Autoencoder-Based Nonlinear Dimension Reduction for Single-Cell RNA-Seq Data: A Comparative Study of t-SNE and UMAP

Xinqi Yang[1], Tianyun Long[1] and Jianjuan Liang[1,2,*]

[1]*Department of Statistics, Beijing Normal-Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, China*

[2]*Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal-Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, China*

Abstract: This paper proposes using an Autoencoder (AE) prior to t-SNE or UMAP visualization for scRNA-seq data. Direct application of t-SNE/UMAP to the raw, sparse expression matrix often yields unstable, poorly separated clusters. To address this, the framework first employs an AE to learn a denoised, compact latent representation. Subsequent t-SNE or UMAP embedding of this latent space produces more robust visualizations with enhanced cluster consistency and structural separability. A real-data-based comparison shows that, when using the same AE-derived latent space, UMAP outperforms t-SNE. It achieves better cluster cohesion, stronger global structure preservation, greater robustness to initialization and data perturbation, and lower computational cost. Statistical validation via a projection *F*-test confirms that clusters in the AE latent space exhibit significant between-group mean differences, quantifying the observed visual improvement. The study concludes that AE-based representation learning creates an effective input space for nonlinear embedding, with the AE-UMAP pipeline emerging as a particularly stable and efficient choice for scRNA-seq exploratory analysis.

Purpose: This study aims to investigate the effectiveness of AE based latent representations in enhancing nonlinear dimension reduction methods, namely t-SNE and UMAP, for single-cell gene expression data analysis. The performance of AE-based UMAP and AE-based t-SNE is systematically evaluated from multiple perspectives, including visualization quality, clustering consistency, structural preservation, and robustness.

Methods: This paper constructs a two-step dimension reduction framework for single-cell gene expression data analysis. First, an AE is employed to compress high-dimensional, sparse, and noisy gene expression data into a low-dimensional latent representation. Subsequently, t-SNE and UMAP are applied to the learned AE latent space for nonlinear embedding and visualization. The performance of different methods is systematically evaluated under multiple experimental conditions using clustering consistency metrics, structure preservation measures, and a project*ed F*-test.

Results: Experimental results indicate that directly applying t-SNE or UMAP to the original expression data fails to stably recover meaningful clustering structures, whereas nonlinear dimension reduction performed on AE latent representations substantially improves visualization quality and clustering stability. Within the same latent space, t-SNE and UMAP exhibit comparable performance in terms of clustering accuracy; however, UMAP demonstrates superior performance with respect to cluster compactness, global structure preservation, stability across repeated experiments, and computational efficiency. Statistical testing further confirms the significance of between cluster differences in the AE latent space.

Contribution: This study systematically reveals the critical role of AE latent representations in stabilizing nonlinear dimension reduction for single cell data and provides a quantitative comparison between t-SNE and UMAP within a unified latent space. The results demonstrate that UMAP applied to AE latent representations achieves superior performance in terms of visualization stability and computational efficiency, offering a more robust two step dimension reduction strategy for exploratory analysis of high dimensional single cell data.

Keywords: scRNA-seq data, Autoencoder, t-SNE, UMAP, Dimension Reduction, Visualization.

1. INTRODUCTION

ScRNA-seq enables transcriptomic profiling at single-cell resolution and has become a fundamental tool for characterizing cellular heterogeneity, identifying cell types, and reconstructing developmental trajectories [1,2]. However, scRNA-seq data are typically high dimensional, sparse, and affected by substantial technical noise, which makes direct visualization and clustering in the original expression space highly challenging [3,4]. Consequently, effective dimension reduction is a critical prerequisite for extracting meaningful biological structure prior to downstream analysis.

Traditional linear dimension reduction methods, such as PCA *(principal component analysis)*, compress data by maximizing global variance and can partially alleviate the curse of dimensionality [**5**]. Nevertheless, scRNA-seq data often reside on complex nonlinear manifolds, limiting the ability of PCA to capture local neighborhood relationships and nonlinear structure [6]. In recent years, nonlinear dimension reduction techniques, particularly t-SNE [6] and UMAP [7], have become the dominant approaches for single-cell visualization. These methods aim to preserve local relationships in low-dimensional embeddings and have been widely adopted in popular analysis frameworks.

Despite their widespread use, previous studies have shown that directly applying t-SNE or UMAP to raw gene expression matrices often leads to unstable

*Address correspondence to this author at the Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal-Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, China; E-mail: jiajuanliang@bnbu.edu.cn

embeddings that are sensitive to random initialization, parameter selection, sample size, and data noise [8,9]. In highly sparse and noisy scRNA-seq settings, such sensitivity can result in ambiguous cluster boundaries, distorted global geometry, and poor reproducibility, limiting the interpretability of visualization results.

With the rapid development of deep learning, AE has been increasingly applied to single-cell data analysis for representation learning and denoising [10-13]. By learning compact and smooth latent representations through nonlinear encoding and decoding, AE can effectively reduce technical noise and redundancy while preserving dominant structural patterns in the data. Previous studies have demonstrated the effectiveness of AE in clustering, denoising, and feature extraction tasks. However, in most existing work, AE is treated as an independent preprocessing or dimension reduction tool, and its latent representations are rarely evaluated as a unified input space for subsequent nonlinear visualization methods.

Notably, although several studies have combined AE with t-SNE or UMAP, systematic quantitative comparisons between these two nonlinear methods within the same AE latent space remain limited [14-17]. Existing evaluations often rely on visual inspection or a small number of metrics, while robustness to noise, initialization, and subsampling, as well as statistical validation of cluster separability, are rarely examined in a unified framework.

Statistically, the integration of autoencoder-based latent representations with nonlinear embeddings transcends mere visualization improvement by establishing a more robust and hypothesis-testable framework for single-cell data analysis. Unlike prior studies that primarily focus on visual separability, our approach explicitly quantifies the enhancement in cluster discriminability through rigorous statistical validation, notably employing the projection-based F-tests to assess between-cluster mean differences in the latent space. This methodological contribution shifts the emphasis from qualitative visual assessment to statistically grounded inference, allowing for objective evaluation of cluster significance and stability. By embedding nonlinear methods within a denoised and structurally coherent latent space, we not only improve visualization reproducibility but also provide a statistically consistent input domain that enhances the reliability of downstream clustering and comparative analyses. Thus, this study articulates a clear statistical advancement: it transforms the latent representation into a stabilized statistical manifold on which nonlinear embeddings operate with greater inferential validity, distinguishing it from earlier works that treated AE merely as a preprocessing step without formal statistical integration.

Motivated by these gaps, this study proposes a two-stage nonlinear dimension reduction framework based on AE latent representations. High-dimensional scRNA-seq data are first compressed into a denoised latent space using AE, followed by nonlinear embedding using t-SNE and UMAP within the same latent representation. Through comprehensive experiments on real single-cell gene expression data, the performance of AE-based t-SNE and AE-based UMAP is systematically evaluated from multiple perspectives, including clustering consistency, structure preservation, robustness, and computational efficiency, with additional statistical validation using projection *F*-tests.

2. METHODOLOGY

2.1. Data Acquisition and Preprocessing

This dataset is a 10X Chromium sample (3994×15716) from peripheral blood mononuclear cells (PBMCs) from a human donor, and the raw data is obtained from the 10X Genomics website. The dataset used in this study can be obtained from:

https://github.com/xzhoulab/DRComparison/blob/master/data/sce_full_Zhengmix8eq.rds

Standard quality control procedures are applied prior to analysis. Genes expressed in fewer than five cells are removed, and cells with fewer than ten detected genes are excluded to eliminate low-quality observations. The remaining expression matrix is log-normalized to correct for differences in sequencing depth across cells.

2.2. Nonlinear Dimension Reduction

An autoencoder (AE) is adopted to perform nonlinear denoising and compression of the high-dimensional gene expression data. The model learns a compact latent representation by encoding the input into a lower-dimensional space and reconstructing it back to the original space, allowing essential structural information to be preserved while reducing noise.

In this study, the AE compresses the original data into a 50-dimensional latent space, which serves as the input for subsequent dimension reduction and clustering analyses. This dimensionality was selected based on established practices in single-cell RNA-seq representation learning, where latent dimensions between 30 and 100 have been shown to capture sufficient biological variance while avoiding overfitting

in datasets of comparable scale and complexity [11]. Statistically, this choice balances the bias–variance trade-off, ensuring that the latent representation retains discriminative power without becoming excessively sparse or noisy.

The model is trained using mean squared error as the reconstruction loss and optimized with the Adam optimizer (learning rate = 1e−3). Training is conducted for 30 epochs on a subset of 1000 samples. The subset size and epoch count were determined through pilot experiments aimed at achieving stable reconstruction loss convergence while maintaining computational feasibility. Training on a representative subset also reduces the risk of overfitting and enhances the generalizability of the learned representation, as the AE learns robust features without memorizing noise from the full dataset. These choices collectively support the statistical goal of deriving a stable, denoised latent space suitable for downstream nonlinear embedding.

2.3. AE Latent Representation, t-SNE, and UMAP

The learned AE latent representations were further reduced to two dimensions using t-SNE and UMAP for visualization and clustering analysis. For t-SNE, different perplexity values were examined to assess sensitivity to local neighborhood size. For UMAP, the number of neighbors and minimum distance were varied to balance local and global structure preservation. Both methods were applied directly to the AE latent space without additional feature engineering. The theory of t-SNE is referred to [6,18,19]. Implementation of t-SNE is carried out using the R package Rtsne (https://cran.r-project.org/web/packages/Rtsne/index.html). UMAP is a nonlinear dimension reduction algorithm grounded in Riemannian geometry and algebraic topology. In the high-dimensional space, UMAP constructs a fuzzy simplicial complex by computing local connectivity. Readers can refer to [7] for more theoretical details. Unlike t-SNE, which minimizes an asymmetric Kullback–Leibler divergence using a Gaussian kernel in the input space, UMAP employs symmetric cross-entropy and an exponential kernel adapted to local density, resulting in superior preservation of global structure and significantly faster optimization. The implementation of UMAP is through the R package umap (https://cran.r-project.org/package=umap).

2.4. Evaluation Metrics

Multiple evaluation metrics are employed to assess clustering quality, structure preservation, and computational efficiency. The Adjusted Rand Index (ARI) is used to measure the agreement between the predicted cluster assignments and reference labels. ARI is a measure of agreement/similarity between two data clusterings, adjusted for chance. It's widely used to compare clustering results against ground truth labels or to compare two clustering algorithms. Some major references on ARI and its application are [19-23]. In this study, the ARI was computed using externally provided reference cell type labels, even though the overall framework is unsupervised. This approach serves as a *benchmark validation* rather than an intrinsic clustering objective, allowing us to quantitatively assess how well the unsupervised embeddings recover biologically meaningful groupings. However, the use of reference labels introduces a potential bias: it presupposes that the “ground truth” labels are both accurate and optimally relevant for the given visualization task. In practice, single-cell clusters may reflect biological states beyond canonical cell types, such as activation states, cycle phases, or transient trajectories, which reference labels might not fully capture. Consequently, a lower ARI does not necessarily indicate poor clustering; it may instead reflect a mismatch between the embedded structure and the provided annotation schema. To mitigate over-reliance on ARI, we complemented it with internal validation measures (Silhouette score, trustworthiness, continuity) that do not depend on external labels. This multi-metric strategy ensures that clustering performance is evaluated both in terms of biological plausibility (via ARI) and intrinsic structural quality (via internal metrics), providing a balanced interpretation of embedding accuracy.

The Silhouette score (Silhouette coefficient) evaluates cluster compactness and separation. Trustworthiness and Continuity are adopted to quantify the preservation of neighborhood relationships between the high-dimensional space and the low-dimensional embedding, with a Silhouette score close to 1 indicating better structural preservation. Silhouette score is an internal clustering validation measure that quantifies how well each data point fits into its assigned cluster based on both cohesion (within-cluster similarity) and separation (between-cluster dissimilarity). Details on the Silhouette coefficient can be referred to references [24-27]. In addition, the M-distance [28] is used to measure the geometric distortion between the original space and the embedded space. Runtime is recorded to evaluate computational efficiency. For M-distance and runtime, lower values indicate better performance.

3. RESULTS

3.1. Direct Application of the t-SNE on the Raw Gene Expression Data

By running the R packages Rtsne and umap on the raw gene expression data and computing the ARI and

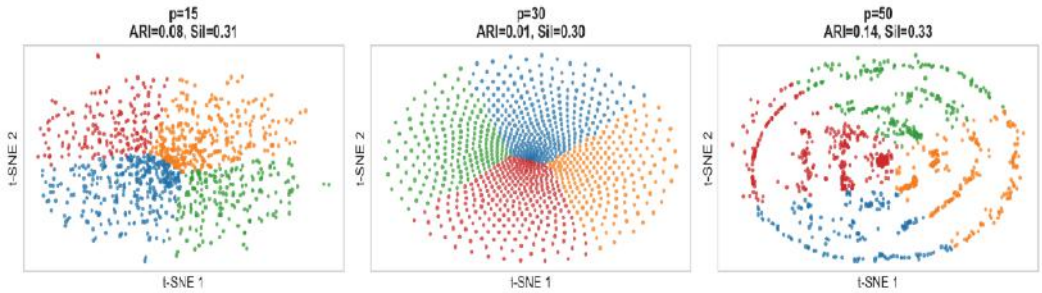

Figure 1: Pure t-SNE Plots under different choices of the perplexity parameter.

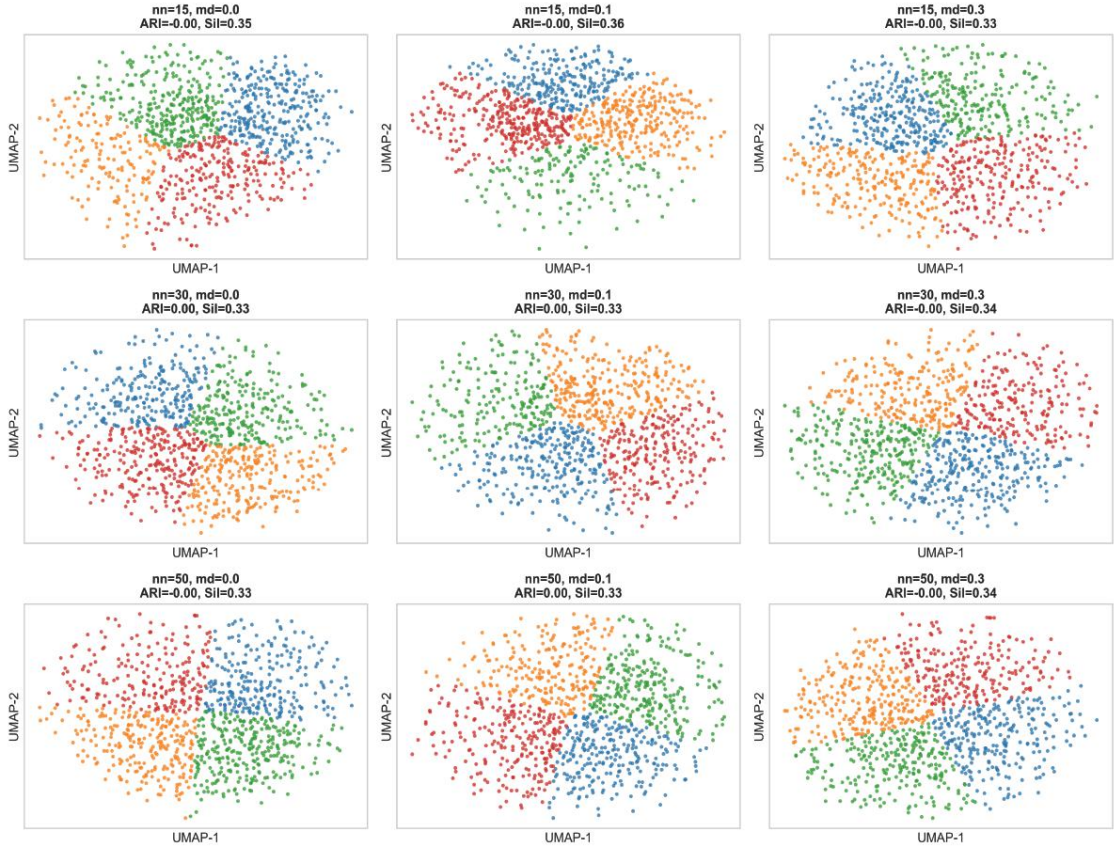

Figure 2: Pure UMAP Plot under different choices of parameters.

Table 1: Pure t-SNE Evaluation Metrics

Perplexity	ARI	Sil.	Trust.	Cont.	M	Time
15	0.078	0.312	0.503	0.731	1.91	2.40
30	0.014	0.303	0.507	0.732	1.86	3.03
50	0.137	0.332	0.510	0.731	1.92	3.82

the Silhouette score, we can obtain the t-SNE plots. As shown in Figures **1-2**, direct application of the t-SNE and UMAP to the original scRNA-seq expression matrix does not reliably recover meaningful clustering structures. The unclear clustering of the t-SNE alone in Figure **1**, and the unclear clustering of the UMAP in Figure **2**, both cannot display desirable clusters under different choices of perplexity (where p=perplexity). Tables **1-2** also provide numerical evidence on the outcomes of the clustering from t-SNE and UMAP based on different parameter configurations and repeated runs. The numerical outcomes in Tables **1-2** also support that fact that the t-SNE in Figure **1** and the UMAP in Figure **2** exhibit substantial variability, blurred cluster boundaries, and distorted global geometry. Changes in initialization and neighborhood-related parameters lead to inconsistent cluster arrangements, indicating that the high dimension, sparsity, and noise inherent in raw expression data hinder stable nonlinear visualization.

Table 2: Pure UMAP Evaluation Metrics

Neighbors	min_dist	ARI	Sil.	Trust.	Cont.	M	Time
15	0.0	-0.002	0.352	0.498	0.553	1.91	10.11
15	0.1	-0.001	0.355	0.501	0.596	1.95	3.19
15	0.3	0.000	0.335	0.499	0.509	1.89	3.26
30	0.0	0.000	0.329	0.503	0.525	1.86	3.49
30	0.1	0.001	0.335	0.502	0.504	1.89	3.43
30	0.3	-0.001	0.341	0.506	0.520	1.89	3.49
50	0.0	0.000	0.333	0.502	0.473	1.89	3.71
50	0.1	0.002	0.332	0.504	0.483	1.87	3.78
50	0.3	-0.002	0.336	0.502	0.451	1.89	3.76

Note: In Tables **1-2**, Trust.=Trustworthiness, Cont.=Continuity, which are two of the most important neighborhood preservation metrics used to quantitatively evaluate the performance of dimensionality reduction techniques like t-SNE and UMAP. They don't evaluate clustering *per se*, but rather how faithfully the low-dimensional embedding preserves the high-dimensional data's neighborhood structure, see [19, 29, 30] for more details.

3.2. Latent Representation from AE

The AE model is used to obtain a low-dimensional latent representation of the data, capturing the key features for clustering analysis. Additionally, the projected *F*-test in [31] is applied to perform multiple mean comparisons, further assessing the differences between clusters in the AE latent space. The projected *F*-test [31] outcomes are summarized in Table **3**, where the sample size n=1000, the data dimension p=15716. The maximum number of projection dimension $r = min(n-1, p) - 1$ in [31] was AE compressed to $r = 50$. The exact null distribution for the projected *F*-test is $F_r \sim F(r, n-1-r)$. The null hypothesis is

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k,$$

where each $\mu_i \ (i = 1, \ldots, k)$ stands for the average cluster level, k is the number of clusters selected from the t-SNE or UMAP. Based on the t-SNE plots in Figure **5** and the UMAP plots in Figure **6**, the projected *F*-test for multiple mean comparisons among the 4 clusters (k=4 in the null hypothesis) is carried out to test the hypothesis. Four choices of projection dimension are indicated in Table **3**, where the notation $[x]$ means the integer part of a real number like [2.1] =2, [2.9] =2. The outcomes in Table **3** show a significant difference exists among the four cluster mean levels in Figures **5-6**.

The projected F-tests in Table **3** provides further partial statistical evidence to support the fact that the four clusters from the AE-UMAP procedure are far apart from one another in the sense of central tendency measured by the mean. Because the F-tests in Table **3** are not independent with each other, the overall statistical significance is not able to be given. This belongs to the big area of multiple mean comparison with dependence. Interested readers can refer to [31, 38, 39] for more statistically justifiable interpretation on the overall statistical significance.

3.3. Determination of the Optimal Number of Clusters

To ensure a fair and reproducible comparison of nonlinear embeddings, the number of clusters is determined based on the structure of the AE latent representations rather than the low-dimensional visualization results. Both elbow analysis and Silhouette evaluation are performed in the AE latent space to identify a stable cluster configuration.

As shown in Figure **3**, the within-cluster sum of squares (WCSS) curve exhibits a clear inflection at k = 4, beyond which further increases in k result in only

Table 3: Projected F-Tests in AE

Projection Dimension	Projected *F*-distribution	*F*-value	*p*-value
$r_1 = \left[\frac{50}{4}\right] = 12$	*F* (12, 987)	89.3484	2.3690e−148
$r_2 = \left[\frac{50}{3}\right] = 16$	*F* (16, 983)	67.3575	1.5761e−145
$r_3 = \left[\frac{50}{2}\right] = 25$	*F* (25,974)	47.2665	4.8151e−145
$r_4 = \left[\frac{3\times 50}{4}\right] = 37$	*F* (37, 962)	34.8906	2.7234e−148

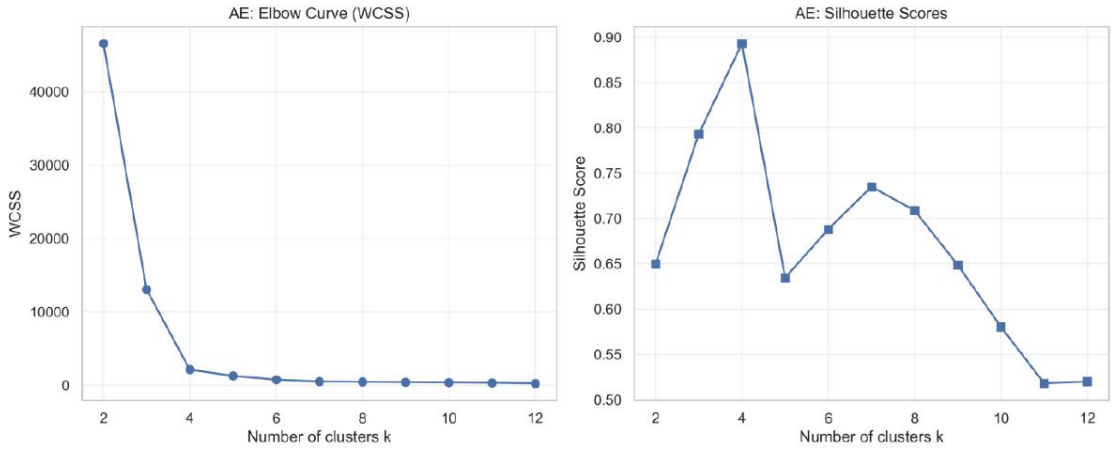

Figure 3: Elbow Plot and Silhouette Score of AE.

marginal reductions in within-cluster variance. Consistently, the Silhouette score reaches its maximum at k = 4, indicating optimal cluster compactness and separation. Based on the agreement between these two criteria, k = 4 is selected as the optimal number of clusters for all subsequent analyses.

3.4. Hierarchical Relationship between Clusters

To further examine the structural relationships among the identified clusters, hierarchical clustering was performed on the centroids of the four clusters obtained in the AE latent space using Ward's linkage method [40]. As illustrated in Figure **4**, the resulting dendrogram reveals a clear multi-level hierarchy, with distinct merging distances reflecting varying degrees of transcriptional similarity between clusters. Specifically, Clusters 1 and 2 merge at a relatively low linkage distance (~2.5), indicating that they represent closely related subpopulations. In contrast, Cluster 3 remains well separated, merging only at a much larger distance (~7.5), which highlights its distinct transcriptional identity. This hierarchical organization statistically reinforces the stability and interpretability of the chosen cluster configuration (k = 4), demonstrating that the AE latent space captures not only discrete cell groups but also biologically meaningful relationships at different levels of granularity. The clear separation between major branches further supports the structural coherence of the learned representation and its suitability for downstream analyses such as trajectory inference or cell-type annotation.

3.5. Evaluation of t-SNE Applied to AE Latent Representations

When the t-SNE is applied to the AE latent representations, clustering performance is markedly improved compared with its direct application to the raw expression data. The ARI increases to approximately 0.52, indicating higher agreement between the resulting cluster assignments and the reference labels. Changes in the perplexity parameter have limited influence on ARI values, whereas higher perplexity values are associated with increased Silhouette scores, reflecting improved within-cluster compactness.

As shown in Figure **5**, samples belonging to different categories form more clearly separated

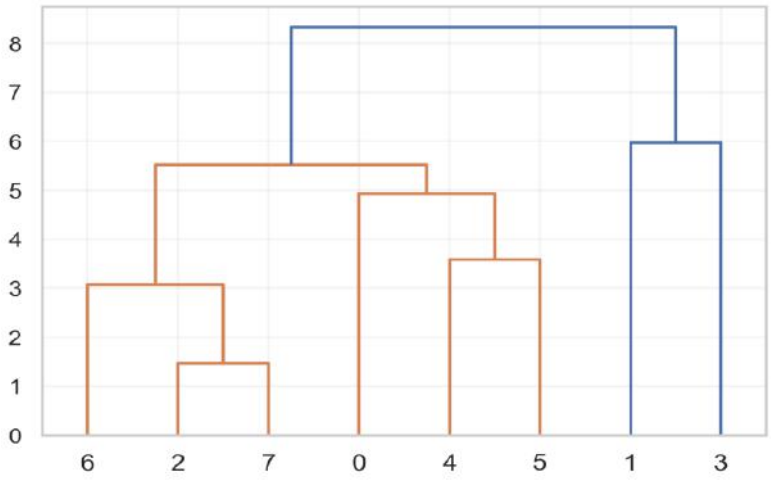

Figure 4: Hierarchical Clustering of True Labels Centroids.

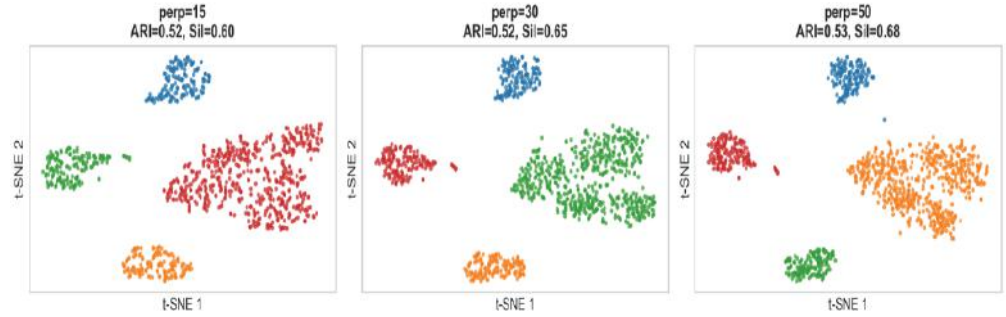

Figure 5: t-SNE Based on AE.

Table 4: Optimal Metrics at Perplexity = 50

Metric	ARI	Sil.	Trust.	Cont.	M	Time
Value	0.526	0.681	0.970	0.970	2.45	2.38

groups in the low-dimensional embedding. Table **4** shows under the optimal parameter configuration with a perplexity of 50, the ARI reaches 0.526 and the Silhouette score reaches 0.681. Trustworthiness and continuity values are both approximately 0.97, indicating that local and global neighborhood relationships are well preserved. The corresponding M-distance remains low at 2.45. The runtime under this configuration is 2.38 seconds.

3.6. Evaluation of UMAP Applied to AE Latent Representations

UMAP applied to AE latent representations produces consistent clustering results across a range of parameter settings. Variations in the number of neighbors and minimum distance lead to only modest changes in the resulting embeddings.

As illustrated in Figure **6**, cluster separation is maintained under different parameter combinations. Quantitative evaluation results are summarized in Table **5**. Across all tested configurations, ARI values remain constant. Silhouette scores decrease gradually as the minimum distance increases, indicating reduced within-cluster compactness. Trustworthiness shows a slight downward trend with increasing minimum distance, whereas continuity remains relatively stable. M-distance values remain low across all parameter

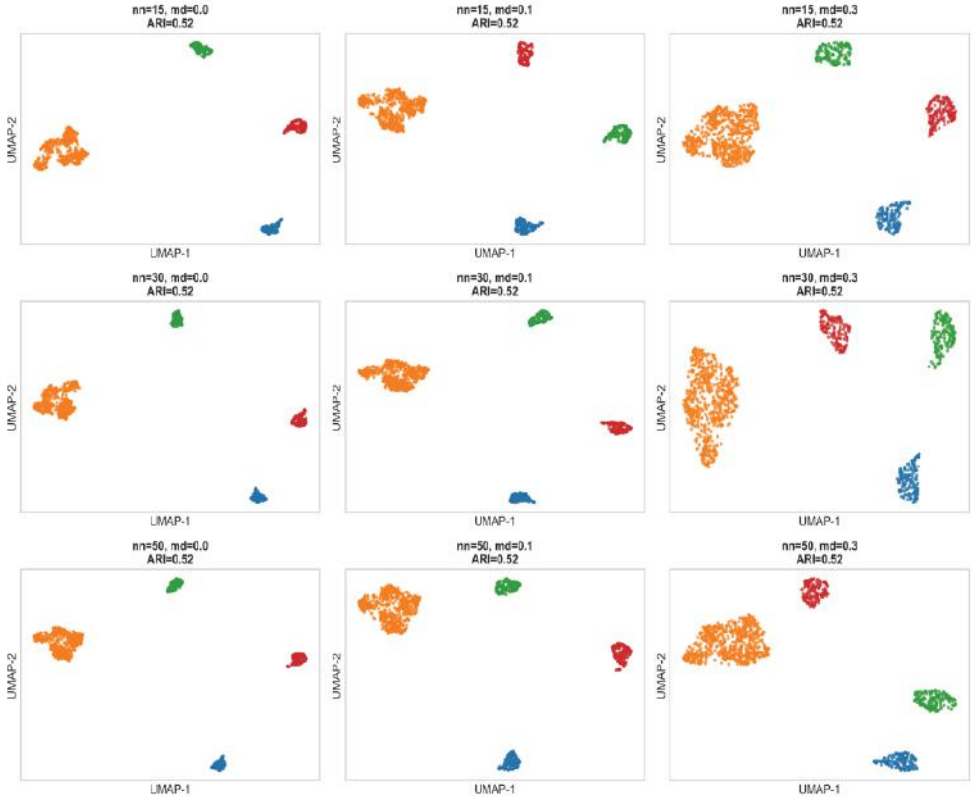

Figure 6: UMAP based on AE.

Table 5: The Comparison Results of UMAP Parameters Combinations

Neighbors	min_dist	ARI	Sil.	Trust.	Cont.	M	Time
15	0.0	0.525	0.893	0.951	0.971	2.50	1.30
15	0.1	0.525	0.860	0.950	0.972	2.52	1.27
15	0.3	0.525	0.807	0.941	0.974	2.47	1.27
30	0.0	0.525	0.889	0.952	0.972	2.52	1.59
30	0.1	0.525	0.879	0.947	0.973	2.54	1.49
30	0.3	0.525	0.821	0.938	0.974	2.32	1.49
50	0.0	0.525	0.891	0.945	0.972	2.54	1.70
50	0.1	0.525	0.850	0.942	0.973	2.50	1.64
50	0.3	0.525	0.744	0.937	0.973	2.46	1.65

settings. From Table **5**, the influence of neighborhood size is limited, with no substantial changes observed in the quantitative metrics as the number of neighbors increases from 15 to 50. The configuration with 15 neighbors and a minimum distance of 0 yields the highest Silhouette score and the lowest M-distance, with a runtime of 1.30 seconds.

3.7. Multi-Stage Embedding

Apart from the AE and UMAP combination, the integration of three methods is also These combinations aimed to explore whether further improvements could be made in dimension reduction and to find a better balance between different methods.

The combination of AE with UMAP and t-SNE is tested, but no significant improvement in clustering performance is observed compared to UMAP based on AE alone.

Since t-SNE handles only 2D or 3D data, AE followed by PCA and UMAP is tested in Figure **8**. While some parameter combinations yield good classification results, this method do not outperform the two-method combinations in terms of evaluation metrics. PCA relies

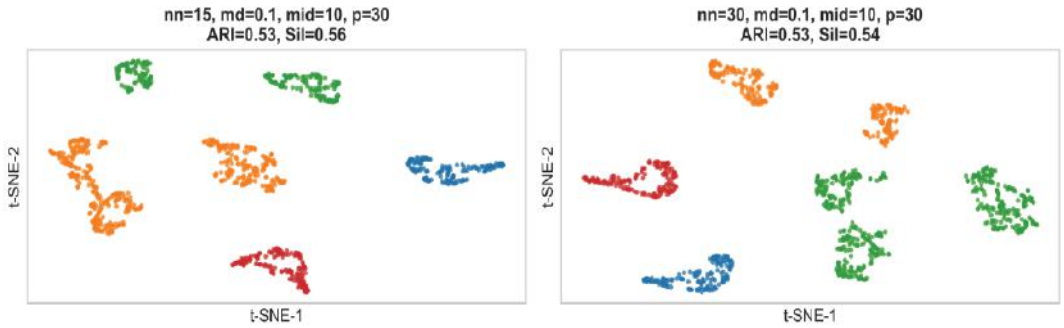

Figure 7: AE followed by UMAP and t-SNE.

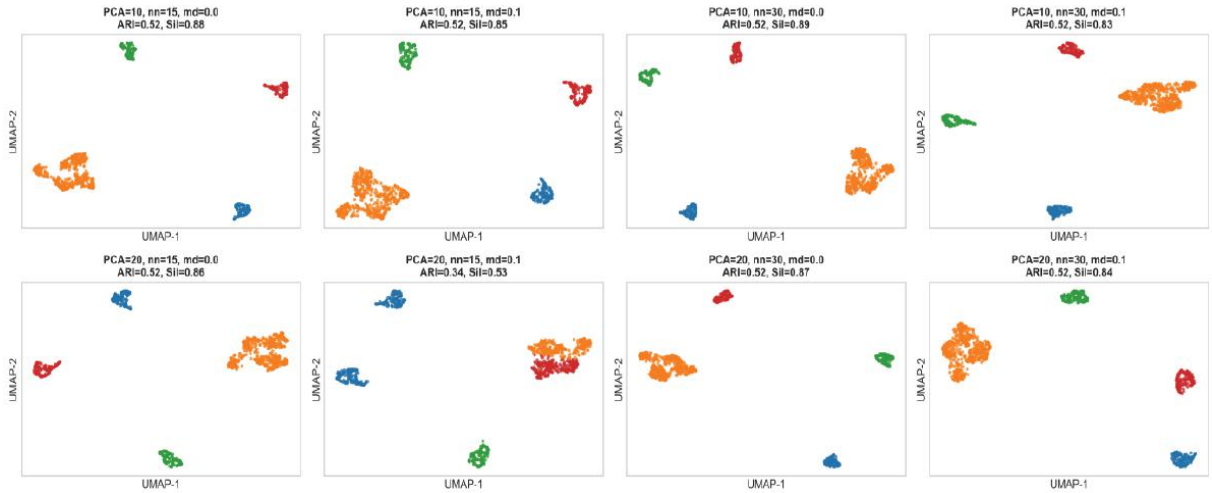

Figure 8: AE followed by PCA and UMAP.

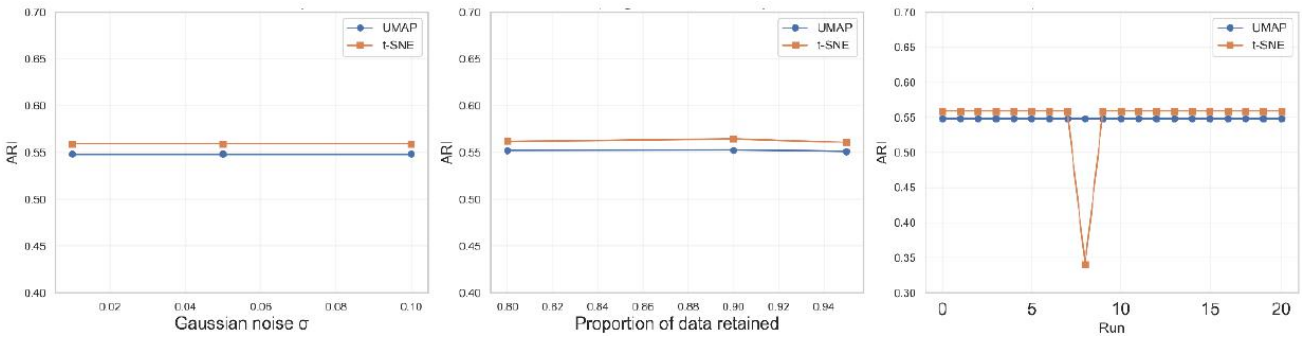

Figure 9: Robustness Analysis.

on linear transformations and cannot capture complex nonlinear relationships as effectively as pure t-SNE and pure UMAP.

3.8. Robustness and Sensitivity Analysis

To further evaluate the reliability of the proposed AE-based nonlinear embedding framework, robustness and sensitivity analyses are conducted under a range of perturbation conditions. Specifically, the stability of clustering performance is examined with respect to noise perturbation, stochastic initialization, subsampling, the number of clusters, and sample size. All evaluations are performed on the AE latent representations using fixed optimal embedding parameters.

Under Gaussian noise perturbation and subsampling, both the t-SNE and UMAP exhibit stable ARI values across different noise levels and retained data proportions, indicating that moderate data perturbations have limited impact on clustering consistency. Repeated experiments with different random initializations further reveal differences in stability between the two methods. As shown in Figure **9**, UMAP produces highly consistent results across runs, whereas t-SNE occasionally exhibits noticeable performance degradation under specific random seeds.

To further examine sensitivity to the number of clusters, both methods are evaluated with cluster numbers ranging from k = 4 to k = 8. As shown in Figure **10**, UMAP achieves its highest Silhouette score at k = 4, indicating well-separated and compact clusters in the latent space, whereas t-SNE exhibits more gradual changes across different cluster numbers. Although ARI values for both methods generally increase with k, the relative performance trends between t-SNE and UMAP remain consistent, and no substantial performance improvement is observed beyond the selected cluster number. These results indicate that k = 4 provides a balanced clustering configuration for this dataset and that the comparative conclusions are robust to moderate variations in the number of clusters.

To further assess the effect of sample size on embedding performance, the t-SNE and UMAP applied to the AE latent representations are evaluated using subsets of different sizes. The analysis focuses on clustering consistency and computational efficiency as the number of samples varies. As shown in Figure **11**, UMAP achieves higher ARI values under small sample conditions, indicating stronger clustering consistency when the number of samples is limited. As the sample size increases, the performance gap between the two methods gradually diminishes, and both methods converge to similar clustering accuracy at medium to large sample sizes. In terms of computational efficiency, Figure **11** shows that UMAP consistently requires less runtime than t-SNE across all sample sizes, with the difference becoming more pronounced as the dataset

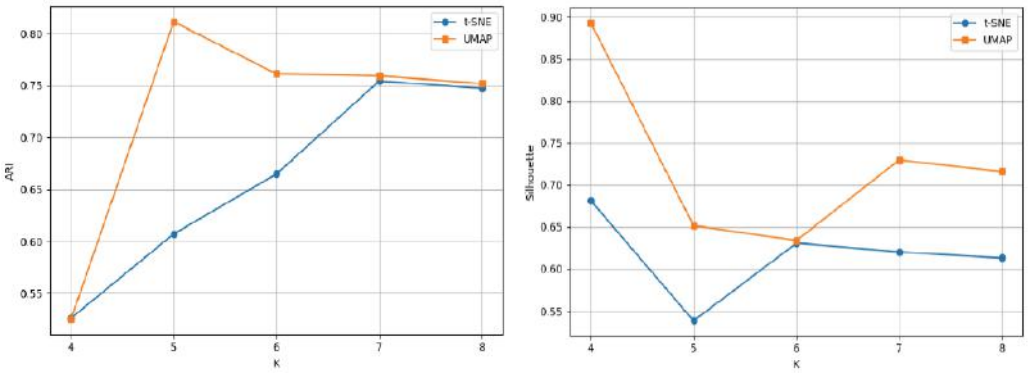

Figure 10: ARI and Silhouette in different k.

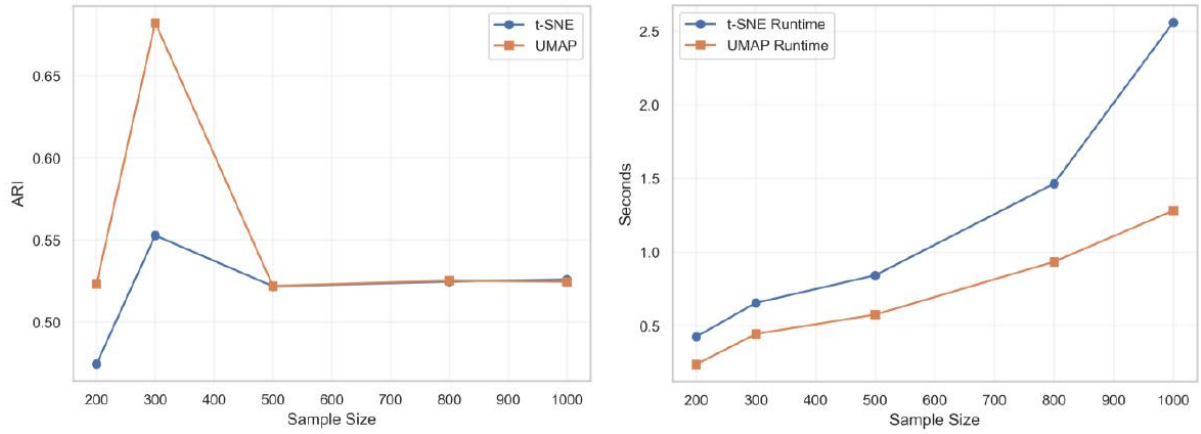

Figure 11: ARI and Runtime in different sample size.

size increases. These results indicate that UMAP offers clear advantages in robustness and scalability, particularly for small to medium-sized datasets.

4. CONCLUDING REMARKS

This study investigated the effectiveness of an autoencoder-based latent representation in stabilizing nonlinear dimension reduction for single-cell RNA-seq data, with a comparative focus on t-SNE and UMAP. Our results demonstrate that the quality of the input representation fundamentally determines the reliability of subsequent nonlinear embedding. Direct application of t-SNE or UMAP to raw expression data yielded unstable and poorly separated clusters, whereas performing nonlinear reduction on AE-learned latent representations substantially improved clustering consistency, visual separability, and structural preservation.

Within the unified AE latent space, both methods succeeded in recovering meaningful cluster structures. However, UMAP exhibited clear advantages in preserving global topology, robustness to parameter variation and initialization, and computational efficiency. These advantages were quantitatively validated through multiple metrics—including higher Silhouette scores, stable ARI under perturbation, and lower runtime—as well as statistical confirmation via projection-based F-tests. Sensitivity analyses further confirmed that UMAP's performance advantages persist across varying sample sizes, noise levels, and cluster numbers.

The framework presented here underscores the importance of representation learning as a prerequisite for reliable nonlinear visualization in high-dimensional, noisy biological data. Rather than treating AE merely as a preprocessing step, this study positions it as a statistically grounded foundation that enhances the inferential validity of downstream embeddings. In practical terms, the AE-UMAP pipeline emerges as a robust, efficient, and reproducible choice for exploratory single-cell analysis, particularly when stability and interpretability are prioritized.

Future work may extend this approach to larger and more diverse datasets, integrate alternative deep learning architectures, and explore the inclusion of biological prior knowledge to further enhance interpretability. Nevertheless, this study provides a systematic, empirically supported framework for combining representation learning with nonlinear dimension reduction, offering clear methodological guidance for the analysis of single-cell transcriptomic data.

While this study demonstrates the effectiveness of AE-based nonlinear embedding, several limitations should be acknowledged. First, our analysis relied on a single PBMC dataset; although it is widely used and representative, further validation across diverse biological contexts and sequencing platforms would strengthen generalizability. Second, the autoencoder architecture and training parameters (e.g., latent dimension, epochs, subset size) were fixed based on pilot experiments and existing conventions; systematic hyperparameter optimization could further refine representation quality. Third, evaluation remained largely visualization- and metric-driven, relying on internal and external clustering indices; future work could benefit from incorporating biological ground truth or functional validation to assess biological relevance more directly. These limitations highlight opportunities for extension rather than invalidating the framework, which remains a statistically principled and practically useful approach for single-cell data exploration.

REFERENCES

[1] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods 2009; 6(5): 377-82. https://doi.org/10.1038/nmeth.1315

[2] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology 2019; 15(6): e8746. https://doi.org/10.15252/msb.20188746

[3] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature methods 2014; 11(7): 740-2. https://doi.org/10.1038/nmeth.2967

[4] Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol 2019; 20(1): 269. https://doi.org/10.1186/s13059-019-1898-6

[5] Ringnér M. What is principal component analysis? Nature biotechnology 2008; 26(3): 303-4. https://doi.org/10.1038/nbt0308-303

[6] Maaten L van der, Hinton G. Visualizing data using t-SNE. Journal of machine learning research 2008; 9: 2579-605.

[7] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction arXiv; 2020. https://doi.org/10.48550/arXiv.1802.03426

[8] Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. Distill 2016; 1(10): e2. http://doi.org/10.23915/distill.00002

[9] Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature communications 2019; 10(1): 5416. https://doi.org/10.1038/s41467-019-13056-x

[10] Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science 2006; 313(5786): 504-7. https://doi.org/10.1126/science.1127647

[11] Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications 2019; 10(1): 390. https://doi.org/10.1038/s41467-018-07931-2

[12] Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nature communications 2018; 9(1): 2002. https://doi.org/10.1038/s41467-018-04368-5

[13] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature methods 2018; 15(12): 1053-8. https://doi.org/10.1038/s41592-018-0229-2

[14] Geddes TA, Kim T, Nan L, Burchfield JG, Yang JYH, Tao D, *et al*. Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis. BMC Bioinformatics [Internet] 2019; 20(S19): 660. https://doi.org/10.1186/s12859-019-3179-5

[15] Xiang R, Wang W, Yang L, Wang S, Xu C, Chen X. A comparison for dimensionality reduction methods of single-cell RNA-seq data. Frontiers in genetics 2021; 12: 646936. https://doi.org/10.3389/fgene.2021.646936

[16] Allaoui M, Kherfi ML, Cheriet A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: El Moataz A, Mammass D, Mansouri A, Nouboud F, editors. Image and Signal Processing Cham: Springer International Publishing; 2020. p. 317-25. https://doi.org/10.1007/978-3-030-51935-3_34

[17] Amir E ad D, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, *et al*. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature biotechnology 2013; 31(6): 545-52. https://doi.org/10.1038/nbt.2594

[18] Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 2000; 290(5500): 2323-6. https://doi.org/10.1126/science.290.5500.2323

[19] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv; 2020. https://doi.org/10.48550/arXiv.1802.03426

[20] Hubert L, Arabie P. Comparing partitions. Journal of Classification 1985; 2(1): 193-218. https://doi.org/10.1007/BF01908075

[21] Rand WM. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 1971; 66(336): 846-50. https://doi.org/10.1080/01621459.1971.10482356

[22] Steinley D. Properties of the Hubert-Arable Adjusted Rand Index. Psychological Methods 2004; 9(3): 386-96. https://doi.org/10.1037/1082-989X.9.3.386

[23] Santos JM, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In: Alippi C, Polycarpou M, Panayiotou C, Ellinas G, editors. Artificial Neural Networks - ICANN 2009 Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 175-84. https://doi.org/10.1007/978-3-642-04277-5_18

[24] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 1987; 20: 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[25] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 2009.

[26] Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics 2005; 21(15): 3201-12. https://doi.org/10.1093/bioinformatics/bti517

[27] Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern recognition 2013; 46(1): 243-56. https://doi.org/10.1016/j.patcog.2012.07.021

[28] Mahalanobis PC. On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-) 2018; 80: S1-7

[29] Venna J, Kaski S. Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study. In: Dorffner G, Bischof H, Hornik K, editors. Artificial Neural Networks — ICANN 2001 Berlin, Heidelberg: Springer Berlin Heidelberg; 2001. p. 485-91. https://doi.org/10.1007/3-540-44668-0_68

[30] Van Der Maaten L, Postma EO, Van Den Herik HJ. Dimensionality reduction: A comparative review. Journal of machine learning research 2009; 10(66-71): 13.

[31] Cao Y, Liang J. Multiple mean comparison for clusters of gene expression data through the t-SNE plot and PCA dimension reduction. International Journal of Statistics in Medical Research 2025; 14: 1-14. https://doi.org/10.6000/1929-6029.2025.14.01

[32] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, *et al*. Comprehensive integration of single-cell data. cell 2019; 177(7): 1888-902. https://doi.org/10.1016/j.cell.2019.05.031

[33] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018; 19(1): 15. https://doi.org/10.1186/s13059-017-1382-0

[34] Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, *et al*. Massively parallel digital transcriptional profiling of single cells. Nature communications 2017; 8(1): 14049. https://doi.org/10.1038/ncomms14049

[35] Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. Nature communications 2021; 12(1): 1873. https://doi.org/10.1038/s41467-021-22008-3

[36] Feng C, Liu S, Zhang H, Guan R, Li D, Zhou F, *et al*. Dimension reduction and clustering models for single-cell RNA sequencing data: a comparative study. International journal of molecular sciences 2020; 21(6): 2181. https://doi.org/10.3390/ijms21062181

[37] Hasan BMS, Abdulazeez AM. A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining 2021; 2(1): 20-30. https://doi.org/10.30880/jscdm.2021.02.01.003

[38] Borenstein M (Ed.), Meta-analysis: A guide to calibrating and combining statistical evidence. Wiley 2024.

[39] Westfall PH, Young SS, Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons 1993.

[40] Ward JH. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. 1963; 58(301): 236-244.

Received on 08-11-2025 Accepted on 13-12-2025 Published on 30-12-2025

https://doi.org/10.6000/1929-6029.2025.14.78

© 2025 Yang *et al*.
This is an open-access article licensed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the work is properly cited.