scNODE: Generative Model for Temporal Single Cell Transcriptomic Data Prediction

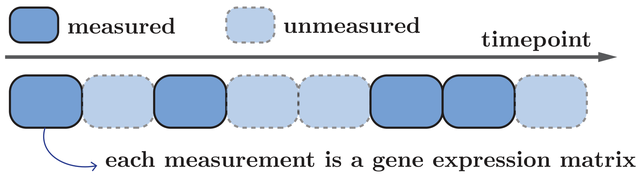
Jiaqi Zhang[1], Erica Larschan[2,3], Jeremy Bigness[2], Ritambhara Singh[1,2]

[1] Department of Computer Science, Brown University [2] Center for Computational Molecular Biology, Brown University

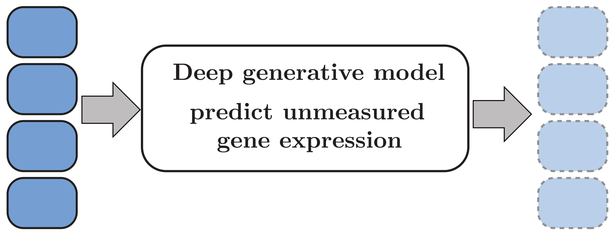
[3] Department of Molecular Biology, Cell Biology and Biochemistry, Brown University

Introduction

Problem: Temporal scRNA-seq data are only profiled at discrete and sparsely spaced timepoints due to laborious and expensive lab experiments



Goal: Predict gene expression at unmeasured timepoints



Limitation of previous works

- Linear dimensionality reduction, incapable to capture compex cell structure
- Fixed latent space obtained from measured timepoints, limiting predictions at unmeasured timepoints that have the distribution shift issue

Solution:

- VAE for complex latent representation learning
- Neural ODE for cellular dynamic modelling
- Dynamic regularization for adjusting latent with overall dynamics

Acknowledgement

This work is supported by National Institute of Health (NIH) award 1R35HG011939-01.

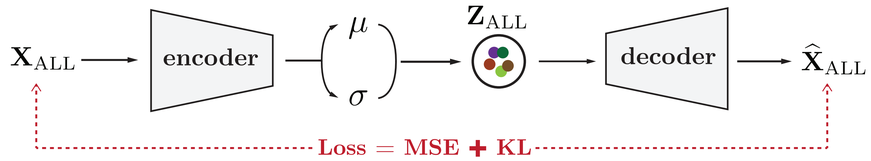
Paper & Codes

github.com/rsinghlab/scNODE

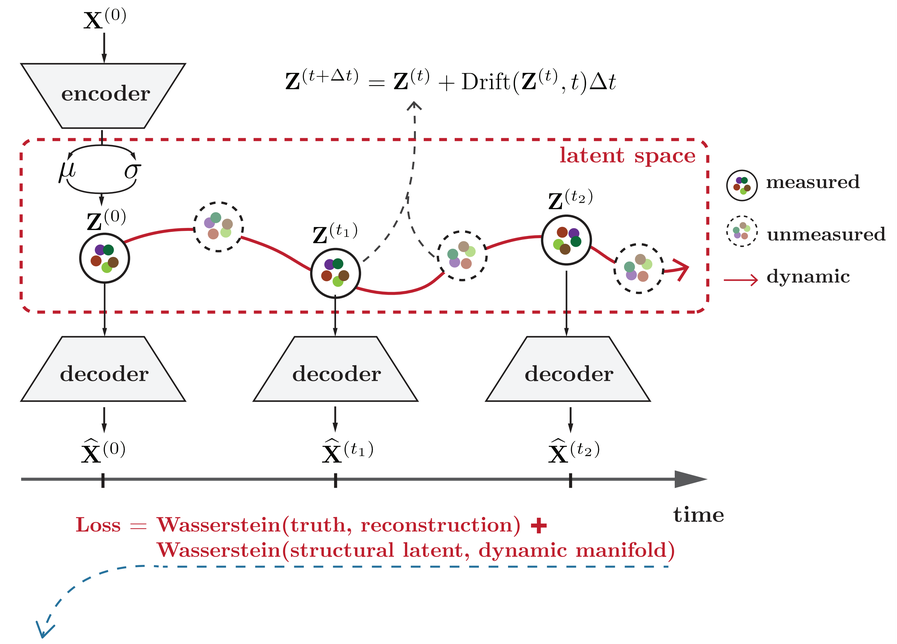
Method: single-cell Neural Ordinary Differnetial Equation (scNODE)

Input: Gene expression $\mathbf{X}^{(t)}$ at measured timepoints $t \in \mathcal{T}$

Stage I: Pre-train VAE to learn a latent space preserving structural relationships



Stage II: Model cell developmental dynamics in the latent space with neural ODE



Dynamic regularization:

- Enforces latent space to incorporate dynamics learnd by neural ODE
- Learns a latent space that is robust to distribution shift

Output: Gene expression at any timepoint

Advantages:

- Non-linearity of VAE captures complex cell structural relationship
- Updating latent space with dynamic regularization improves generalizability and robustness against distribution shifts

Results

Dataset & Preprocessing

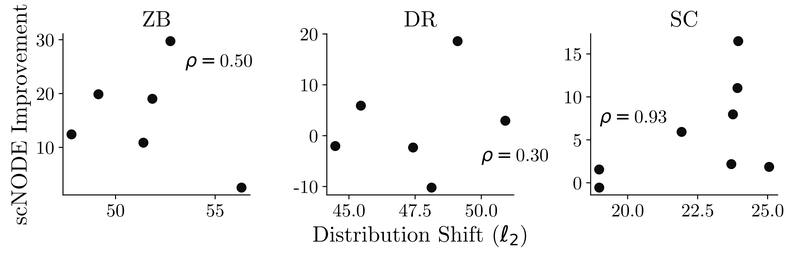
- three real-world scRNA-seq datasets of various tissues and # of timepoints
- 2000 HVGs → cell total count normalization → log-transformation
- preprocessing based on training timepoints to avoid data leakage

scNODE accurately predicts expression at unmeasured timepoints

leave-out middle timepoints (**interpolation**) and last few timepoints (**extrapolation**)

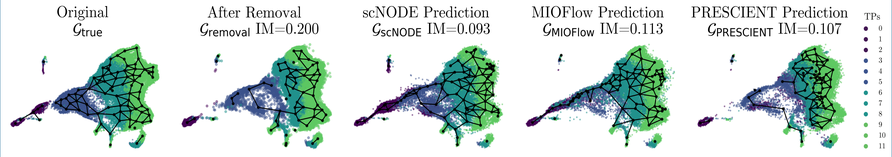
Method	Wasserstein Distance (↓)							
	Interpolation				Extrapolation			
	$t=5$	$t=7$	$t=9$	$t=11$	$t=15$	$t=16$	$t=17$	$t=18$
scNODE	55.22	**59.89**	**103.26**	**140.81**	**132.86**	**148.89**	**137.90**	**151.13**
MIOFlow	**55.07**	61.80	108.72	156.51	162.12	191.40	189.39	215.74
PRESCIENT	85.36	87.47	114.16	142.03	150.53	161.59	147.23	155.06

scNODE is robust aginst distribution shifts



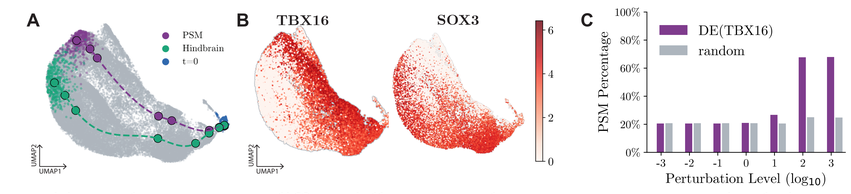
(Distribution Shift: distance between testing and training data)

scNODE predictions help recover cell trajectories



Ipsen-Mikhailov distance (↓) : $\text{IM}(\mathcal{G}_{\text{true}}, \mathcal{G}_{\text{pred}}) < \text{IM}(\mathcal{G}_{\text{true}}, \mathcal{G}_{\text{removal}})$

scNODE assists with perturbation analysis



transition path → differentially expressed genes → simulation