Combined Statistical and Machine Learning Approaches for Effective sc-RNA-seq Imputation and Denoising

Group 7

Yufei Deng, Xinyi Di, Yihan Liu

Background & Introduction

Handling of noisy data is vital to many computational problems. Although technological advancement in single-cell RNA expression sequencing (scRNA-seq) has enabled the analysis of gene expression at the refined cellular level, the extensive noise introduced by "dropouts"--characterized by false zero counts in the data–remains a major challenge in the analysis of scRNA-seq data.

Different methods for denoising and imputation have been proposed to help recover the data from technical variants while preserving the inherent biological variability across cell types, subsequently improving downstream analyses such as cell type annotation, clustering, and classification. Some of these methods rely on the probabilistic assumptions for the data distribution–such as zero-inflation negative binomial (ZINB) and multinomial–some use deep learning such as autoencoder to learn the complex interactions among genes in an unsupervised fashion, and some combine the two[1][2[3]], offering a flexible. Therefore, our primary goal is to find the optimal imputation method for scRNA-seq data by exploring various approaches and evaluating them upon the performance on downstream tasks such as cell type classification, utilizing both real and simulated gene expression data.

Specific Aim 1: Comparison of Imputation Methods Against Baseline

We will utilize different imputation approaches on both real gene expression data and simulated data, comparing against the baseline method, which imputes dropout counts with the average expression value of the gene across all cells. This comparison will provide insights into the strengths and limitations of each imputation and denoising algorithm, aiding in the selection of the optimal imputation approach.

Specific Aim 2: Evaluate Performance on Downstream Functional Analyses

Our final aim is to evaluate the effectiveness of our chosen imputation method on downstream analyses, particularly cell-type classification. Leveraging the cell annotation created by the data curator or those generated through simulation, we will perform multi-class classification using machine learning algorithms and derive the evaluation metric using the predicted cell labels. By comparing the classification accuracy and robustness across different imputation methods, we aim to ascertain the impact of imputation on downstream analyses and validate the utility of our chosen approach.

Specific Aim 3: Validation of Previous Results through Simulation Study

To further validate the effectiveness of different scRNA-seq imputation methods, we will conduct a simulation study using existing software (e.g., Splatter in R). By simulating gene expression data from a Poisson-gamma statistical model and synthetically generating cell labels for subsequent classification tasks, we aim to evaluate the performance of each method in a controlled setting. This simulation study not only validates our findings from real datasets but also elucidates the impact of noise on downstream analyses.

Research Strategy

Significance

Our project is crucial in discovering the optimal imputation method for scRNA-seq data and enhancing the performance of downstream analyses. By systematically evaluating various approaches and validating our findings across real and simulated datasets, we aim to provide valuable insights into handling noise in scRNA-seq data, ultimately advancing our understanding of cell heterogeneity and function.

Innovation

The innovation of our study lies in our comprehensive approach to addressing the challenge of noise in scRNA-seq data. While existing methods have made significant efforts in denoising and imputation, our project innovates by systematically comparing and evaluating these methods across real and simulated datasets. The scope of methods encompasses diverse imputation techniques, including those based on statistical models, deep learning, and combined. We also aim to design our own algorithm which uses a model-free iterative imputation method to explore the potential advantages of a method free of restraints from statistical assumptions and is not as complex as deep neural networks in terms of implementation. Furthermore, our utilization of simulation provides validation of our findings from real datasets and insights into the impact of noise on downstream analyses. This holistic approach provides more accurate and reliable utilization of single-cell data, leading us toward a better understanding of cellular heterogeneity and function.

Specific Aim 1: Denoising using different Imputation Methods

Hypothesis

We propose using an autoencoder approach for scRNA-seq data imputation and denoising. In particular, we suggest substituting the reconstruction loss in traditional autoencoders with the likelihood of noise models, which we hypothesize will tailor the neural network toward learning the sparse count data seen in scRNA-seq tasks. Subsequently, the autoencoder will effectively compress the high-dimensional data and recover the true gene UMI counts of every single cell, ultimately helping to uncover the masked biological heterogeneity in the dataset.

Rationale

An autoencoder is an artificial neural network that learns the compression of data in an unsupervised manner by minimizing a loss function through an encoder-decoder architecture. The intuition of using the autoencoder architecture is given by the following fundamental assumptions: (1) The high-dimensional gene expression count data can be represented by a lower-dimensional manifold; (2) Autoencoders can effectively learn the non-linear mapping functions between layers, taking into account the dynamic relationships among features and observations. In the scenario of scRNA-seq denoising, the lower-dimensional cell manifold is represented by the output of the hidden bottleneck layer, which is then projected back to the original dimensions by the decoder to obtain the reconstructed cell count data.

Another reason for choosing an autoencoder framework is its flexibility in the construction of the loss function and scalability to handle larger datasets easily. We will explore a range of different loss functions for building the autoencoder and implement them on both real-world and simulation datasets. This will allow us to comprehensively evaluate the multiple existing algorithms for scRNA-seq data imputation and denoising.

Experimental Approach

We deployed a few existing variations of the autoencoder algorithm and compared them against the baseline method, where we imputed all zero entries in the data with the average of a gene's expression counts across all cells. Our main proposed approach is to combine autoencoders with statistical models by applying probability distributional assumptions to the gene expression data. These include the ZINB distribution and the Multinomial-Bernoulli distribution. We also implemented two variations of the mean squared error (MSE) loss for the autoencoder, which is the weighted MSE and masked MSE. Finally, we included a baseline method that imputes all the 0 values by taking the average count of a given gene across all cells to serve as a reference for the rest of the methods.

ZINB. Many existing methods have applied the ZINB distribution for modeling scRNA-seq data[1]. The ZINB distribution is parameterized by a Negative Binomial (NB) distribution (μ, θ) and a parameter (π) that represents the weight of a point mass at zero (δ_0).

$$f_{ZINB}(x; \pi, \mu, \theta) = \pi\,\delta(x) + (1 - \pi) f_{NB}(x;\ \mu, \theta)\,,$$

where $f_{NB}(x; \mu, \theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x}$

The point mass function models true dropout events, while the NB distribution captures the non-negativity of the UMI count matrix. Together, the parameters π, μ, θ jointly define the likelihood function of the noise model, which will also be the loss function for the deep autoencoder.

Multinomial-Bernoulli. We also adopted a Multinomial distribution to model the UMI count data[2]. We assume that the gene counts of the ith cell, $X_i = (X_{i1}, ..., X_{im})$, can be modeled by a multinomial distribution with parameters $p_i = (p_{i1}, p_{i2}, ..., p_{im})$, representing the probability of the jth gene in the ith cell. Suppose $n_i = \sum_{j=1}^{m} X_{ij}$ is the total UMI counts in the ith cell:

$$f_i(X_i) = \frac{n_i!}{X_{i1}!X_{i2}!...X_{im}!}\prod_{j=1}^{m} p_{ij}^{X_{ij}}$$

Moreover, due to excessive zero counts in the data, we model the dropout events with a binary random variable, Uij~Bern(πij), with Uij = 0 representing that the jth gene drops out in the ith cell. Hence, the probability pij is in fact $\pi_{ij}V_{ij} / \Sigma\pi_{ij}V_{ij}$, where Vij is the masked true UMI count of the jth gene of the ith cell. The parameters πij and Vij jointly define the likelihood function of the noise model, which will also serve as the loss upon which the autoencoder model optimizes its parameters.

MSE Methods. The MSE loss is commonly used for training artificial neural network models[3]. To tailor it toward denoising scRNA-seq data, where a large proportion of the data are false zero values, we used two variations of the MSE loss–the weighted MSE and masked MSE. The former weights the squared error by the raw count of the jth gene of the ith cell.

$$L = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}(X_{ij} - \hat{X}_{ij})^2$$

The latter computes the MSE only on entries that are originally non-zero.

$$L = \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_{ij} > 0)(X_{ij} - \hat{X}_{ij})^2$$

Finally, we included a baseline denoising method, where we treated all zero entries as dropout events and imputed each zero value with the average non-zero expression counts of that given gene across all cells.

Autoencoder Architecture. We similarly used hidden layer sizes of (64, 32, 64) for data compression and reconstruction across ZINB, Multinomial-Bernoulli, and MSE methods. However, we used different output layer structures for ZINB and Multinomial-Bernoulli models due to their variations in the parameter space. For the former, we outputted the parameters (μ, θ) using an exponential activation function and the parameter π with a sigmoid activation function. For the latter, we outputted the parameter V with an exponential activation function and π with a sigmoid activation function.

We used a learning rate of 0.001, a batch size of 64, and 100 training epochs. The hyperparameter setting was shared across the methods during neural network training.

Specific Aim 2: Evaluate Performance on Downstream Functional Analyses

Hypothesis

Denoising methods utilizing a statistical model with an autoencoder outperform both the baseline method and methods without any denoising.

Rationale

Single-cell RNA sequencing (scRNA-seq) technologies are crucial for understanding cellular heterogeneity but are complicated by significant noise issues due to amplification and dropout in large, sparse datasets. Traditional denoising methods, which typically use MSE or NB distributions, often fail to capture the true complexity and variability of gene expression data. By integrating a deep autoencoder with a multinomial statistical model, this approach more accurately represents the underlying data structure. The method enhances denoising effectiveness and data interpretability by leveraging soft clustering within a robust low-dimensional latent space, offering superior performance in handling the challenges of large-scale scRNA-seq datasets compared to existing methods.

Experimental Approach

We implemented two primary methods to evaluate the effectiveness of our denoising technique: clustering and classification. For clustering, we utilized the k-means algorithm to assess how well the denoised data could be segmented into distinct groups[3], reflecting underlying biological variations. The metrics we use to evaluate the clustering results are Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Silhouette coefficient (SC). We leveraged Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and visualization of the clustering results. In classification, we applied logistic regression to determine the accuracy with which the processed data could predict the annotated cell type, thereby testing the preservation of critical information after denoising. We evaluate the classification results by F1 score, recall, precision, AUC score, and accuracy. In addition, we fitted a Random Forest Classifier to determine the feature importance of genes, aiming to uncover those obscured by noise. In tree-based classification systems, feature importance is gauged from information gain—the more a gene informs about the cell type, the greater its expression within that cell.

Interpretation of Results

For clustering, the results presented in **Table 1 (Appendix) and Figure 1(Appendix)** demonstrate that the denoising method utilizing multinomial loss consistently outperforms other methods across all evaluation metrics. **Figure 2 (Appendix)** shows the UMAP clustering results of different data with or without denoising. We can observe that the cellular landscape of the raw data demonstrates an extensive overlap between the H1 and H9 stem cells and their progenitor cell types, indicating the presence of noise. The denoised data using the baseline method did not improve the results, even adding additional noise to the data, as the TB subtype is mapped closer to the other clusters. The autoencoder with Multinomial and ZINB loss functions, on the other hand, showed marked improvement in clustering results. Particularly, the Multinomial method successfully separated the H1 and H9 cell types, while other denoising methods failed to achieve this, indicating its superior performance in handling dropouts and discerning covered biological variability.

Similarly, the classification results, as detailed in **Table 2 (Appendix)** and **Figure 3 (Appendix)**, further confirm the superiority of the denoising method with multinomial loss over competing approaches. However, other denoising methods don’t surpass non-denoise methods on this dataset, which may be caused by specific gene expression properties of endoderm. The bar charts in **Figure 4 (Appendix)** and **Figure 5 (Appendix)** illustrate a significant discrepancy in the top 50 gene ranking between the original data and the optimally denoised data using multinomial loss (based on classification results). The results indicate technical noise in the data introduced in cell preparation and RNA sequencing steps can mask important biological variability in the data. Imputation is essential for recovering true RNA expression counts and improving downstream analyses such as clustering, pathway analysis, etc.

Potential Problems and Alternative Approaches

The effectiveness of our algorithm is currently confined to datasets derived from endoderm cells, limiting its general applicability. We intend to refine our model across a more varied set of datasets, including both additional real-world biological datasets and controlled simulated environments. Moreover, the model’s interpretability remains opaque, which poses challenges in terms of scientific transparency. Enhancing model interpretability is crucial as the next step.

Finally, the practical application of our denoising methods in the broader field of biology has yet to be realized. To bridge this gap, we may apply them in biological studies to demonstrate their utility in uncovering biologically relevant insights.

Specific Aim 3: Validation of Previous Results through Simulation Study

Hypothesis

Similar results that denoising methods utilizing a statistical model with an autoencode outperform both the baseline methods and methods without any denoising will be detected using simulated data.

Rationale

To further validate the effectiveness of different scRNA-seq imputation methods, we conducted a simulation study using the Splatter package in R[4]. Using simulated data enables a direct comparison between the imputed data with the ground truth. Since the true underlying values are known in simulated data, it provides a benchmark for evaluating the accuracy of imputation methods. Also, simulated data enables control of various parameters such as dropout rates, number of groups, and missing data patterns. This control is essential for systematically assessing the performance of different imputation methods under diverse conditions.

Experimental Approach

We conducted a simulation study using the Splatter package in R. Splatter is a comprehensive simulation tool designed for replicating key characteristics of real single-cell RNA sequence data. It employs a gamma-Poisson hierarchical model to simulate gene expression levels, incorporating features such as high expression outliers, varying library sizes across cells, trended gene-wise dispersion, and zero-inflation. Specifically, gene means are initially sampled from a Gamma distribution, with the addition of high expression outliers based on a specified probability. Library sizes are modeled using a log-normal distribution, and a mean-variance trend is enforced by simulating biological coefficient of variation (BCV) values. Zero-inflation is modeled through a logistic function, determining the probability of zero counts.

We used the following 6 combinations of parameters to generate simulated data with 200 genes and 2000 cells: group numbers = 2, dropout midpoint = 1; group numbers = 2, dropout midpoints = 3; group numbers = 2, dropout midpoints = 5; group numbers = 5, dropout midpoint = 1; group numbers = 5, dropout midpoints = 3; group numbers = 5, dropout midpoints = 5. And all 5 denoising methods including baseline, multinomial, DCA, ZINB, Weight MSE and Mask MSE were performed on the 6 simulated data.

Interpretation of results

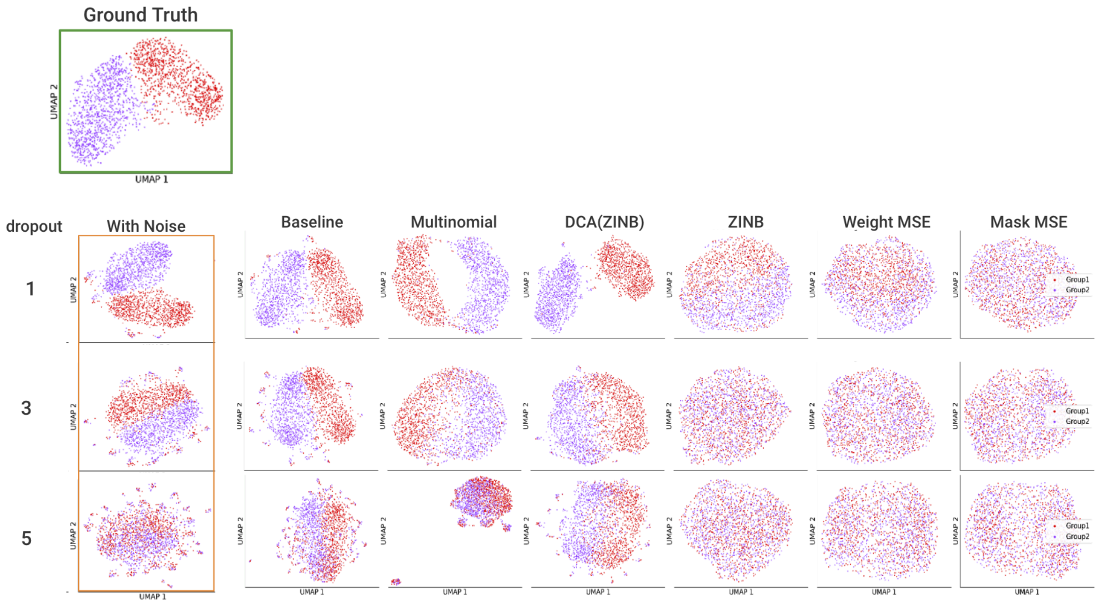


Figure 1. The UMAP plot of different denoising methods across various noise conditions (group numbers = 2)

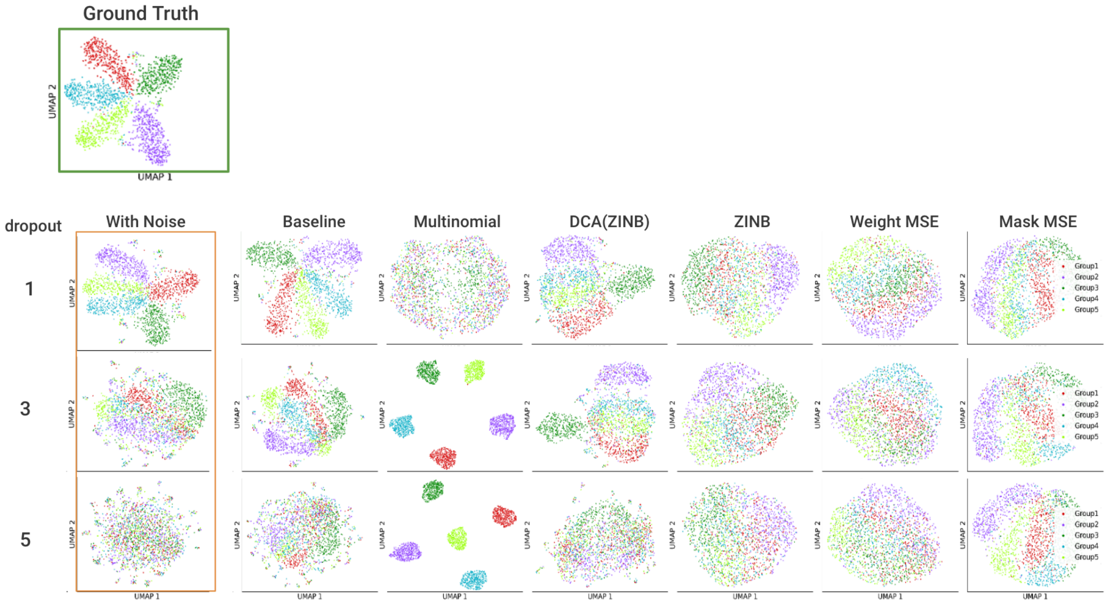


Figure 2. The UMAP plot of different denoising methods across various noise conditions (group numbers = 5)

From the UMAP plot of different denoising methods implemented on the simulated data with two groups, we can evaluate the performance of different denoising methods under diverse conditions. In the UMAP plots of simulated data with 2 groups (**Figure. 1**), the ground truth shows two distinct clusters in the UMAP space, representing the two simulated groups. It provides a baseline expectation for how the denoised data should ideally appear if the denoising methods are effective. The UMAP plots for data with different dropout midpoints show how varying dropout midpoints (1, 3, and 5) increasingly disperse the cluster. As the dropout midpoint increases, the separation between the two groups diminishes, illustrating the impact of noise on data separation. From the UMAP plots for various denoising methods, we can see that the baseline method shows moderate recovery of the original cluster shapes, particularly when dropout midpoint is high. While the multinomial and DCA(ZINB) both showed relatively good improvement in cluster definition compared to the baseline. DCA, which utilizes zero-inflated NB distribution modeling, appears slightly more effective in maintaining cluster integrity against noise especially in high-noise scenarios. ZINB, Weight MSE and Mask MSE all show poor recovery of the original cluster shapes.

Similar results are observed in simulated data with 5 groups (**Figure. 2**). The baseline method continues to show moderate effectiveness. It fails to completely recover the distinct cluster of five groups, especially as dropout midpoint increases. The multinomial method and the DCA(ZINB) method demonstrate relatively strong capability to preserve group integrity and separation in different noise scenarios. While the multinomial methods have problems recovering the precise distinction among all five groups under lower dropout conditions, the DCA(ZINB) method struggles to fully recover the separation as noise increases. The ZINB, Weight MSE and Mask MSE all show limited success in reconstructing the original clusters across both two-group and five-group setups, especially in scenarios with higher dropout midpoints.

To sum up, the simulation study highlights the importance of choosing methods including multinomial and DCA(ZINB) for scRNA-seq data denoising, particularly when dealing with complex datasets that include multiple groups and high levels of dropout. These methods not only improve the clarity of clustering in UMAP visualizations but also enhance the reliability of downstream analyses by better preserving the integrity of the original data. This is crucial for accurate biological interpretation and discovery in studies relying on single-cell RNA sequencing data.

Potential Problems and Alternative Approaches

While the simulation study provides valuable insights into the effectiveness of various scRNA-seq denoising methods, several potential problems could affect the reliability and generalizability of the results. Firstly, the splatter package, while comprehensive, might incorporate assumptions and simplifications that do not fully capture the complexity and variability of real scRNA-seq data. We can consider using multiple simulation tools or to introduce real dataset benchmarks alongside simulated data to validate the denoising methods under more varied and realistic conditions. Secondly, from our results of performance evaluation of different denoising methods, we can see that when there are many groups, distinguishing between closely related groups becomes challenging, and some denoising methods may fail to make good separations. In the future, we may consider working on developing or incorporating advanced denoising methods to enhance the distinction between groups.

Discussion

Imputation and denoising are essential for removing technical noise in the RNA expression data and improving downstream analyses, ultimately contributing to the dissection of complex biological systems. According to our results, denoising approaches that combine statistical models with an autoencoder framework performed the best in downstream tasks including clustering and classification, suggesting their superior capability in handling dropouts and discerning biological signals from the raw noisy input.

However, caution is needed when generalizing these methods to a broader application. For example, autoencoders with a Multinomial loss perform denoising rather than imputation–this is because, while showcasing a prominent ability reducing the noise, its recovered matrix values are much smaller in scale than the original input. ZINB is also argued to fail at representing the true underlying RNA expression in cases where there are fewer dropouts (e.g., UMI count) and that alternative distributions such as NB should be considered.

In addition, although the MSE loss demonstrated an overall poor performance on downstream analysis, there is still potential to utilize it by adding regularizations to the loss functions to constrain the sparsity of imputed values in matrix completion.

Finally, to conclude, some of the possible future extensions to our work include (1) incorporating more datasets to enhance the robustness of denoising and imputation methods, (2) evaluating the performance upon more downstream tasks, such as differentially expressed genes, to enrich the understanding of the impact of denoising on data clarity, (3) refining models with regularization, hyperparameter search, and neural network settings such as adaptive learning rate and early stopping mechanisms, and (4) structuring our analysis pipeline into an end-to-end solution–for instance, adding a few more stacked layers to the denoising autoencoder for automatic cell type classification/ annotation. These implementations will allow us to generate more rigorous findings and facilitate reproducible research, which can help better address the problem of technical noise in biological data.

Reference

[1] Eraslan, G., Simon, L.M., Mircea, M. et al. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun 10, 390 (2019). https://doi.org/10.1038/s41467-018-07931-2

[2] Chen L, Wang W, Zhai Y and Deng M (2020) Single-Cell Transcriptome Data Clustering via Multinomial Modeling and Adaptive Fuzzy K-Means Algorithm. Front. Genet. 11:295. doi: 10.3389/fgene.2020.00295

[3] Talwar D, Mongia A, Sengupta D, et al. AutoImpute: autoencoder based imputation of single-cell RNA-seq data. Sci Rep 2018;8(1):1–11. https://doi.org/10.1038/s41598-018-34688-x

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

Appendix

Figure 1. Clustering Results by Denoise Methods

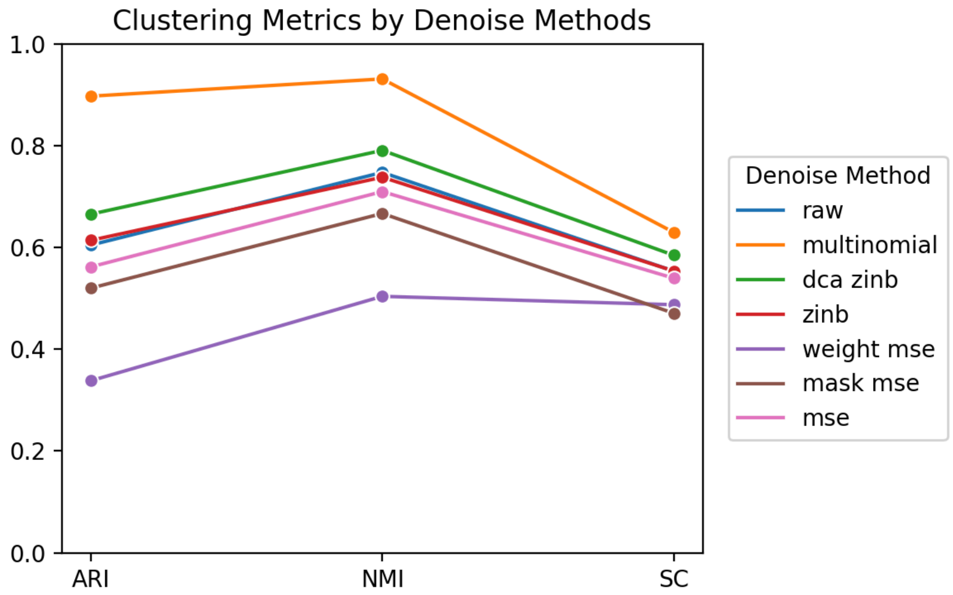
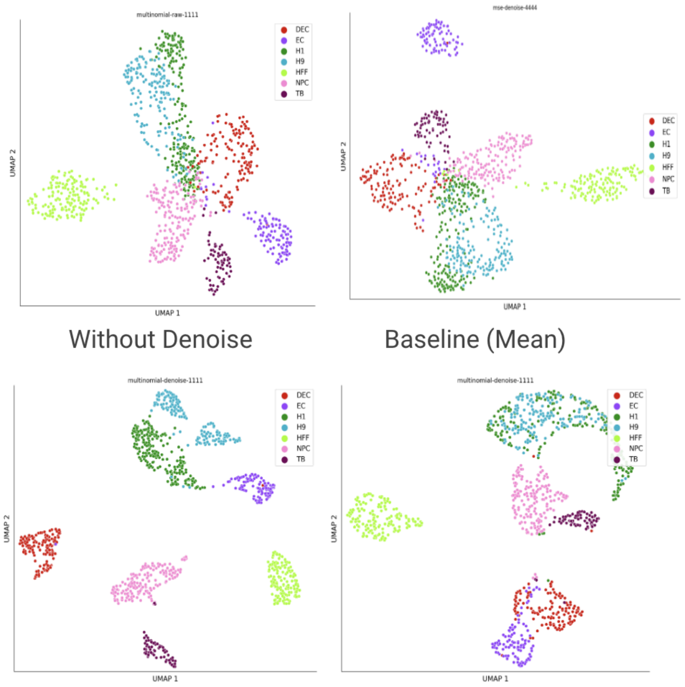


Figure 2. Low dimensional denoised data using UMAP



Without Denoise

Baseline (Mean)

Multinomial

ZINB

Figure 3. Classification Results by Denoise Methods

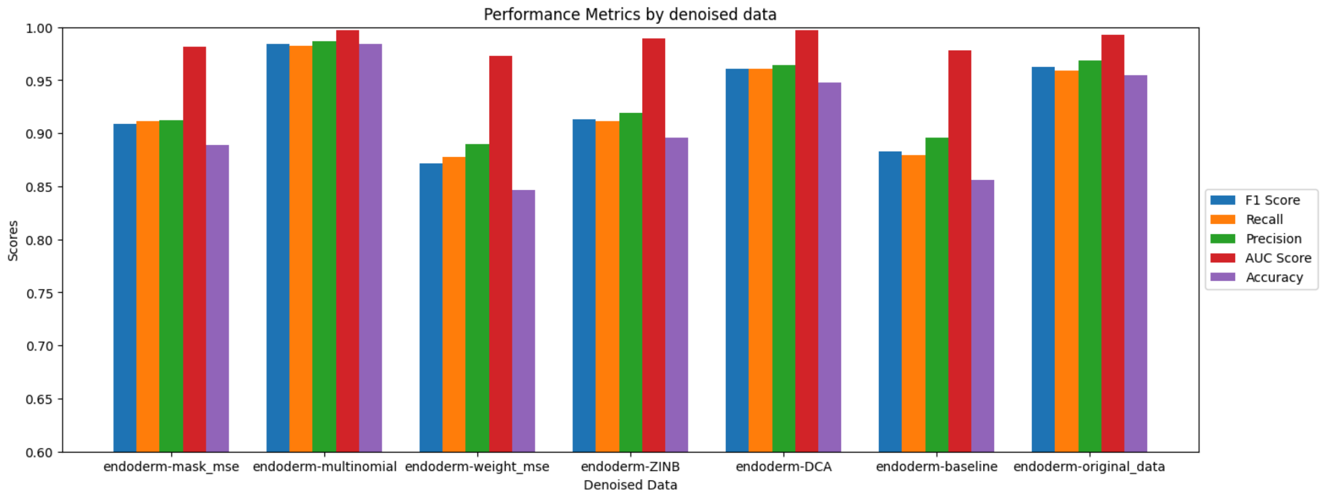

Performance Metrics by denoised data
Scores
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
endoderm-mask_mse
endoderm-multinomial
endoderm-weight_mse
endoderm-ZINB
endoderm-DCA
endoderm-baseline
endoderm-original_data
Denoised Data
F1 Score
Recall
Precision
AUC Score
Accuracy

Figure 4. Feature Importance from Raw Data (Top 50)

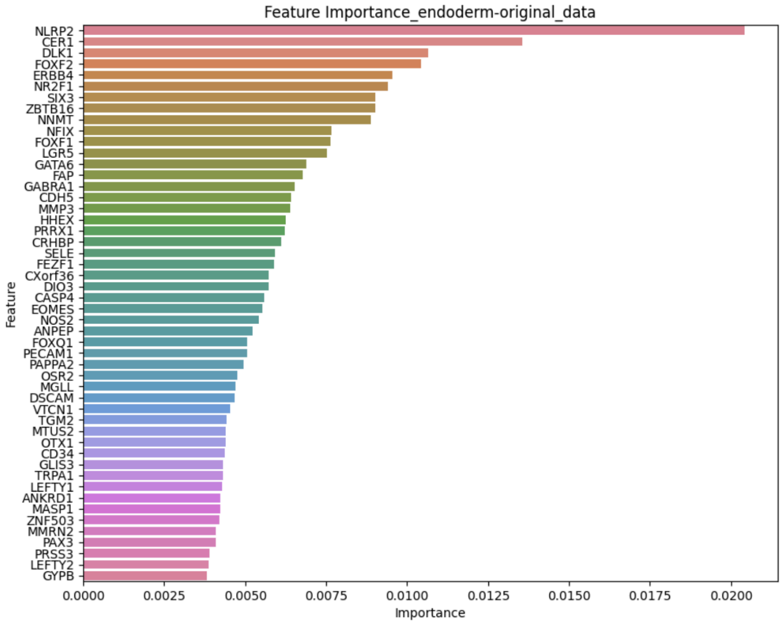


Figure 5. Feature Importance from Denoised Data using Multinomial (Top 50)

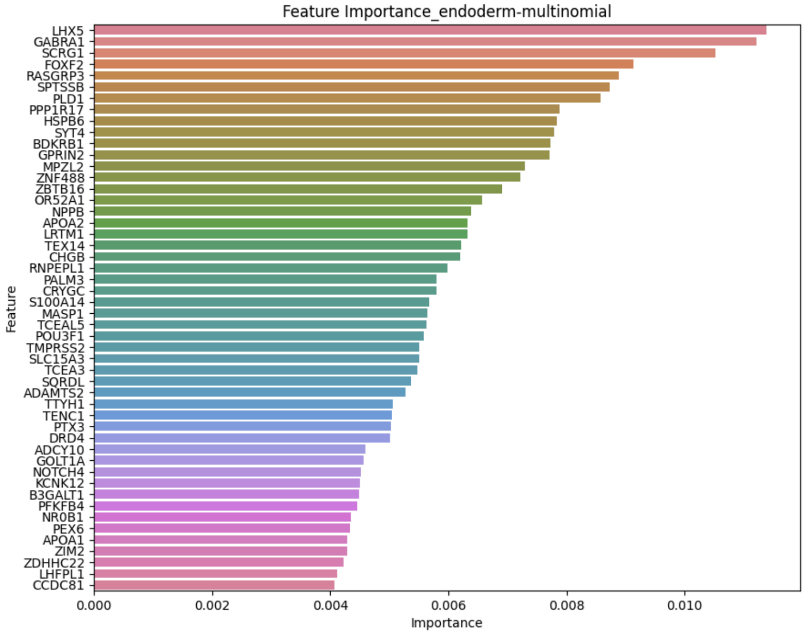


Table 1. Clustering Results by Denoise Methods

	ARI	NMI	SC

Raw	0.6045	0.7472	0.5534
Multinomial	0.8970	0.9308	0.6295
DCA ZINB	0.6652	0.7905	0.5845
ZINB	0.6138	0.7377	0.5535
Weight MSE	0.3380	0.5037	0.4873
Mask MSE	0.5199	0.6665	0.4706
MSE	0.5617	0.7095	0.5394

Table 2. Classification Results by Denoise Methods

	F1	Recall	Precision	AUC score	Accuracy
Raw	0.9620	0.9588	0.9681	0.9925	0.9542
Multinomial	0.9842	0.9824	0.9866	0.9970	0.9837
DCA ZINB	0.9604	0.9603	0.9644	0.9973	0.9477
ZINB	0.9134	0.9115	0.9187	0.9889	0.8954
Weight MSE	0.8717	0.8780	0.8896	0.9726	0.8464
Mask MSE	0.9090	0.9110	0.9121	0.9817	0.8889
MSE	0.8830	0.8793	0.8957	0.9781	0.8562