scNODE: Generative Model for Temporal Single Cell Transcriptomic Data Prediction
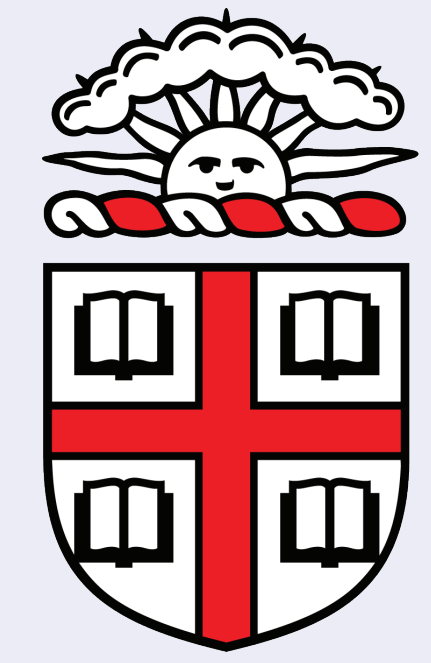
Jiaqi Zhang[1], Erica Larschan[2,3], Jeremy Bigness[2], Ritambhara Singh[1,2]

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

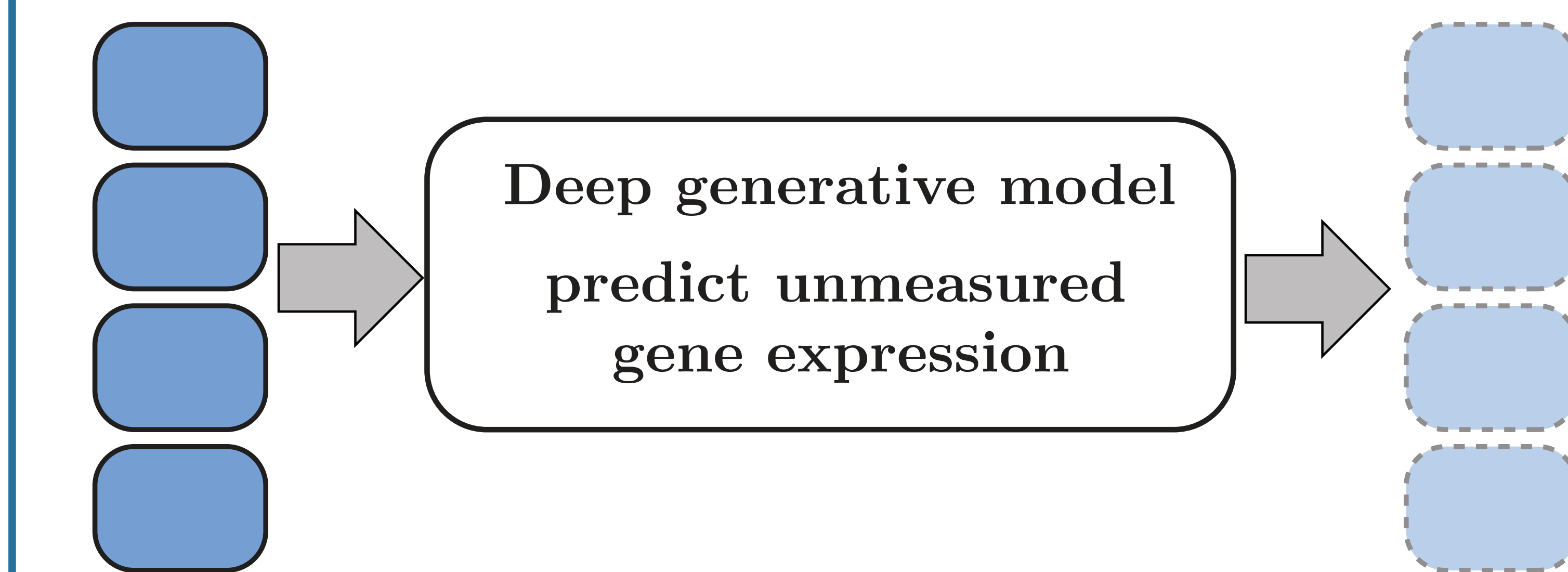[3] Department of Molecular Biology, Cell Biology and Biochemistry, Brown University

BROWN

CENTER FOR Computational Molecular Biology

Introduction

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

measured   unmeasured   timepoint

↳ each measurement is a gene expression matrix
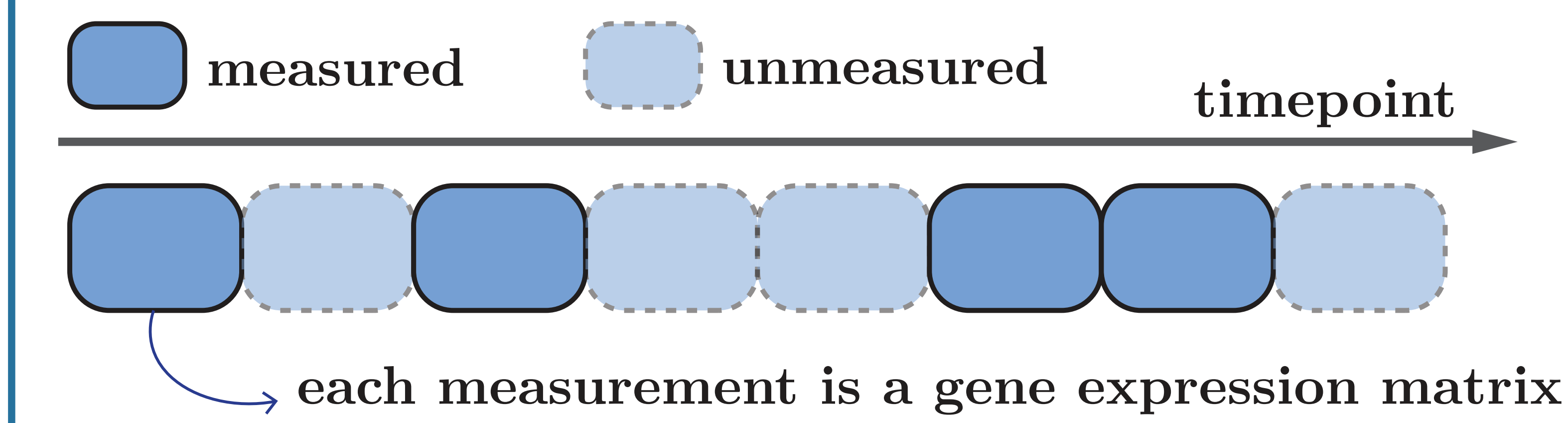
Goal: Predict gene expression at unmeasured timepoints

Deep generative model predict unmeasured gene expression

Limitation of previous works
- Linear dimensionality reduction, incapable to capture complex cell structure
- Fixed latent space obtained from measured timepoints, limiting predictions at unmeasured timepoints

Solution: VAE for latent representation learning + ODE for dynamic modelling + regularization for adjusting latent with overall dynamics
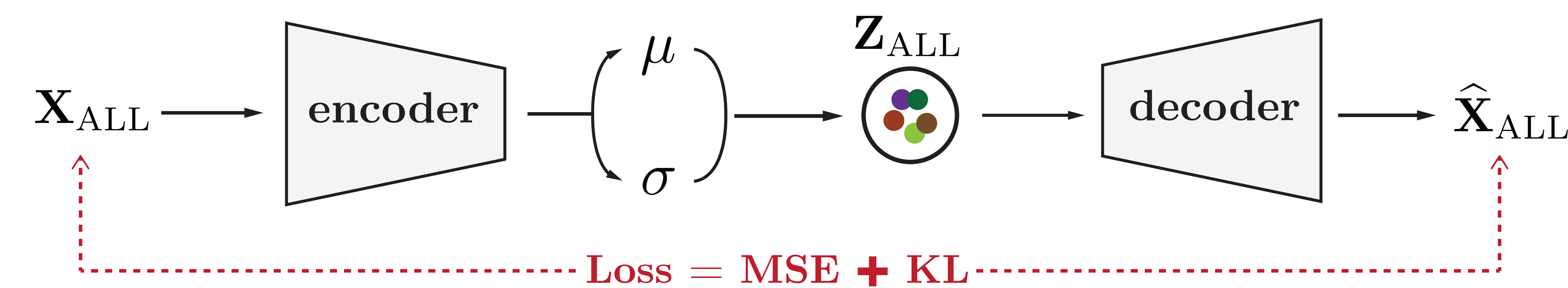
Acknowledgement

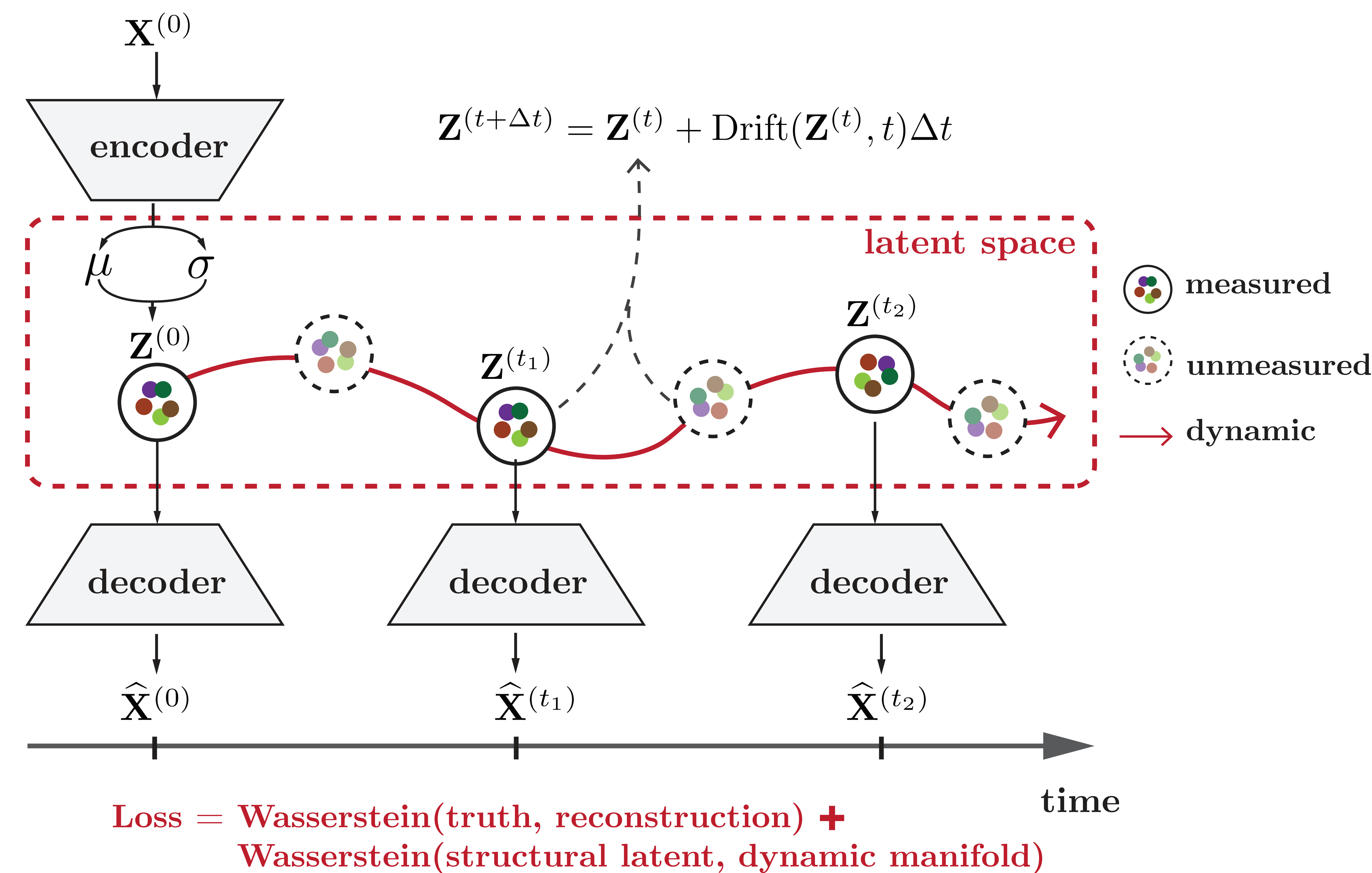### Paper & Codes

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 of observed cells

$$\mathbf{X}_{\text{ALL}} \rightarrow \text{encoder} \rightarrow \binom{\mu}{\sigma} \rightarrow \mathbf{Z}_{\text{ALL}} \rightarrow \text{decoder} \rightarrow \widehat{\mathbf{X}}_{\text{ALL}}$$

Loss = MSE + KL

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

$\mathbf{X}^{(0)}$

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

encoder

$\mu \quad \sigma$

latent space

$\mathbf{Z}^{(0)}$ $\mathbf{Z}^{(t_1)}$ $\mathbf{Z}^{(t_2)}$

● measured
○ unmeasured
→ dynamic

decoder   decoder   decoder

$\widehat{\mathbf{X}}^{(0)}$ $\widehat{\mathbf{X}}^{(t_1)}$ $\widehat{\mathbf{X}}^{(t_2)}$

time

Loss = Wasserstein(truth, reconstruction) + Wasserstein(structural latent, dynamic manifold)

**Output:** Gene expression at any timepoint

Advantages:
- Non-linearity of VAE captures complex cell structural relationship
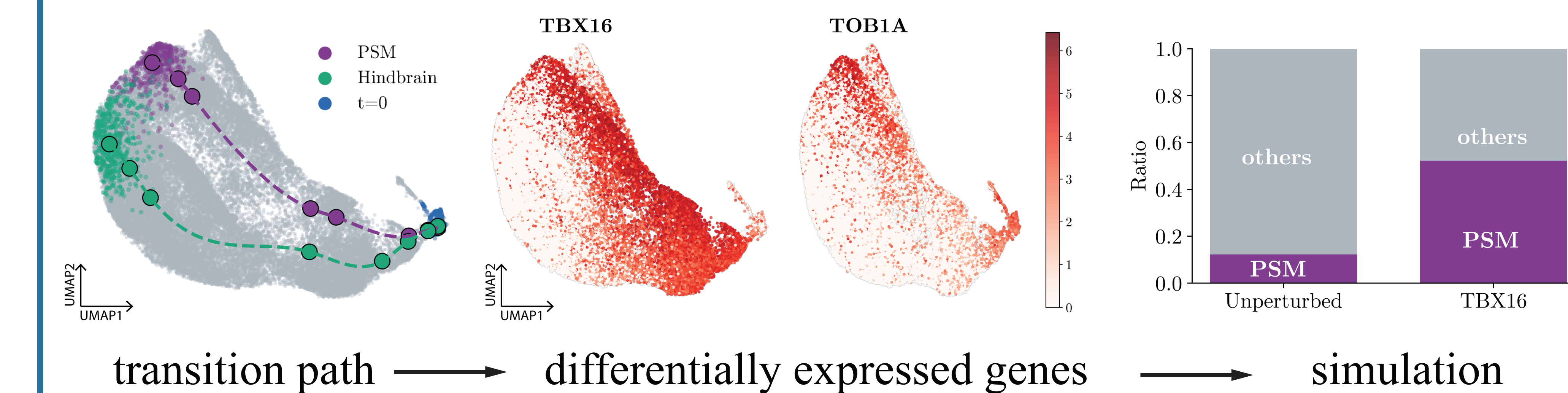- Adjusting latent space with cellular dynamics, improving generalizability

Results

Dataset & Preprocessing
- six 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**) or last few timepoints (**extrapolation**)

scNODE   PRESCIENT   WOT   MIOFlow   Naive
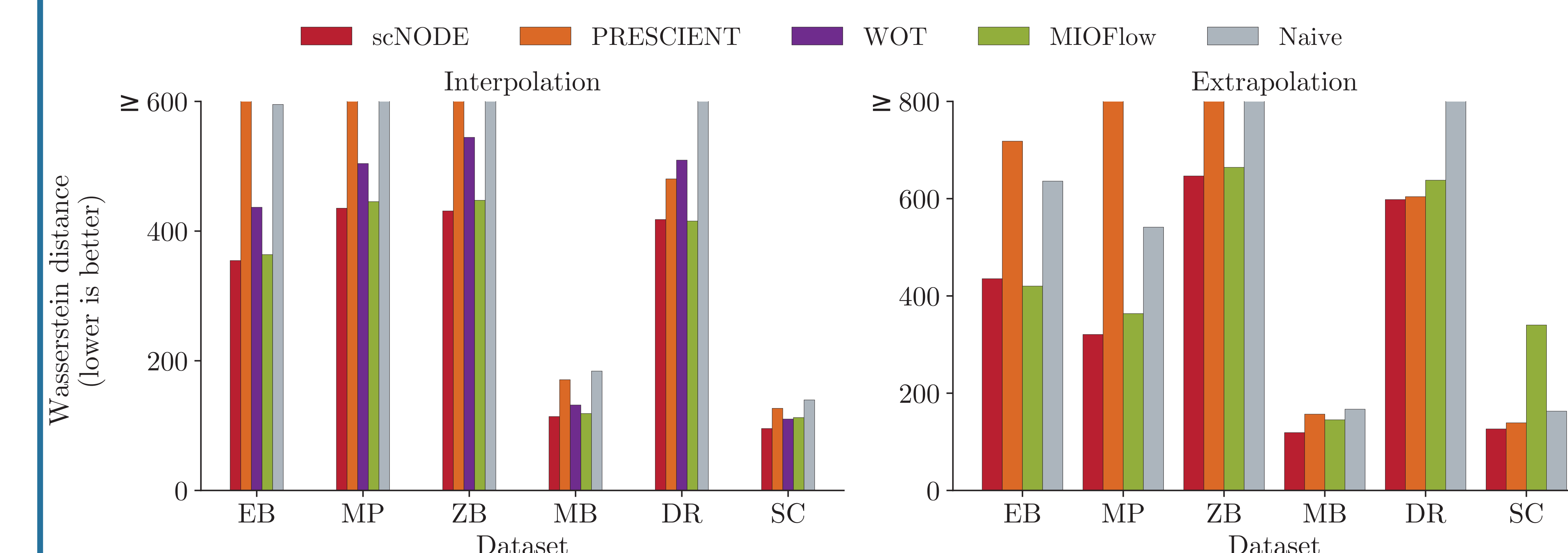
scNODE assists with perturbation analysis

transition path ⟶ differentially expressed genes ⟶ simulation

scNODE predictions help recover cell trajectories

$\mathcal{G}_{\text{true}}$ $\mathcal{G}_{\text{removal}}$ $\mathcal{G}_{\text{pred}}$

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