Journal Title Here, 2023, pp. 1–5

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

scCST: A continuous Spatial-temporal Transformer predicts cellular trajectories

Yuqi Lei, Zitian Tang, Yu Zhong, Shangyang Min and Yuyang Luo

Abstract

Single-cell sequencing techniques have unveiled new dimensions in gene expression analysis at the cellular level, offering insights into cell differentiation and development. We introduce scCST, a Continuous Spatial-Temporal Transformer, specifically designed to interpolate cellular trajectories between sampled time points, providing value information on the unknown data. This model outperforms conventional Transformers, delivering a more coherent and detailed representation of cellular progression across time points.

Our study encompasses a comprehensive evaluation of various model architectures. For projecting cells into latent space, we experimented with PCA, scVAE, and ZINB-WaVE. In cell alignment, we compared the Optimal Transport (OT) Algorithm for continuous alignment and the Jonker-Volgenant (JV) Algorithm for linear alignment. Further, we assessed the effectiveness of incorporating noise in our CST model against a standard Transformer setup. The results indicate that scVAE, combined with batch effect correction and discrete cell alignment, and employing a CST model without noise amplification, yields the most accurate tracking of cellular development. This optimal configuration significantly enhances the understanding of cellular dynamics, marking a substantial advancement in single-cell temporal analysis.

Introduction

Single-cell RNA sequencing (scRNA-seq) is a high-throughput-based technology that quantifies the expression level of each individual gene across different cells. It allows the study of the diversity of cell types within a tissue or sample and facilitates recovering the unknown cell type from sub-populations of cells. scRNA-seq has made it possible to investigate developmental landscapes through the examination of gene expression in individual cells collected at various timestamps during the differentiation process. Adding to this, integrating temporal information significantly enriches the understanding of cell differentiation by unveiling the dynamic transitions and trajectories that cells undergo over time. This temporal dimension provides a continuous perspective, enabling the capture of transient states and the unraveling of the sequential order of gene expression changes that govern cell fate decisions.

The recent computational methods involved with modeling how cells evolve stochastically and in physical time, have emerged as essential for a better understanding of the way that cells are driven into different states *in vivo*. In this study, we propose a transformer-based neural network named scCST that leverages latent representations from scRNA-seq datasets to interpolate expression profiles queried by the timestamp of interest. Through this approach, scCST harnesses both the rich gene expression data and the temporal metadata from scRNA-seq datasets, aiming to provide a more nuanced and accurate depiction of cellular dynamics across different developmental stages or conditions.

Related works

scRNA-seq datasets are often characterized by a high degree of zeros or missing values in the expression matrix. It is often due to the technical limitation of sequencing methods, such as genes that are expressed but fail to be detected. Such data sparsity issues might take negative effects on the downstream analysis.

ZINB-WaVE [7] is a novel method designed for the analysis of high-dimensional zero-inflated count data, such as those generated by single-cell RNA sequencing (scRNA-seq) assays. It provides a low-dimensional representation of the data, accounting for zero inflation, over-dispersion, and the count nature of the data. In the tasks of cell trajectory analysis, ZINB-WaVE plays a crucial role in preprocessing the scRNA-seq data before trajectory inference. The low-dimensional representation obtained from ZINB-WaVE serves as an input for trajectory inference tools like Slingshot. This preprocessing step is critical as it can enhance the accuracy and robustness of trajectory inference.

scVAE [5] proposed a Bayesian-based approach to model latent variables in scRNA-seq data. It was mainly composed of two parts: Generative Model and Inference Model. Generative Models aimed to generate observed expressed counts in scRNA-seq, and Inference Models took control of how each latent variable was learned based on observed variables. scVAE was optimized to learn a condition distribution of latent variables given observed data, which was done by maximizing the ELBO objective function (the evidence of lower bound). It leveraged the benefit of neural networks to parameterize the free

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

parameters to train the model accordingly. scVAE visualized the latent space of expression profiles in scRNA-seq and demonstrated that the model performance was better than other current methods.

A previous related work in single-cell temporal modeling is the Prescient model [9], a generative model approach that is designed to harness the time-based dynamics of scRNA-seq to anticipate cellular outcomes and declares can handle the perturbation well while other solutions can't. By using the continuous-time process to capture each cell's gene expression dynamics, and analyzing all cells together, Prescient can reduce a comprehensive range of potential cellular states for any future moment, Its application has led to the discovery of novel cellular states and intricate patterns in gene expression, underscoring its profound impact in the field [10].

Datasets

Currently, there are many useful scRNA-seq about different animals, organs, and time periods. Some previous work includes: PBMC[6], Schiebinger2019[8], Cao2019[1], and Chen2022[2] offer unique insights into single-cell sequencing, each with its own advantages and limitations. The PBMC dataset, derived from peripheral blood mononuclear cells, is widely recognized and easy to access, yet it lacks a temporal dimension essential for tracking cellular transitions over time. The Cao2019 dataset, stemming from a study on mammalian organogenesis, provides a rich, spatially-resolved single-cell atlas but may not provide the temporal resolution desired for long-period analysis. The Chen2022 (Live-seq) dataset introduces a novel technology for temporal transcriptomic recording in single cells, offering the temporal dimension but might still be in nascent stages with limited benchmark results. The Schiebinger2019 dataset stands out for its well-structured temporal data collection across two different time-course experiments, making it particularly suited for studying dynamic biological processes. It also accompanied by benchmark results, facilitating easier comparison and validation of analytical methods. The temporal relevance, completeness, and readily available benchmark results make the Schiebinger2019 dataset a robust and reliable choice for diving deeper into temporal relationships in single-cell data, thereby justifying its selection for further study.

It provides a comprehensive 18-day cellular profile, with the first experiment documenting 65,781 cells over 10 time points and the second 259,155 cells over 39 time points.

Methods

Cell differentiation can be represented as a series of matrices $X_1, X_2, \cdots, X_n \in \mathbb{R}^{C\times G}$, where C is the number of different cells and G is the number of genes. X_i is expressed count matrix at timestamp t_i, where we have $t_1, t_2, \cdots, t_n$ form an increasing time sequence. In this work, we model the dynamics of cell differentiation. Specifically, given the matrices and timestamps above, we aim to interpolate the matrices at another set of timestamps $t'_1, t'_2, \cdots, t'_m$ within $[t_1, t_n]$. However, The matrices X_i's are large and sparse. Instead of directly modeling the dynamics of these matrices, we first develop a structure model to compress each X_i into a low-dimensional matrix $x_i \in \mathbb{R}^{C\times d}$, where each row of x_i is a d-dimensional latent representation of the corresponding cell at timestamp t_i. Then we develop a dynamics model to predict the latent cell representations at timestamps $t'_1, t'_2, \cdots, t'_m$.

Data processing

We collected scRNA-seq datasets in Mouse Embryonic Cells from the previous study [8], across 16 different timepoints. We down-sampled cells (1948) across timepoints such that one specific timepoint had a comparable number of cells from others, and only selected the top 1479 highly variable genes across cell by gene count matrix, denoted as $\mathbb{R}^{C\times G\times T}$. For each timepoint, we have a cell by gene count matrix, denoted as $\mathbb{R}^{1948\times 1479}$.

Structure model

We proposed to use three different dimension reduction strategies to learn the embedding space for sparse count matrices: Principle Component Analysis (PCA), ZINB-WaVE model, and scVAE model. The strategy that achieves the best performance for cell clustering will be utilized in this study and followed by the downstream modeling.

ZINB-WaVE has the characteristics of robust performance on zero-inflated negative binomial distributions in single-cell sequencing data, and we want to use its ability to facilitate the extraction of latent space. Different from previous methods, we used the data that is not normalized in the status of the raw format while keeping the same data choices by filtering the quality data and making sure ZINB-WaVE is running in an ideal environment. In the actual experiment, we set the number of latent variables k number to 2, and the regularization parameter to 1000, which is considered as a balanced choice between its accuracy and other negative effects like over-fitting may cause. In ZINB-WaVE model, given an input Y denotes the count of the genes in cells, Y_{ij} represents the count of the gene j ($j = 1, \cdots, J$) for cell i ($i = 1, \cdots, n$). We assume Y_{ij} follow the ZINB distribution with parameters μ_{ij}, θ_{ij}, and π_{ij}, and model the whole problem as a regression task for the parameters.

Fig. 7 demonstrates the process of the ZINB-WaVE model. After the model learns the parameters, it can be used to fit the input data with noise reduction.

In scVAE generative model, $\mathbf{x}_m$ represents a gene count vector in the m_{th} cell. $\mathbf{z}_m$ embeds the latent space for each cell. y_m is a categorical latent variable indicating the cluster for cell types (4). scVAE parameterized θ through Multiple Layer Perception (MLP) to represent latent embedding for the cell m (5). The choice of likelihood function can consider any meaningful discrete function (6), such as Poisson distribution or Negative Binomial distribution. scVAE utilized Variational Inference to estimate the parameters during the inference process due to intractable marginal likelihood of $\mathbf{x}_m$. The approximate posterior distribution of $\mathbf{z}$ should follow a Multivariate Gaussian distribution (7).

$$p_\theta(\mathbf{z}|y) = \mathcal{N}(\mathbf{z}; \mu_\theta(y), \sigma^2_\theta(y)\mathbf{I}) \tag{1}$$

$$\lambda_\theta(\mathbf{z}) = h(\mathbf{W}\mathbf{z} + b) \tag{2}$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = f(\mathbf{x}; \lambda_\theta(\mathbf{z})) \tag{3}$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})\mathbf{I}) \tag{4}$$

We selected Negative Binomial distribution as the likelihood function for modeling expression counts, and the dimension of latent space, $\mathbf{z}_m$, to be 100. K, the number of components in

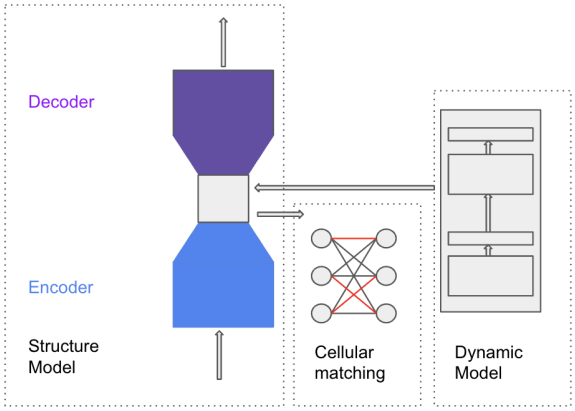

Fig. 1. Overview of our model. The structure model takes in cell by gene matrices and outputs their embedding. Then a cell alignment model matches the cells with the lowest overall cost, and feeds into the Dynamic model for interpolation prediction.

the Gaussian Mixture Model, was manually chosen to be 10 due to the unavailability of cell type labels. We also indicated batch labels corresponding to different timepoints for batch effects correction. Other settings were chosen by default as their original implementation.

Cell Alignment

After cells are projected into the latent space, we need to match cells from different time points, each match will create a cell trail sample that CST will use to learn. To accomplish this, we claim the alignment rules that each cell will be matched to only one cell in the next time point, and 2 cells will not matched to the same one. To solve this, we addressed 2 methods, using Sinkhorn Algorithm to minimize the Optimal Transport loss (OT), and using Jonker-Volgenant (JV) Algorithm to minimize the Mean Square Error (MSE) to find the biporate matching solution. The OT loss treats each cell as a probability distribution and finds a coupling matrix to move the distribution to the latter time. After the coupling matrix is obtained, we then apply linear assignment algorithm to find the best alignment. The JV algorithm directly characterizes the loss between 2 points using mean square error and solves the linear assignment problem directly. In reality, a cell can differentiate during different time phases, so one cell could be connected to multiple cells between time points, but different cells cannot map to the same cells. The OT loss can capture this non-linearity but may also introduce incorrect connections, so to avoid this issue, we still need to find a one-to-one mapping scheme. The JV algorithm instead, doesn't take the cell differentiation into account. On the other hand, the CST model must learn the cell differentiation process separately, so only one cell trail is sampled at one time. There is a trade-off between minimizing multiple transport loss and minimizing single transport loss, and we set up experiments to find out which can generate a more balanced cell trajectory.

Dynamics model

Once the cells are aligned, we utilize Continuous Spatio-temporal Transformers (CST) [4] to model the dynamics over the representation space. At each training step, we randomly sample a cell j and regard its latent representation as a function with respect to time. With a series of representations $x_{1,j}, x_{2,j}, \cdots, x_{n,j}$ and their corresponding timestamps $t_1, t_2, \cdots, t_n$ in hand, we randomly sample some $t'_1, t'_2, \cdots, t'_m$ in $[t_1, t_n]$ as dummy points and linearly interpolate their representation values. After adding noise, we feed the timestamps and corresponding function values into a transformer to predict a series of denoised and smooth function values.

Structure	Matching	Dynamic	Sobolev Loss	MSE
scVAE	OT	CST	**0.56**	2.23
PCA	OT	CST	2.56	50.80
scVAE	JV	CST	0.59	2.46
scVAE	OT	Transformer	0.92	**1.46**

Table 1. Model performances with various structure models, matching methods, and dynamic models.

CST uses Sobolev loss between the model prediction and ground truth value. Sobolev loss restricts the norm of high-order derivatives of the model output with respect to the input. Optimizing Sobolev loss helps the model learn to give smooth predictions when the given timestamp changes continuously. In specific, given the model output y and ground truth $\hat{y}$, the Sobolev loss is defined as

$$\mathcal{L}(y, \hat{y}) = \|y - \hat{y}\|_p^p + \mu \cdot \sum_{|q|=1}^{k} \|D^q(y)\|_p^p, \tag{5}$$

where $D^q(y)$ is a $|q|$-th partial derivative with respect to the input x and q is a vector indicating the number of order of the derivative on x's each dimension. p, k, μ are constants here.

In practice, we have $n = 16$ known timestamps and sample $m = 500$ dummy points during training. We use $p = 3$, $k = 3$, and $\mu = 0.01$ in the Sobolev loss. In our dataset, we have $C = 1948$ cells at each timestamp. We randomly split them into training and validation sets in a ratio of 7:3. We train the model for 100 epochs and the checkpoint achieving the lowest validation loss is used as the interpolation model for inference.

Results

Benchmark study of different structure models

We benchmarked three different structure models and evaluated the latent information of gene expression profiles in Mouse Embryonic Cells [8]. As expected, the regular PCA method was limited to capturing meaningful latent information due to the prevalent issue of data sparsity in scRNA-seq datasets. ZINB-WaVE model mainly extracted clusters that were biologically relevant across timepoints, while it was computationally demanding especially for the expression matrix with tens of thousands of cells. For example, ZINB-WaVE only permitted learning a limited dimension of latent space by using a small sampling proportion of original datasets. scVAE illustrated as an memory-efficient, accurate method that captured latent space from expression profiles and handled batch effects across timepoints in our study.

In scVAE model, we demonstrated that by taking into account batch effects, the structure model embedded latent information that generally mixed up across different timepoints, while preserving timepoints-specific signatures across different batches. For example, Timepoint 0 was separated from Timepoint 1 in latent space without batch effects correction (Fig. 2 left), but they were close to each other after controlling batch effects (Fig. 3 left). We also labeled latent embedding

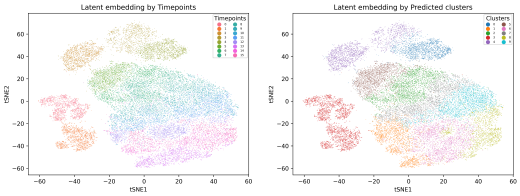

Fig. 2. Latent embedding from structure model (scVAE) without batch effects correction. left: Colored by batches, right: Colored by predicted clusters.

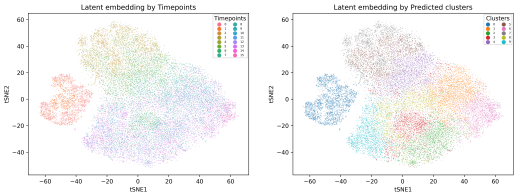

Fig. 3. Latent embedding from structure model (scVAE) with batch effects correction. left: Colored by batches, right: Colored by predicted clusters.

of cells with predicted assignment, suggesting that some sub-populations of cells shared similar expression profiles across some timepoints (timepoint 0 vs timepoint 1), while others had timepoints-specific expression profiles (timepoint 0 vs timepoint 2). In ZINB-WaVE model, we also visualize the latent space with and without batch effects corrections, which is shown in Fig. 8

In Table 1, we show the model interpolation performances when scVAE and PCA are used as the structure model. PCA leads to significantly higher Sobolev loss and MSE than scVAE. From Fig. 4(c) and (d), we find that the function formed by the latent embeddings of PCA is not smooth. This is the reason that our model performs poorly on PCA embeddings and indicates that PCA is not an effective structure model for our task.

Benchmark study of cell matching methods and different dynamic models

In the Method section, we introduce both OT and JV algorithms to match the cells. Here, we investigate their influence on our interpolation method. In Table 1, we find that OT achieves slightly lower losses than JV. This observation is interesting because the matching loss of OT is worse than that of JV. This experiment suggests that OT is a more suitable matching method for our task.

To verify the interpolation capability of CST, we compare it to a vanilla transformer baseline. We ignore the second term in Equation (8) in the vanilla transformer so that the smoothness of the approximated function is not restricted. As shown in Table 1, CST achieves lower Sobolev loss than transformer, indicating that the interpolation given by CST is more smooth. Besides, although the MSE of transformer is lower, it is only evaluated on the ground truth data points. So it cannot fairly represent the quality of the interpolations.

Cellular Trajectory Visualization

Previous benchmark experiments presented the best setting of the model. For the structure model, scVAE is used to reduce the dimension of the original gene expression matrices; for the cellular matching, we optimal transport loss to create a distribution and use Hungarian Algorithm to find the one-to-one matching; for the Dynamic model, we use the scCST

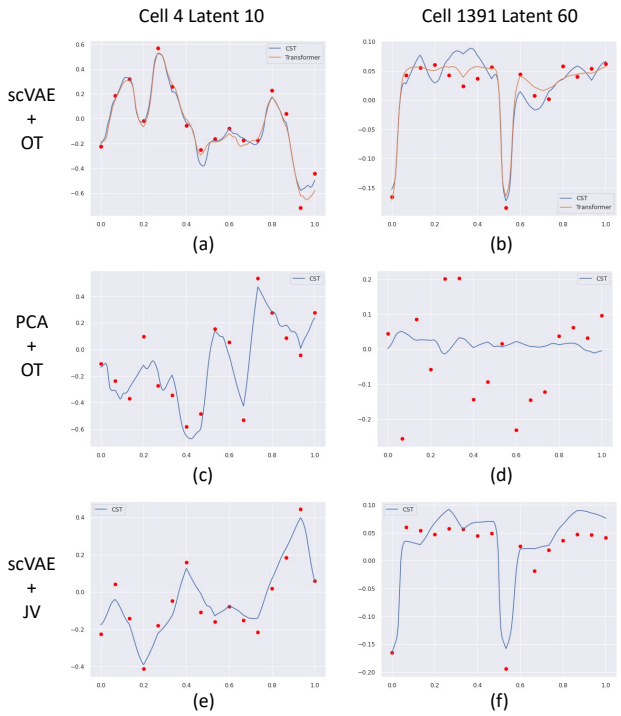

Fig. 4. Interpolation results of different models. The red dots are the ground truth values at the 16 provided timestamps.

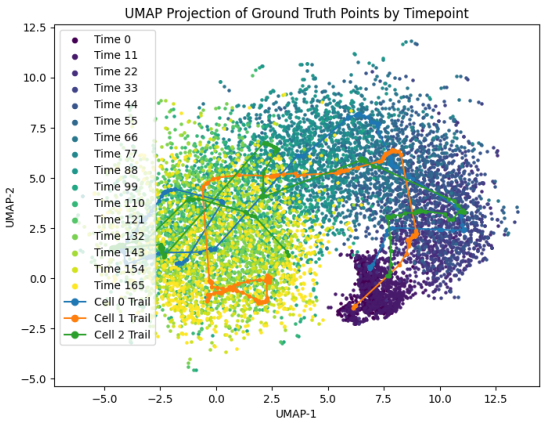

Fig. 5. Visualization of our scCST output, and show 3 cellular trails

to generate a smooth and informative trajectory. In the previous experiments, we output the 2D trajectory of cells using UMAP (Uniform Manifold Approximation and Projection). The background points are the ground truth of those 16 timepoints, and 3 randomly select cellular trajectories are presented on the projected graph. In the Fig. 5. As we can see, for the points that are interpolated, our model generates a very smooth trajectory. Upon that, our model still generate a very representative heterogeneous cell development trail. Further study to confirm the biological representation of those trails is expected.

Conclusion

Prediction of single-cell gene expression developmental trajectories has consistently presented significant challenges in

genomics research [8]. We take this research journey aiming to not only overcome the challenges but also gain more insights from the experiments, we hope the experiments we did can enhance our understanding of genomics. One of the significant challenges in genomics and sequencing data is the noise inherent in the measurements. [3] Our methods, aim to mitigate the noise and extract more meaningful factors from the data. A better outcome would be our expectation. Our study also sucessfully trained a continuous spatial-temporal Transformer that learn the cell trajectory in a smooth manner, and our benchmark experiments revealed the best configuration of the model. The final result is visualized and analyzed to prove that our model can generate the smooth trajectory we expected.

References

1. Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.
2. Wanze Chen, Orane Guillaume-Gentil, Pernille Yde Rainer, Christoph G Gäbelein, Wouter Saelens, Vincent Gardeux, Amanda Klaeger, Riccardo Dainese, Magda Zachara, Tomaso Zambelli, et al. Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608(7924):733–740, 2022.
3. Nils Eling, Michael D Morgan, and John C Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, 2019.
4. A. H. D. O. Fonseca, E. Zappala, J. O. Caro, and D. V. Dijk. Continuous spatiotemporal transformers. *International Conference on Machine Learning*, 2023.
5. Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 05 2020.
6. Roy Oelen, Dylan H de Vries, Harm Brugge, M Grace Gordon, Martijn Vochteloo, single-cell eQTLGen consortium, BIOS Consortium, Chun J Ye, Harm-Jan Westra, Lude Franke, et al. Single-cell rna-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nature Communications*, 13(1):3267, 2022.
7. Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, 9(1):284, 2018.
8. Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019.
9. Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature Communications*, 2021.
10. Grace Hui Ting Yeo, Sachit D Saksena, and David K Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):3222, 2021.

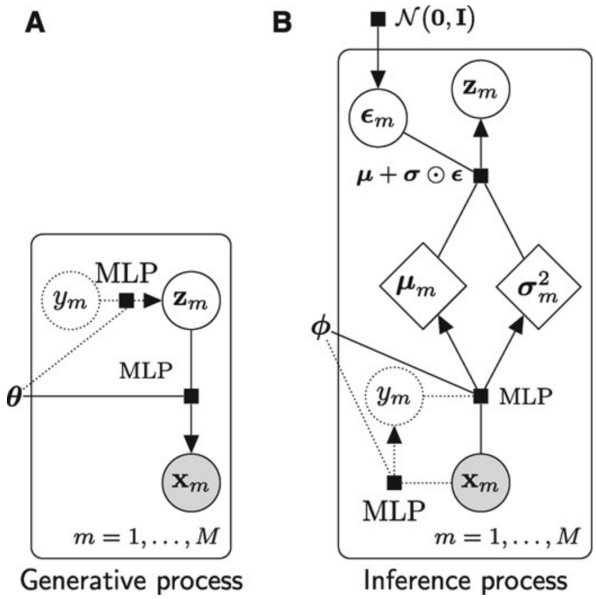

Fig. 6. The overview of scVAE model [5]. (**A**) Generative Model. (**B**) Inference Model.

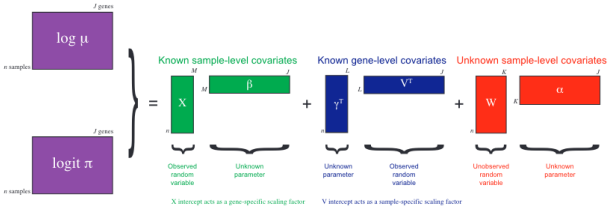

Fig. 7. Schematic view of ZINB-WaVE model [7]. We use it to denoise the input and lower its dimensions.

Contribution

Yuqi Lei: Main idea, pre-processing data, cellular mapping, design experiments, and visualize the trajectory result.

Zitian Tang: Implemented CST and ran all the interpolation experiments, visualize interpolation result.

Yu Zhong: Data processing; Structure model (scVAE) implementation. Results interpretation; Manuscript draft.

Shangyang Min: Implement ZINB-WaVE, address batch effect, visualize ZINB-WaVE result.

Yuyang Luo: Implement and visualize the ZINB-WaVE method, writing relevant part of the paper.

Thanks to Dr. Ritambhara Singh and Dr. Jeremy Bigness who gave advices to our project!

Supplymentary material

We illustrated the overall model architecture of scVAE (Fig. 6) and ZINB-WaVE as the reference (Fig. 7). The latent results of ZINB-WaVE were also shown here (Fig. 8).

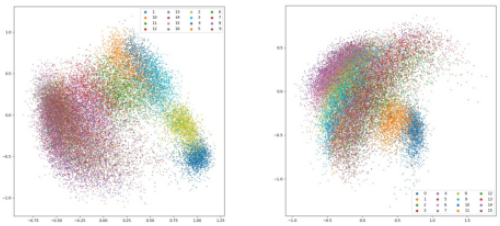

Fig. 8. Latent embedding from structure model (ZINB-WaVE). left: Without batch effects correction. right: With batch effects correction.