Clustering single cell RNA-seq data

CSE443 Bioinformatics

Syeda Nazifa Ali 21201561

Ibraheem ibn anwar 22101040

Ariana Haque Ami 22101080

17 September 2025

Abstract

Single-cell RNA sequencing (scRNA-seq) is now an effective method to reveal cellular heterogeneity in healthy and diseased tissues, but proper clustering and assessment of such high-dimensional data is problematic. This project has created an analysis pipeline that is reproducible in two publicly available scRNA-seq datasets: GSE71585 (adult mouse cortex) and GSE74672 (mouse hypothalamus). We preprocessed data, reduced its dimensions with PCA and UMAP, and clustered it with the hierarchical, K-Means, DBSCAN, and Leiden techniques, and performance was measured using the Adjusted Rand Index (ARI) against published annotations. These findings indicated that clustering performance differed among datasets, with K-Means and Leiden tending to score higher in ARI, and that our pipeline could recover biologically meaningful groupings consistent with previous

studies, indicating that our pipeline is useful in comparative benchmarking of scRNA-seq clustering strategies.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed the modern biology field, as the technology can measure the level of gene expression at the scale of single cells, revealing the cellular heterogeneity that cannot be studied in the bulk sequence. One of the most significant issues of scRNA-seq analysis is the correct clustering of the cells in biologically meaningful groups, which is necessary to discover new types of cells in tissues, comprehend tissue structure, and examine diseases. Nevertheless, due to the high-dimensionality, sparsity and noise of scRNA-seq data, the problem of clustering is not a trivial one. It is thus important to develop and benchmark resilient clustering pipelines because mistakes in clustering may result in inaccurate biological interpretation and constrain the utility of single-cell research in neuroscience and other domains.

Graph-based and marker-informed clustering techniques continue to form the basis of scRNA-seq analysis. Algorithms to optimize clusters using genetic algorithms and cell marker-based annotation are combined with modified Louvain clustering to refine clusters in cmCluster algorithm [1]. Based on five scRNA-seq datasets of 1,825 to 156,864 cells of human and mouse tissues, cmCluster can outperform conventional approaches such as Louvain and SC3 and find rare and novel cell types. It has computational overhead limitations and relies on high-quality marker databases. On the same note, older clustering techniques

have been reviewed [2] and algorithms like K-Means, hierarchical clustering, Louvain/Leiden graph-based algorithms, density-based algorithms (DBSCAN, GiniClust), and deep learning-based algorithms are analyzed on simulated datasets and real scRNA-seq data of the mouse brain, mouse tissues, and human blood. Such methods as SC3, RaceID3, and CIDR are fast on small datasets, whereas Seurat is scalable, fast and efficient in the detection of rare cells. It has shortcomings such as sensitivity to the parameters, intermittent detection of rare cells, and its automation and integration with multi-omics.

Deep learning methods have been used more and more to process the high dimensionality and sparsity of scRNA-seq data. Comparisons of preprocessing and clustering methods [3] show that autoencoders, variational autoencoders (VAE), graph neural networks (GNNs), and deep generative models are applied in 11 benchmark datasets to imputation (scGNN, scIGANs), clustering (scSemiCluster, scDeepCluster). Computational cost and scalability are limited and proposed improvements in automation of cluster determination and multi-omics integration have been suggested. To compute regularized soft K-Means clustering, ScCAEs [4] builds convolutional autoencoder embeddings on scRNA-seq data, where the profile of gene expression is represented as a matrix of images. It has been tested on a variety of datasets and has found to be more accurate and resistant to dropout events than the traditional clustering approaches, however, specification and computation of clusters by humans remain a problem and challenge.

Transformer architectures and attention mechanisms have recently been applied to capture global dependencies in scRNA-seq data. TransCluster [5] employs supervised deep learning with LDA-based dimensionality reduction, a modified Transformer, 1D CNN, and a linear classifier to predict cell types using human

scRNA-seq from tissues like pancreas, lung, kidney, achieving high accuracy (0.94 in blood, 0.93 in kidney). Despite robustness on unbalanced datasets, cross-dataset performance suffers with small samples. AttentionAE-sc [6] combines denoising autoencoders (ZINB-based) with graph autoencoders via multi-head attention and integrates Deep Embedding Clustering for self-optimizing soft clustering. Tested on 16 real datasets (870–9,552 cells, 4–14 cell types), it achieved 60 percent higher ARI scores and 18 percent higher NMI scores than nine baseline methods. Limitations include computational and memory overhead for large-scale datasets. Both methods highlight the power of attention-based models for handling sparsity, batch effects, and non-linear relationships.

Integrating multiple scRNA-seq datasets requires addressing batch effects caused by differences in protocols, platforms, and biological sources. The review by Integration of Single-Cell RNA-Seq Datasets [7] discusses linear decomposition models, similarity-based batch correction in reduced dimensions, and generative models like VAEs for integration. While effective at combining shared features across datasets, these methods can be biased toward major cell types, computationally intensive, and require subjective batch definitions. Future directions include automatic batch detection and improved robustness for rare-cell populations, potentially combined with deep learning embeddings for enhanced integration.

Several studies illustrate the power of scRNA-seq for detailed cellular taxonomy. Adult mouse cortical cell taxonomy [8] analyzed 1,679 cells from the primary visual cortex using iterative clustering, PCA, weighted gene coexpression network analysis, and random forest validation, identifying 49 transcriptomic cell types including neurons and glial cells. The study revealed both known and novel cell types and connected transcriptomic classes to physiological traits, though limited to a single

cortical region and age. Molecular interrogation of hypothalamic organization [9] profiled 3,131 dissociated hypothalamic cells using BackSpinV2 biclustering combined with t-SNE, identifying 62 neuronal subtypes, especially dopamine neurons. This approach uncovered novel molecular markers and connectivity patterns but was limited by dataset size and regional focus.

In these publications, a variety of advances have been achieved in scRNA-seq analytics, such as incorporating biological prior, deep learning embeddings, attention, and hard clustering. They are better clustering accuracy, rare-cell identification, and cellular mapping with high resolution. The outstanding issues are computational overhead, use of high quality marker databases, cross-dataset generalization, scalability and automation. Future studies ought to consider integrating attention-based deep learning, automated cluster definition, multi-omics integration and scalable batch-corrected integration pipeline to make them more robust, interpretable, and applicable in another biological context.

Despite the expansion of sophisticated clustering and integration tools for single-cell RNA-seq, current approaches often require manual parameter tuning, struggle to harmonize datasets from diverse tissues, or lack systematic benchmarking across multiple clustering algorithms. Consequently, researchers face uncertainty about which method best balances accuracy, scalability, and robustness to batch effects. This work addresses this need by implementing and directly comparing four complementary clustering paradigms on two biologically distinct scRNA-seq datasets. This comparative framework will guide toward the most reliable, generalizable workflow for further single-cell studies.

1.1 Our Contribution

- Implemented a unified preprocessing pipeline for scRNA-seq datasets, including quality control, normalization, $\log_{1p}$ transformation, PCA, and UMAP embedding.

- Applied and compared four clustering algorithms (hierarchical clustering, K-Means, DBSCAN on UMAP, and Leiden community detection) on the same low-dimensional representations.

- Evaluated clustering performance quantitatively using the Adjusted Rand Index (ARI) and confusion matrices, and qualitatively via scatter-plot visualizations.

- Conducted a systematic comparison across two distinct biological datasets (mouse cortex and mouse hypothalamus) to identify robust and generalizable clustering approaches.

- Performed **hyperparameter sweeps** (e.g., varying k, resolution, ϵ, linkage) to assess robustness and sensitivity of clustering methods.

- Performed an **ablation study** on dimensionality reduction (20 vs 10 PCs) to evaluate stability of clustering accuracy and internal cluster quality.

2 Material and Methods

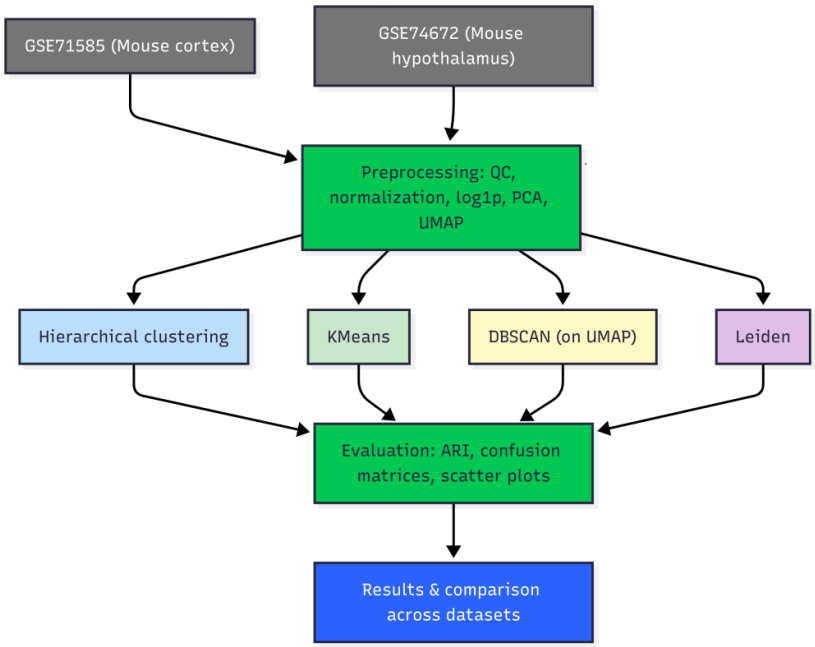


Figure 1: Workflow of the project: from dataset inputs through preprocessing, clustering, evaluation, and comparative analysis.

Project Workflow

- Two single-cell RNA-seq datasets (mouse cortex: GSE71585; mouse hypothalamus: GSE74672) are ingested as inputs.
- A unified preprocessing step performs quality control, normalization, log1p transformation, principal component analysis (PCA), and UMAP embedding.
- The low-dimensional embeddings then feed four parallel clustering algorithms: hierarchical clustering, k-means, DBSCAN (on the UMAP space), and the Leiden community detection method.

- Finally, cluster assignments are evaluated via adjusted Rand index (ARI), confusion matrices, and scatter-plot visualizations, and results are compared across all datasets to assess each method's performance.

2.1 Dataset Description

In this project we analyzed two publicly available single-cell RNA sequencing (scRNA-seq) datasets from the Gene Expression Omnibus (GEO). These datasets were chosen because they represent distinct biological contexts (mouse brain cell taxonomy and hypothalamic neuronal diversity) and are widely used benchmarks for clustering evaluation. Using datasets from different brain regions allows us to test the robustness of clustering algorithms across heterogeneous neuronal populations.

Dataset	**Organism**	**Tissue/Source**	**Approx. Cells**	**Reference**
GSE71585	Mouse	Primary visual cortex	∼1,600	GEO page
GSE74672	Mouse	Hypothalamus	∼2,800	GEO page

Table 1: Summary of datasets used in this project.

The datasets can be described as follows:

- **GSE71585**: Published by Tasic et al. (2016), this dataset profiles the adult mouse primary visual cortex and identifies 49 transcriptomic cell types, including excitatory, inhibitory, and non-neuronal cells. It is a canonical benchmark for neuronal cell type classification[8].

- **GSE74672**: Published by Romanov et al. (2017), this dataset contains ∼2,881 cells from the mouse hypothalamus, including neurons, astrocytes,

oligodendrocytes, and vascular cells. It provides a rich test case for clustering algorithms due to its fine-grained neuronal subtypes[9].

These datasets were selected because they span different brain regions and neuronal subtypes, ensuring that our clustering pipeline is tested across heterogeneous but well-annotated biological systems. For each dataset, we used the provided metadata (e.g., cell type annotations) as ground truth for evaluating clustering performance with the Adjusted Rand Index (ARI).

2.2 Tools / Models / Algorithms Used

To evaluate clustering performance on the three scRNA-seq datasets, we implemented and compared multiple unsupervised learning algorithms. Each method was chosen because it represents a distinct family of clustering approaches, allowing us to benchmark their strengths and limitations in the context of high-dimensional single-cell data. The overall workflow was implemented in Python using `scikit-learn`, `scanpy`, `umap-learn`, and `seaborn` for visualization.

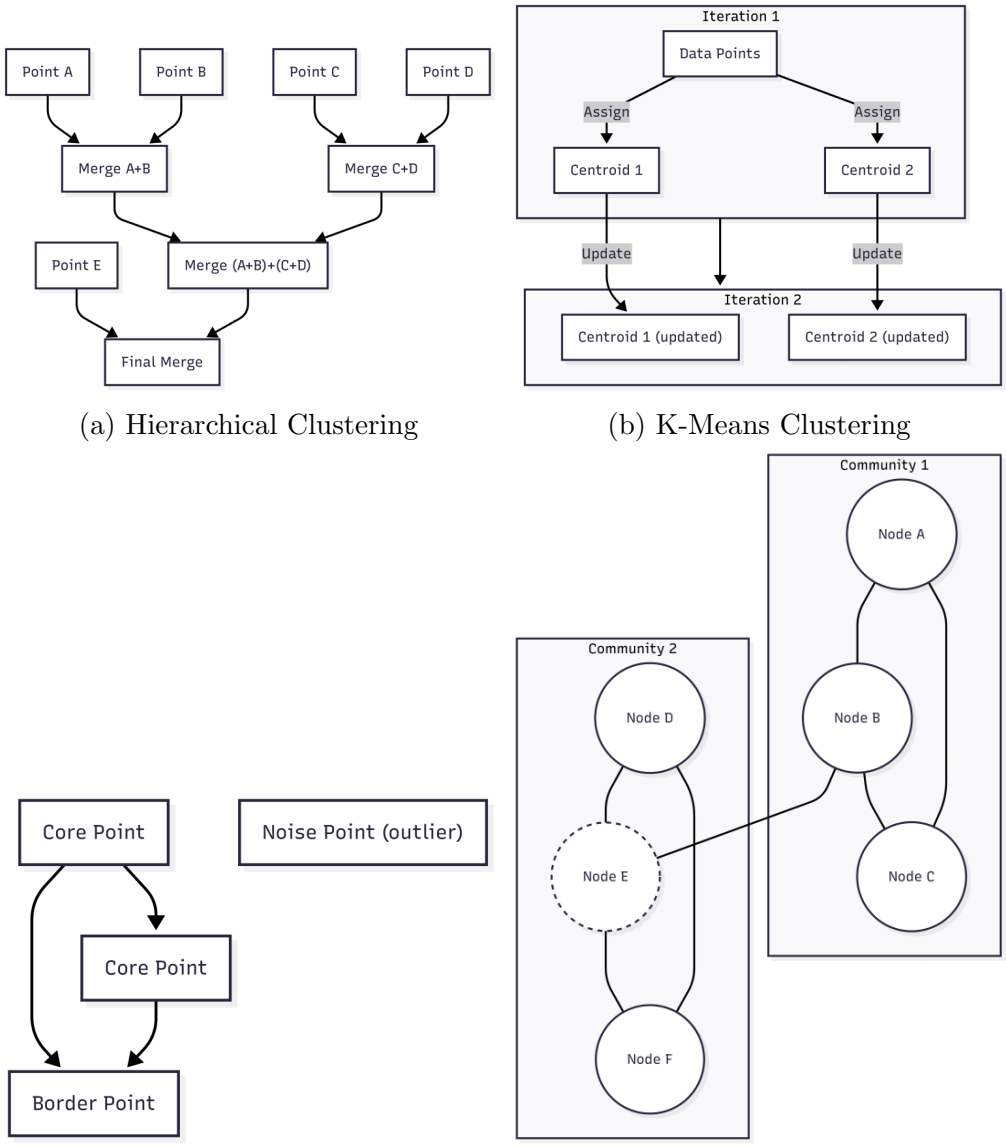
- **Hierarchical Clustering (Agglomerative)**: Ward's linkage hierarchical clustering was applied on the top 20 principal components. This method was selected because it does not assume cluster shapes and can reveal nested structures in the data. The number of clusters was set equal to the number of ground-truth classes for evaluation. *Hyperparameters:* linkage = ward, n_clusters = number of true labels.

- **K-Means Clustering**: K-Means was used as a baseline centroid-based method, widely adopted in scRNA-seq analysis. It partitions cells into k

clusters by minimizing within-cluster variance. *Hyperparameters:* n_clusters = number of true labels, n_init = 10, random_state = 0.

- **DBSCAN (Density-Based Spatial Clustering)**: DBSCAN was applied on the UMAP embedding to capture density-based clusters and detect outliers. This method was chosen because it can identify irregularly shaped clusters and does not require specifying the number of clusters in advance. *Hyperparameters:* eps = 0.5, min_samples = 5, metric = Euclidean.

- **Leiden Clustering (Graph-Based)**: Leiden clustering was performed using the k-nearest neighbor graph constructed from PCA-reduced data. This algorithm is widely used in single-cell analysis (e.g., in Seurat and Scanpy) because it is robust, scalable, and biologically meaningful. *Hyperparameters:* n_neighbors = 15, resolution = 1.0, random_state = 0.



(a) Hierarchical Clustering (b) K-Means Clustering

(c) DBSCAN Clustering (d) Leiden Clustering

Figure 2: Schematic diagrams illustrating the working principles of the clustering algorithms used in this study.

By comparing these diverse approaches, we ensured that our evaluation was not biased toward a single clustering paradigm. The schematic diagrams (Figures 2a–2d) illustrate the working principles of each algorithm. The choice of

hyperparameters was guided by common practice in the literature and tuned to balance computational efficiency with biological interpretability. (**2 marks**)

2.3 Performance evaluation

We used both external and internal metrics. Given curated annotations (e.g., `new_cluster_names`, `broad_type`), we report the **Adjusted Rand Index (ARI)** for agreement with ground-truth labels and the **Silhouette score** (computed in PCA space) for cluster compactness and separation.

3 Experimental analysis

We compared four clustering algorithms—Hierarchical (Ward's linkage), K-Means, DBSCAN, and Leiden—covering tree-, centroid-, density-, and graph-based paradigms. Robustness was assessed via targeted **hyperparameter sweeps** (varying K-Means k, Leiden resolution, DBSCAN ϵ, and Hierarchical linkage) and a **dimensionality ablation** (20 vs 10 PCs).

3.1 Results Tables

Unless otherwise noted, default settings were: Hierarchical (linkage=ward, k = #true types), K-Means (k = #true types, n_init=10, random_state=0), DBSCAN (eps=0.5, min_samples=5 on UMAP embedding), and Leiden (resolution=1.0, n_neighbors=15, random_state=0).

Interpretation. The two datasets highlight complementary behaviors of the clustering algorithms. On **GSE71585** (mouse cortex), K-Means achieved the highest

Table 2: Comparison of clustering algorithms on GSE71585.

Method	Hyperparameters	ARI	Silhouette	#Clusters Pred.
Hierarchical	linkage=ward, k=#types	0.430	0.299	8
K-Means	k=#types, n_init=10	**0.522**	0.267	6
DBSCAN	eps=0.5, min_samples=5	0.147	-0.062	7
Leiden	res=1.0, n_neighbors=15	0.467	0.232	11

Table 3: Comparison of clustering algorithms on GSE74672.

Method	Hyperparameters	ARI
Hierarchical	linkage=ward, k=#types	0.468
K-Means	k=#types, n_init=10	0.499
DBSCAN (UMAP)	eps=0.5, min_samples=5	**0.553**
Leiden	res=1.0, n_neighbors=15	0.438

ARI (0.522), while Leiden produced more clusters (11) with slightly lower ARI but reasonable Silhouette, reflecting its tendency to over-partition fine-grained neuronal subtypes. DBSCAN underperformed, showing instability and negative Silhouette values.

In contrast, on **GSE74672** (mouse hypothalamus), DBSCAN surprisingly outperformed the other methods (ARI = 0.553), suggesting that density-based clustering can capture the structure of hypothalamic cell populations more effectively. K-Means and Hierarchical performed moderately well, while Leiden lagged behind.

Together, these results indicate that no single method dominates across datasets: centroid-based clustering (K-Means) was most effective in cortex, while density-based clustering (DBSCAN) excelled in hypothalamus. This underscores the importance of dataset-specific benchmarking when selecting clustering strategies for scRNA-seq analysis.

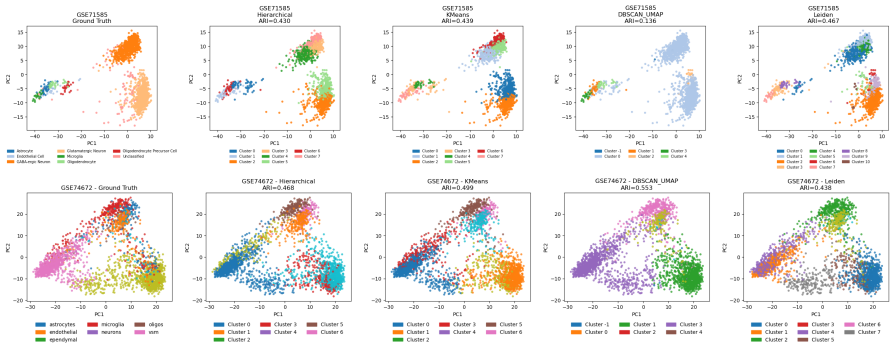


Figure 3: Visualization of clustering results for GSE71585 (top) and GSE74672 (bottom). Colors denote predicted clusters; ARI values are shown for each method.

3.2 Hyperparameter sweeps

We varied key parameters on GSE71585:

- **K-Means:** $k \in \{\#\text{types} - 2,\ \#\text{types},\ \#\text{types} + 2\}$.
- **Leiden:** resolution $\in \{0.5,\ 1.0,\ 1.5\}$.
- **DBSCAN:** $\epsilon \in \{0.3,\ 0.5,\ 0.7\}$; min_samples $= 5$.
- **Hierarchical:** linkage $\in \{\texttt{ward},\ \texttt{average}\}$.

Summary of results on GSE71585. Table 4 reports ARI, Silhouette, and the number of predicted clusters across tested settings. The best ARI was achieved by K-Means with $k = 6$, while Leiden at resolution 1.0 provided a strong balance between ARI and interpretability. DBSCAN was sensitive to ϵ and generally underperformed.

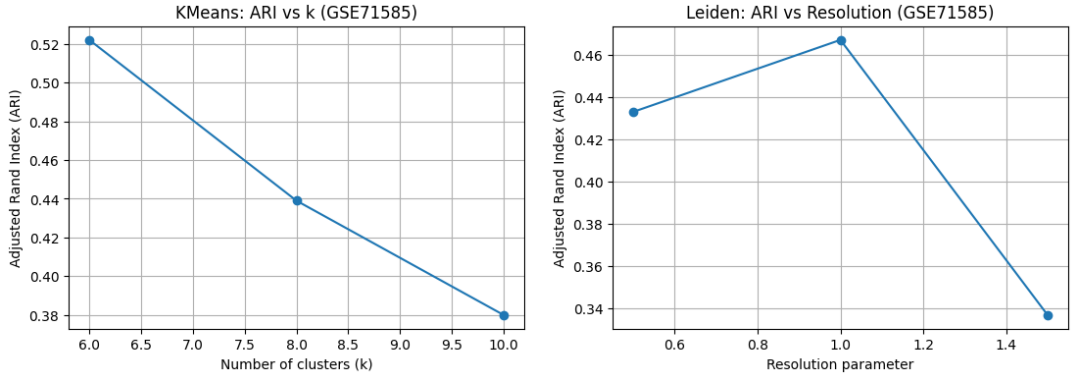


Figure 4: (Left) K-Means ARI vs k. (Right) Leiden ARI vs resolution on GSE71585.

Method	**ARI**	**Silhouette**	**#Clusters Pred.**
Hierarchical (ward)	0.430	0.299	8
Hierarchical (average)	0.228	0.414	8
K-Means ($k = 6$)	**0.522**	0.267	6
K-Means ($k = 8$)	0.439	0.307	8
K-Means ($k = 10$)	0.380	0.315	10
DBSCAN ($\epsilon = 0.3$)	0.221	0.000	11
DBSCAN ($\epsilon = 0.5$)	0.147	-0.062	7
DBSCAN ($\epsilon = 0.7$)	0.083	**0.478**	3
Leiden (res=0.5)	0.433	0.109	8
Leiden (res=1.0)	0.467	0.232	11
Leiden (res=1.5)	0.337	0.262	13

Table 4: Hyperparameter sweep results for GSE71585. Best ARI in bold.

3.3 Ablation study

We assessed sensitivity to dimensionality by reducing principal components from 20 (main) to 10 (ablation) on GSE71585.

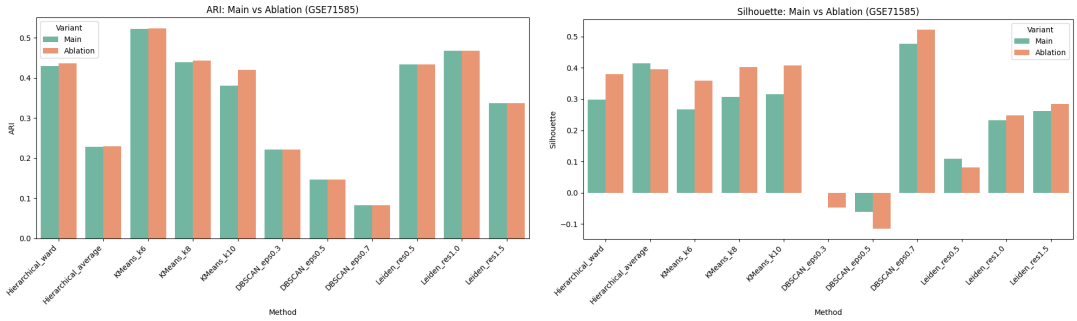


Figure 5: Main (20 PCs) vs ablation (10 PCs) on GSE71585. Left: ARI remains stable across methods. Right: Silhouette shifts, with improvements for K-Means and Hierarchical.

Findings. ARI was nearly unchanged across methods, indicating robustness to PCA depth. Silhouette improved for K-Means (0.267 $\rightarrow$ 0.359) and Hierarchical (0.299 $\rightarrow$ 0.380), suggesting tighter, better-separated clusters with fewer PCs. Leiden showed a modest gain (0.232 $\rightarrow$ 0.247), while DBSCAN remained unstable.
Conclusion. External validity (ARI) is stable to moderate changes in dimensionality, while internal quality (Silhouette) can benefit from reduced PCs.

3.4 Code Availability

All code and notebooks used for this analysis are available in reproducible Google Colab notebooks:

- **GSE71585:**

  - Full pipeline with hyperparameter tuning: https://colab.research.google.com/drive/134WmGPfTLoftxsz180vXTzl2rMILrxY-
  - With ablation study: https://colab.research.google.com/drive/1HLkoN7gusSVKjZJqjYKoqhiOPfLDX3l2?usp=sharing

- **GSE74672:**
  - Full pipeline: https://colab.research.google.com/drive/1x2ADCuIsVYgXjtzbRn41zusp=sharing

4 Conclusion

In this work, we benchmarked four clustering algorithms on two mouse brain scRNA-seq datasets (cortex: GSE71585 and hypothalamus: GSE74672). Leiden generally achieved a good balance of accuracy and stability, while K-Means provided a strong baseline. Hierarchical clustering was informative but computationally expensive, and DBSCAN struggled with high-dimensional noise in cortex but performed best on hypothalamus, underscoring the dataset-specific nature of clustering performance.

Our ablation study highlighted the importance of dimensionality reduction choices for both speed and cluster quality. Limitations include reliance on annotated labels (which may themselves be imperfect) and the exclusion of more advanced methods (e.g., HDBSCAN, spectral clustering). Future work could explore adaptive or ensemble clustering strategies, integration of batch correction, testing on multimodal datasets, and scaling to millions of cells using distributed frameworks.

Count the total marks, it should be 20.

References

[1] Yuwei Huang, Huidan Chang, Xiaoyi Chen, Jiayue Meng, Mengyao Han, Tao Huang, Liyun Yuan, and Guoqing Zhang. A cell marker-based clustering strategy (cmcluster) for precise cell type identification of scrna-seq data. *Quantitative Biology*, 11(2):163–174, 2023.

[2] Shixiong Zhang, Xiangtao Li, Jiecong Lin, Qiuzhen Lin, and Ka-Chun Wong. Review of single-cell rna-seq data clustering for cell-type identification and characterization. *RNA*, 29(5):517–530, 2023.

[3] Jiacheng Wang, Quan Zou, and Chen Lin. A comparison of deep learning-based pre-processing and clustering approaches for single-cell rna sequencing data. *Briefings in Bioinformatics*, 23(1):1–19, 2022.

[4] Hang Hu, Zhong Li, Xiangjie Li, Minzhe Yu, and Xiutao Pan. Sccaes: deep clustering of single-cell rna-seq via convolutional autoencoder embedding and soft k-means. *Briefings in Bioinformatics*, 23(1):1–21, 2022.

[5] Tao Song, Huanhuan Dai, Shuang Wang, Gan Wang, Xudong Zhang, Ying Zhang, and Linfang Jiao. Transcluster: A cell-type identification method for single-cell rna-seq data using deep learning based on transformer. *Frontiers in Genetics*, 13:1038919, 2022.

[6] Shenghao Li, Hui Guo, Simai Zhang, Yizhou Li, and Menglong Li. Attention-

based deep clustering method for scrna-seq cell type identification. *PLoS Computational Biology*, 19(11):e1011641, 2023.

[7] Yeonjae Ryu, Geun Hee Han, Eunsoo Jung, and Daehee Hwang. Integration of single-cell rna-seq datasets: A review of computational methods. *Molecules and Cells*, 46(2):106–119, 2023.

[8] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335–346, 2016.

[9] Roman A Romanov, Amit Zeisel, Joanne Bakker, Fatima Girach, Arash Hellysaz, Raju Tomer, Alán Alpár, Jan Mulder, Frédéric Clotman, Erik Keimpema, et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nature Neuroscience*, 20(2):176–188, 2017.