eccb2024

scNODE: **Generative Model for Temporal Single Cell Transcriptomic Data Prediction**

Jiaqi Zhang, Erica Larschan, Jeremy Bigness, and Ritambhara Singh

@ ECCB 2024 (Single Cells Session)
Sep. 19 2024, Turku, Finland

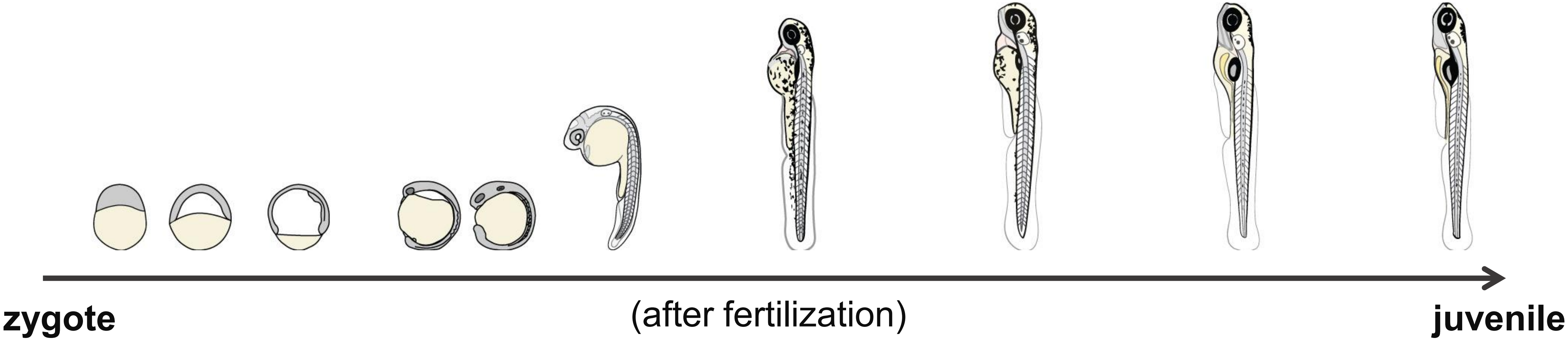
Jiaqi Zhang
Department of Computer Science
Brown University

BROWN

CENTER FOR
Computational
Molecular Biology

Understanding Dynamical Biological Processes is Crucial for Life Science

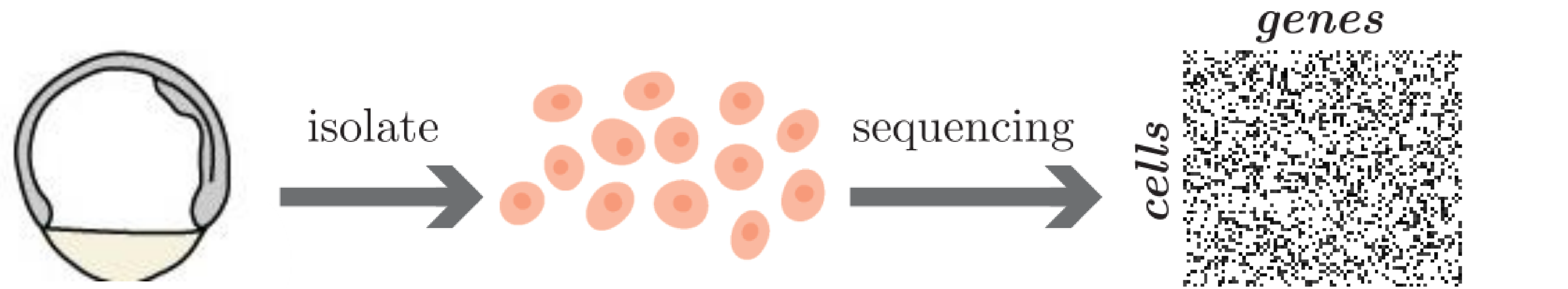
- A biological system is inherently dynamic at different levels
- Cellular dynamics reveals how cells grow, divide, and differentiate
- Understanding cell-level dynamics is key to analyze biological systems



(Sur, et.al., Dev. Cell, 2023)

Temporal scRNA-seq Offers High-Resolution Insights about Cellular Dynamics

- Single-cell RNA sequencing (scRNA-seq) technique measures gene expression levels within individual cells



isolate

sequencing

genes

cells

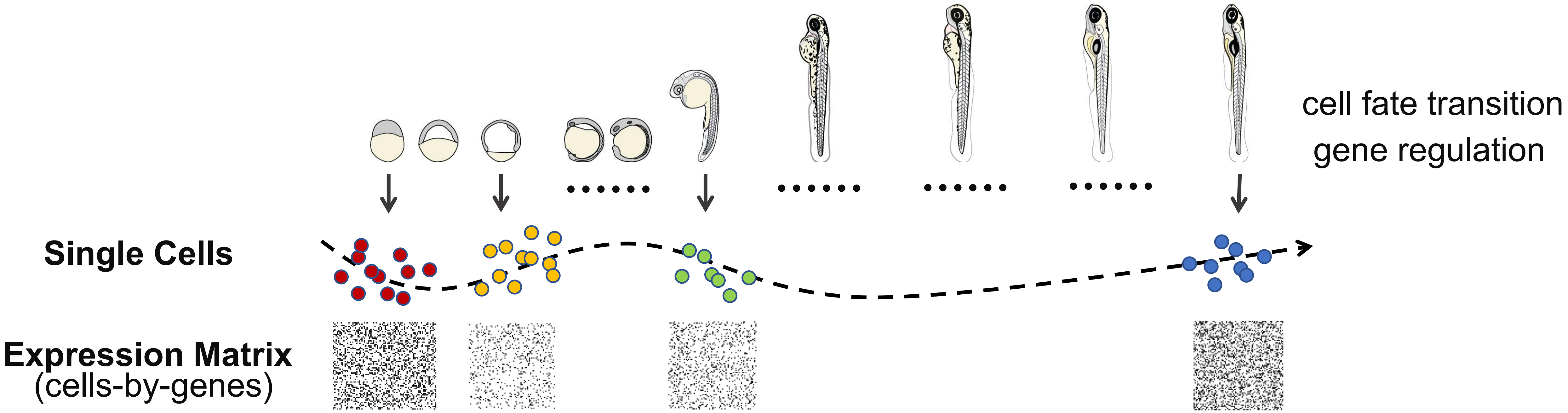
Epiboly
(heterogeneous)

Single Cells

Expression Matrix
(cells-by-genes)

Temporal scRNA-seq Offers High-Resolution Insights about Cellular Dynamics

- Collecting scRNA-seq data at multiple timepoints/stages allows us to observe gene expression dynamics



Single Cells

Expression Matrix
(cells-by-genes)

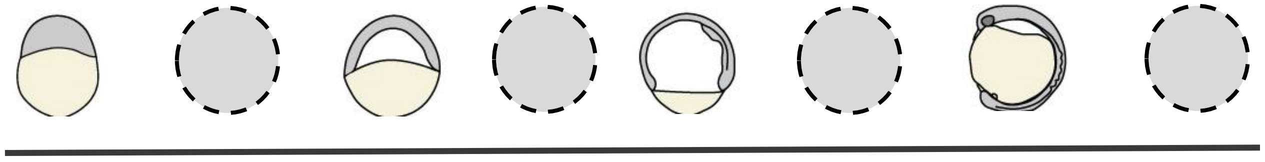
cell fate transition
gene regulation

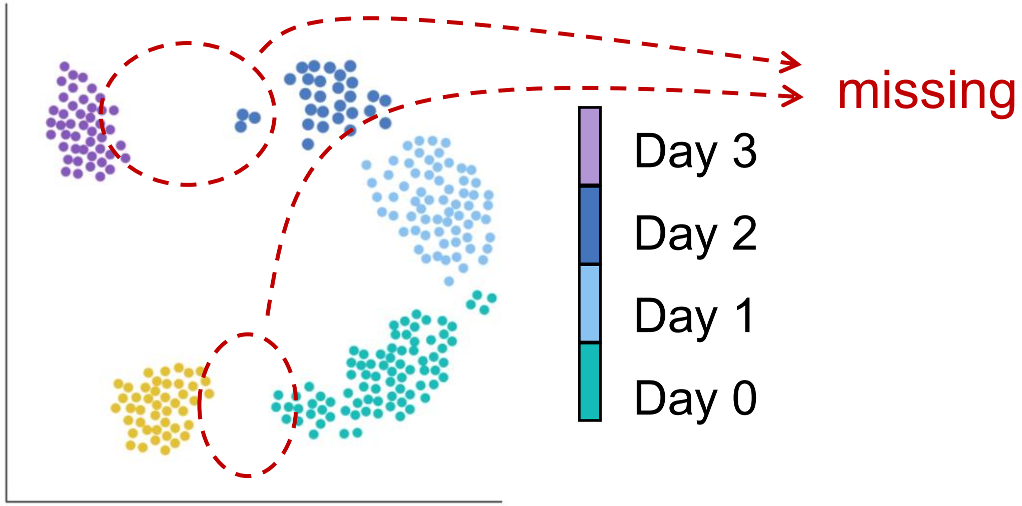
*figure adopted from (Sur, et.al., Dev. Cell, 2023)

But Temporal Data Have Limitations Due to Expensive and Laborious Experiments

- Because expenditures of time/labor/money, researchers generally profile gene expression at **sparsely spaced discrete time**

- So existing datasets can lose information between two consecutive discrete timepoints





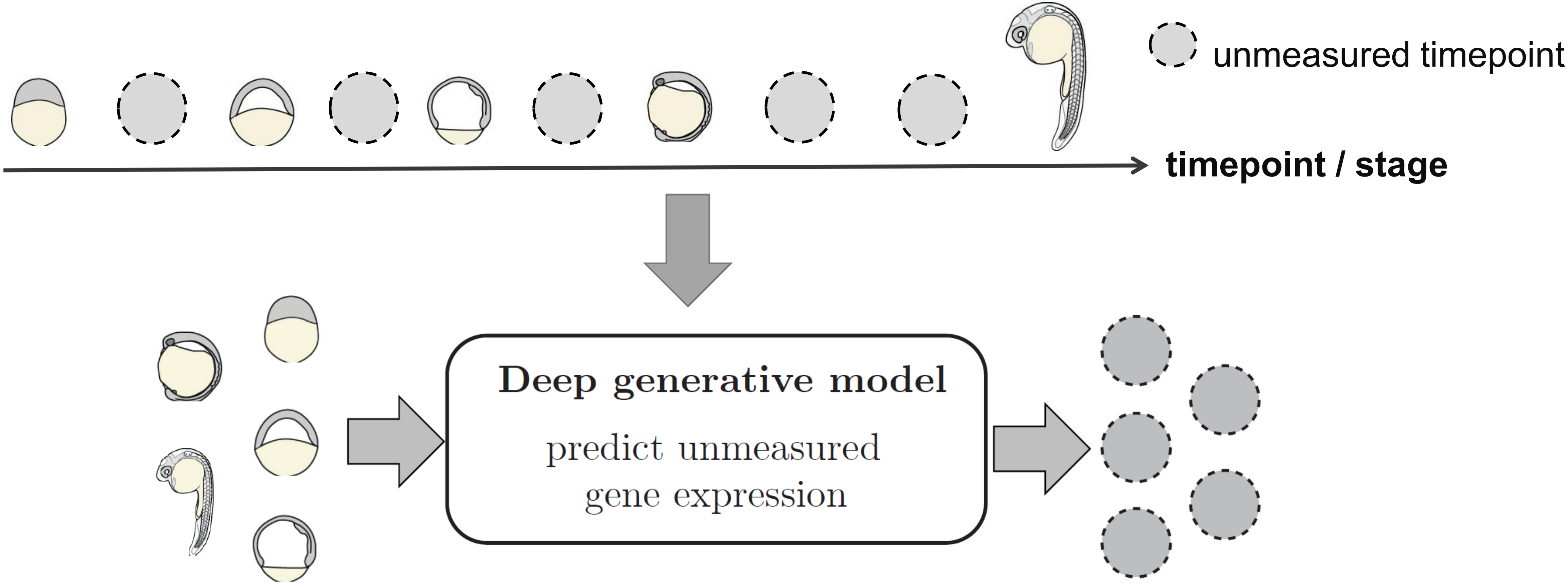
(Saelens, et.al., Nat. Biotechnol, 2019)

(Ding, et.al., Nat. Rev. Genet, 2022)

inaccurate representation & misleading conclusions

But Temporal Data Have Limitations Due to Expensive and Laborious Experiments

- **Goal:** predict realistic samples at any timepoint to enable & improve temporal downstream analysis



unmeasured timepoint

timepoint / stage

Deep generative model

predict unmeasured gene expression

Developing Such a Generative Model has Several Challenges

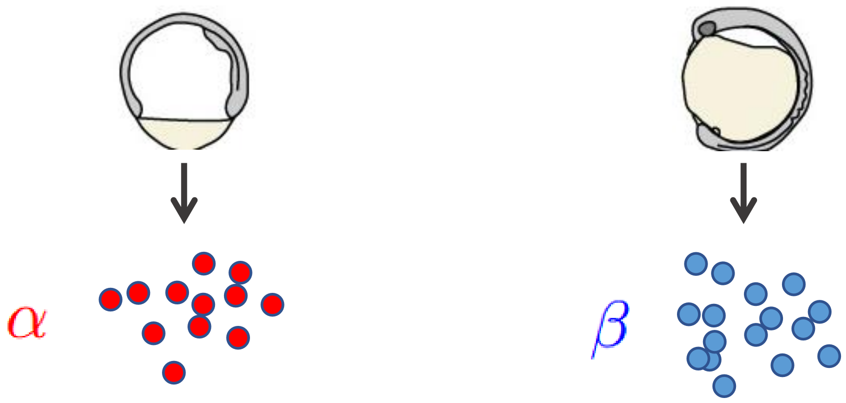
- **Challenge I**: lack of cell correspondence between timepoints

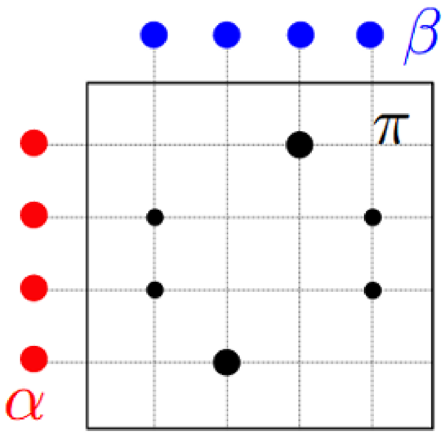
- **Challenge II**: noisy and high-dimensional data

- **Challenge III**: capture cellular dynamics when distribution shifts exist

Challenge I: Lack of Cell Correspondence between Timepoints

- Different set of cells are measured at each timepoint (destruction of cells during scRNA)
- **Solution: cell alignment with optimal transport**





Transport cost $\mathbf{D}$

Pair-wise distance between masses of two distributions

$$\mathbf{D}_{ij} = \| i - j \|_2 \ \text{ with } i \in \alpha \text{ and } j \in \beta$$

Transport plan π

Mapping masses of two distributions

- Optimal transport find the best cell correspondence between two set of cells

(Schiebinger, et.al., Cell, 2019) (Forrow and Schiebinger, Nat. Commun., 2021)

(https://www.wias-berlin.de/people/dvureche/)

Developing Such a Generative Model has Several Challenges

- ~~**Challenge I**: lack of cell correspondence between timepoints~~

  Solution: cell alignment with optimal transport

- **Challenge II**: noisy and high-dimensional data

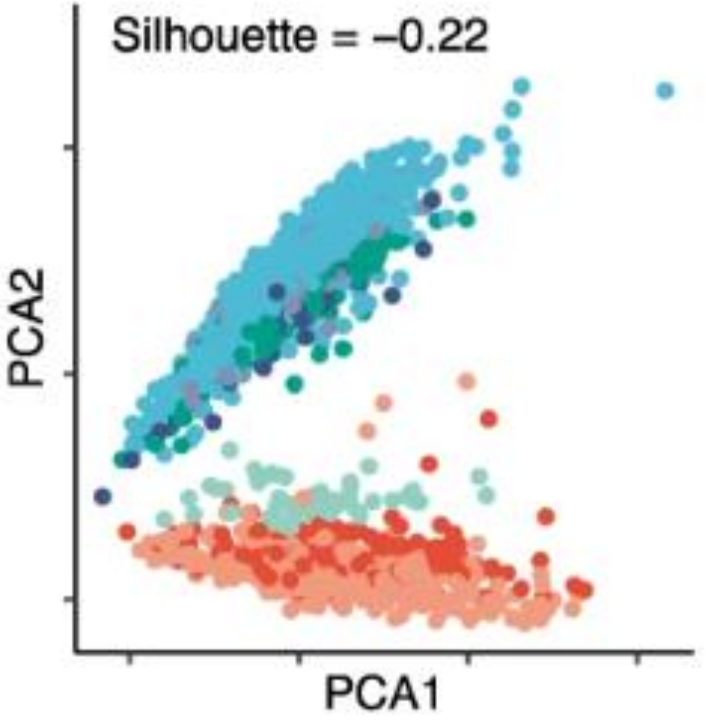
- **Challenge III**: capture cellular dynamics when distribution shifts exist

Challenge II: Noisy and High-Dimensional Data

- Due to high sparsity and high dimensionality of scRNA-seq data, we always model cell dynamics in low-dimensional space

- Many previous works use Principal Component Analysis (PCA), but it has the overcrowding issue

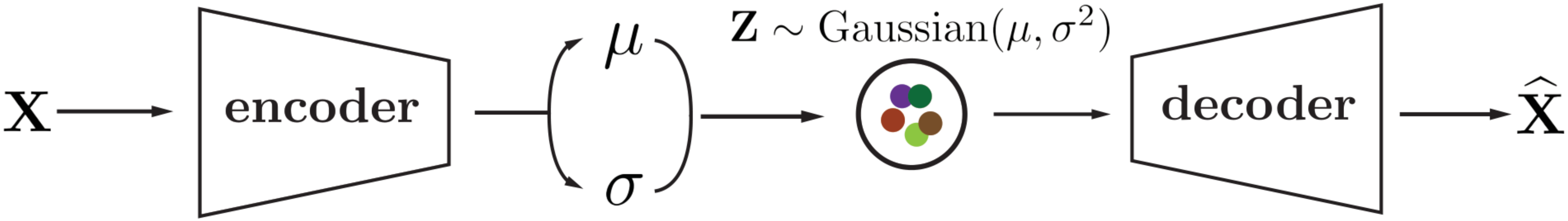
- **Solution: use Variational Auto-Encoder (VAE) to capture complex cell relationships**



(Tran, et.al., Genome Biol., 2020)

Challenge II: Noisy and High-Dimensional Data (cont.)

- Recent works use VAE to capture complex cell relationships
  - ○ $\mathbf{X} \in \mathbb{R}^{n \times p}$: gene expression of n cells and p genes
  - ○ learn d-dimensional latent variables $\mathbf{Z} \in \mathbb{R}^{n \times d}$ $(d \ll p)$



- VAE has superior performance on capturing cell type variations

(Tong, et. al., ICML, 2020)

(Yeo, et. al., Nat. Commun., 2021)

(Huguet, et. al., NeurIPS, 2022)

Developing Such a Generative Model has Several Challenges

- ~~**Challenge I**: lack of cell correspondence between timepoints~~

  Solution: cell alignment with optimal transport

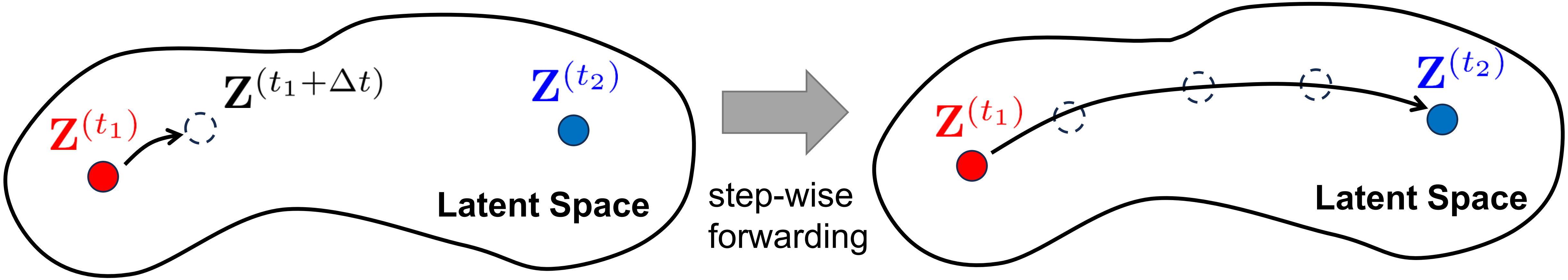
- ~~**Challenge II**: noisy and high-dimensional data~~

  Solution: use VAE for dimensionality reduction

- **Challenge III**: capture cellular dynamics when distribution shifts exist

Challenge III: Capture Cellular Dynamics when Distribution Shifts Exist

- Previous works adopts differential equation in VAE latent space to capture cell dynamics



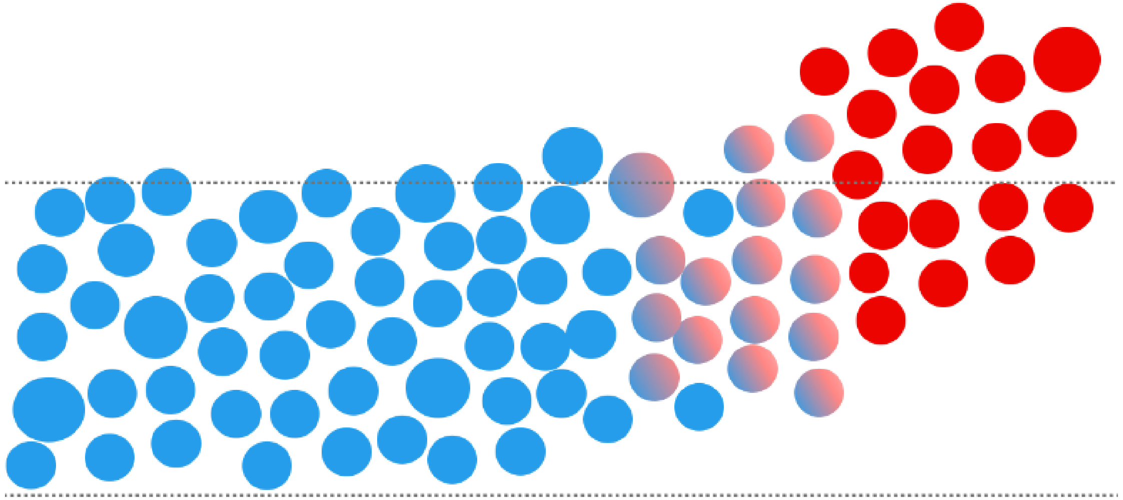
$$\mathbf{Z}^{(t_1+\Delta t)} = \mathbf{Z}^{(t_1)} + f\left(\mathbf{Z}^{(t_1)}\right)\Delta t$$

- However, the cell path/cellular dynamics are not naturally defined in VAE latent space

(Connor et.al., ICML, 2021)

Challenge III: Capture Cellular Dynamics when Distribution Shifts Exist (cont.

- Latent space ignores cellular dynamic → struggle to deal with distribution shift
  - especially when predicting timepoints beyond the measured range (i.e., extrapolations)



(credit to Evidently AI)

- **Unsolved problem: fails on extrapolations & interpolation w/ large shifts**
- **Our solution: adjust the latent space with cellular dynamics captured in modelling**

Developing Such a Generative Model has Several Challenges

- ~~**Challenge I**: lack of cell correspondence between timepoints~~

  Solution: cell alignment with optimal transport

- ~~**Challenge II**: noisy and high-dimensional data~~

  Solution: use VAE for dimensionality reduction

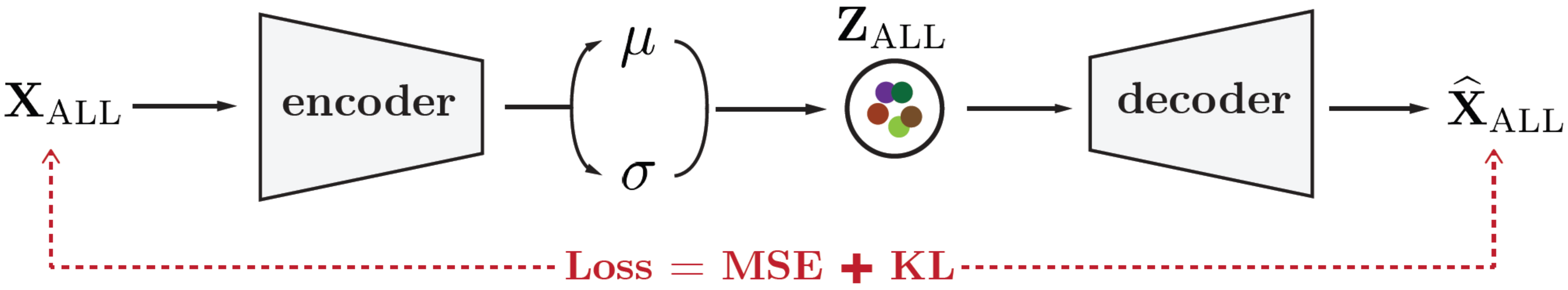
- **Challenge III**: capture cellular dynamics when distribution shifts exist

  Unsolved in previous works

  Solution in our work: adjust the latent space with cellular dynamics

Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

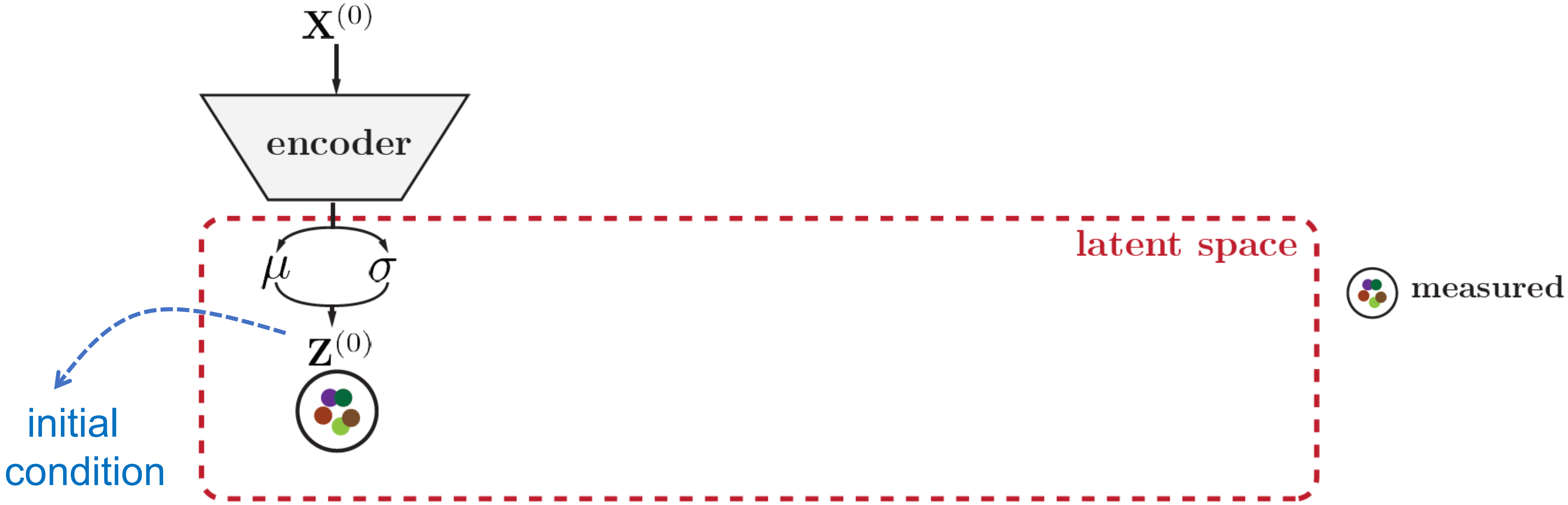
- **Step I:** uses VAE to learn complex low-dimensional space
  - gene expression $\mathbf{X}^{(t)}$ at measured timepoints $t \in \mathcal{T}$
  - learn latent space with all observed cells $\mathbf{X}_{\text{ALL}} = \text{CONCAT}(\mathbf{X}^{(t)} \mid t \in \mathcal{T})$


$\mathbf{X}_{\text{ALL}}$
encoder
μ
σ
$\mathbf{Z}_{\text{ALL}}$
decoder
$\widehat{\mathbf{X}}_{\text{ALL}}$
Loss = MSE + KL

  - pre-train a low-dimensional latent space to capture complex cell relationships

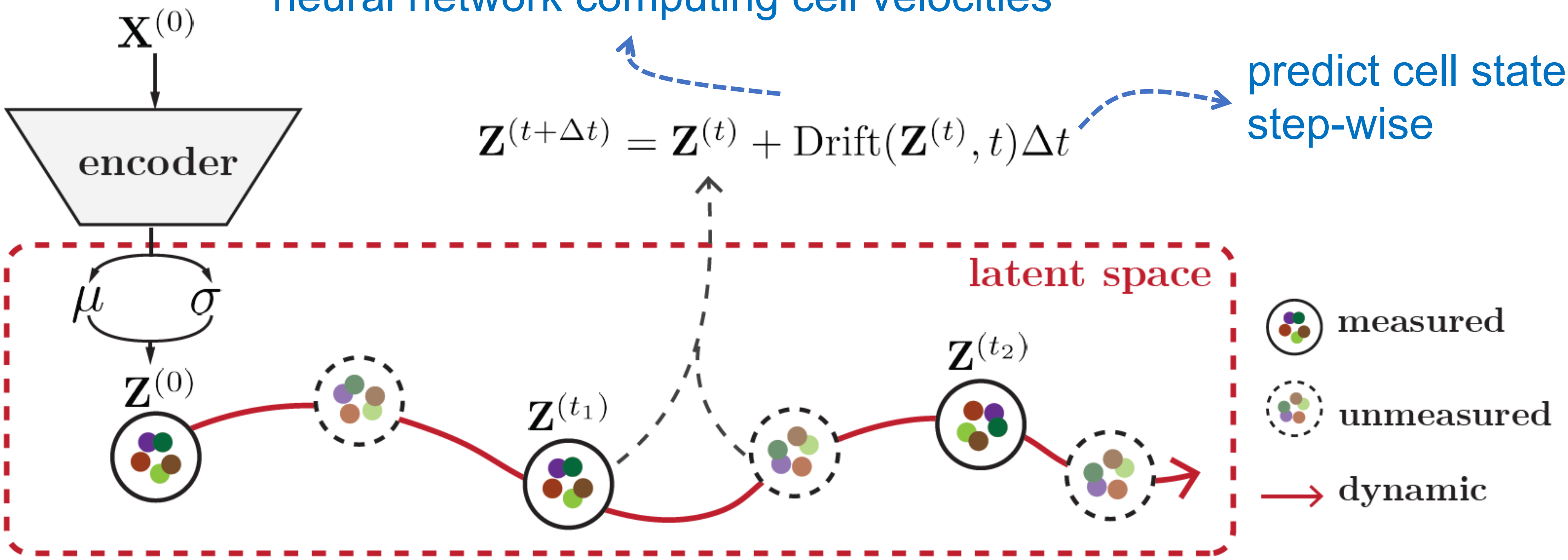
Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

- **Step II:** uses neural Ordinary Differential Equation (ODE) to model cell dynamics



Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

- **Step II:** uses neural Ordinary Differential Equation (ODE) to model cell dynamics



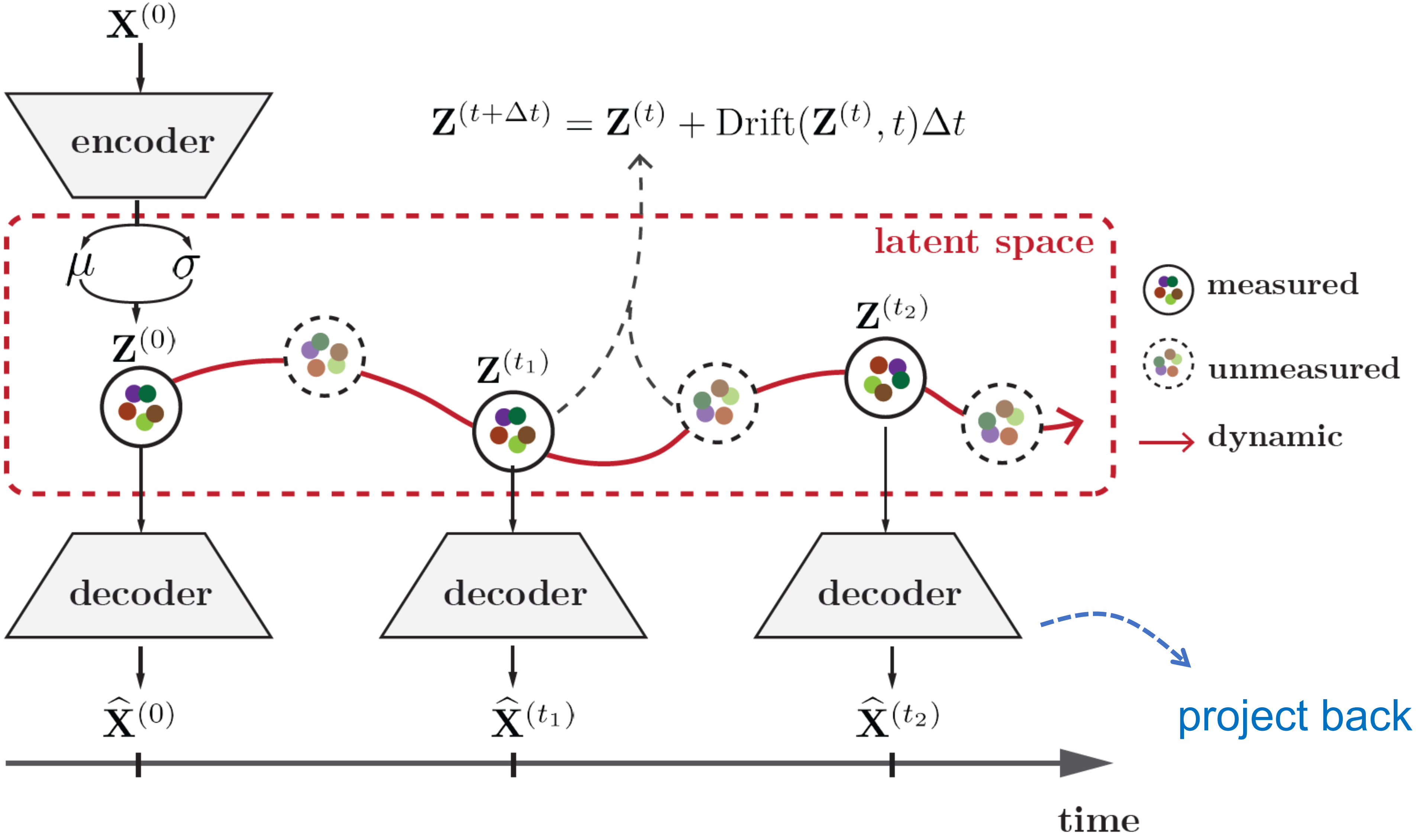
neural network computing cell velocities

$$\mathbf{Z}^{(t+\Delta t)} = \mathbf{Z}^{(t)} + \mathrm{Drift}(\mathbf{Z}^{(t)}, t)\Delta t$$

predict cell states step-wise

Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

- **Step II:** uses neural Ordinary Differential Equation (ODE) to model cell dynamics


$\mathbf{X}^{(0)}$
encoder
$\mathbf{Z}^{(t+\Delta t)} = \mathbf{Z}^{(t)} + \text{Drift}(\mathbf{Z}^{(t)}, t)\Delta t$
latent space
μ
σ
$\mathbf{Z}^{(0)}$
$\mathbf{Z}^{(t_1)}$
$\mathbf{Z}^{(t_2)}$
measured
unmeasured
dynamic
decoder
decoder
decoder
$\widehat{\mathbf{X}}^{(0)}$
$\widehat{\mathbf{X}}^{(t_1)}$
$\widehat{\mathbf{X}}^{(t_2)}$
time

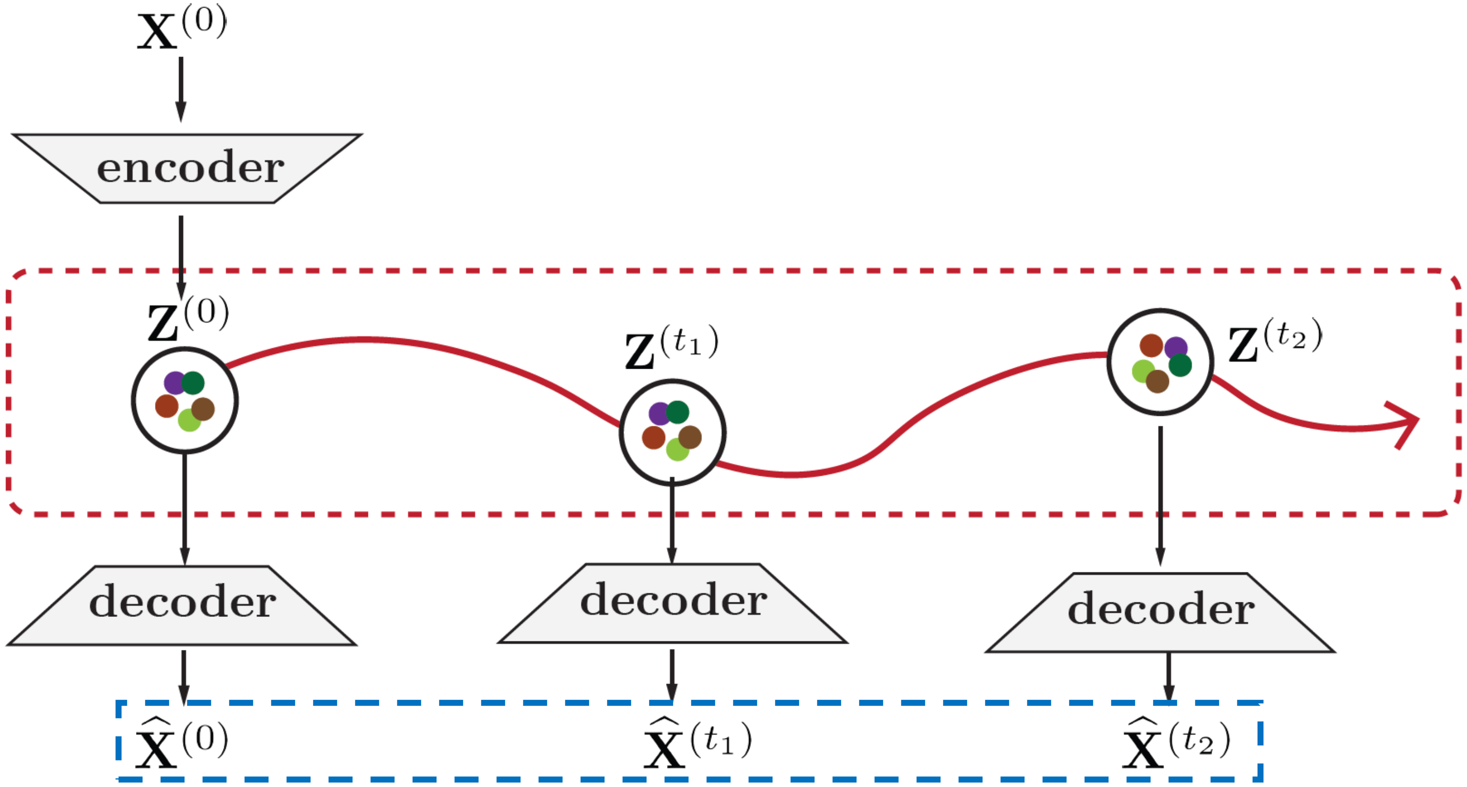
project back to gene space

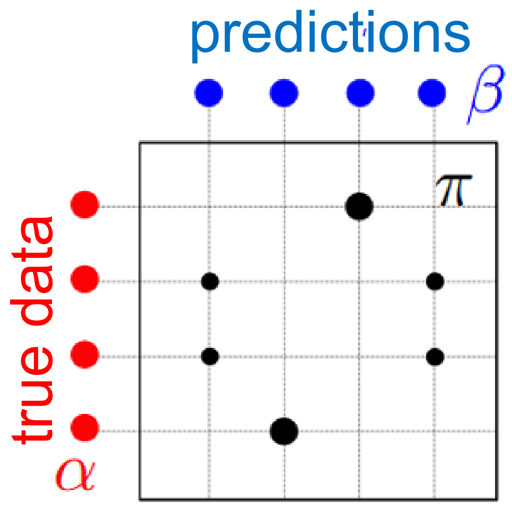
Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

- **Loss function:** reconstruction loss + dynamic regularization

- Reconstruction loss:
  - ○ Use optimal transport distance as reconstruction loss
  - ○ Wasserstein distance between ground truth & predictions

$$\sum_{t \in \mathcal{T}} \text{Wasserstein}(X^{(t)}, \widehat{X}^{(t)})$$



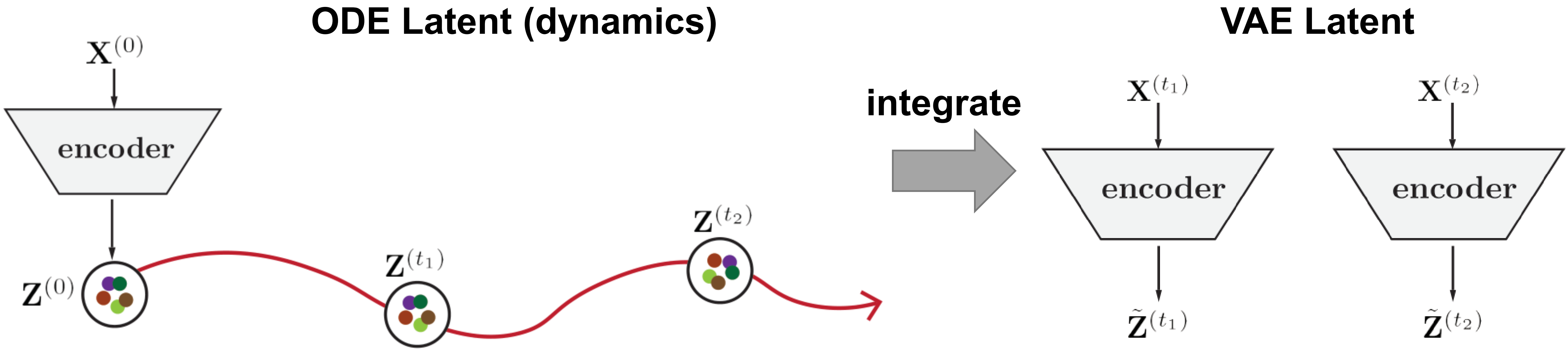


$$\text{Wass}(\alpha, \beta) = \left(\min_{\pi} \sum_{i,j} D_{ij}^2 \pi_{ij} \right)^{1/2}$$

Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

- **Loss function:** reconstruction loss + dynamic regularization
- Dynamic regularization:
  - Enforces latent space to incorporate dynamics learned by neural ODE

$$\text{Wasserstein}(\text{VAE latent},\ \text{ODE latent}) \rightarrow \text{Wasserstein}(\tilde{\mathbf{Z}}^{(t)},\ \mathbf{Z}^{(t)})$$



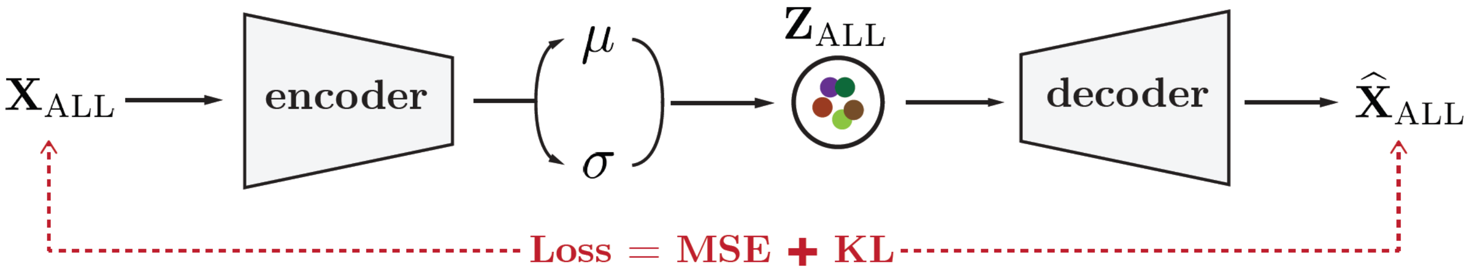
ODE Latent (dynamics)

integrate

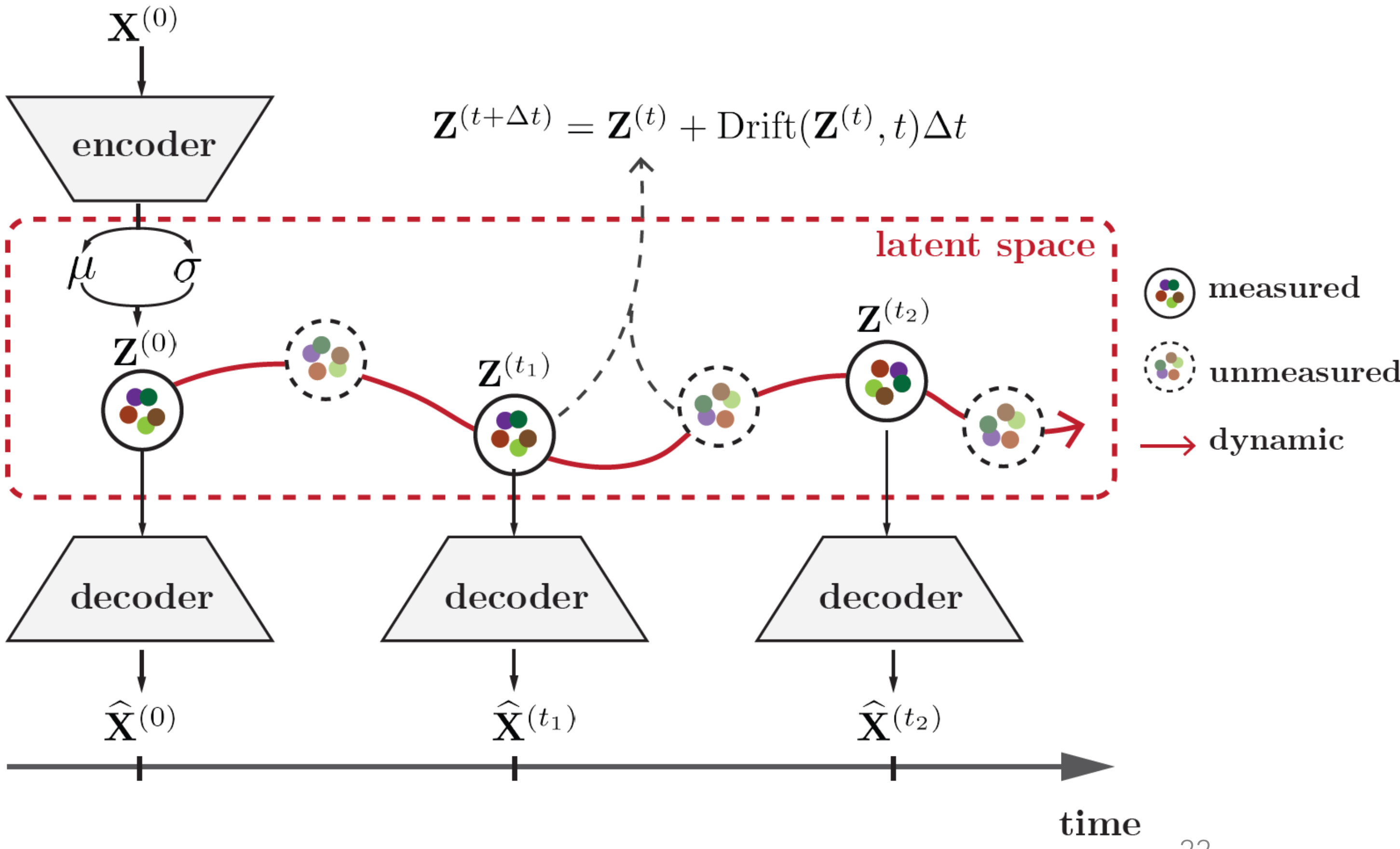
VAE Latent

Our Method: single-cell Neural Ordinary Differential Equation (scNODE)

- **Step I:** VAE captures complex cell relationships

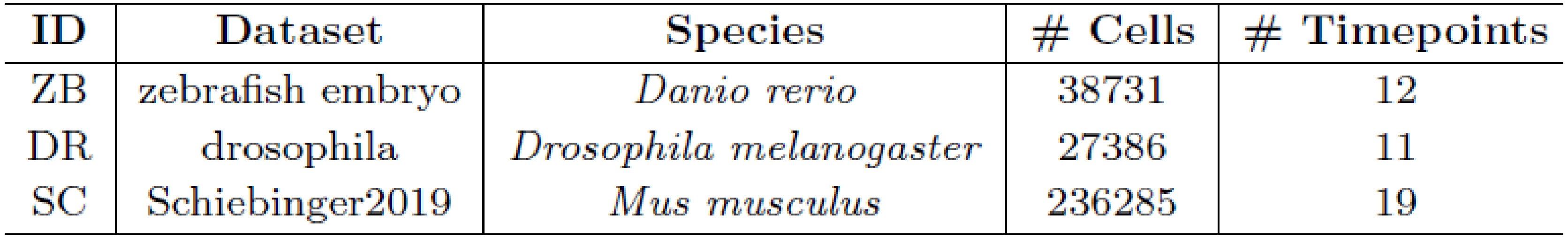

$\mathbf{X}_{\text{ALL}}$
encoder
μ
σ
$\mathbf{Z}_{\text{ALL}}$
decoder
$\widehat{\mathbf{X}}_{\text{ALL}}$
Loss = MSE + KL

- **Step II:** ODE models cell dynamics
  - ○ dynamic regularization
  - ○ capture long-term dynamics
  - ○ robust against distribution shifts


$\mathbf{X}^{(0)}$
encoder
$\mathbf{Z}^{(t+\Delta t)} = \mathbf{Z}^{(t)} + \text{Drift}(\mathbf{Z}^{(t)}, t)\Delta t$
latent space
μ
σ
$\mathbf{Z}^{(0)}$
$\mathbf{Z}^{(t_1)}$
$\mathbf{Z}^{(t_2)}$
measured
unmeasured
dynamic
decoder
decoder
decoder
$\widehat{\mathbf{X}}^{(0)}$
$\widehat{\mathbf{X}}^{(t_1)}$
$\widehat{\mathbf{X}}^{(t_2)}$
time

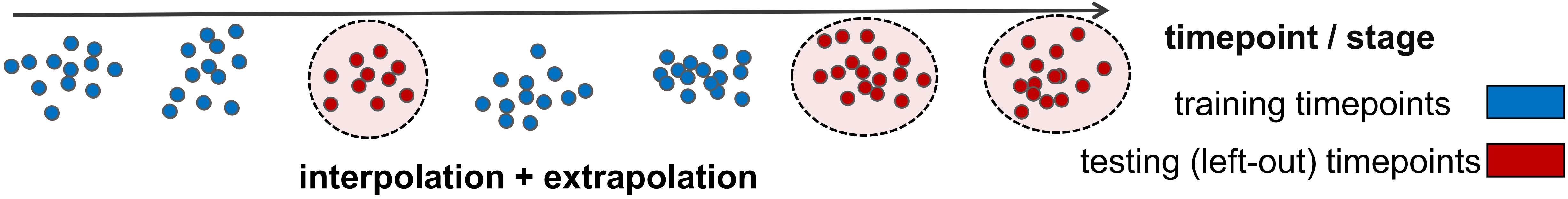
Experiment Setup

- **Dataset**: three scRNA-seq datasets



ID	Dataset	Species	# Cells	# Timepoints
ZB	zebrafish embryo	*Danio rerio*	38731	12
DR	drosophila	*Drosophila melanogaster*	27386	11
SC	Schiebinger2019	*Mus musculus*	236285	19

- **Setup**: remove several timepoints → recover these left-out observations

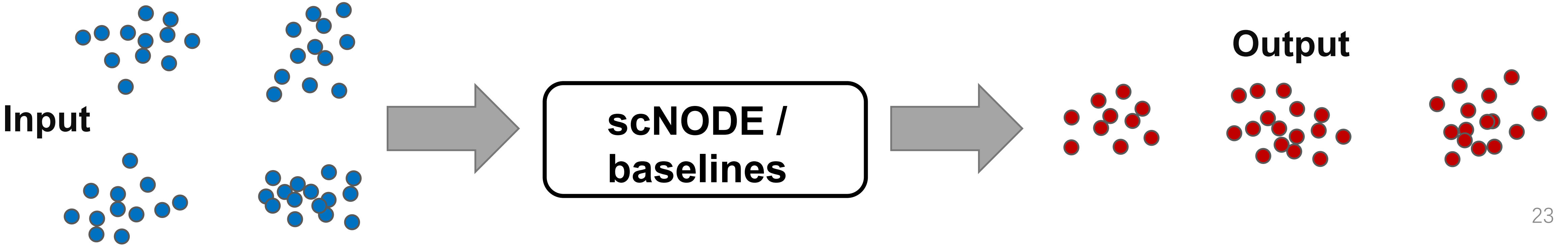


timepoint / stage

training timepoints 🟦

testing (left-out) timepoints 🟥

interpolation + extrapolation



Input

scNODE / baselines

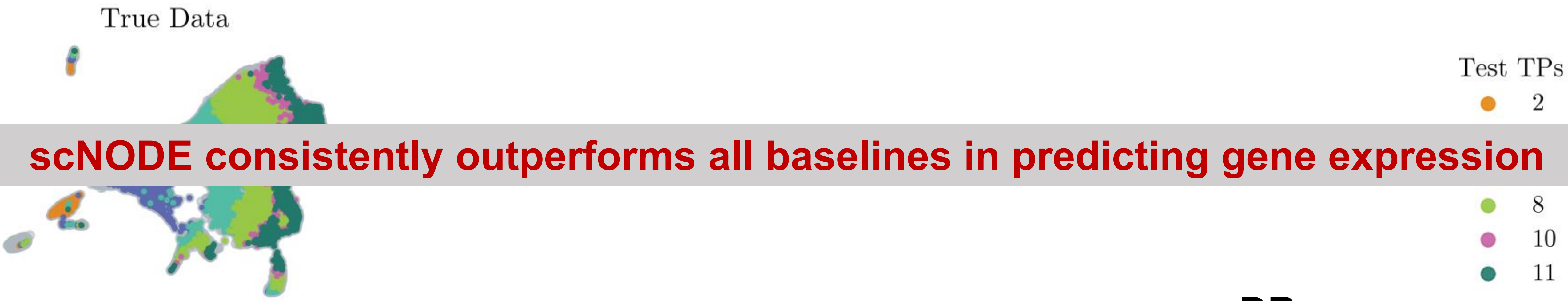
Output

Experiment Setup (cont.)

- **Metric**: Wasserstein distance between predictions and ground truth (lower is better)

- **Baselines**: two state-of-the-art methods
  - PRESCIENT (Yeo, et. al., Nat. Commun., 2021)
  - MIOFlow (Huguet, et. al., NeurIPS, 2022)

Experiment I: scNODE can Accurately Predict Gene Expression at Unobserved Timepoints



scNODE consistently outperforms all baselines in predicting gene expression

ZB

Method	Left-out Timepoints					
	Interpolation				Extrapolation	
	$t = 2$	$t = 4$	$t = 6$	$t = 8$	$t = 10$	$t = 11$
scNODE	579.10	508.55	440.92	517.81	652.36	707.10
MIOFlow	580.18	516.59	453.61	536.35	671.23	734.42
PRESCIENT	1381.96	1002.62	730.974	701.29	916.51	973.17

DR

Method	Left-out Timepoints					
	Interpolation			Extrapolation		
	$t = 2$	$t = 4$	$t = 6$	$t = 8$	$t = 9$	$t = 10$
scNODE	445.82	464.78	535.78	600.18	585.60	718.20
MIOFlow	443.56	469.51	532.93	617.48	680.41	852.02
PRESCIENT	524.38	511.61	539.38	621.31	575.45	718.56

SC

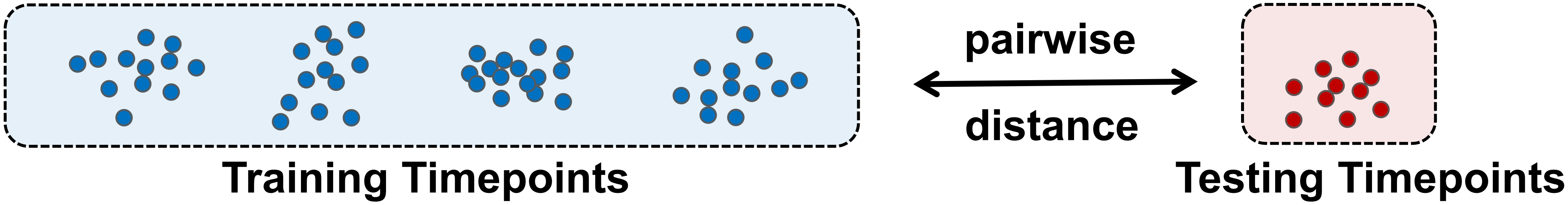
Method	Left-out Timepoints							
	Interpolation				Extrapolation			
	$t = 5$	$t = 7$	$t = 9$	$t = 11$	$t = 15$	$t = 16$	$t = 17$	$t = 18$
scNODE	55.22	59.89	103.26	140.81	132.86	148.89	137.90	151.13
MIOFlow	55.07	61.80	108.72	156.51	162.12	191.40	189.39	215.74
PRESCIENT	85.36	87.47	114.16	142.03	150.53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

best performance

second best performance

Experiment II: scNODE is More Robust Against Distribution Shift

- **Distribution shift**: averaged pairwise Euclidian distance between training & testing tps
  - ○ higher value indicates a more significant distribution shift



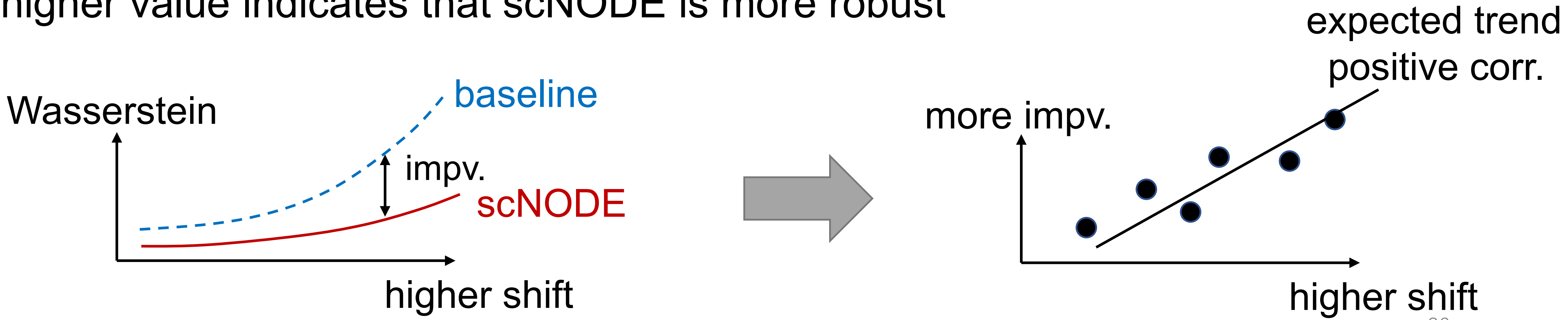
pairwise

distance

Training Timepoints

Testing Timepoints

- **scNODE improvement:** diff. between performance of scNODE & second-best baseline
  - ○ higher value indicates that scNODE is more robust



Wasserstein

baseline

impv.

scNODE

higher shift

expected trend

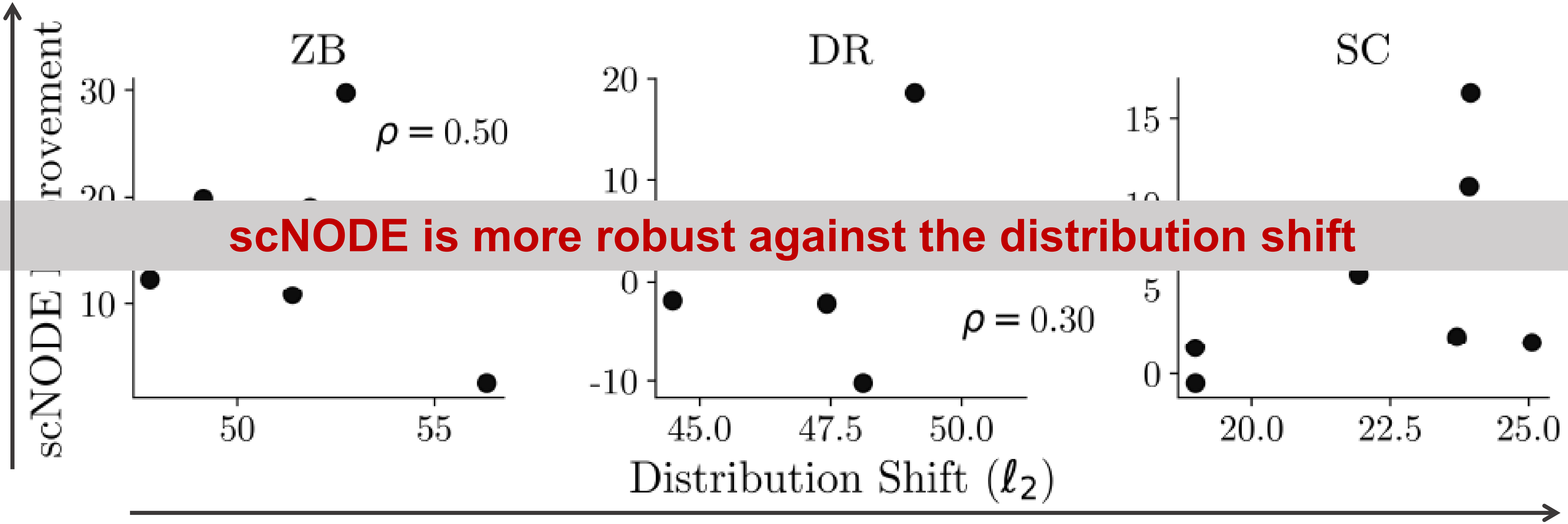
positive corr.

more impv.

higher shift

Experiment II: scNODE is More Robust Against Distribution Shift

more impv.



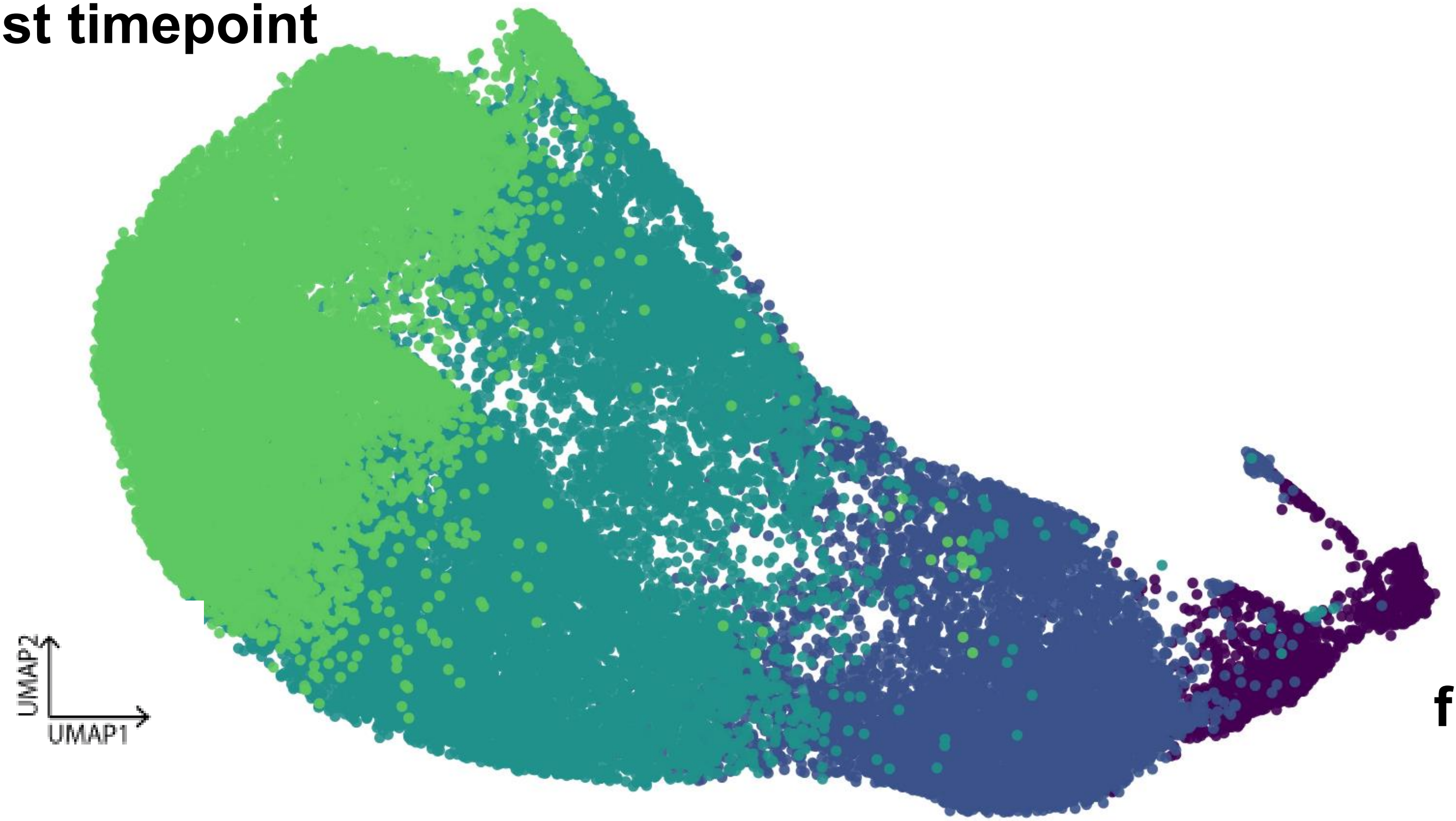
scNODE is more robust against the distribution shift

higher shift

Experiment III: scNODE’s Interpretable Latent Space Assists with Analysis

- We take the latent space learned by scNODE on ZB dataset

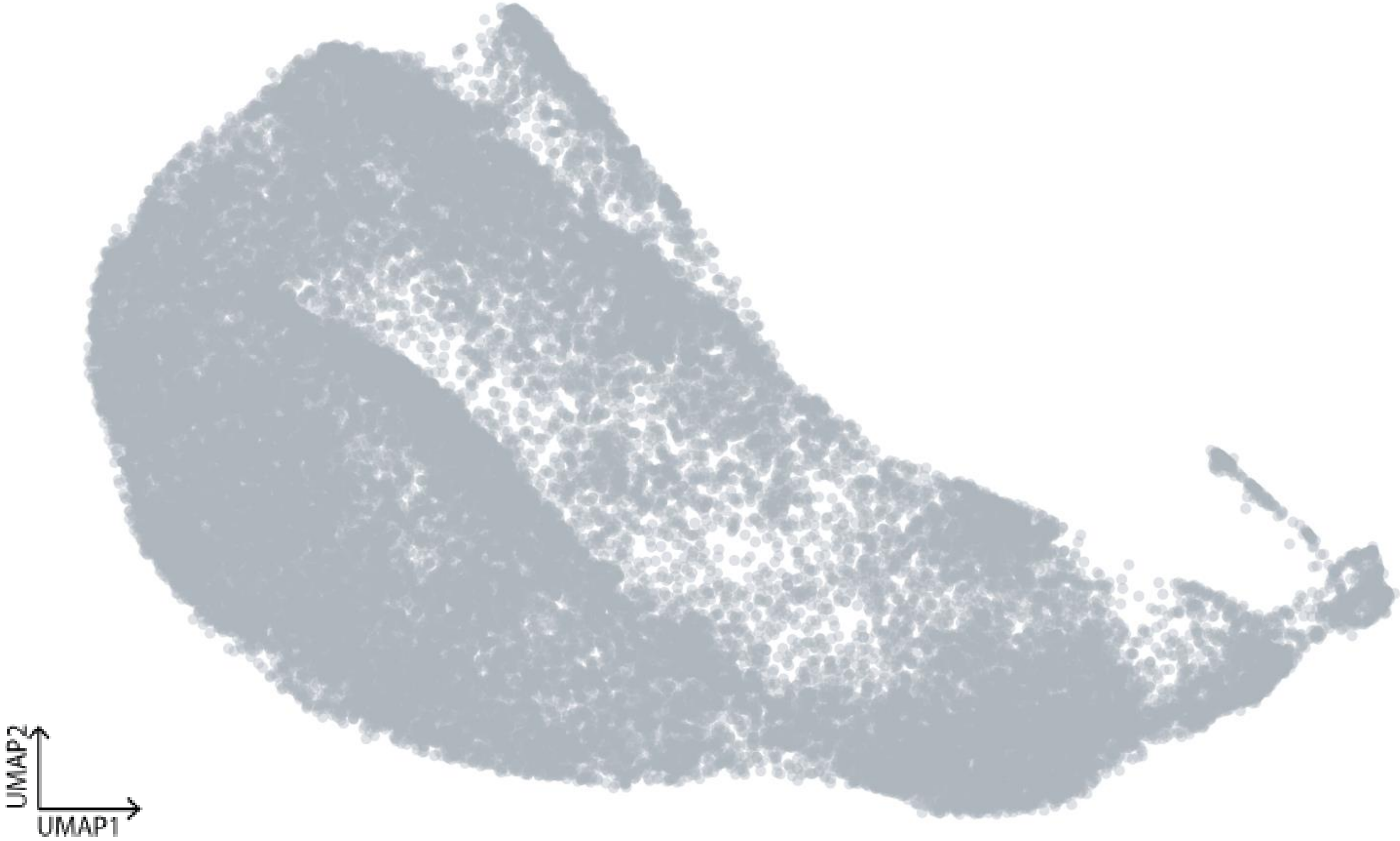
last timepoint



first timepoint

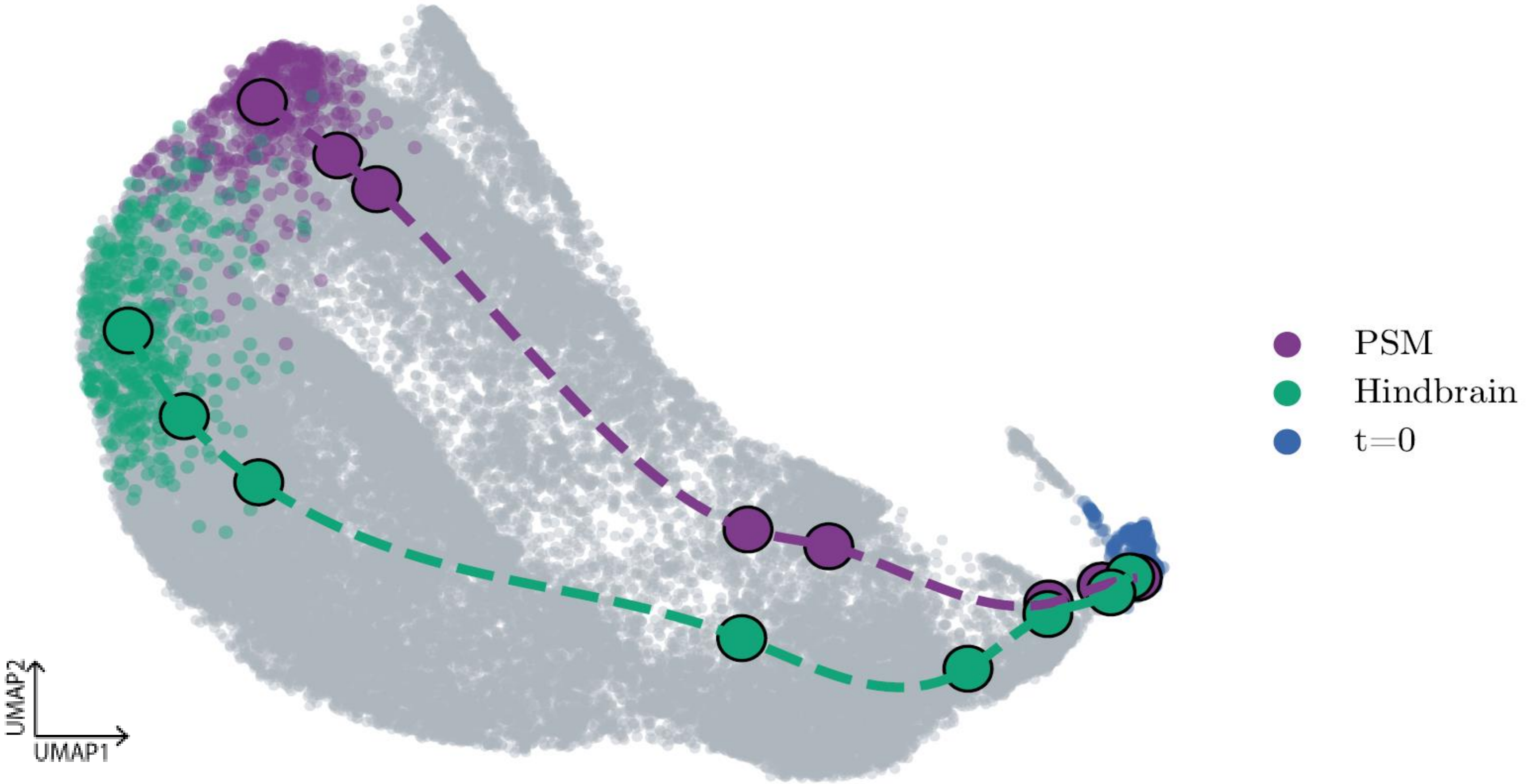
Experiment III: scNODE's Interpretable Latent Space Assists with Analysis

- We take the latent space learned by scNODE on ZB dataset



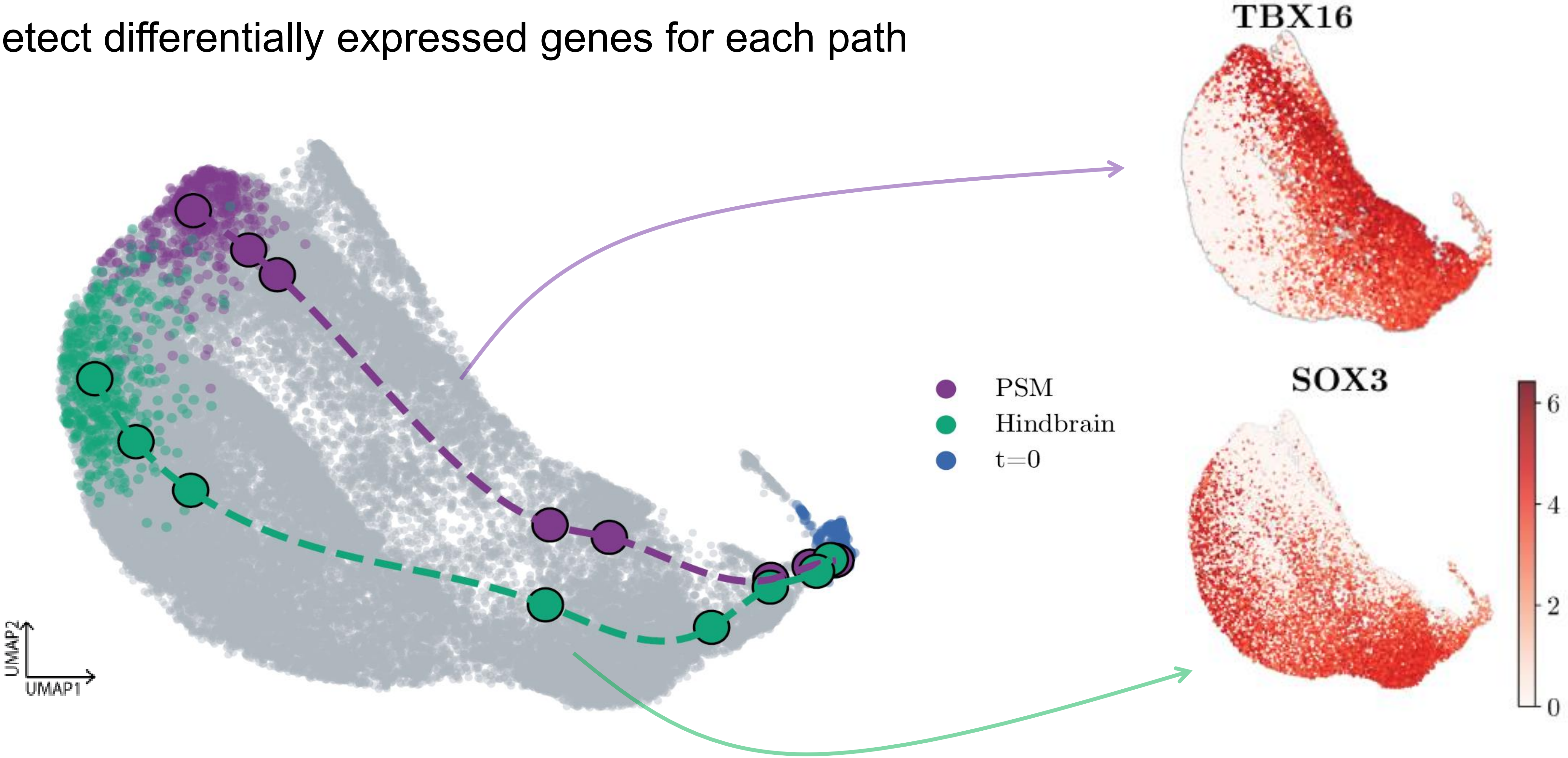
Experiment III: scNODE's Interpretable Latent Space Assists with Analysis

- Construct cell transition path



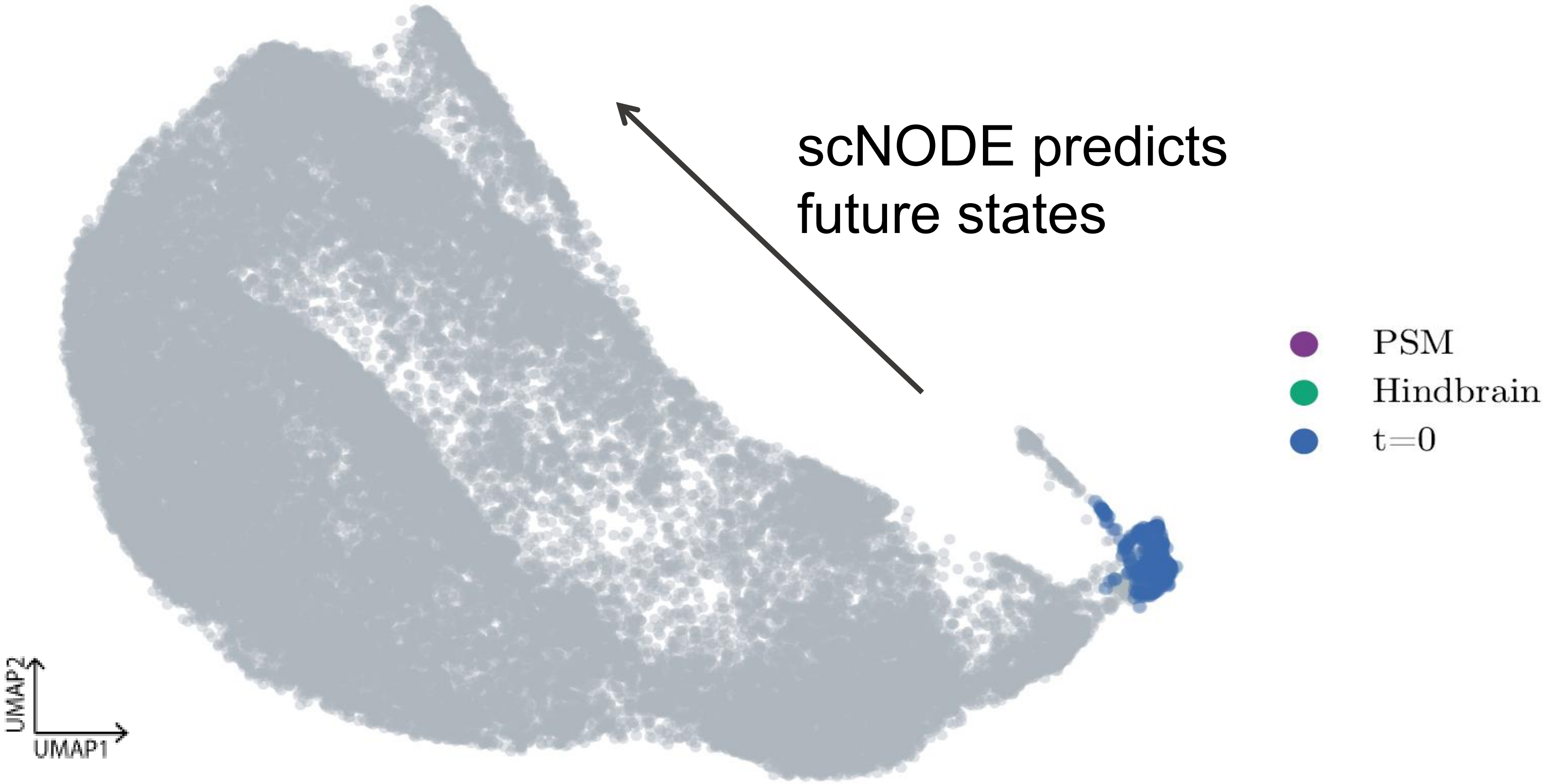
Experiment III: scNODE’s Interpretable Latent Space Assists with Analysis

- Detect differentially expressed genes for each path



Experiment III: scNODE’s Interpretable Latent Space Assists with Analysis

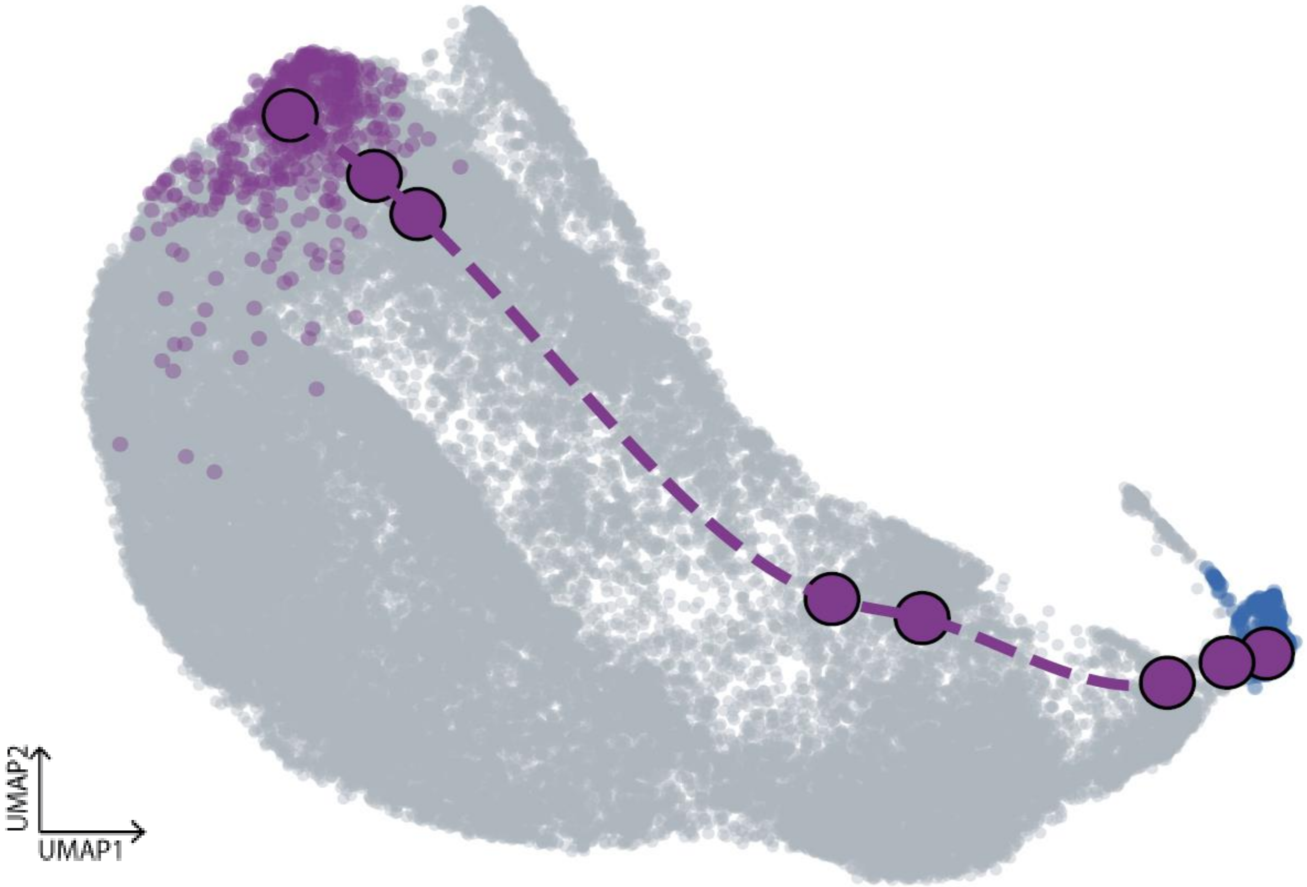
- Conduct *in silico* perturbation

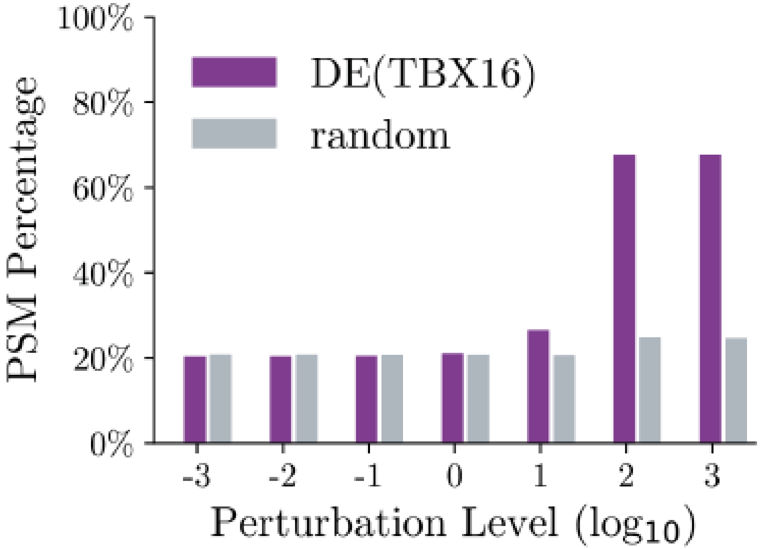


scNODE predicts future states

Experiment III: scNODE’s Interpretable Latent Space Assists with Analysis

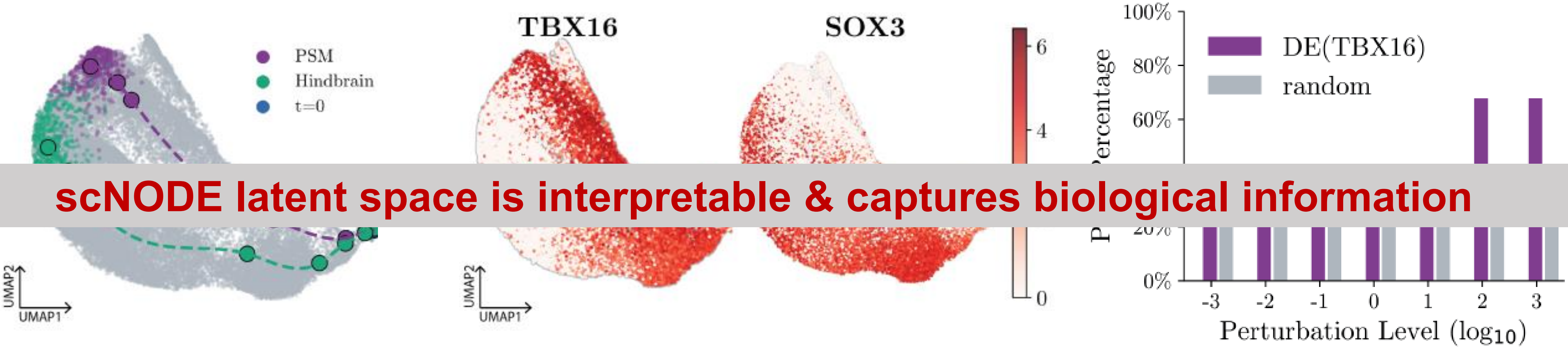
- Conduct *in silico* perturbation





Experiment III: scNODE's Interpretable Latent Space Assists with Analysis

- We take the latent space learned by scNODE on ZB dataset
- Construct cell transition path
- Detect differently expressed genes for each cell transition path
- *In silico* perturbation



scNODE latent space is interpretable & captures biological information

Conclusion

- scNODE is robust against distribution shifts
- scNODE accurately predicts gene expression
- scNODE assists with temporal downstream analysis

github.com/rsinghlab/scNODE

- Extension:
    - Model dynamics from temporal multi-omic data (e.g., transcriptomic and chromatin accessibility)
    - Translate between two omics at any timepoint

COME BY OUR POSTER

(Poster Session 1, P353)

Acknowledgement

Ritambhara Singh
(Brown CS and CCMB)

Erica Larschan
(Brown Mol. Biol. and CCMB)

Jeremy Bigness
(CCMB)

github.com/rsinghlab/scNODE

Singh Lab @ Brown

BROWN

CENTER FOR
Computational
Molecular Biology

Funding

NIH National Institutes of Health

1R35HG011939-01

Travel Fellowship

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Conclusion

- scNODE is robust against distribution shifts
- scNODE accurately predicts gene expression
- scNODE assists with temporal downstream analysis

github.com/rsinghlab/scNODE

https://doi.org/10.1093/bioinformatics/btae393

- Extension:
  - Model dynamics from temporal multi-omic data (e.g., transcriptomic and chromatin accessibility)
  - Translate between two omics at any timepoint

COME BY OUR POSTER

(Poster Session 1, P353)