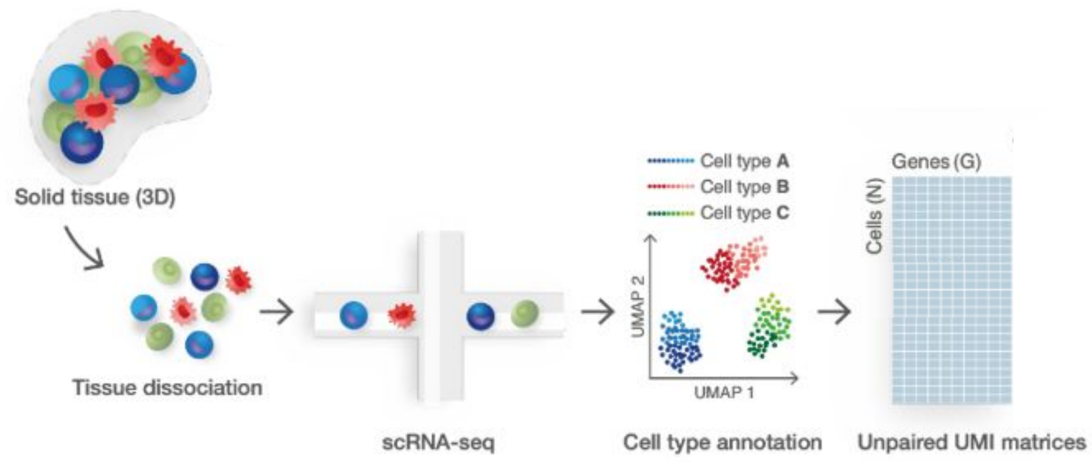
Distance-Preserving Generative Modeling of Spatial Transcriptomics

Wenbin Zhou, Jin-Hong Du

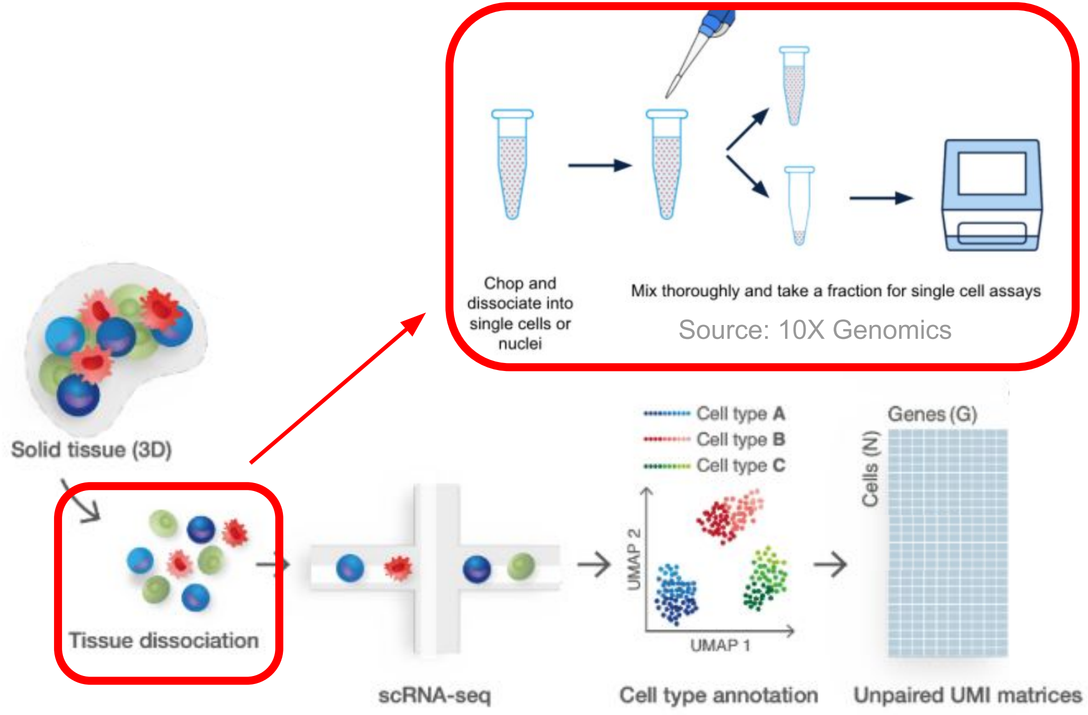
Heinz College of Information Systems and Public Policy
Department of Statistics and Data Science
Machine Learning Department, School of Computer Science
Carnegie Mellon University

Introduction: spatial transcriptomics



[Lopez et al. 2022]

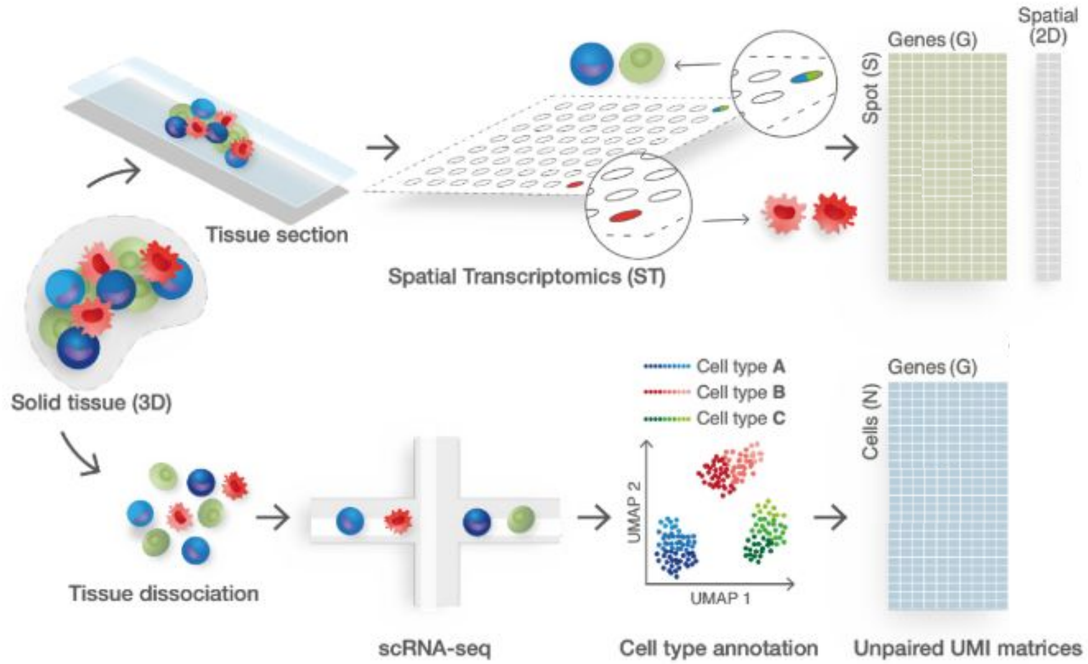
Introduction: spatial transcriptomics



Lose spatial information!

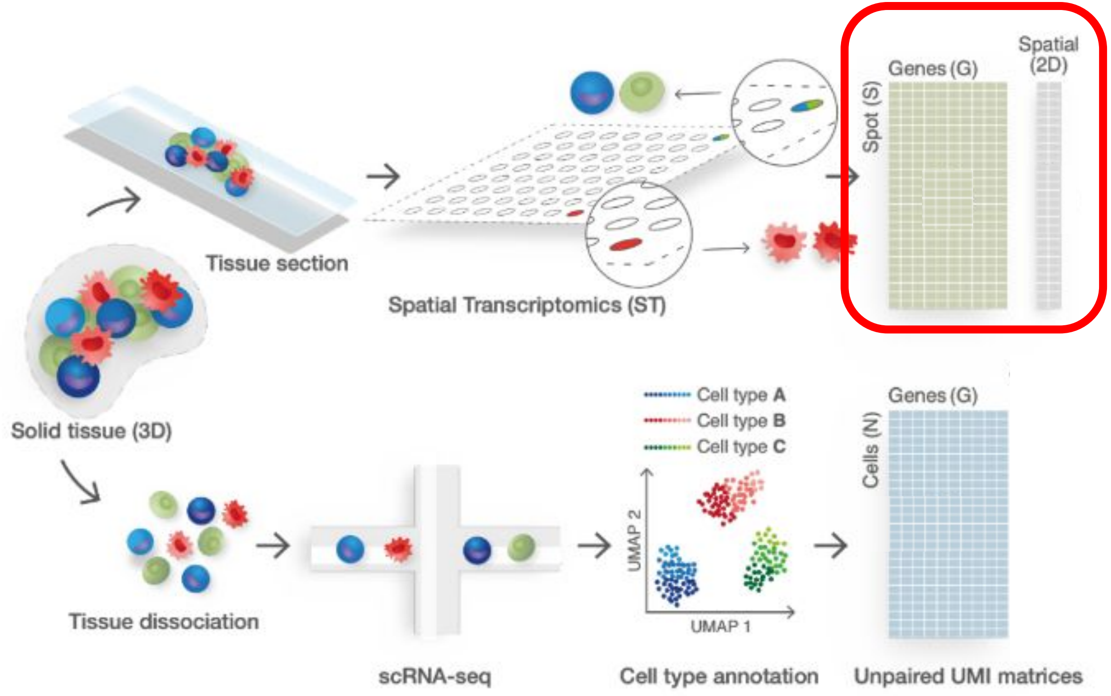
[Lopez et al. 2022]

Introduction: spatial transcriptomics



[Lopez et al. 2022]

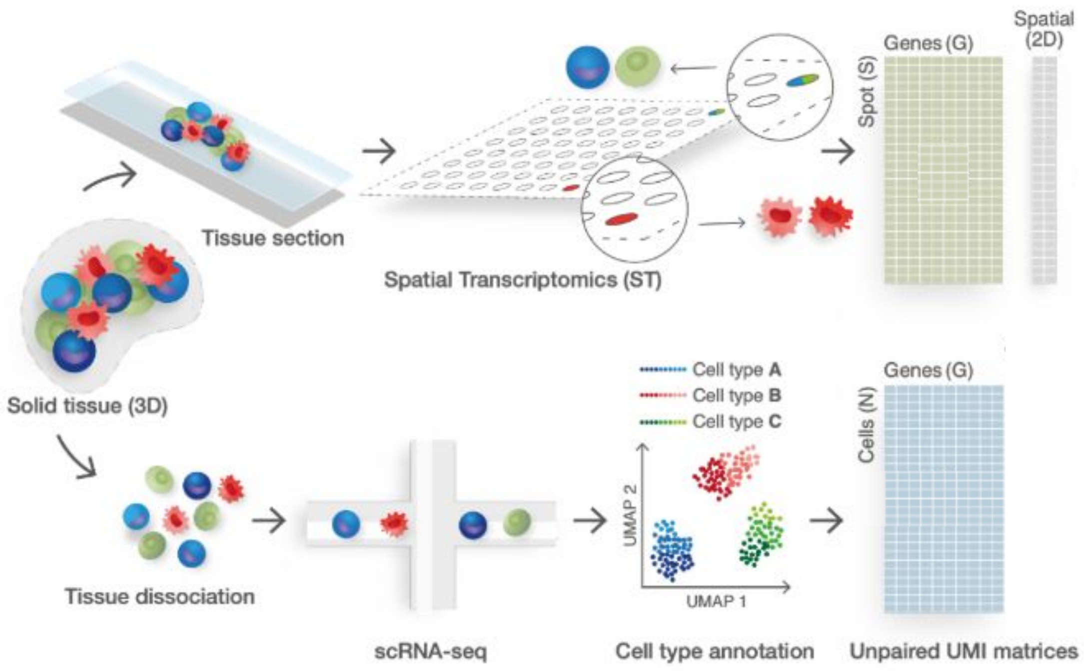
Introduction: spatial transcriptomics



[Lopez et al. 2022]

- ❑ Gene expression matrix
- ❑ Spatial information matrix

Introduction: spatial transcriptomics



[Lopez et al. 2022]

- ❑ Gene expression matrix
- ❑ Spatial information matrix

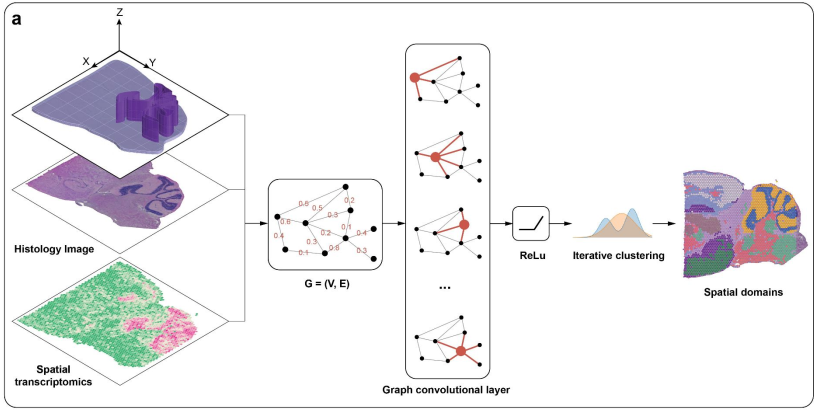
Efficiently utilize the spatial information…

Related works

1. Machine Learning for Spatial Transcriptomics
    - ❑ Graph-based [Zhu et al. 2020, Hu et al. 2021, ...]
    - ❑ Generative Models [Lopez et al. 2018, ...]

Related works

1. Machine Learning for Spatial Transcriptomics
   - ❑ Graph-based [Zhu et al. 2020, Hu et al. 2021, ...]
   - ❑ Generative Models [Lopez et al. 2018, ...]



[Hu et al. 2021]

***Con:** They have limited usage for different downstream analysis*

Related works

1. Machine Learning for Spatial Transcriptomics
   - ❑ Graph-based [Zhu et al. 2020, Hu et al. 2021, ...]
   - ❑ Generative Models [Lopez et al. 2018, ...]

Pro: *Flexible for different downstream tasks: identify cell types or subtypes, batch correction, visualization, clustering, and differential expression…*
Con: *How to encode spatial Information?*

Related works

1. Machine Learning for Spatial Transcriptomics
   - ❑ Graph-based [Zhu et al. 2020, Hu et al. 2021, ...]
   - ❑ Generative Models [Lopez et al. 2018, ...]
   - ❑ ***Challenge****: How to efficiently model spatial information when building generative models for spatial transcriptomics?*

Related works

2. Geometry-preserving Generative Model
   - ❑ Isometry [Beshkov et al. 2022, ...]
   - ❑ Constrained-optimization [Chen et al. 2022, ...]

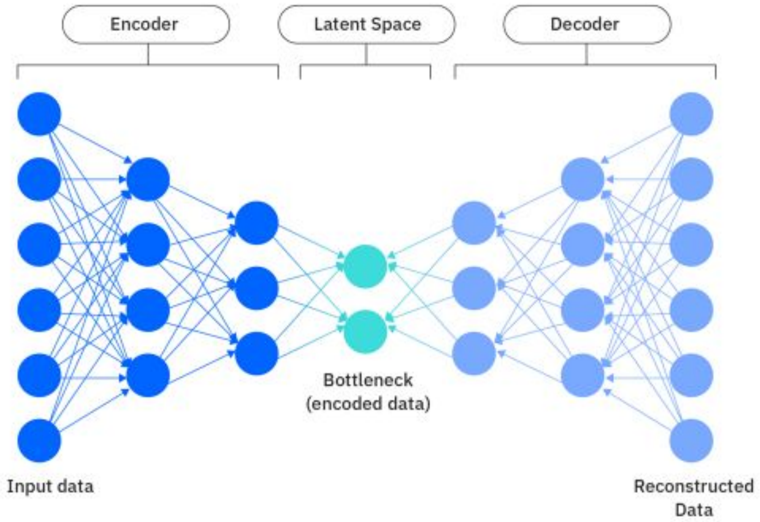
Related works

2. Geometry-preserving Generative Model
   - ❑ Isometry [Beshkov et al. 2022, ...]
   - ❑ Constrained-optimization [Chen et al. 2022, ...]
   - ❑ Existing works have adopt the idea of geometry-preserving generation in computer vision tasks, while direct application to spatial data is not straightforward.
   - ❑ ***Questions:*** *How to incorporate the idea of geometry-preserving in studying spatially-resolved gene expressions?*

This paper

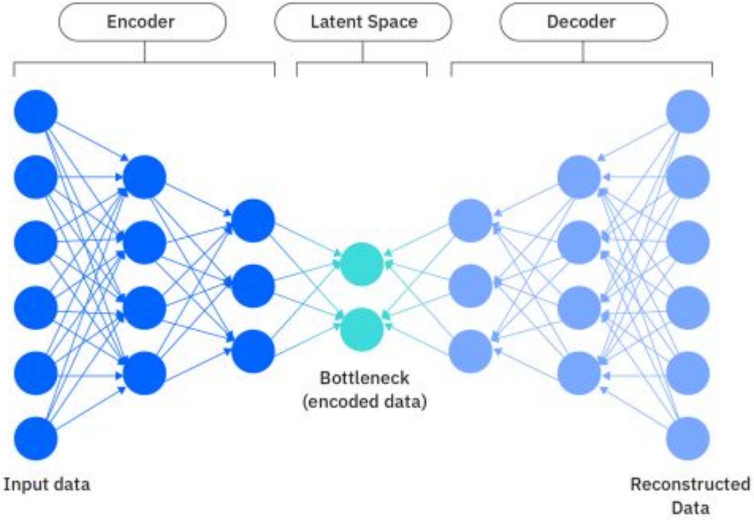
- ❏ **Background**
- ❏ Introduce distance-preserving generative model
- ❏ Deriving loss function and model specification
- ❏ Experiment on mouse brain tissues Visium dataset

Background: Variational Autoencoders



Source: IBM

Background: Variational Autoencoders



Source: IBM

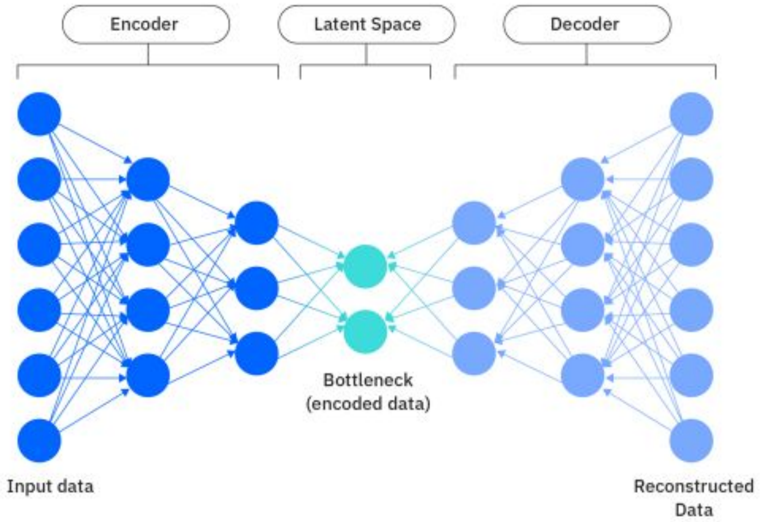
$$\ell_{\mathrm{VAE}} = \sum_{i=1}^{N} -\log p_\phi(y_i|z_i) + D_{\mathrm{KL}}\left(q_\theta(z_i|y_i) \| p(z_i)\right).$$

Loss function of VAE (evidence lower bound)

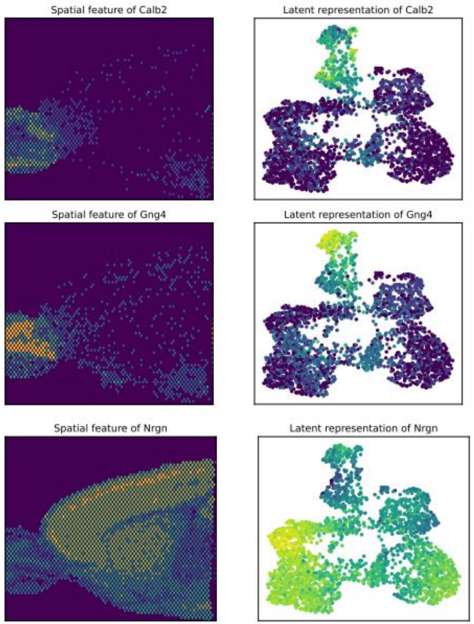
Notation:

- ❑ y denotes gene expression
- ❑ z denotes latent representation
- ❑ p_ϕ, q_θ denotes encoder and decoder network (distributions!)

Background: Variational Autoencoders

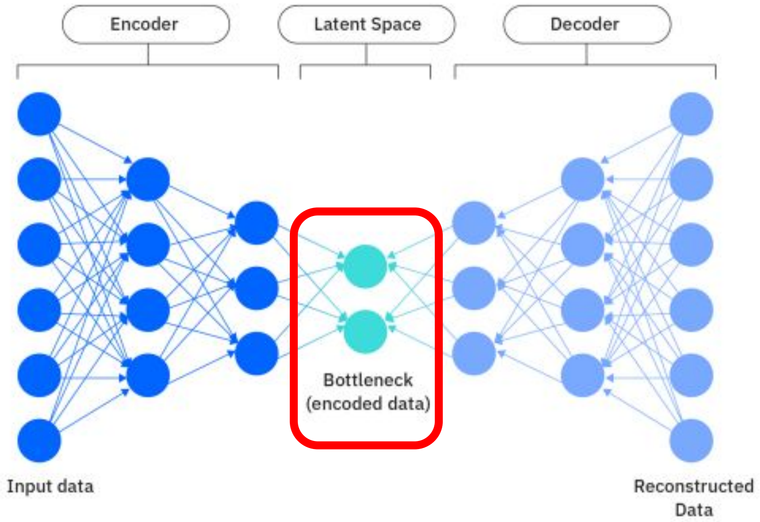


Source: IBM

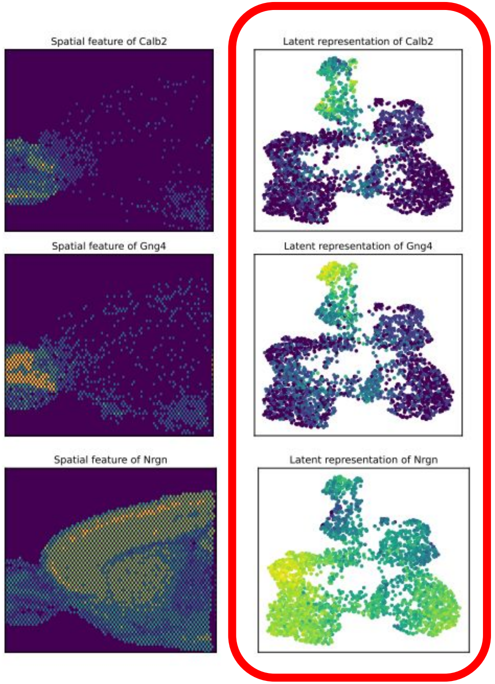


Spatial feature of gene Calb2, Gng4, Nrgn (left) and their latent representation extracted using VAE (right)

Background: Variational Autoencoders



Source: IBM

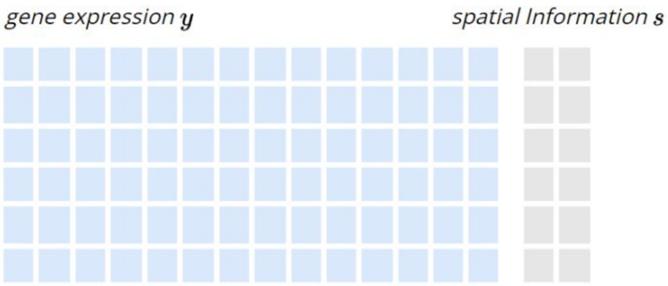


Spatial feature of gene Calb2, Gng4, Nrgn (left) and their latent representation extracted using VAE (right)

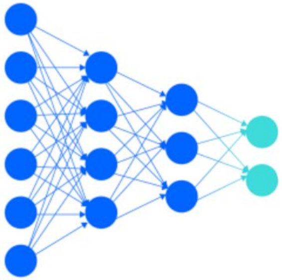
(Without spatial information)

Idea

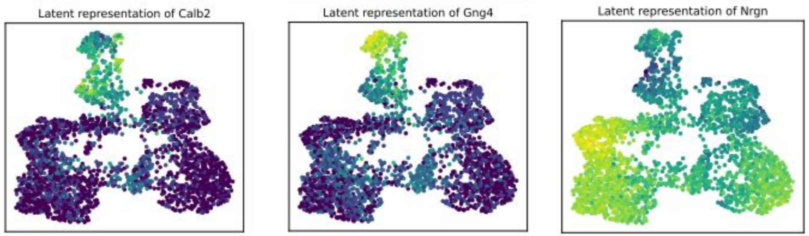
Data



Encoder

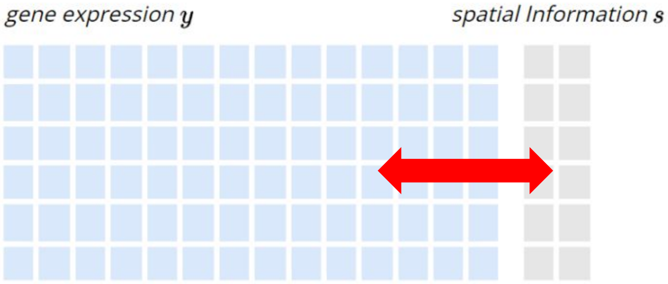


Latent Representation

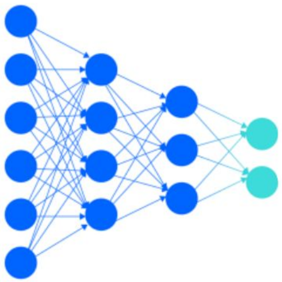


Idea

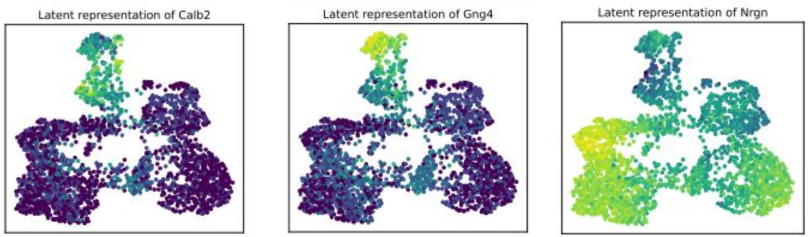
Data

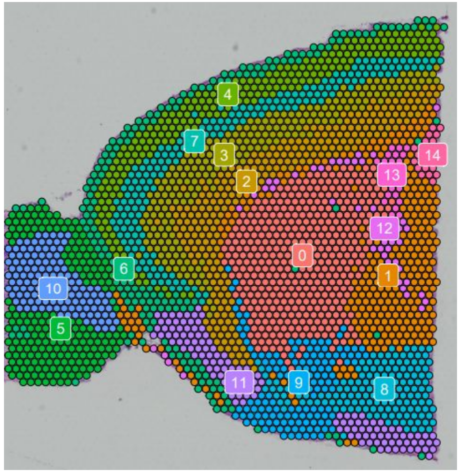


Encoder



Latent Representation

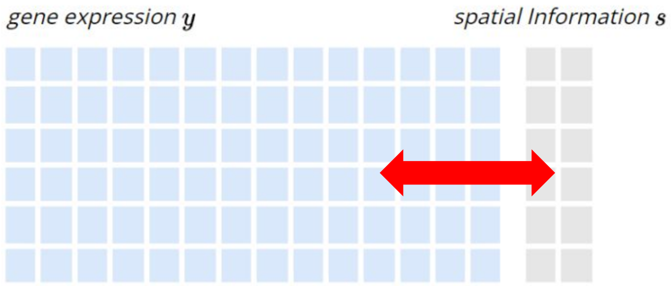




Within each indexed spatial domains, gene expressions exhibit similar properties

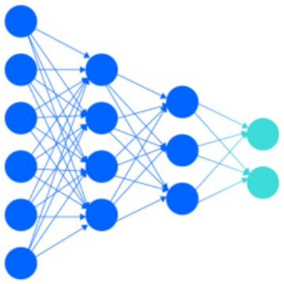
Idea

Data

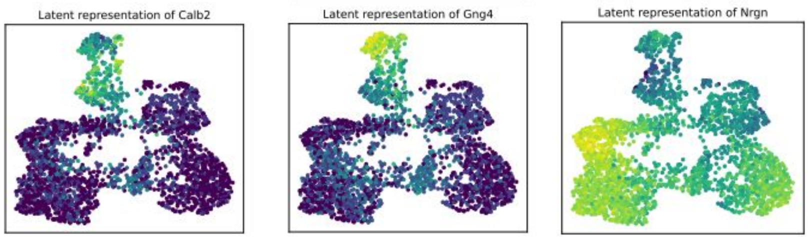


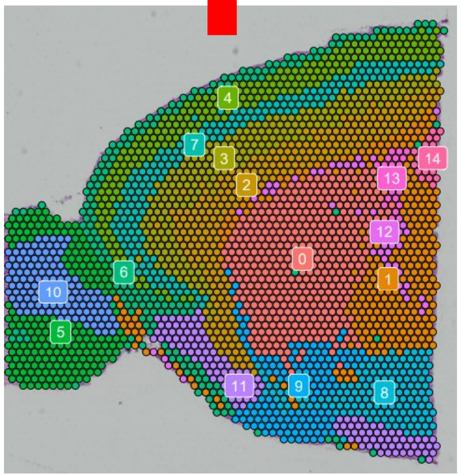
(Locally) geometric-preserving

Encoder



Latent Representation





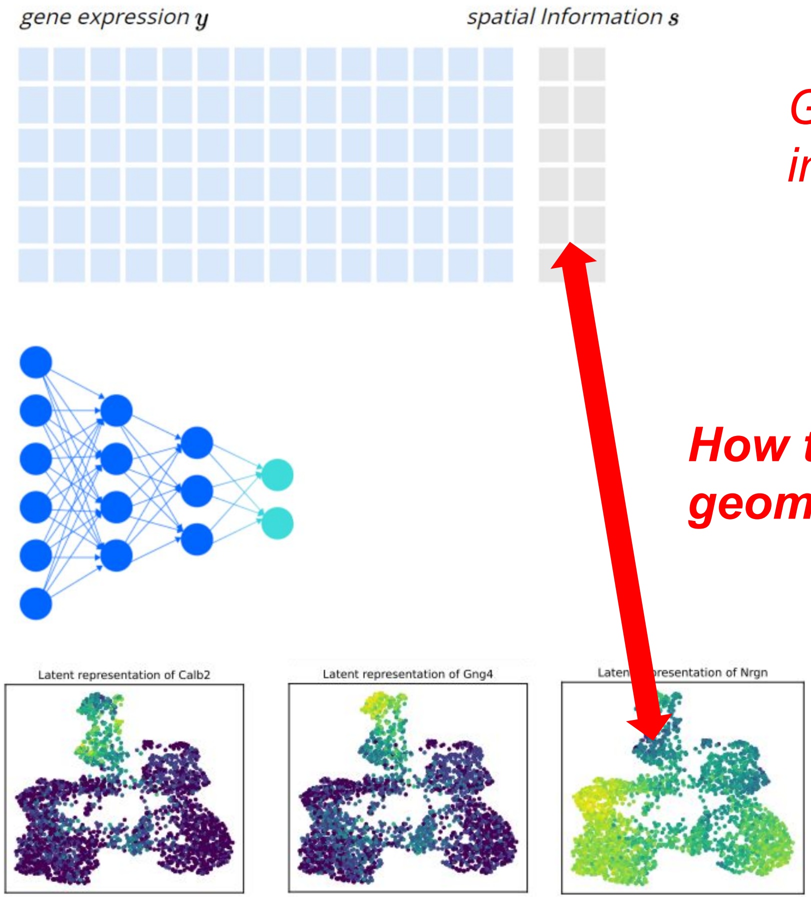
Within each indexed spatial domains, gene expressions exhibit similar properties

Idea

Data

Encoder

Latent Representation



Geometric similarity in the original space

How to encode spatial geometric properties?

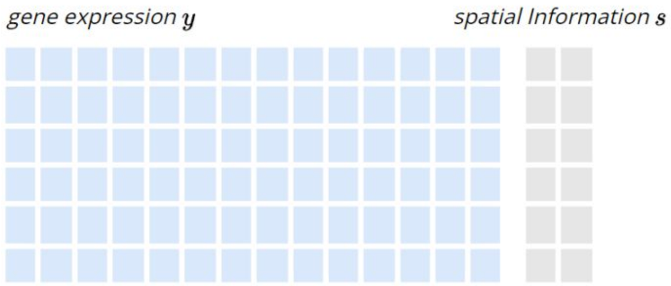
Geometric similarity in the latent space

This paper

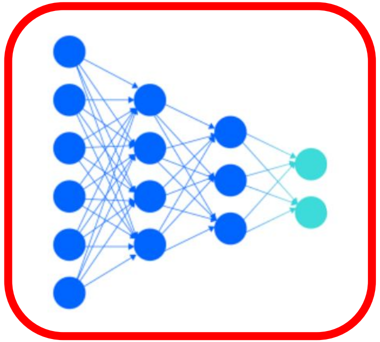
- ❏ Background
- ❏ **Introduce distance-preserving generative model**
- ❏ Deriving loss function and model specification
- ❏ Experiment on mouse brain tissues Visium dataset

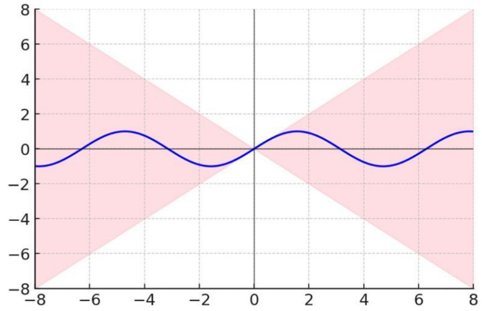
Idea

Data



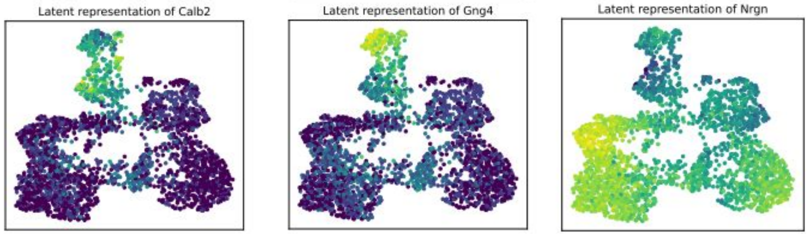
Encoder





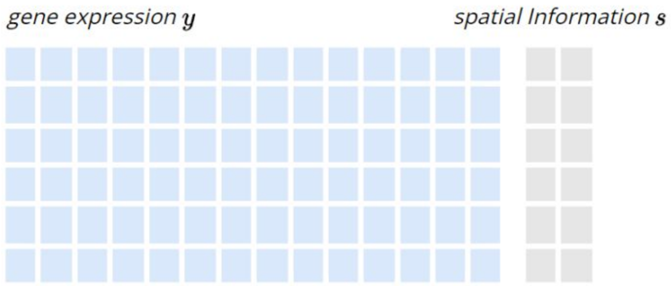
Ex. A Lipschitz function

Latent Representation

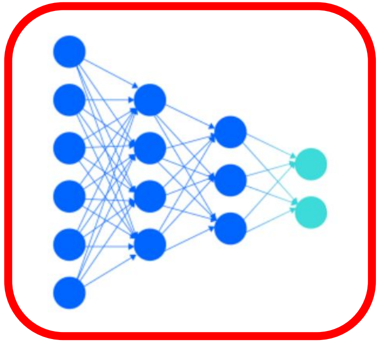


Idea

Data

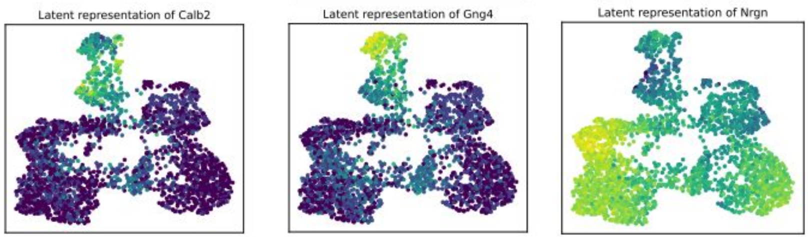


Encoder



How to define “smooth” probabilistic encoder networks?

Latent Representation



Distance-preserving generative model

Definition (simplified)

A **distance-preserving generative model** satisfies:

$$\mathbb{P}\left(\lambda d_{\mathcal{S}}(s, s') \leq d_{\mathcal{Z}}(z, z') \leq L \cdot \lambda d_{\mathcal{S}}(s, s')\right) \geq 1-\epsilon,$$

where:

- $d_{\mathcal{S}}, d_{\mathcal{Z}}$ denote the **spatial distance** and **latent distance** metric
- s, s', z, z' denotes the generation process of the generative model
- λ is some arbitrary constant
- L is the distortion constant
- ϵ is the error parameter

Distance-preserving generative model

Definition (simplified)

A **distance-preserving generative model** satisfies:

$$\mathbb{P}\left(\lambda d_{\mathcal{S}}(s, s') \leq d_{\mathcal{Z}}(z, z') \leq L \cdot \lambda d_{\mathcal{S}}(s, s')\right) \geq 1-\epsilon,$$

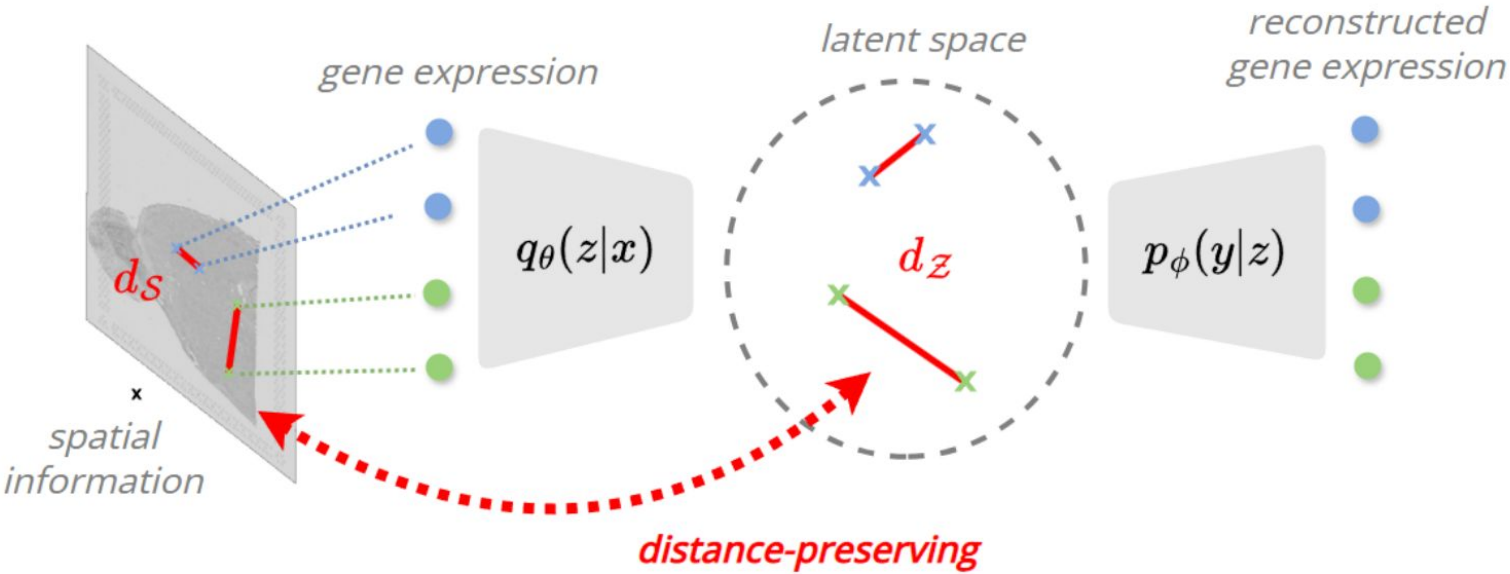
where:

- ❑ $d_{\mathcal{S}}, d_{\mathcal{Z}}$ denote the **spatial distance** and **latent distance** metric
- ❑ s, s', z, z' denotes the generation process of the generative model
- ❑ λ is some arbitrary constant
- ❑ L is the distortion constant
- ❑ ϵ is the error parameter

- ❑ *Measures how unsmooth the probabilistic encoder function is.*
- ❑ *Measures the maximum allowance of the ratio of outliers*

Distance-preserving generative model

Consider 4 cells/samples indicated by dots:



An illustration of the definition of distance-preserving generative models

Distance-preserving generative model

Definition (simplified)

A **distance-preserving generative model** satisfies:

$$\mathbb{P}\left(\lambda d_{\mathcal{S}}(s, s') \leq d_{\mathcal{Z}}(z, z') \leq L \cdot \lambda d_{\mathcal{S}}(s, s')\right) \geq 1-\epsilon,$$

where:

- ❑ $d_{\mathcal{S}}, d_{\mathcal{Z}}$ denote the **spatial distance** and **latent distance** metric
- ❑ s, s', z, z' denotes the generation process of the generative model
- ❑ λ is some arbitrary constant
- ❑ L is the distortion constant
- ❑ ϵ is the error parameter

How to tractably learn a distance-preserving generative model?

This paper

- ❑ Background
- ❑ Introduce distance-preserving generative model
- ❑ **Deriving loss function and model specification**
- ❑ Experiment on mouse brain tissues Visium dataset

Model: distortion loss

Theorem (simplified)

Define the following **(population) distortion loss**:

$$\mathcal{L}_{\text{DIS}} = \mathbb{E}\left[\left|d_{\mathcal{Z}}(z, z') - \lambda \cdot d_{\mathcal{S}}(s, s')\right|\right],$$

where the expectation is taken w.r.t. the randomness of the generation process. Given some fixed error parameter ϵ, the *distortion constant* can be bounded as:

$$L \leq C + \mathcal{O}\left(\frac{\mathcal{L}_{\text{DIS}}}{\epsilon}\right),$$

where C is some constant that depends on the probabilistic structure of the generation process.

Takeaway: To minimize L, it is equivalent to minimize $\mathcal{L}_{\text{DIS}}$.

Model

The proposed objective is given by:

$$\min_{\theta,\phi,\lambda} \ell := \ell_{\text{VAE}} + \alpha \bar{\ell}_{\text{DIS}}, \qquad \bar{\ell}_{\text{DIS}} = \frac{1}{N^2} \|G \odot D_z - \lambda \cdot G \odot D_s\|_1 ,$$

Model

The proposed objective is given by:

weighted matrix (e.g. adjacency matrix)

$$\min_{\theta,\phi,\lambda} \ell := \ell_{\mathrm{VAE}} + \alpha \bar{\ell}_{\mathrm{DIS}},$$

$$\bar{\ell}_{\mathrm{DIS}} = \frac{1}{N^2} \| G \odot D_z - \lambda \cdot G \odot D_s \|_1 ,$$

pairwise distance matrices
in the latent and spatial spaces

Model

The proposed objective is given by:

$$\min_{\theta,\phi,\lambda} \ell := \ell_{\text{VAE}} + \alpha \bar{\ell}_{\text{DIS}},$$

weighted matrix (e.g. adjacency matrix)

$$\bar{\ell}_{\text{DIS}} = \frac{1}{N^2} \|G \odot D_z - \lambda \cdot G \odot D_s\|_1 ,$$

pairwise distance matrices
in the latent and spatial spaces

- **Tractability**:
  - Unconstrained optimization problem
  - The distortion loss is decomposable
- **Flexibility**: allowing arbitrary VAE architecture and models

This paper

- ❑ Background
- ❑ Introduce distance-preserving generative model
- ❑ Deriving loss function and model specification
- ❑ **Experiment on mouse brain tissues Visium dataset**

Distance-preserving generative modeling improve representation learning.

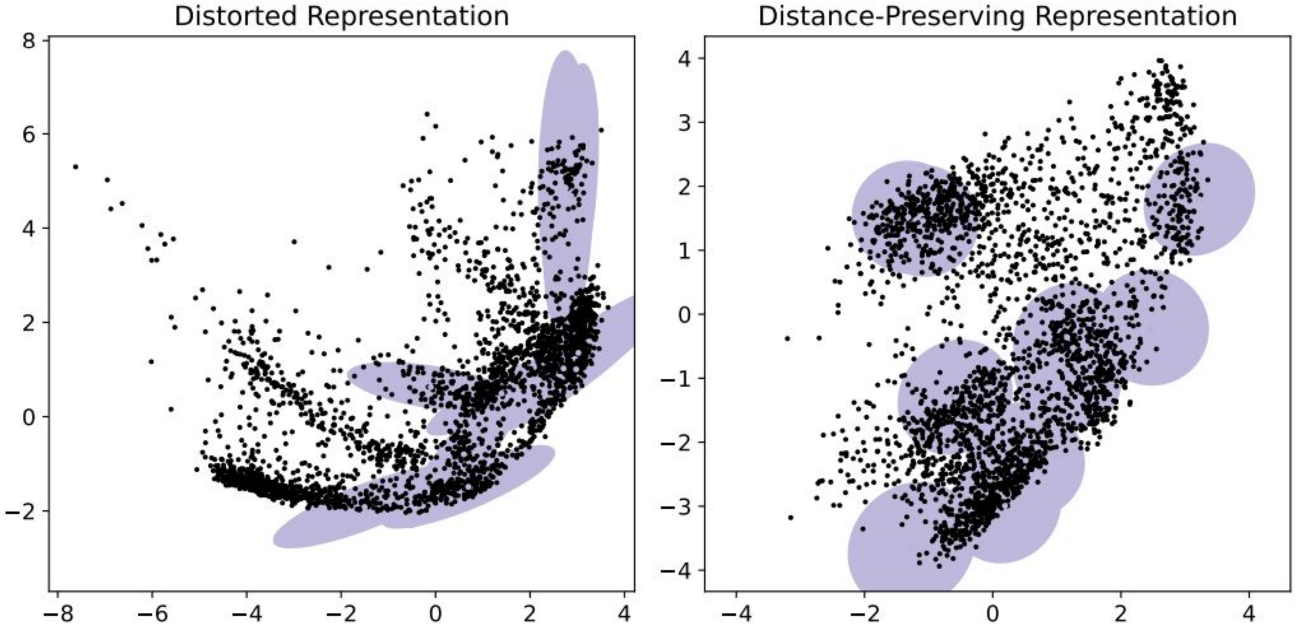


Figure 4. Visualization of latent representation space obtained from scVI (left) and scVI regularized with distortion loss (right). More isotropic and homogeneous ellipses indicate more distance-preserving.

Spatial correlations preserved in the latent space

Table 1. Moran's I and Geary's C of the latent representation extracted by scVi and VAE on 4 test datasets, with and without distance-preserving penalty, averaged over 5 repeated trials. Enforcing distance-preserving property induces stronger spatial autocorrelations.

	Method	Moran's I				Geary's C			
		A2	A1	P2	P1	A2	A1	P2	P1
Distorted	VAE	0.62(0.07)	0.55(0.05)	0.52(0.05)	**0.52(0.03)**	0.36(0.06)	0.41(0.03)	0.49(0.05)	**0.43(0.03)**
	scVI	0.43(0.03)	0.52(0.04)	0.37(0.02)	0.45(0.04)	0.57(0.03)	0.48(0.04)	0.62(0.02)	0.55(0.04)
Distance Preserving	VAE	**0.64(0.02)**	**0.60(0.03)**	**0.56(0.06)**	0.49(0.06)	**0.35(0.02)**	**0.37(0.02)**	**0.45(0.07)**	0.46(0.05)
	scVI	**0.45(0.04)**	0.52(0.04)	**0.43(0.02)**	**0.47(0.03)**	**0.55(0.05)**	0.48(0.04)	**0.57(0.02)**	**0.53(0.03)**

Smaller reconstruction errors on several baseline models

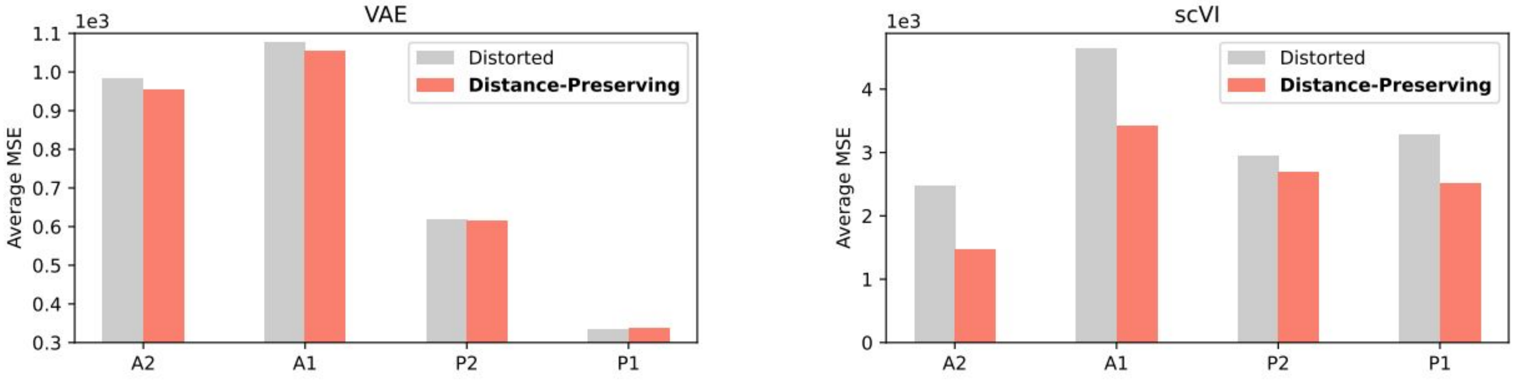


Figure 5. Mean squared error (MSE) of scVi and VAE on 4 test datasets, with and without distance-preserving penalty, averaged over 5 repeated trials. Enforcing distance-preserving property induces smaller reconstruction errors of log-normalized and library-size-adjusted data.

Stability and robustness

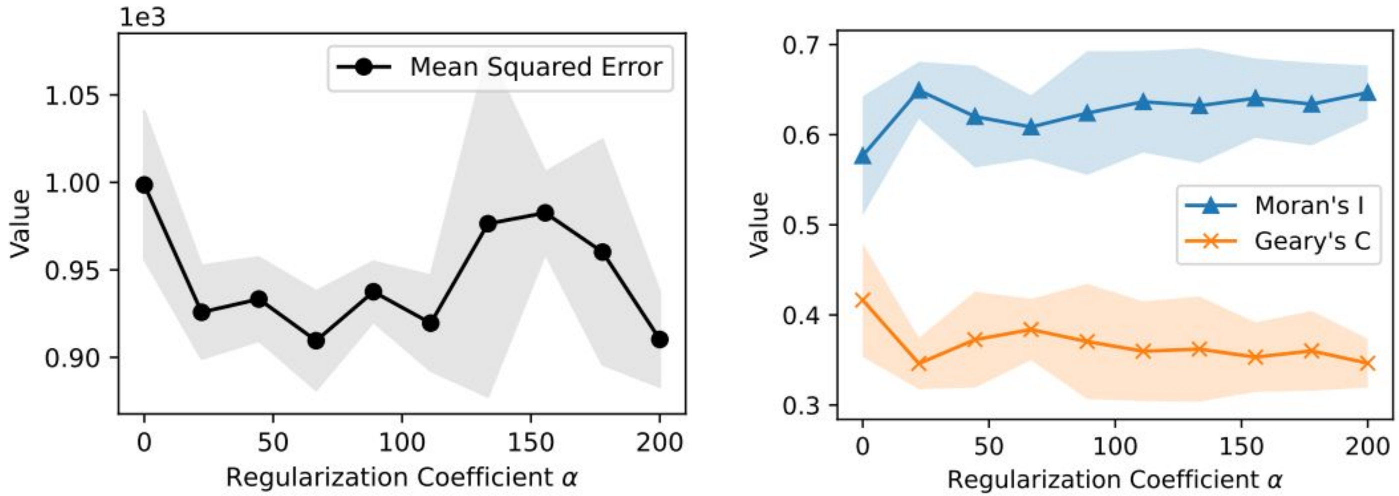


Figure 6. Sensitivity analysis of distance-preserving regularization strength in the performance.

Acknowledgement & contacts

Part of the work was done while Wenbin is working as a research intern at Argonne National Laboratory.

Feel free to contact us via email:

- ❏ Wenbin Zhou: wenbinz2@andrew.cmu.edu
- ❏ Jin-Hong Du: jinhongd@andrew.cmu.edu