1038/ni.3368.

[31] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S.P. Panula, S. Codeluppi, et al., Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos, Cell 165 (2016) 1012–1026, https://doi.org/10.1016/j.cell.2016.03.023.

[32] Mouse tissue microarray in 3x3 layout with 2 mm edge to edge spacing, 10XgenomicsCom (2023), 2023, pp. 1–3. https://www.10xgenomics.com/datasets/mouse-tissue-microarray-in-3x3-layout-with-2-mm-edge-to-edge-spacing-ffpe-2-standard, 10xgenomics.com (Accessed 14 November 2023).

[33] 110xgenomics.com. SC3_v3_NextGem_Neurons_5K. 10XgenomicsCom (2020), 2020, pp. 1–3. https://cf.10xgenomics.com/samples/cell-exp/6.0.0/. https://cf.10xgenomics.com/samples/cell-exp/6.0.0/SC3_v3_NextGem_DI_Neurons_5K_SC3_v3_NextGem_DI_Neurons_5K/SC3_v3_NextGem_DI_Neurons_5K_SC3_v3_NextGem_DI_Neurons_5K_web_summary.html (accessed October 10, 2023).

[34] F.H. Biase, X. Cao, S. Zhong, Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing, Genome Res. 24 (2014) 1787–1796, https://doi.org/10.1101/gr.177725.114.

[35] 10xgenomics.com, 1k human PBMCs with TotalSeq™-B human TBNK antibody cocktail, 3' LT v3.1. 10xgenomicsCom 2021. https://www.10xgenomics.com/resources/datasets/1-k-human-pbm-cs-with-total-seq-b-human-tbnk-antibody-cocktail-3-lt-v-3-1-3-1-low-6-0-0. (Accessed 12 October 2023).

[36] 10xgenomics.com, 5k Adult mouse heart nuclei isolated with chromium nuclei isolation kit, 10xgenomicsCom (2022), 2022, pp. 1–3. https://www.10xgenomics.com/resources/datasets/5k-adult-mouse-heart-nuclei-isolated-with-chromium-nuclei-isolation-kit-3-1-standard (Accessed 12 November 2023).

[37] 10xgenomics.com. 5k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 Cells, Chromium NextGEM Single Cell 3'. 10XgenomicsCom, 2024, pp. 1–3. https://www.10xgenomics.com/datasets/5k-hgmm-3p-ne. (Accessed 11 August 2024). https://www.10xgenomics.com/datasets/5k-hgmm-3p-nextgem.

[38] L. Zappia, B. Phipson, A. Oshlack, Splatter: simulation of single-cell RNA sequencing data, Genome Biol. 18 (2017) 174, https://doi.org/10.1186/s13059-017-1305-0.

[39] W. Chung, H.H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, et al., Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer, Nat. Commun. 8 (2017) 15081, https://doi.org/10.1038/ncomms15081.

[40] Li Wang, Kinnebrew Liu, Stover Zhang, et al., Identification of alternatively-activated pathways between primary breast cancer and liver metastatic cancer using microarray data, Genes 10 (2019) 753, https://doi.org/10.3390/genes10100753.

[41] A. Nilchian, J. Johansson, A. Ghalali, S.T. Asanin, A. Santiago, O. Rosencrantz, et al., CXADR-mediated formation of an AKT inhibitory signalosome at tight junctions controls epithelial–mesenchymal plasticity in breast cancer, Cancer Res. 79 (2019) 47–60, https://doi.org/10.1158/0008-5472.CAN-18-1742.

[42] S. Donzelli, E. Milano, M. Pruszko, A. Sacconi, S. Masciarelli, I. Iosue, et al., Expression of ID4 protein in breast cancer cells induces reprogramming of tumour-associated macrophages, Breast Cancer Res. 20 (2018) 59, https://doi.org/10.1186/s13058-018-0990-2.

[43] M. Voglstaetter, A.R. Thomsen, J. Nouvel, A. Koch, P. Jank, E.G. Navarro, et al., Tspan8 is expressed in breast cancer and regulates E-cadherin/catenin signalling and metastasis accompanied by increased circulating extracellular vesicles, J. Pathol. 248 (2019) 421–437, https://doi.org/10.1002/path.5281.

[44] Y. Murakami-Nishimagi, K. Sugimoto, M. Kobayashi, K. Tachibana, M. Kojima, M. Okano, et al., Claudin-4-adhesion signaling drives breast cancer metabolism and progression via liver X receptor β, Breast Cancer Res. 25 (2023) 41, https://doi.org/10.1186/s13058-023-01646-z.

[45] G. Encinas, V.Y. Sabelnykova, E.C. de Lyra, M.L. Hirata Katayama, S. Maistro, P.W.M. de Vasconcellos Valle, et al., Somatic mutations in early onset luminal breast cancer, Oncotarget 9 (2018) 22460–22479, https://doi.org/10.18632/oncotarget.25123.

[46] N. Nanashima, K. Horie, T. Yamada, T. Shimizu, S. Tsuchida, Hair keratin KRT81 is expressed in normal and breast cancer cells and contributes to their invasiveness, Oncol. Rep. 37 (2017) 2964–2970, https://doi.org/10.3892/or.2017.5564.

[47] J. Tian, V. Wang, N. Wang, B. Khadang, J. Boudreault, K. Bakdounes, et al., Identification of MFGE8 and KLK5/7 as mediators of breast tumorigenesis and resistance to COX-2 inhibition, Breast Cancer Res. 23 (2021) 23, https://doi.org/10.1186/s13058-021-01401-2.

[48] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson Correlation Coefficient, 2009, pp. 1–4, https://doi.org/10.1007/978-3-642-00296-0_5.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[50] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR, 2019 2019.
[51] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res. 30 (2005), https://doi.org/10.3354/cr030079.
[52] P. Veličković, A. Casanova, P. Liò, G. Cucurull, A. Romero, Y. Bengio, Graph attention networks, in: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018, pp. 1–12, https://doi.org/10.1007/978-3-031-01587-8_7.
[53] A. Veltman, D.W.J. Pulle, R.W. De Doncker, The transformer, Power Systems (2016) 47–82, https://doi.org/10.1007/978-3-319-29409-4_3.
[54] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, 2014, pp. 1–14.