# Single-cell RNA-seq data feature extraction using dual-depth model

Xiuxiu Su, Shuran Mo, Yang Zhao, Faning Long*
School of Computer Science and Engineering, Yulin Normal University, Yulin 537000, Guangxi, China

## ABSTRACT

In the realm of single-cell transcriptomic sequencing, deep generative models have proven invaluable in capturing gene expression features. Nevertheless, technical challenges have introduced a notable presence of missing values in the data, leading to the observed "dropout" phenomenon within the gene expression matrix. This phenomenon is characterized by numerous technical zero values, potentially stemming from data noise. To address this issue, interpolation algorithms leverage known values to infer and fill in these "dropout" occurrences, effectively mitigating data incompleteness and aiding in the preservation of biological information within the samples. Relevant studies suggest that interpolation algorithms play a crucial role in enhancing the reliability and completeness of data in the context of feature extraction within deep models. To contribute to this area, this research introduces scDIVAE, a framework encompassing two deep generative models. The first model is dedicated to interpolating gene data, sharing information among similar cells to eliminate noise and the "dropout" phenomenon. The second model employs a natural language topic model for data feature extraction. This methodology not only improves the clustering accuracy of deep generative models but also effectively eliminates batch effects.

**Keywords:** Single-cell RNA-seq, interpolation algorithm, deep generative model, feature extraction

## 1. INTRODUCTION

Single-cell sequencing technology holds promising prospects in the biomedical domain, aiding a comprehensive understanding of cellular functionalities and heterogeneity[1]. The utilization of scRNA-seq has significantly advanced research in biological processes and human diseases, leading to a paradigm shift in genomics. In experiments, scRNA data often represent high-dimensional sparse genomic data, where over 80% of the values are missing (zero values). These undetectable zero values are termed as "dropout"[2]. The "dropout" phenomenon is prevalent in single-cell RNA sequencing data, stemming from detection limitations or instrument malfunctions, resulting in a plethora of missing values, posing challenges to data integrity and reliability. During the training of deep learning models, these missing values can potentially affect the stability and predictive capabilities of the models. Hence, specialized modeling techniques and data handling strategies are essential to rectify data incompleteness, consequently enhancing the models' performance and interpretability.

## 2. RELATED WORK

In recent years, the academic community has proposed multiple imputation algorithms to tackle the prevalent "dropout" phenomenon observed in single-cell RNA sequencing (scRNA-seq) data. An influential contribution comes from van Dijk et al., who introduced MAGIC[3]. This method relies on a Markov affinity matrix and exhibits exceptional performance in reconstructing gene relationships and other structural aspects within scRNA-seq data. Another noteworthy imputation algorithm is DrImpute[4], which leverages clustering techniques. DrImpute excels in imputing scRNA-seq data, particularly in distinguishing between missing zero values and genuine zero values. These imputation algorithms, including MAGIC and DrImpute, focus on various aspects of cellular interaction or model inference to fill in missing values within scRNA-seq datasets. While these algorithms have demonstrated effectiveness in enhancing the quality of original datasets and preserving biological differences, a significant challenge arises due to their higher time complexity. This complexity limits their practical application, especially given the continuously expanding volume of single-cell sequencing data. The Deep Count Autoencoder network (DCA)[5] presents a tailored imputation algorithm specifically designed for scRNA-seq data. DCA concentrates on addressing the "dropout" phenomenon occurring

*longfaning@163.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133954M · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049218

between genes and cells by employing a zero-inflated negative binomial noise model. By addressing issues related to data count distributions, over-dispersion, and sparsity, DCA aims to reconstruct omitted values within large-scale datasets. Consequently, it provides a more precise depiction of the characteristics and correlations within single-cell RNA sequencing datasets.

While imputation algorithms play a crucial role in mitigating noise and addressing the "dropout" phenomenon in data, relying solely on these techniques for the original data fails to significantly enhance model performance. This limitation arises from challenges in model transferability, interpretability, scalability issues, and difficulties associated with batch effects. In recent years, researchers have proposed various deep models based on the Variational Autoencoder (VAE)[6] to conduct large-scale comprehensive analyses of single-cell RNA sequencing (scRNA-seq) data. Models such as scVI[7] consider library size and batch effects, while scVAE-GM[8] modifies the prior distribution of latent variables in VAE and introduces classification latent variables for cell clustering. LDVAE[9] incorporates a linear decoding layer during model training to enhance interpretability. Auto-cell[10] combines graph embedding and probabilistic deep Gaussian mixture models for inferring the distribution of high-dimensional sparse scRNA-seq data. Despite their ability to effectively extract features and identify patterns within the data, VAE-based models lack interpretability, necessitating further analysis to decipher the significance of model parameters.

This paper introduces scDIVAE, a dual-depth model that combines a Variational Autoencoder (VAE) with a deep neural network-based imputation algorithm, leveraging the effectiveness of VAE in extracting features and identifying patterns from high-dimensional sparse genomic data. The core architecture comprises two key components: a deep neural network-based imputation model for interpolation and recovery of data, addressing missing or sparse occurrences, and a deep generative model based on VAE for feature extraction and cell clustering. Initially, scDIVAE employs a divide-and-conquer approach, constructing multiple sub-neural networks with dropout layers to learn patterns within the data. This process aims to eliminate noise from the cell count matrix and fill in "dropout" data, enhancing data quality for subsequent input into deep models. Following preprocessing involving noise elimination, the data is fed into a transferable neural network-based encoder and an interpretable linear decoder. These components compute embedding vectors for cells, enabling downstream tasks such as clustering, differential expression analysis, and enrichment analysis. By utilizing a dual-depth model, this paper addresses deficiencies in traditional models' feature extraction capabilities, enhancing the generalization of deep learning models and overcoming limitations in local feature extraction abilities typical in conventional deep models. The model inputs interpolated gene expression data, obtaining highly interpretable feature embedding vectors, thereby improving clustering precision within the deep model and effectively removing batch effects.

## 3. METHODS

Before delving into the intricacies of the complexity inherent in our proposed model, scDIVAE, it is necessary to recognize the significant contributions made by various existing deep learning models in addressing the complexities inherent in the analysis of single-cell RNA sequencing (scRNA-seq) data. These models have laid a solid foundation for our approach, which strategically employs a dual-depth architecture to seamlessly integrate imputation and generative modeling, thus overcoming inherent limitations in traditional methodologies. To gain a thorough understanding of our innovative framework, we will now proceed to provide a detailed exposition of the scDIVAE architecture.

### 3.1 Overall design

The overall architecture of scDIVAE encompasses two pivotal models. The first key model is the DCA (Deep Count Autoencoder) model, specifically designed for imputing single-cell RNA sequencing data. This model aims to address the "dropout" phenomenon, where a substantial number of missing values exist in the gene expression matrix due to data noise. The DCA model employs a zero-inflated negative binomial distribution model, intending to reveal latent relationships between genes and cells while effectively handling the data's excessive sparsity and sparseness. The second key model of scDIVAE is based on the Variational Autoencoder (VAE), with the topic modeling section utilizing a neural network to model samples in the latent space of single-cell RNA sequencing data and extract the probability distribution of topics. In single-cell RNA sequencing data, selecting Highly Variable Genes (HVGs) is a common strategy to identify genes with significant variations in the dataset. scDIVAE employs a criterion based on the variance-to-mean ratio exceeding 0.5 for HVG selection.

Figure 1 provides a clear depiction of the workflow of scDIVAE. Firstly, it identifies HVGs meaningful for subsequent analysis and undergoes data preprocessing, including regularization. Then, the DCA model is employed to complete the imputation task for single-cell data, addressing issues of excessive sparsity and the "dropout" phenomenon while

handling inconsistencies resulting from missing values. The imputed data fills in the missing values, thereby enhancing data integrity for more comprehensive and reliable downstream analyses. To handle the data more effectively, the DCA model divides genes into N random subsets, each containing S genes referred to as "target genes." Each subset establishes a four-layer neural network model, including genes related to the target genes as the input layer, a fully connected hidden layer with 256 neurons, a 20% dropout layer to prevent overfitting, and an output layer for the target genes. Subsequently, scDIVAE inputs the imputed data into the VAE's topic model. This model employs a word embedding method, inspired by the model design of scETM, to further decompose the gene distribution matrix into topic embedding and gene embedding matrices. This allows for a deeper exploration of the relationships between all genes and topics in the embedding space, where "topics" represent latent features or patterns in the data, reflecting specific characteristics or behaviors of cells in gene expression. The encoder in the topic model of scDIVAE is responsible for dimensionality reduction and calculating the topic distribution for each cell, which is valuable for tasks like cell clustering. Once the model is trained, the encoder effectively extracts features and performs tasks such as cell clustering. Simultaneously, the decoder attempts to reconstruct the topic distribution into an expression matrix in an interpretable manner, including topics and gene embeddings, as well as parameters for batch effect correction. The topic modeling section utilizes the softmax function to represent the topic distribution and employs the cross-entropy loss function to measure the cross-entropy between the generated topic distribution and the true data distribution. Experimental results indicate a significant impact on the performance of the topic model before and after data imputation.

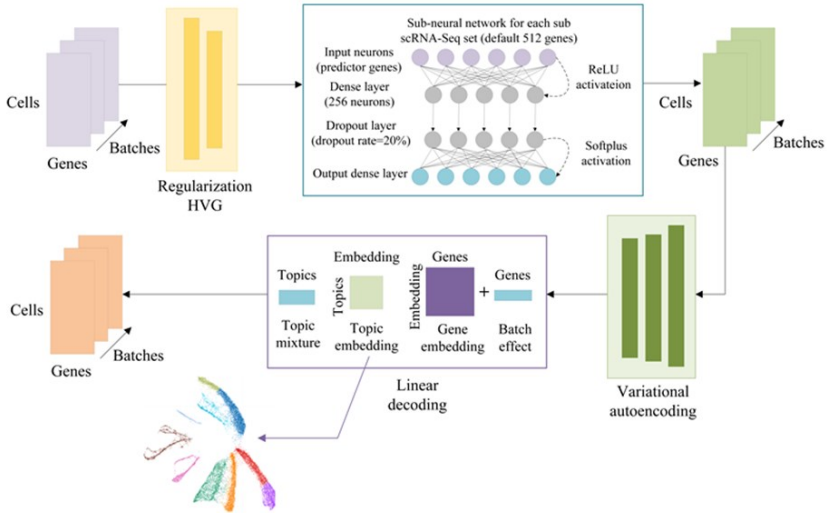


Figure 1. Overview of the scDIVAE framework.

### 3.2 Eliminate batch effects

scDIVAE utilizes a Variational Autoencoder as its foundational model, designed to acquire latent representations of the data. Its primary advantage lies in efficiently capturing the inherent distributional characteristics of the dataset. By projecting data onto a shared latent space, scDIVAE effectively mitigates batch effects, thereby enhancing both the interpretability and comparability of the data. This methodology retains biological variances within the dataset while minimizing noise stemming from batch effects. Compared to the scETM[11] model, scDIVAE typically demonstrates superior clustering quality and consistency.

### 3.3 Clustering and evaluation

scDIVAE conducts cell clustering by extracting latent features and employing the Leiden clustering method[12]. Subsequently, it assesses the accuracy of the obtained clustering results in comparison to models such as scVI, LDVAE, auto-cell, scETM, and scETM+HVG. The evaluation of clustering performance in experiments relies on adjusted rand index (ARI) and normalized mutual information (NMI) scores, precisely measuring the performance of each model.

### 3.4 Noise simulation evaluation

In this paper, 10%, 30%, 50% non-zero data values are randomly set to zero following a Gaussian distribution. In this noise simulation evaluation, scDIVAE shows excellent robustness. Even under the extreme conditions simulated, the

model maintains excellent performance. These results strongly demonstrate that the model has an excellent ability to cope with highly sparse and noisy data, can effectively maintain its performance level and accurately capture the characteristics of the data. This further emphasizes the reliability and robustness of scDIVAE when dealing with real single-cell transcriptome data, which strongly supports its application in complex environments.

# 4. RESULTS

We conducted extensive experimental validations on scDIVAE, with a primary focus on clustering and batch effect removal. Through these experiments, we assessed the model's performance in handling single-cell RNA sequencing (scRNA-seq) data and examined its accuracy at the cellular level in terms of clustering. Clustering analysis is a crucial task aimed at identifying potential cell subtypes or expression patterns within the data. By comparing scDIVAE's performance in clustering, we validated its outstanding capability in extracting intrinsic structural features and cell subtypes from the data.

Another key aspect of our investigation was batch effect removal, a common challenge in the analysis of scRNA-seq data. Batch effects can introduce inconsistencies in the data, affecting the accurate capture of real biological variations. Through experiments, we evaluated the effectiveness of scDIVAE in mitigating batch effects, ensuring that the embedding representations generated by the model exhibit improved transferability and robustness for subsequent analyses.

These experimental validations were conducted to comprehensively understand the performance of scDIVAE in processing scRNA-seq data, providing a solid foundation for its application in biological research. We emphasize the model's clustering capabilities and its resilience against batch effects, making it a powerful tool for handling complex single-cell datasets.

### 4.1 Batch effect elimination results

This study conducted batch effect removal experiments on seven datasets, including mouse pancreatic islets (MP), human pancreatic islets (HP), mouse liver (ML), human spleen (HS), human kidney (HK), mouse lung immune cells (MLI), and mouse synovial joints (MJ). It compared scDIVAE against five state-of-the-art models: scETM, scETM+HVG, scVI, LDVAE, and auto-cell. Additionally, UMAP visualization[13] was applied to the dimensionality reduction results of scETM, scETM+HVG, and scDIVAE models. (The description of the dataset is detailed in Table 1)

Table 1. Dataset description.

| ID | Datasets | Sequencing protocol | Sequencing protocol | Number of cells | Number of gene | Reference |
|---|---|---|---|---|---|---|
| GSE84133 | MP | scRNA-seq | 13 | 1886 | 14878 | 14 |
| GSE81076 | HP | scRNA-seq | 14 | 8569 | 20125 | 15 |
| GSM5009539 | ML | scRNA-seq | 10 | 2703 | 9776 | 16 |
| GSE151302 | HK | snRNA-seq | 20 | 19985 | 27146 | 16 |
| GSE194058 | MLI | scRNA-seq | 11 | 13172 | 32292 | 17 |
| GSE164430 | HS | scRNA-seq | 6 | 7505 | 52025 | 18 |
| GSE151985 | MJ | scRNA-seq | 7 | 2500 | 31065 | 19 |

Figure 2 visually demonstrates the scenarios of cells of the same type across different batches. The results illustrate that the scDIVAE model exhibits the highest consistency among cells of the same type across different batches. In terms of cell type identification and batch correction, scDIVAE demonstrates outstanding performance by more effectively sharing information between similar cells, restoring connections between genes, and reconstructing associations among genes. Hence, it possesses a more significant ability for cell type characterization compared to the scETM and scETM+HVG models.

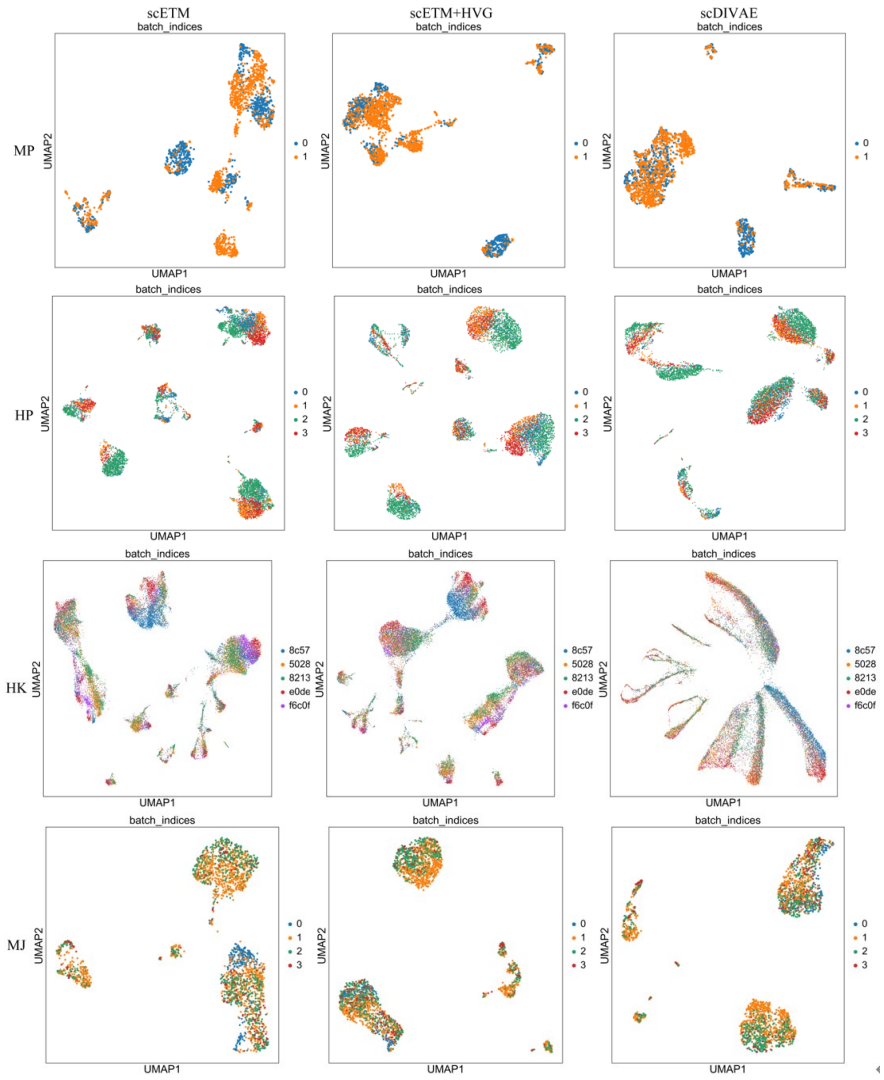


Figure 2. Batch effect elimination comparison.

### 4.2 Unsupervised clustering analysis results

During the experiments, the Leiden algorithm was employed to compare the cell embeddings generated by each model for cell clustering analysis. Multiple resolution values were attempted during the clustering process, and the best results for each model in terms of ARI and NMI were recorded. The title of each subplot in Figure 3 corresponds to the resolution yielding the highest ARI. The Leiden clustering results vividly demonstrate that utilizing the scDIVAE model for clustering output in the human pancreas dataset highly aligns with the predetermined cell type annotations (NMI=0.9503, ARI=0.9219). Its clustering performance significantly outperforms other models.

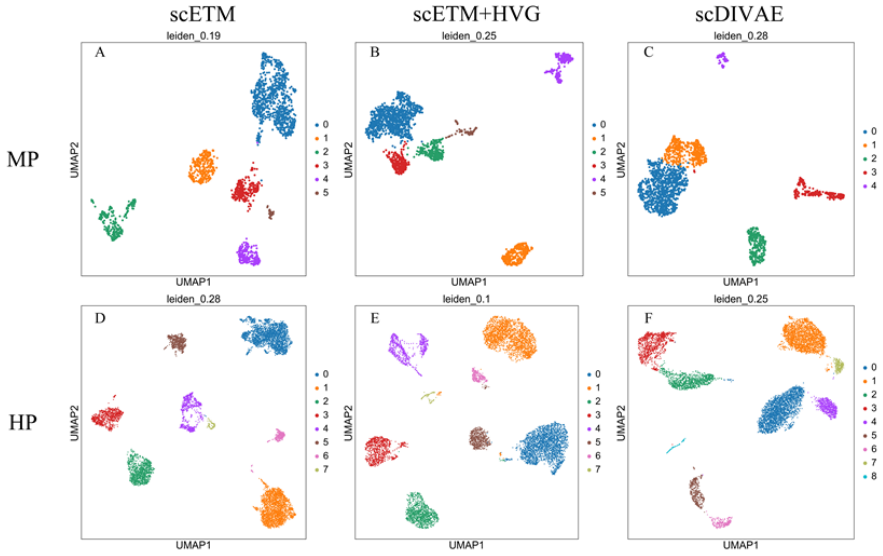


Figure 3. Unsupervised cluster comparison.

This experiment conducted ten unsupervised clustering analyses for the six models across all datasets, and recorded the average of the highest ARI and NMI scores from the ten experiments as the final data in Table 2. The results indicate that, across the seven datasets, the scDIVAE model achieved the highest ARI and NMI scores in six datasets. Its performance on the human kidney dataset significantly outperformed the other five models, demonstrating its high-level clustering performance. Additionally, the experiment assessed the separability between different clusters to ensure each cluster distinctly represents a unique subset in the data. The clustering results of scDIVAE exhibited exceptional separability, with very low similarity between clusters.

In the mouse pancreas dataset, both scVI and LDVAE models exhibited higher scores, whereas their performance was average on other datasets. The auto-cell model showed lower scores across all datasets and took the most time when training on larger datasets. The scETM+HVG model initially preprocessed data by extracting highly variable genes (HVG), and then trained the data using the scETM model. This method enhanced clustering accuracy and yielded the best results on the mouse synovial joint progenitor dataset. Although the improved scETM model achieved higher ARI scores than the original scETM model on most datasets, it failed to converge when handling the mouse liver dataset (recorded as NA in Table 2).

Table 2. Unsupervised clustering performance results.

| **Models/Metrics** | | **Datasets** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **MP** | **HP** | **ML** | **HK** | **MLI** | **HS** | **MJ** |
| scVI | ARI | 0.931 | 0.759 | 0.540 | 0.567 | 0.615 | 0.693 | 0.746 |
| LDVAE | ARI | 0.876 | 0.655 | 0.534 | 0.566 | 0.645 | 0.655 | 0.753 |
| auto-cell | ARI | 0.540 | 0.750 | 0.530 | 0.583 | 0.685 | 0.547 | 0.597 |
| scETM | ARI | 0.860 | 0.936 | 0.486 | 0.625 | 0.706 | 0.802 | 0.958 |
| | NMI | 0.756 | 0.901 | 0.508 | 0.805 | 0.781 | 0.701 | 0.923 |
| scETM+HVG | ARI | 0.903 | 0.931 | NA | 0.625 | 0.837 | 0.672 | 0.985 |
| | NMI | 0.849 | 0.892 | NA | 0.804 | 0.842 | 0.568 | 0.955 |
| scDIVAE | ARI | 0.923 | 0.940 | 0.819 | 0.628 | 0.843 | 0.759 | 0.988 |
| | NMI | 0.865 | 0.907 | 0.637 | 0.805 | 0.845 | 0.687 | 0.937 |

### 4.3 Analysis results of noise simulation evaluation

In the dataset of mouse synovial joint progenitors, the models scETM, scETM+HVG, and scDIVAE demonstrated notable performance, detailed in Table 2. To assess the model's data imputation capabilities, a noise simulation evaluation was conducted using a dataset comprising seven cell types of mouse synovial joint progenitor cells. Given the high-dimensional sparsity inherent in the single-cell dataset, 10%, 30%, and 50% of the non-zero data values were randomly converted to zeros following a Gaussian distribution to simulate real data loss scenarios.

The first column of Figure 4 presents the Leiden clustering plot based on the highest ARI score, while the second column displays the distribution of cells within each batch in the clustering results. The third column exhibits the clustering effects and cell type distributions. Experimental results indicate that, regardless of a 10%, 30%, or 50% dropout rate, the scDIVAE model consistently achieved the highest scores, displaying imputations closer to true expression values. Comparing the results of the three models (Figure 4 and Table 3), it's evident that scDIVAE excels in interpolation performance and feature extraction, displaying slightly superior overall performance compared to other models.

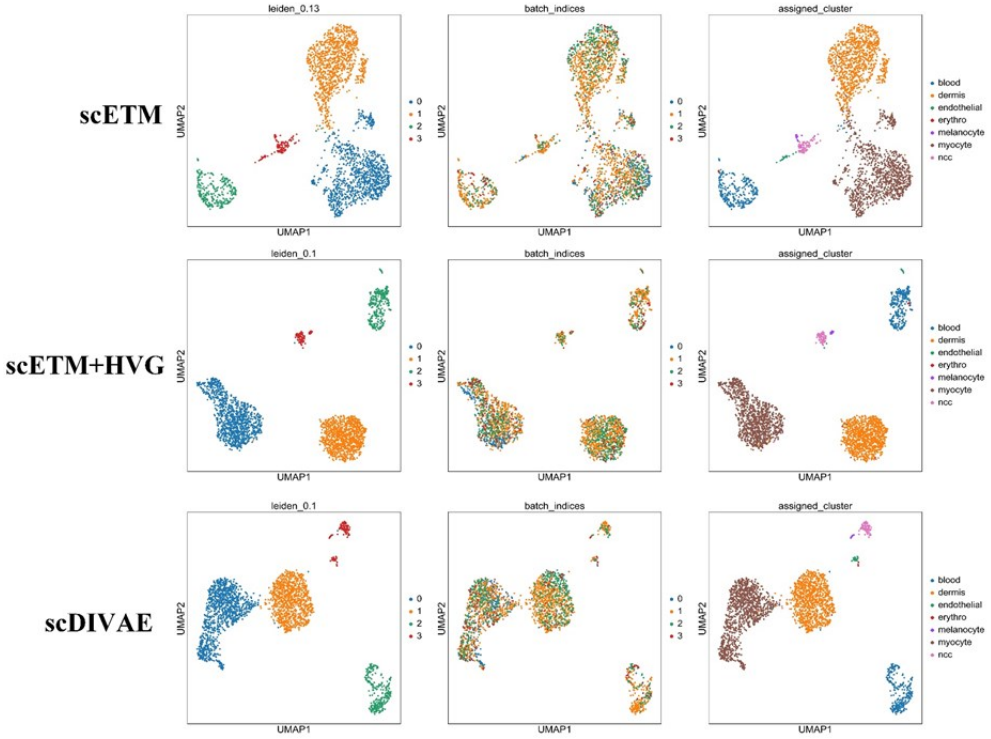


Figure 4. Model results with dropout rate of 30%.

Table 3. The score of the model at three dropout rates.

| Dropout rate | | scETM | scETM+HVG | scDIVAE |
|---|---|---|---|---|
| 10% | ARI | 0.8456 | 0.9457 | 0.9632 |
| | NMI | 0.8327 | 0.9146 | 0.9301 |
| 30% | ARI | 0.9289 | 0.9670 | 0.9684 |
| | NMI | 0.8934 | 0.9263 | 0.9286 |
| 50% | ARI | 0.7816 | 0.9545 | 0.9580 |
| | NMI | 0.7273 | 0.9145 | 0.9242 |

### 4.4 The results of differential expression analysis

The experimental visualization of differential gene expression comprehensively portrays the gene expression levels across different datasets and distinct cellular clusters. Within the human pancreatic dataset, as depicted in Figure 5, the visualized outcomes revealed a pronounced enrichment of differentially expressed genes within cellular clusters such as beta, ductal, endothelial, and macrophage. This representation not only delineates the gene expression variations among these cellular clusters but also provides deeper insights into their substantial disparities in biological characteristics and functionalities. Specifically, the experiment visually showcases the presence of pivotal differentially expressed genes, emphasizing their specificity in expression across distinct cellular clusters.

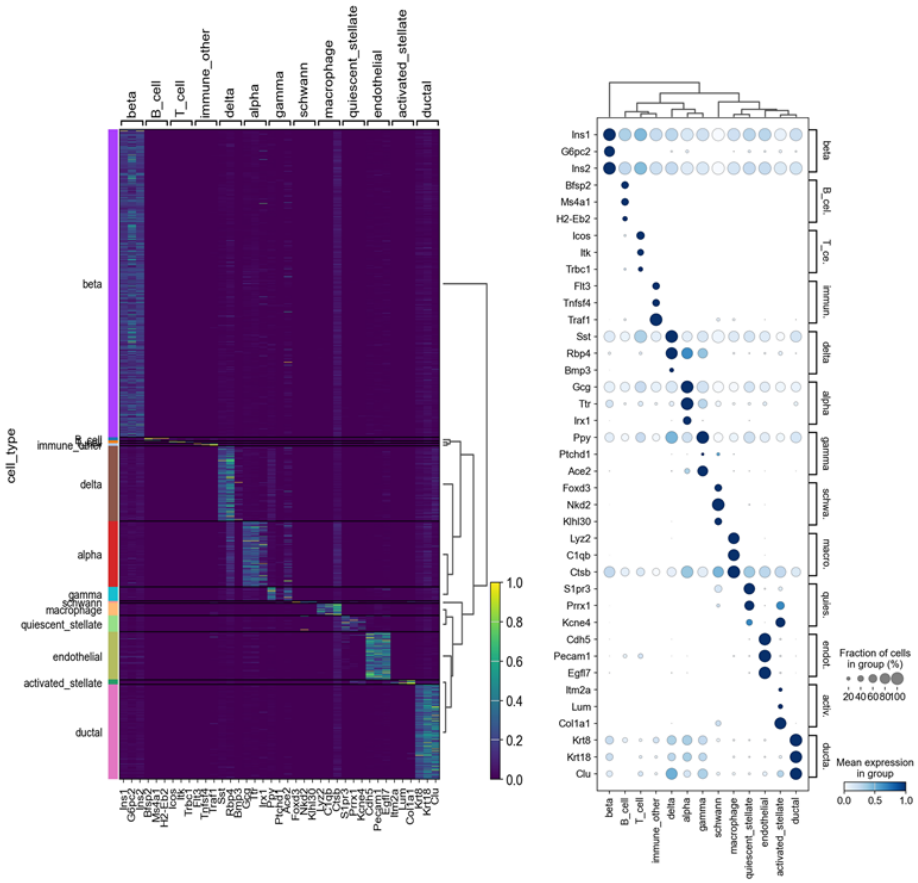


Figure 5. Differential expression results in the human pancrea dataset.

These visualized outcomes augment our understanding of the gene expression characteristics within cellular subpopulations, offering an intuitive and credible basis. By identifying these highly enriched differentially expressed genes across various cellular clusters, the experiment delves deeper into the functional differentiation of human pancreatic cell populations and the molecular features among cellular subgroups, which holds significant implications for understanding cellular differentiation, developmental processes, and mechanisms underlying disease progression.

However, despite the insights gained from the differential expression analysis, further functional validation and biological experiments are imperative to ascertain their specific biological relevance within pancreatic cell populations and to validate these findings in a broader context.

## 5. CONCLUSION

This study, based on single-cell sequencing data, explores the clustering analysis and batch effect removal performance following single-cell feature extraction using a dual-depth model. According to the experimental results, scDIVAE

demonstrates not only higher overall accuracy across various metrics and extensive validation methods, but also exhibits faster computational speed, and requires less computer memory.

In terms of eliminating batch effects, the introduction of an adversarial loss function aims to mitigate the adverse impact of batch effects in data analysis, while maintaining the model's ability to correct for these effects. The scDIVAE model, employing this adversarial learning mechanism, more effectively restores the correlations among genes. It demonstrates a more discriminative cellular type description compared to the models scVI, LDVAE, auto-cell, scETM, and scETM+HVG. Importantly, it retains the capability for batch effect correction, showcasing its robust performance in both cell type discrimination and batch correction benchmarks.

In terms of denoising, the scDIVAE model employs reparameterized Gaussian distributions for sampling latent variables, allowing for an estimation of noise within the variational expectation. This approach effectively mitigates noise influence while maximizing the variational lower bound expectation. The optimization process adjusts model parameters, such as encoder weights, topics, and gene embeddings, through gradient backpropagation to enhance noise robustness in the data. This refinement enables the model to accurately capture genuine data features, free from the interference of noise.

In conclusion, scDIVAE effectively restores missing gene expressions within single-cell data. Its neural network-based interpolation algorithm efficiently mitigates batch effects and noise, resulting in excellent preprocessing outcomes. This enhancement in data quality elevates the model's performance in single-cell feature extraction tasks.

## ACKNOWLEDGMENTS

This work was supported by the Innovation and Entrepreneurship Training Program for college students (Grant No. 202410606017).

## REFERENCES

[1] Brennecke, P., Anders, S. and Kim, J., "Accounting for technical noise in single-cell RNA-seq experiments," Nat. Methods 10(12), 1093-1095 (2013).
[2] Nitish, S., Geoffrey, H. and Alex, K., "Dropout: a simple way to prevent neural networks from overfitting," J. Mach. Learn. Res. 15, 1929-1958 (2014).
[3] Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P. and Carr, A. J., "Recovering gene interactions from single-cell data using data diffusion," Cell 174, 716-729.e27 (2018).
[4] Gong, W., Kwak, I. Y. and Pota, P., "DrImpute: imputing dropout events in single cell RNA sequencing data," BMC Bioinformatics 19, 220 (2018).
[5] Eraslan, G., Simon, L. M. and Mircea, M., "Single-cell RNA-seq denoising using a deep count autoencoder," Nat. Commun. 10(1), 390-414 (2019).
[6] Kingma, D. P. and Welling, M., "Auto-Encoding Variational Bayes," arXiv arXiv:1312.6114v11, (2013).
[7] Lopez, R., Regier, J. and Cole, M. B., "Deep generative modeling for single-cell transcriptomics," Nat. Methods 15, 1053-1058 (2018).
[8] Grønbech, C. H., Vording, M. F. and Timshel, P. N., "scVAE: variational auto-encoders for single-cell gene expression data," Bioinformatics 36(16), 4415-4422 (2020).
[9] Valentine, S., Adam, G., Nir, Y. and Lior, P., "Interpretable factor models of single-cell RNA-seq via variational autoencoders," Bioinformatics 36, 3418-3421 (2020).
[10] Xu, J., Xu, J. and Meng, Y., "Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data," Cell Rep. Methods 3(1), 100382 (2023).
[11] Zhao, Y., Cai, H. and Zhang, Z., "Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data," Nat. Commun. 12, 5261 (2021).
[12] Traag, V. A., Waltman, L. and van Eck, N. J., "From Louvain to Leiden: guaranteeing well-connected communities," Sci. Rep. 9, 5233 (2019).
[13] Mcinnes, L. and Healy, J., "Umap: uniform manifold approximation and projection for dimension reduction," The Journal of Open-Source Software 3(29), 861 (2018).

[14] Stuart, T., Butler, A. and Hoffman, P., "Comprehensive integration of single-cell data," Cell 177(7), 1888-1902.e21 (2019).
[15] Baron, M., Veres, A. and Wolock, S. L., "A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure," Cell Systems 3(4), 346-360 (2016).
[16] Lee, Y., Bogdanoff, D. and Wang, Y., "XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment," Sci. Adv. 7(17), eabg4755 (2021).
[17] Muto, Y., Wilson, P. C. and Ledru, N., "Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney," Nat. Commun. 12, 2190 (2021).
[18] MacLean, A. J., Richmond, N. and Koneva, L., "Secondary influenza challenge triggers resident memory B cell migration and rapid relocation to boost antibody secretion at infected sites," Immunity 55(4), 718-733.e8 (2022).
[19] Bian, Q., Cheng, Y. H. and Wilson, J. P., "A single cell transcriptional atlas of early synovial joint development," Development (Cambridge, England) 147(14), dev185777 (2020).