Producing Informative Cell-specific Data using Generative Artificial Intelligence

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Eric Xie

Spring, 2024

Technical Project Team Members

Hyun Jae Cho

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Aidong Zhang, Department of Computer Science

ABSTRACT

Single cell RNA sequencing, or scRNA-seq, acts as a potent analytical tool that allows for the comprehensive examination of gene expression profiles at a single-cell level. This methodology has many key applications, allowing biological researchers to better understand the specifics of the states of various cell types within different biological tissues. This, in turn, enables the precise identification of the specific cell types and the associated genetic profiles that underlie pathological conditions. The main drawback to this analytical tool is that scRNA-seq tends to be cost prohibitive and yields a relatively limited quantity of samples, especially in the context of human disease investigations. In contrast, there exists a wealth of an easily accessible alternative, bulk RNA-seq. However, bulk RNA-seq does not include any of the cell type specific information that is found within scRNA-seq data.

To address this divide and harness the potential of the abundant amount of bulk RNA-seq data, in this research endeavor, we introduce an innovative computational framework that capitalizes on the capabilities of generative AI techniques to effectively transform bulk RNA-seq data into scRNA-seq data. Our model, the “bulk to single cell” (Bulk2SC) variational autoencoder, is trained to deconvolute the aggregated bulk RNA-seq data into their individual single-cell transcriptomes by learning the specific distributions and proportions of each cell type. The potential implications of the Bulk2SC approach are particularly significant when applied to extensive human disease bulk RNA-seq datasets. Providing insights at the single cell level into the underlying mechanisms behind the disease processes is essential to furthering our understanding of diseases.

INTRODUCTION

The integration of artificial intelligence (AI) with biological data analysis is reshaping the cutting edge of research and clinical diagnostics. Our potential to decode complex biological systems has continued to grow alongside the capabilities of AI. This convergence of technology and biology shows promise for enhancing our biological understandings.

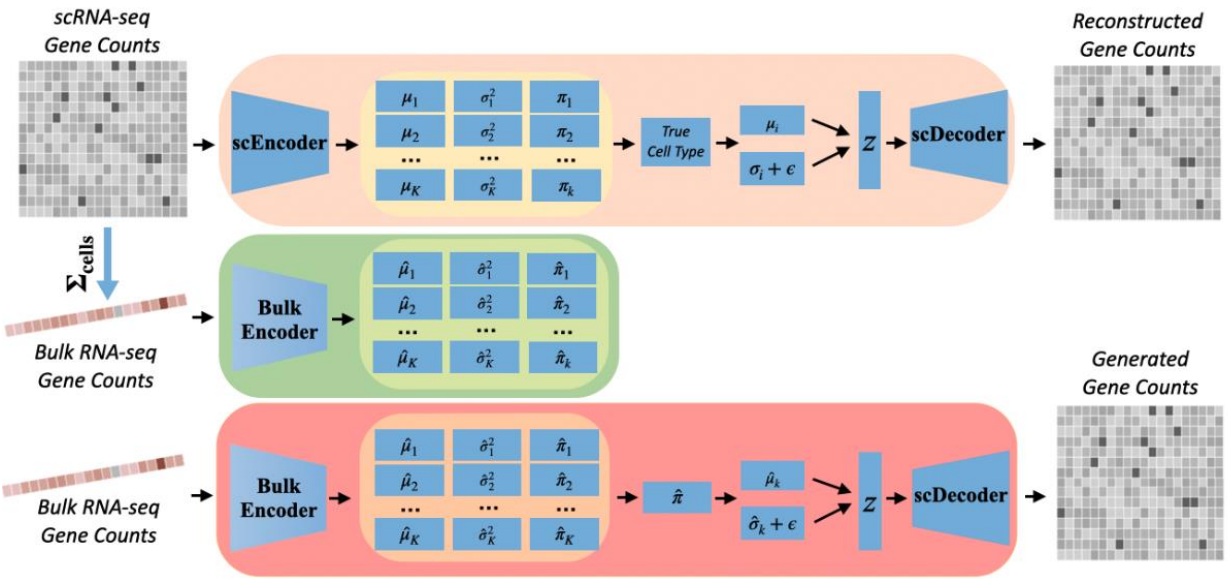
Single-cell RNA sequencing (scRNA-seq) is a data analysis technique that has revolutionized our ability to study cellular diversity and activity in health and disease (Haque et al., 2017). By providing gene expression profiles at the level of individual cells, scRNA-seq facilitates a deeper understanding of cellular functions and interactions within tissues. However, the immense benefits that scRNA-seq data can bring is limited by its high generation costs and technical complexity that limit its accessibility, especially in larger scale research (Kharchenko, 2021). In contrast, bulk RNA sequencing (bulk RNA-seq) offers a far more accessible, yet less detailed, alternative (Li & Wang, 2021). While it provides valuable insights into the overall gene expression of tissue samples, it fails to distinguish the contributions of individual cell types in the overall gene expression profiles. This limitation is a significant barrier for in-depth analysis of these tissues.

To address these challenges, this technical project introduces a novel computational approach that uses generative AI to transform the widely available bulk RNA-seq data into synthetic, but realistic scRNA-seq data. We use a “bulk to single cell” variational autoencoder (Bulk2SC), a model that learns to transform the mixed gene expression data and separate it into discrete single-cell profiles. This model architecture has proven effective in learning cell patterns (Grønbech et al., 2020; Xu et al., 2023). This approach extends the utility of existing bulk RNA-seq data while democratizing access to single-cell insights to aid discoveries in cellular biology.

By providing a method to generate detailed cellular insights from bulk RNA-seq data, Bulk2SC aims to enhance our understanding of diseases and foster the development of more targeted therapies.

METHODOLOGY

In the development of our computational model to transform bulk RNA-seq data into synthetic single-cell RNA-seq data, we employed a model architecture with three major components, each integrating several advanced machine learning techniques. The first component is the Single Cell Gaussian-Mixture Variational Autoencoder (scGMVAE). This component learns the distributional and Gaussian mixture parameters in the latent space, capturing the intrinsic cellular variability and heterogeneity. These parameters are essential for the accurate reconstruction of a single-cell profile (Xu et al., 2023). The second component is the Bulk RNA-seq Encoder. The Bulk Encoder can create an accurate compressed representation of any given bulk RNA-seq data by learning the cell type-specific estimates of proportions, means, variance as a function of bulk RNA-seq data. These representations are then passed to the final component, genVAE, which takes the compressed versions of bulk RNA-seq data as input and outputs the final synthesized scRNA-seq data. By integrating these three components, this model effectively bridges the gap between the deep insights offered by scRNA-seq data and the practicality of the bulk RNA-seq datasets.



Overall Model Architecture

Data preprocessing is another key step in the implementation of this model. Prior to its training, the raw data undergoes several preprocessing steps, such as normalization and scaling, to ensure that the input data is suitable for processing by the algorithms used later. These processing steps must ensure that no vital information is lost throughout the transformation process. These steps are essential for mitigating batch effects and any other potential discrepancies that may exist within the data, resulting in a more accurate and reliable analysis. Ensuring the data maintains its integrity and quality is the first step in maximizing the accuracy and consistency of the model's outputs. Following the data preprocessing, the training process of the model is designed to optimize the learning of specific distributions and proportions of each cell type. This is achieved using a set of scRNA-seq data as the training input, where ethe model learns to capture unique gene expressions patterns exhibited by individual cell types. The Bulk Encoder is required to accurately learn these parameters to interpret and deconvolve the bulk

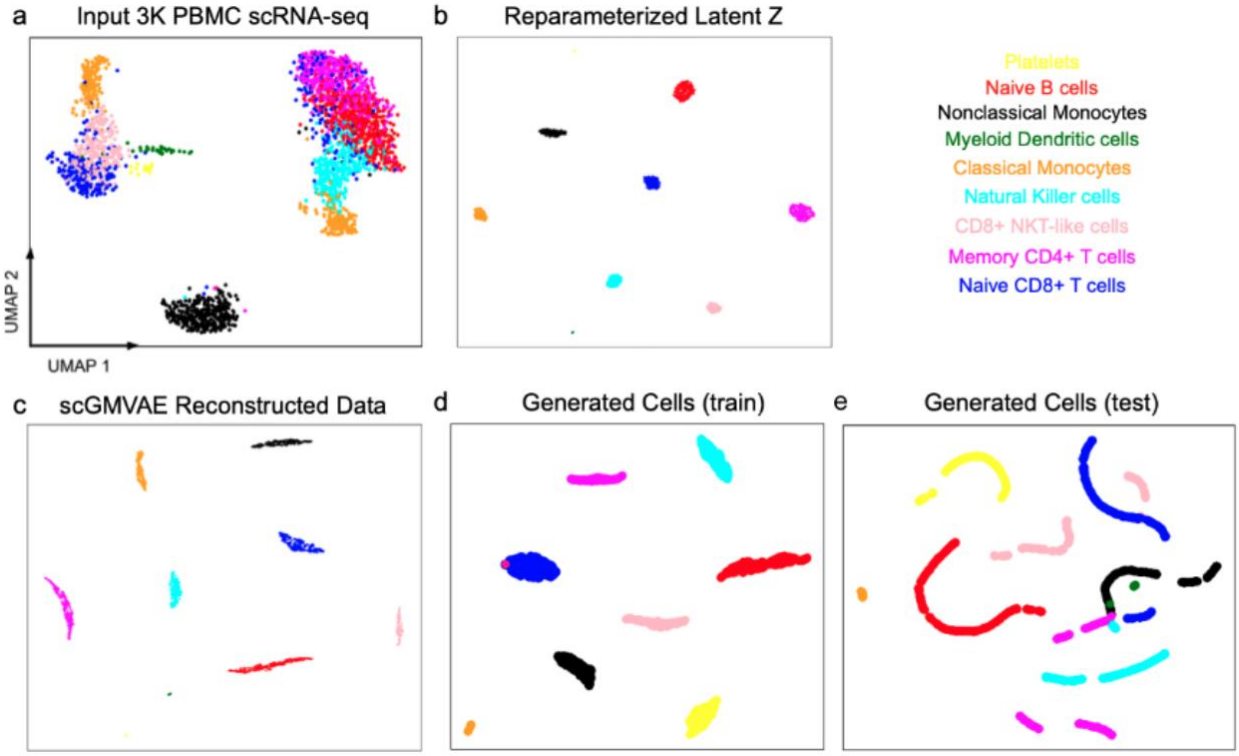
RNA-seq datasets. Once learned, the overall model is able to effectively reconstruct single-cell profiles from the bulk RNA-seq data.

The evaluation and validation of the model's performance are conducted through qualitative and quantitative measures. Qualitatively, we use Uniform Manifold Approximation and Projection (UMAP) plots to visually analyze the clustering of the different cell types in the generated single-cell data. Any patterns that are observed are then compared to those in actual scRNA-seq data. Quantitatively, this approach computes the cosine similarity and Pearson's correlation coefficients, measuring the resemblance between the gene interactions within the scRNA-seq data and the original scRNA-seq datasets. Both cosine similarity and Pearson's correlation coefficients are widely used to compare the overall structure in gene expression data (Chen et al., 2023; Jaskowiak et al., 2014). This provides insight into whether or not the relationships between each gene are preserved after transformation. These metrics help confirm that the model not only replicates the statistical distribution of the original data, but also maintains the biological information within the gene expression profiles.

RESULTS

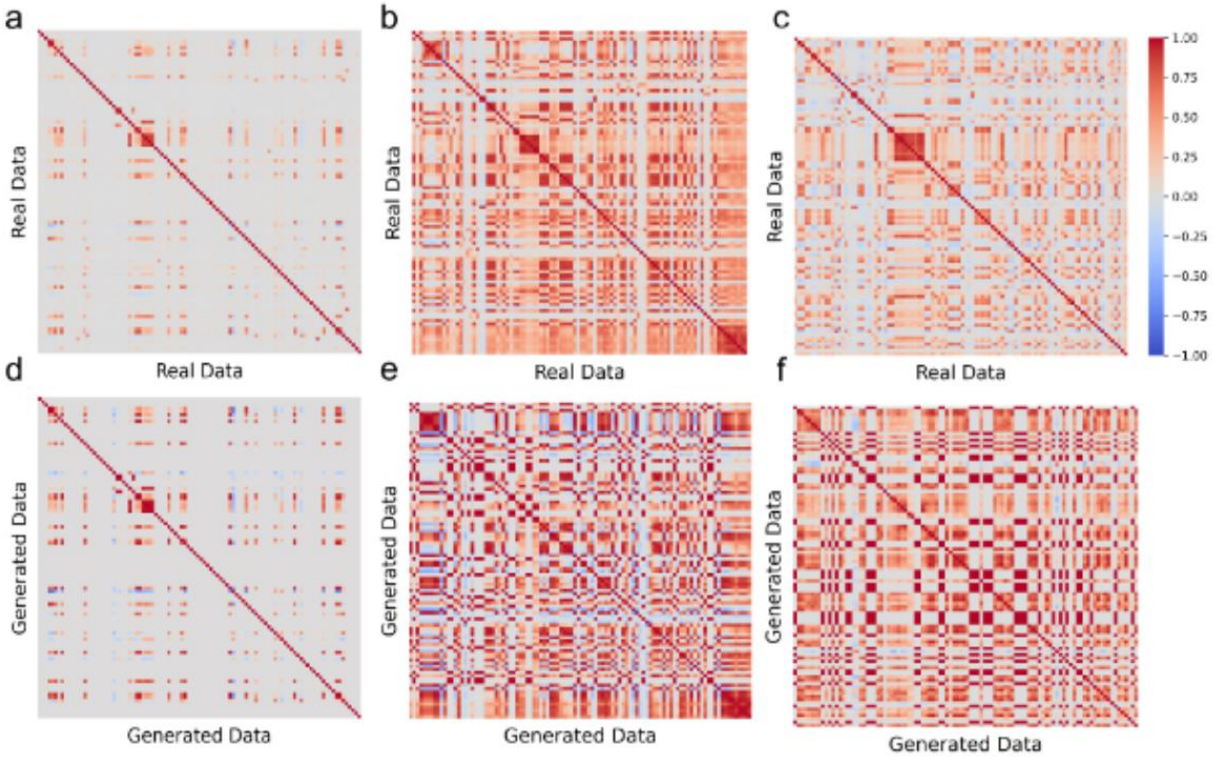
In this study, we establish a few metrics for evaluating the quality of reconstructed scRNA-seq data from bulk RNA-seq data. First, for our qualitative evaluations, we use Uniform Manifold Approximation and Projection (UMAP) visualizations, which effectively demonstrates our model's ability to accurately maintain the specific cell-type clusters that would normally be observed in authentic scRNA-seq data. This allows us to better understand how well the model can reconstruct the similarities across each cell type, the variation within each type, and the

number of counts for each. To further explore how well genetic relationships between cells are mimicked within synthetic data, we establish Pearson correlation heatmaps as an appropriate metric to show the overall trends and patterns of gene-to-gene relationships.



UMAP Visualizations

In this research, we also find the Pearson correlations between the 100 most variable genes. These genes were selected by isolating the top 10% genes with the highest mean expression levels, followed by identifying the 100 genes with the highest variance-to-mean ratio from within this subset. The level of similarity between the heatmaps for the generated data and the original data demonstrates the level to which gene-to-gene relationships are preserved.



Pearson Correlation Heatmaps

The first quantitative metric we use for measuring the effectiveness of the synthetic data is cosine similarity score. Higher values of cosine similarity act as a general indication of a high level of congruency between the generated data and the original scRNA-seq data. The downside of this metric is that it is influenced by the inherent sparsity within gene expression data, however, our model was still able to maintain extremely high cosine similarity scores in all comparisons and datasets.

PBMC	Input & Recon	Recon & Gen	Input & Gen
3K	0.99	0.99	0.98
10K	0.99	0.99	0.98
68K	0.99	0.97	0.96

Cosine Similarity Scores

To measure the preservation of linear relationships in the generated data, we use the average Pearson correlation coefficient across all genes. We use this metric to compare the three groups of data input and reconstructed data, the reconstructed and generated data, and the input and generated data. These scores range from [-1, 1], where scores closer to 1 indicates a better preservation of linear relationships between genes.

PBMC	Input & Recon	Recon & Gen	Input & Gen
3K	0.81	0.91	0.81
10K	0.82	0.77	0.82
68K	0.83	0.99	0.83

Average Pearson Correlation Coefficients

Similar to the Pearson correlation values, we find the correlation discrepancy (CD) between each set (Heydari et al., 2022). This metric ranges from [0, 198], where lower values indicates less discrepancy between the data. CD's underlying metric is the Spearman's rank correlation, which assesses monotonic relations rather than Pearson's linear relations. This makes the CD value sensitive to both linear and non-linear associations. This value is especially important in evaluating gene expression data since expression patterns tend to be complex and follow nonlinear trends.

PBMC	Input & Recon	Recon & Gen	Input & Gen
3K	48.82	45.41	45.33
10K	38.36	55.06	43.78
68K	72.38	93.62	54.01

Correlation Discrepancy Values

Lastly, we use the Integration Local Inverse Simpson's Index (iLISI) score for each dataset pairing to measure the level of integration between each dataset (Korsunsky et al., 2019). The iLISI score itself measures the level of integration. The capacity denotes the iLISI score that would imply maximal integration, which is determined by the ratio between the dataset sizes. "# PCs" lists the number of principal components used in each analysis, while the "% Var" describes the percentage of the total variance explained by the selected principal components. This metric gives insight into a few details regarding the generation quality. First, the overall iLISI score demonstrates how similar each generated cell type cluster is to its authentic counterpart. Furthermore, the number of principal components needed to reach our variance threshold of 90% describes the overall level of similarity between the data.

	iLISI	Capacity	% Var	# PCs
3k Train & Original	1.048	1.969	90.00	1638
3k Train & Test	1.114	1.368	92.62	4
3k Test & Original	1.065	1.290	90.00	1724
10k Train & Original	1.016	1.883	90.00	3667
10k Train & Test	1.642	1.647	93.52	3
10k Test & Original	1.035	1.890	90.00	3993
68k Train & Original	1.019	1.541	90.00	3993
68k Train & Test	1.471	1.471	92.95	3
68k Test & Original	1.023	1.146	90.00	4761

iLISI Score for Assessing Dataset Integration Quality

CONCLUSIONS

The extraction of real scRNA-seq data is costly, leading to overall limited dataset availability. This research presents a technique that leverages the abundance of bulk RNA-seq

data by converting it into scRNA-seq data. Our results indicate that bulk2sc is capable of successfully generating scRNA-seq data from bulk RNA-seq data with a sufficient degree of similarity to the real data in terms of cell type distributions and gene expression relationships. This research direction is known to be an extremely difficult and ambitious task, yet we are confident that this bulk2sc framework provides a strong foundation for building models capable of generating single-cell data from bulk RNA-seq datasets.

REFERENCES

Chen, J., Ng, Y.K., Lin, L., Zhang, X., Li, S.: On triangle inequalities of correlation-based distances for gene expression profiles. BMC Bioinformatics 24(1) (Feb 2023)

Grønbech, C.H., Vording, M.F., Timshel, P.N., Sønderby, C.K., Pers, T.H., Winther, O.: scvae: variational autoencoders for single-cell gene expression data. Bioinformatics 36(16), 4415–4422 (2020)

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome medicine, 9, 1-12.

Heydari, A.A., Davalos, O.A., Zhao, L., Hoyer, K.K., Sindi, S.S.: Activa: realistic single-cell rna-seq generation with automatic cell-type identification using introspective variational autoencoders. Bioinformatics 38(8), 2194–2201 (2022)

Jaskowiak, P.A., Campello, R.J., Costa, I.G.: On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics 15(S2) (Jan 2014)

Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. Nature methods, 18(7), 723-732.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.r., Raychaudhuri, S.: Fast, sensitive and accurate integration of single-cell data with harmony. Nature Methods 16(12), 1289–1296 (Nov 2019)

Li, X., & Wang, C. Y. (2021). From bulk, single-cell to spatial RNA sequencing. International Journal of Oral Science, 13(1), 36.

Xu, J., Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., ... & Cheng, F. (2023). Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. Cell Reports Methods, 3(1).