인공지능 기반 scRNA-seq Data 분석

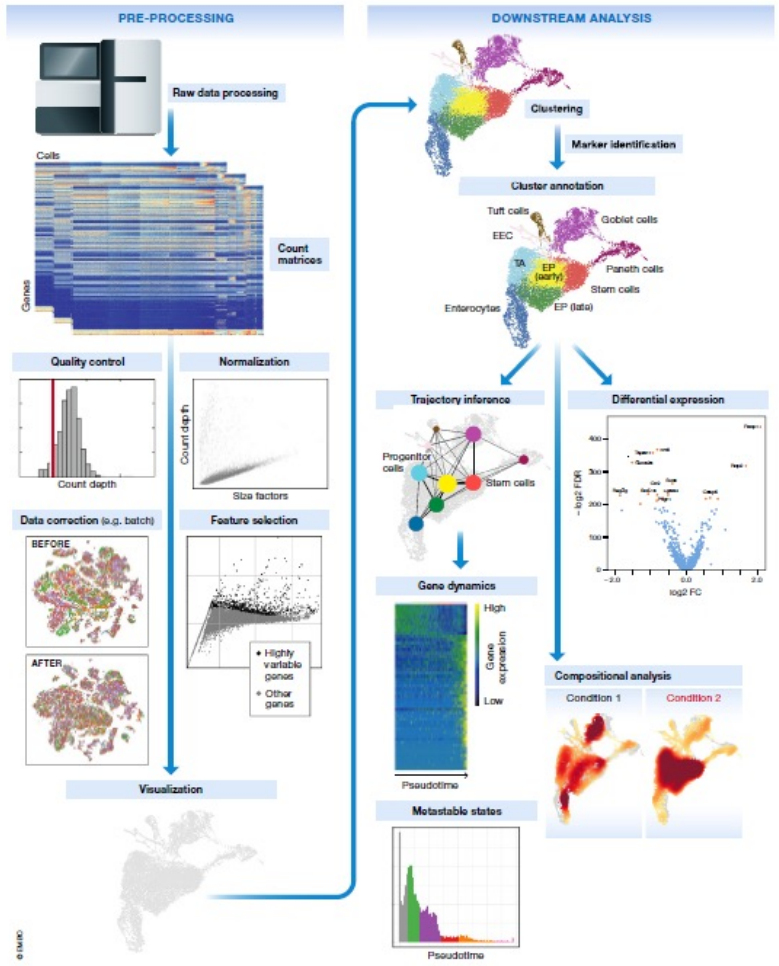
박찬영

Korea Advanced Institute of Science and Technology (KAIST)

DSAIL Data Science & Artificial Intelligence

BACKGROUND Single Cell RNA-sequencing (scRNA-seq)

Pipeline of scRNA-seq Data analysis



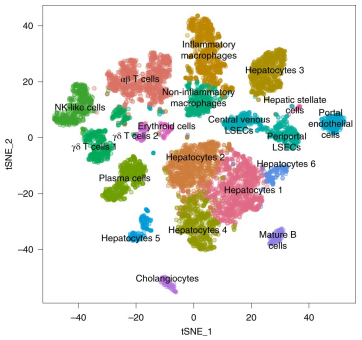
1) Generate Raw data using sequencing machine

2) Pre-processing

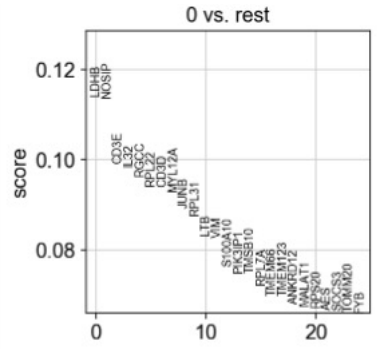
- Single-cell RNA Sequencing Data Imputation Using Bi-level Feature Propagation, **Briefings in Bioinformatics 2024**
- Single-cell RNA-seq data imputation using Feature Propagation, **ICML 2023 CompBio Workshop (Best Paper Award)**

3) Downstream Analysis

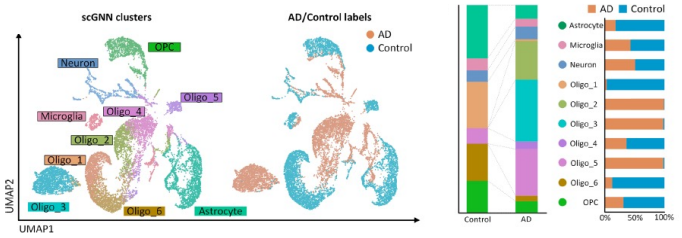
- Cell Annotation (Clustering)
  - Deep Single-cell RNA-seq Data Clustering with Graph Prototypical Contrastive Learning, **Bioinformatics 2023**
- Differentially expression gene analysis
- Trajectory inference
- Compositional analysis



Cell Annotation (Clustering)



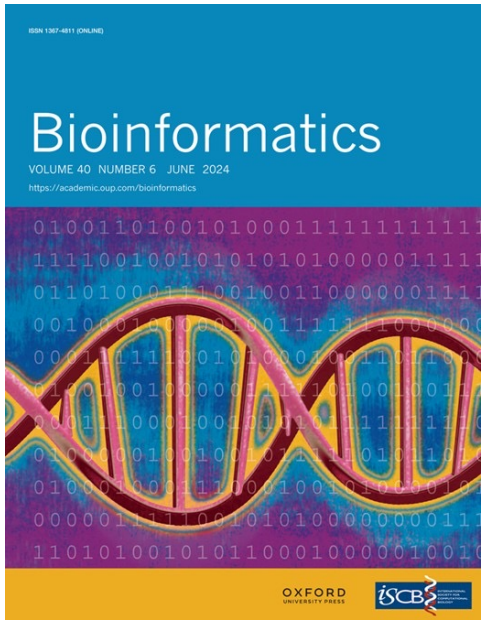
Differentially expression gene analysis



Compositional analysis

[Molecular Systems Biology 2019] Current best practices in single-cell RNA-seq analysis: a tutorial

Bioinformatics, 2023, **39(6)**, btad342
https://doi.org/10.1093/bioinformatics/btad342
Advance access publication 26 May 2023
Original Paper



Gene expression

Deep single-cell RNA-seq data clustering with graph prototypical contrastive learning

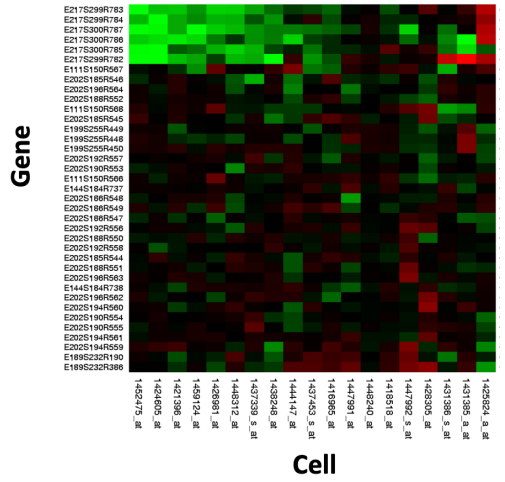
Junseok Lee [1], Sungwon Kim[1], Dongmin Hyun[2], Namkyeong Lee[1], Yejin Kim [3], Chanyoung Park[1,*]

[1]Department of Industrial and Systems Engineering, KAIST, Daejeon 34141, Republic of Korea
[2]Institute of Artificial Intelligence, POSTECH, Pohang 37673, Republic of Korea
[3]Center for Safe Artificial Intelligence for Healthcare, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, United States

*Corresponding author. Department of Industrial and Systems Engineering, 4104, E2-2, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. E-mail: cy.park@kaist.ac.kr

Associate Editor: Valentina Boeva

CHALLENGES



Target task **Clustering on scRNA-seq Data**

Challenge 1 High dimensionality

- Traditional clustering methods (e.g., K-means clustering) suffer from curse of dimensionality

Challenge 2 Dropout phenomena → Noisy feature

	Gene1	Gene2	Gene3	Gene 4
Cell 1	18	1010	0	22
Cell 2	0	506	49	2
Cell 3	0	49	0	0

Ground Truth

Sequencing Machine

	Gene1	Gene2	Gene3	Gene 4
Cell 1	18	1010	0	22
Cell 2	0	506	49	0
Cell 3	0	0	0	0

Observed

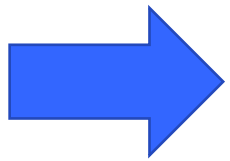
- Do zero counts reflect a 1) true absence of expression? or 2) technical issues (dropout)?

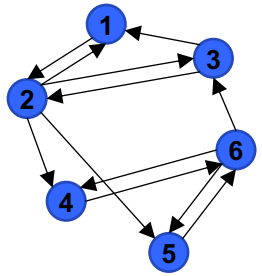
MOTIVATION PERFORMANCE DEGRADATION OF CELL-CELL GRAPH WITH SPARSE DATA

Problem: The quality of Cell-Cell graph highly depends on the quality of given scRNA-seq data

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Cell 1	5	0	18	0	0	0	0
Cell 2	10	29	0	5	0	0	0
Cell 3	0	40	5	0	10	0	0
Cell 4	0	0	0	19	0	37	0
Cell 5	9	0	0	0	0	0	3
Cell 6	0	0	0	97	7	0	8

Calculate
Cell-Cell Similarity



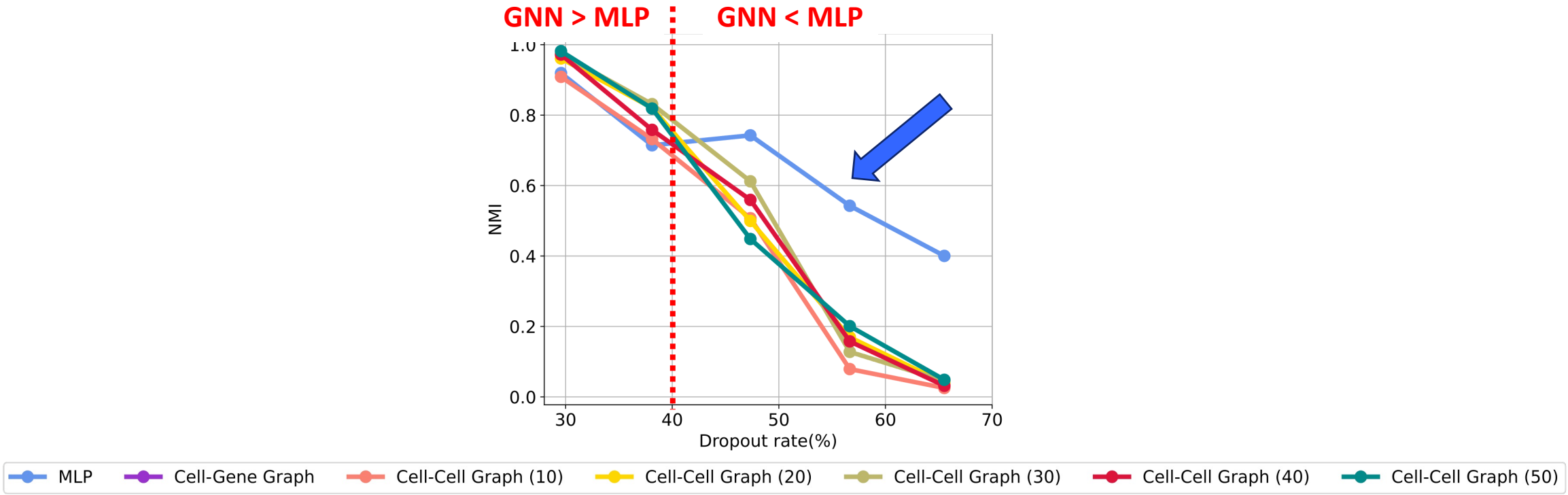


KNN Graph

K (=2) Nearest Neighbor Construction

위와 같이 K-NN 기반으로 graph가 생성되기 때문에 sparse하고 noisy한 matrix가 주어지면
필연적으로 low-quality graph가 생성

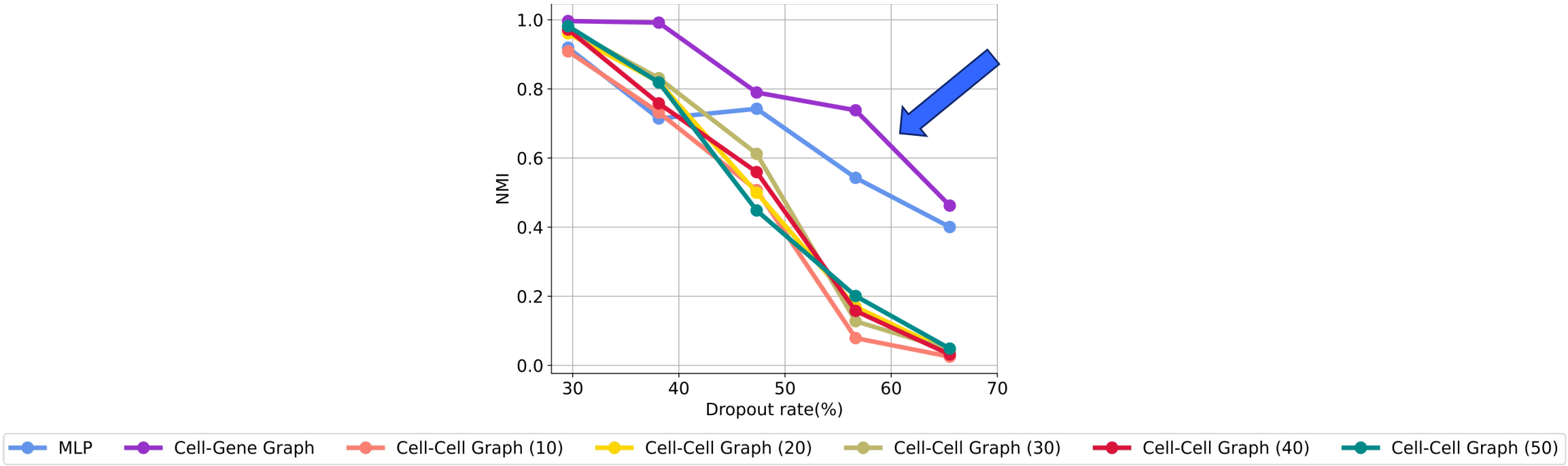
MOTIVATION PERFORMANCE DEGRADATION OF CELL-CELL GRAPH WITH SPARSE DATA



Clustering Performance over various dropout rates on cell-gene matrix

MLP performs better than GNNs as dropout rate increases

MOTIVATION PERFORMANCE DEGRADATION OF CELL-CELL GRAPH WITH SPARSE DATA



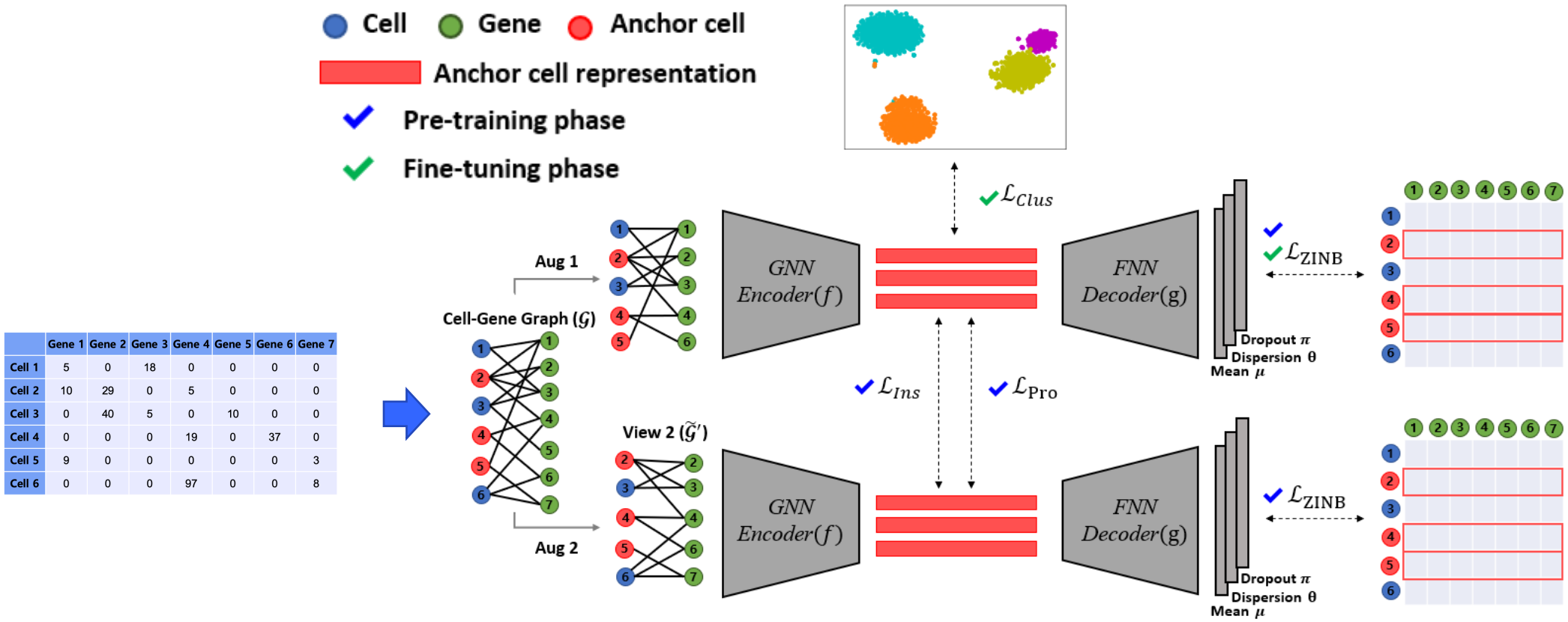
Clustering Performance over various dropout rates on cell-gene matrix

Our Idea Leverage the co-expression based **cell-gene graph** that preserves the natural relationship

Single-cell Graph Prototypical Contrastive Learning (scGPCL)

Key Idea

- Leverage the relationship information inherent in the scRNA-seq data using bipartite cell-gene graph
- Do not depend on the reconstruction-based loss that highly depends on the data quality (Use Contrastive loss)



	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Cell 1	5	0	18	0	0	0	0
Cell 2	10	29	0	5	0	0	0
Cell 3	0	40	5	0	10	0	0
Cell 4	0	0	0	19	0	37	0
Cell 5	9	0	0	0	0	0	3
Cell 6	0	0	0	97	0	0	8

EXPERIMENTS

Simulated Data Assume three situations in which learning cell representations may be challenging

- Case 1 : Gene-expression matrix is highly sparse due to the dropout phenomena
- Case 2 : Gene-expression values contain relatively low signal strength required for clustering
- Case 3 : Subgroups of the cells are imbalanced in number

Real Data Conduct experiments on real scRNA-seq datasets over various sequencing platforms

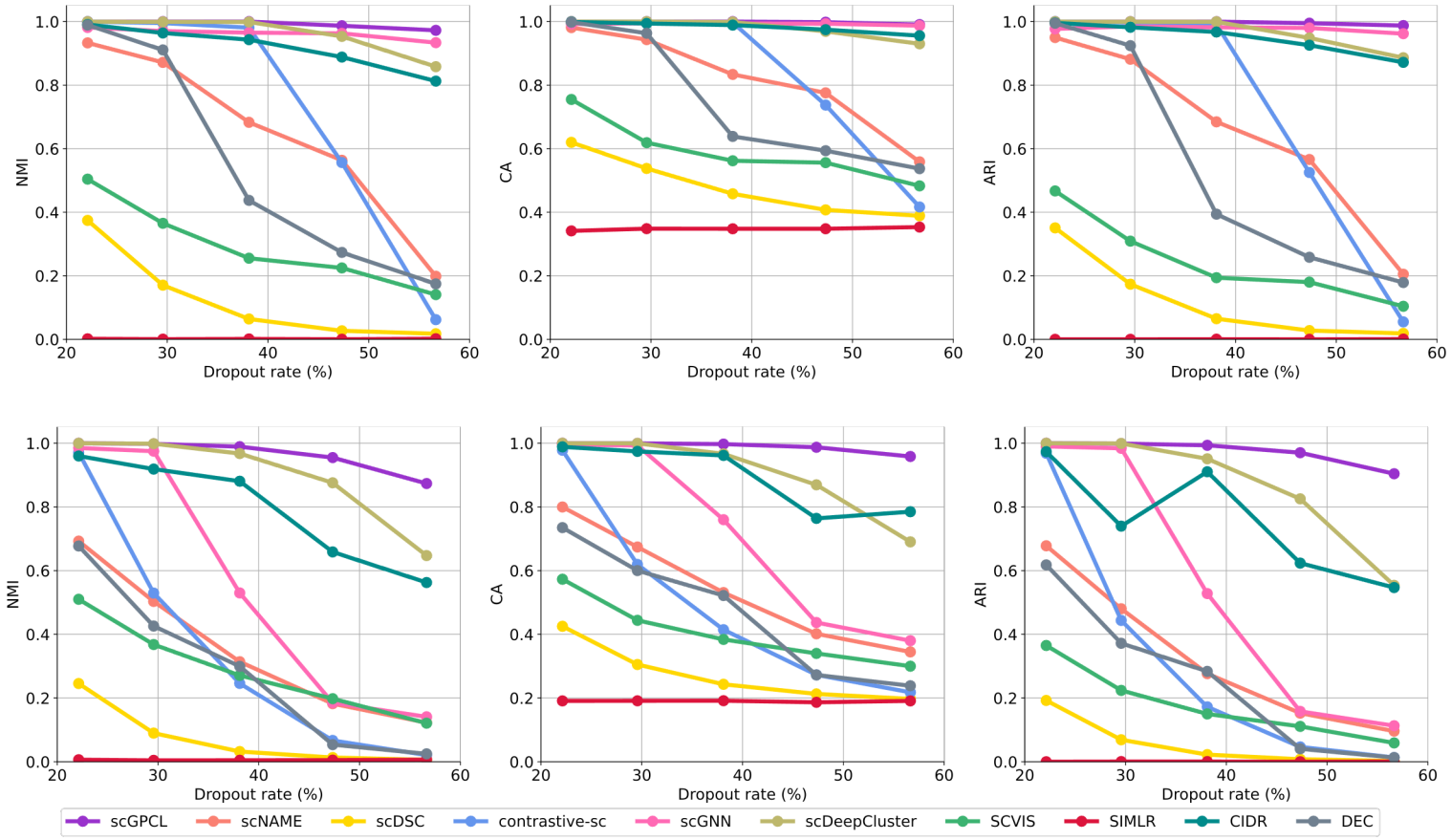
Data	Sequencing platform	# of Cells	# of Genes	# of Subgroups
Mouse ES cells	Droplet barcoding	2,717	24,047	4
Mouse bladder cells	Microwell-seq	2,746	19,771	16
Zeisel	STRT-seq UMI	3,005	19,972	9
Worm neuron cells	sci-RNA-seq	4,186	13,488	10
10X PBMC	10X	4,340	19,773	8
Human kidney cells	10X	5,685	25,215	11
Shekhar mouse retina cells	Drop-seq	27,499	13,166	19

EXPERIMENTS Simulated data

Case 1 Evaluation under Dropout phenomena

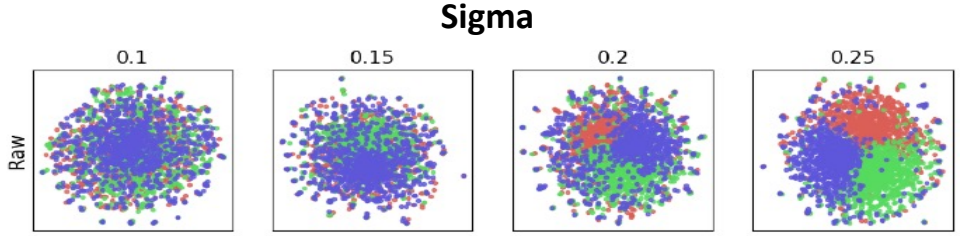
3 subgroups

6 subgroups



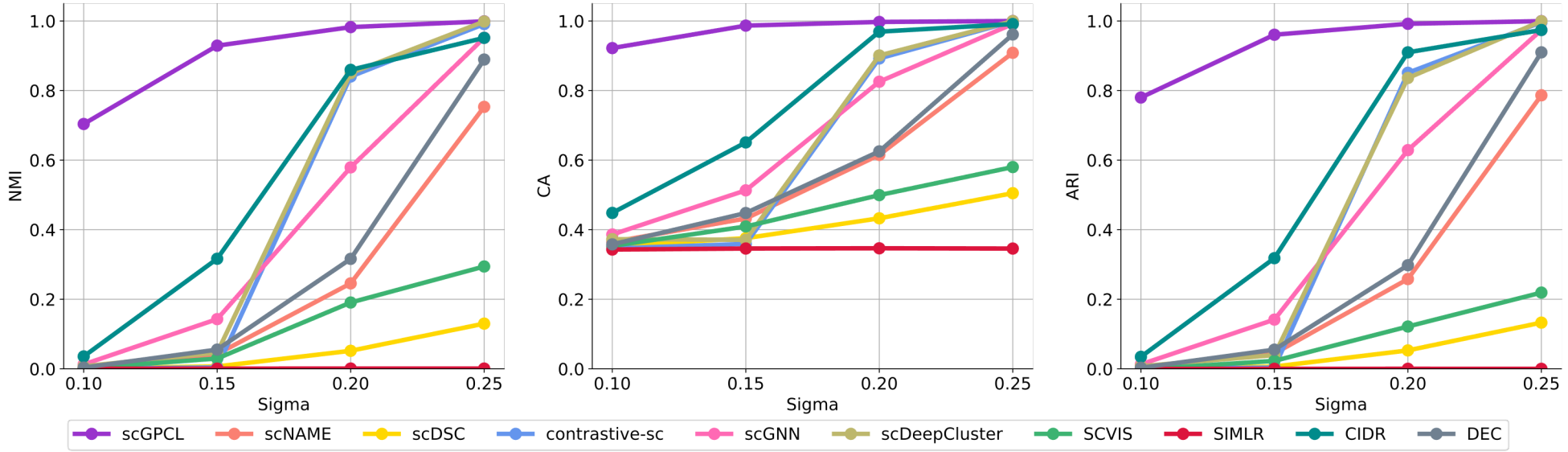
- scGPCL robustly separates the subgroup of the cells even if the gene expression matrix suffers from severe dropout phenomena

EXPERIMENTS Simulated data



Low sigma → Noisy features

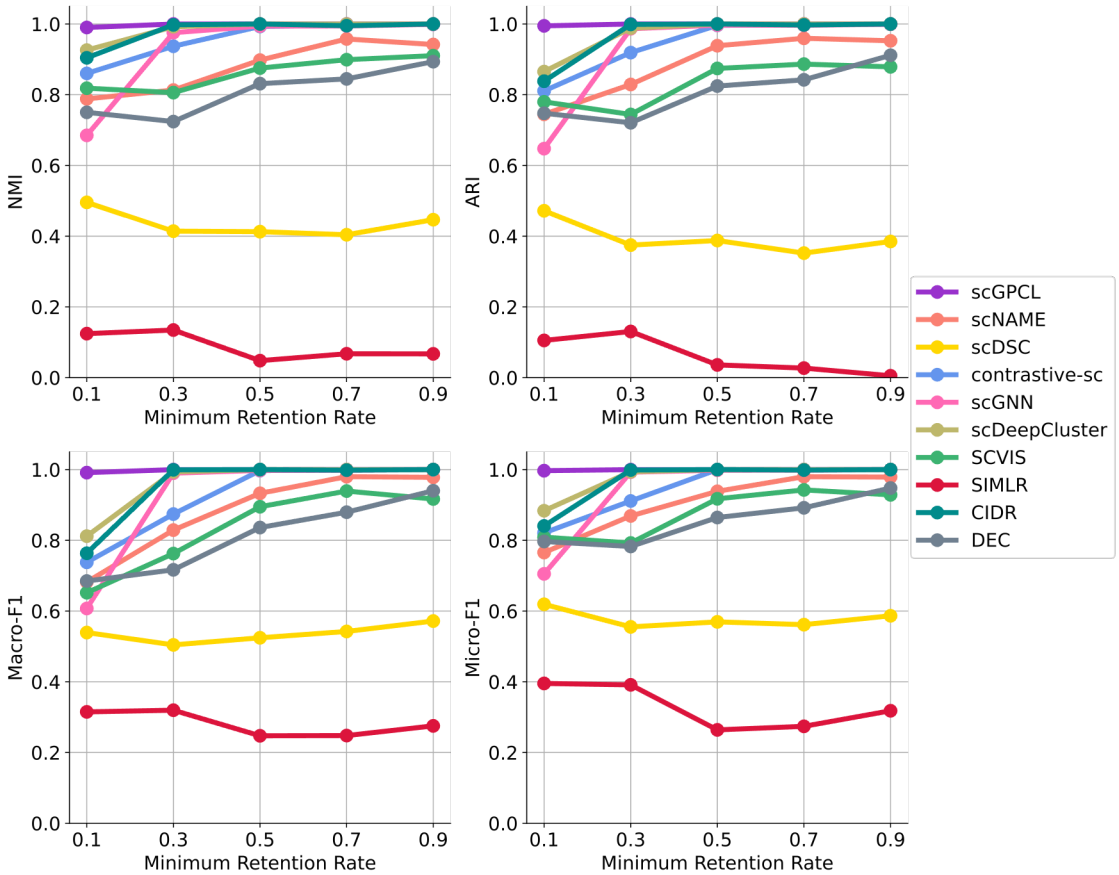
Case 2 Evaluation under Low signal

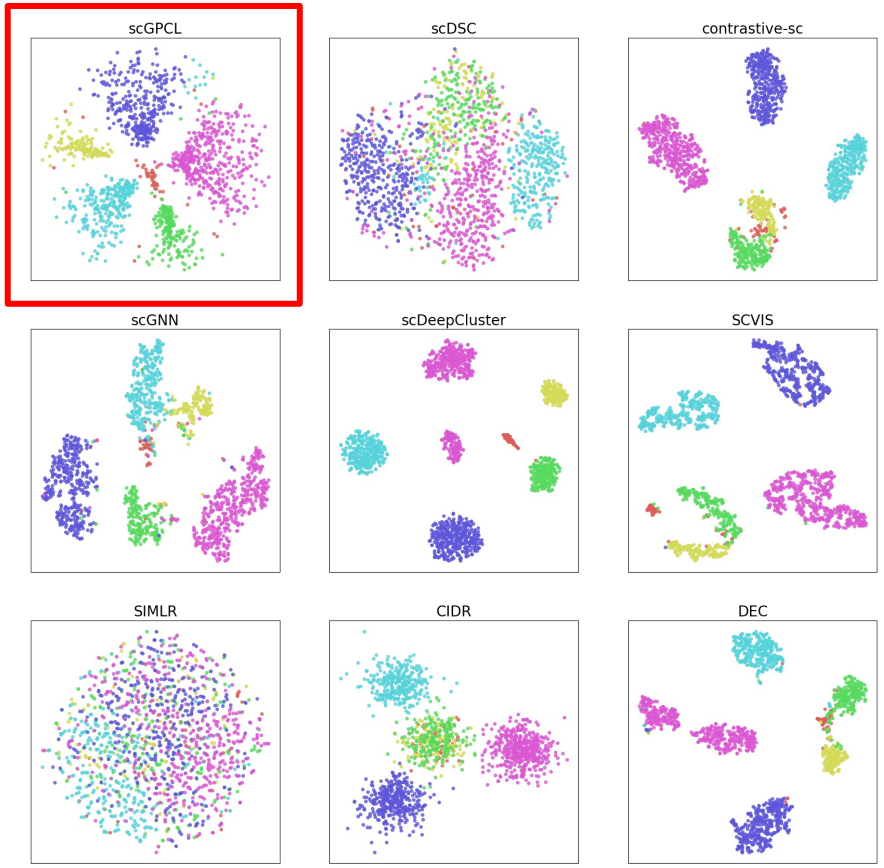


- Baseline methods fall short of robustly learning cell representations when the input matrix is not informative
- scGPCL robustly achieves accurate cluster assignments even if the information from the input feature is not sufficient

EXPERIMENTS Simulated data

Case 3 Evaluation under Imbalanced Subgroups of Cells

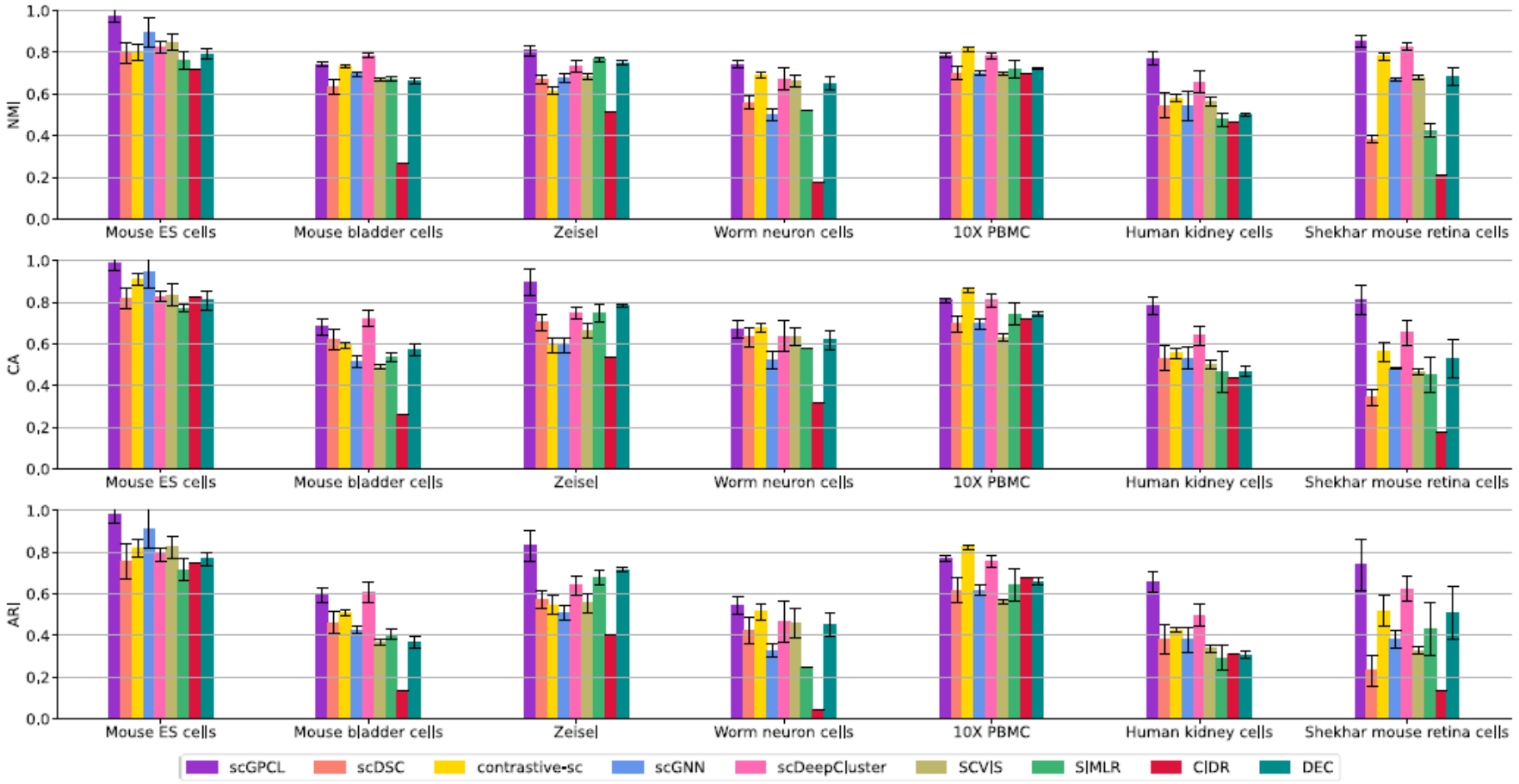




- scGPCL can separate the cells that belong to the minority subgroup

EXPERIMENTS Real scRNA-seq data

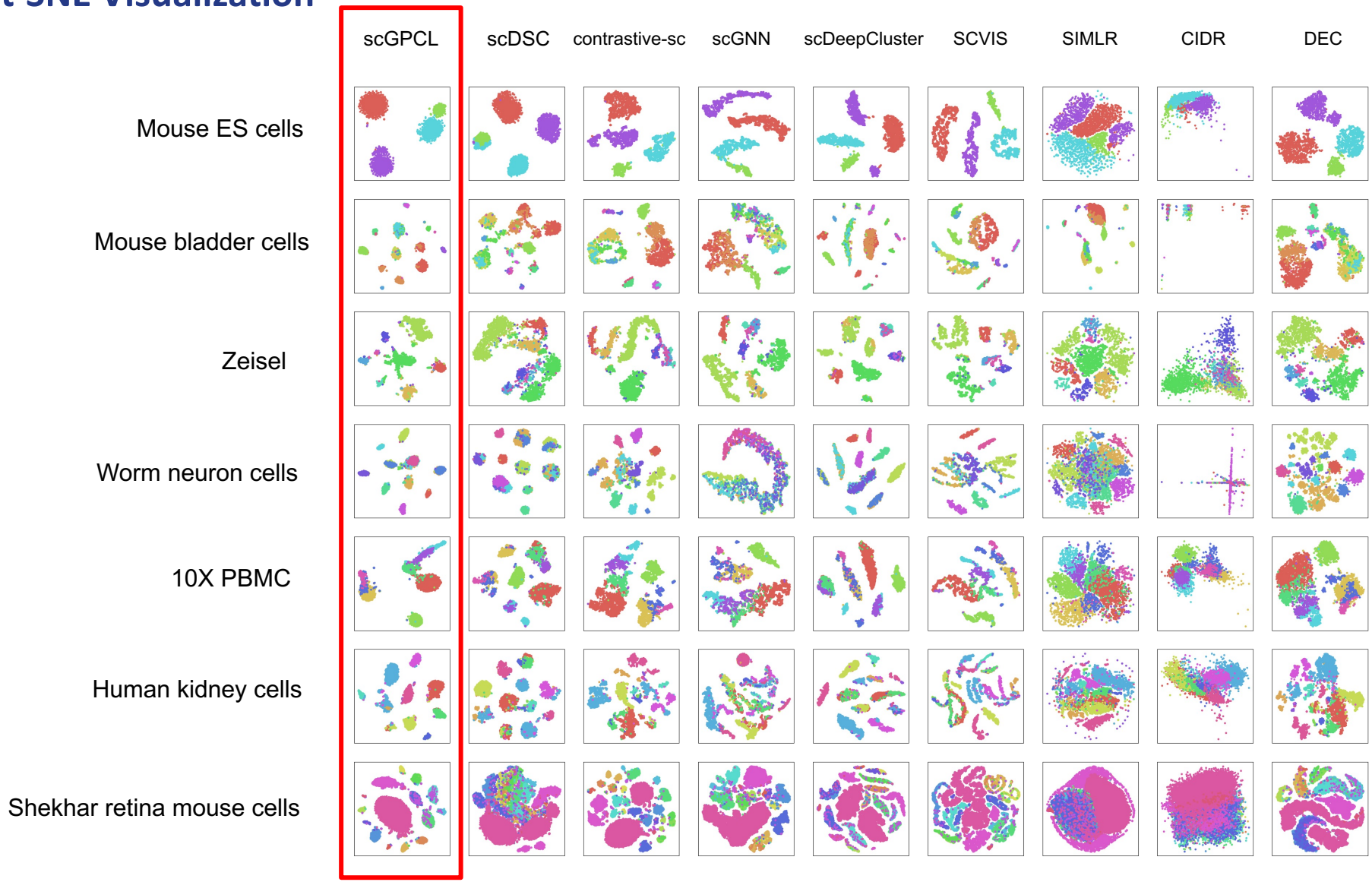
Comparison on real data



- scGPCL consistently outperforms the state-of-the art baselines on five datasets and achieves competitive scores on the two remaining datasets

EXPERIMENTS Real scRNA-seq data

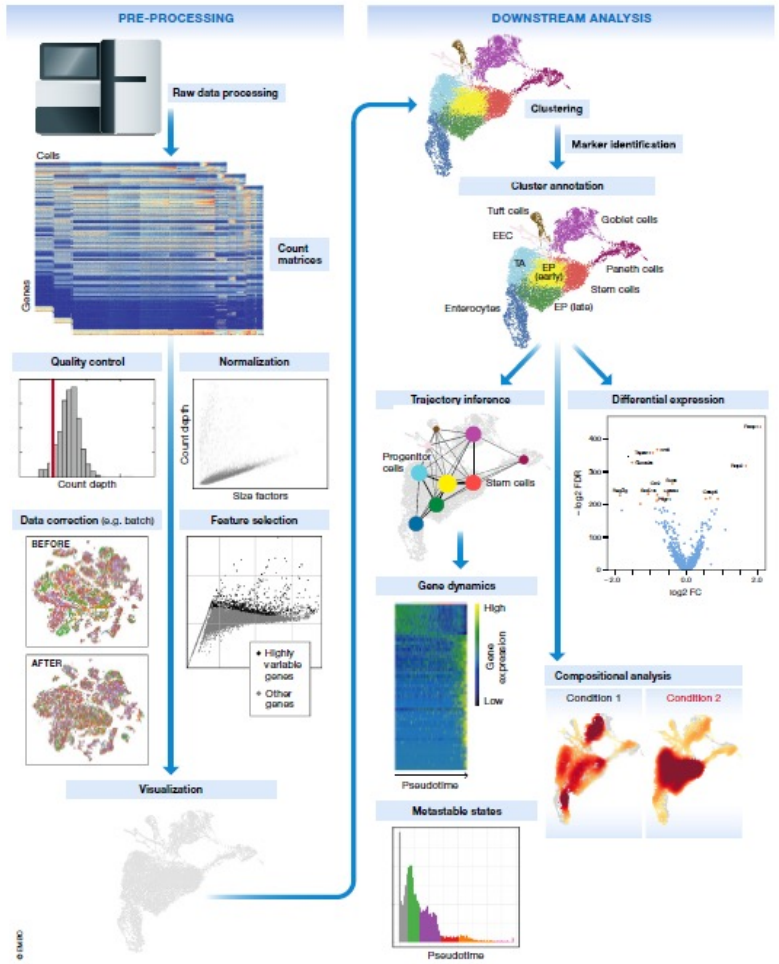
t-SNE Visualization



BACKGROUND Single Cell RNA-sequencing (scRNA-seq)

We want to handle the pre-processing phase to enhance the performance of various downstream tasks

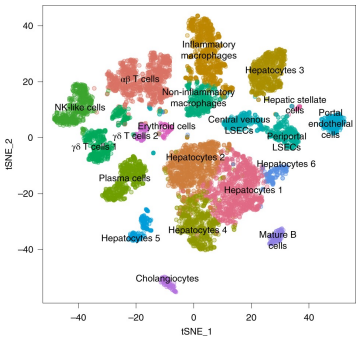
Pipeline of scRNA-seq Data analysis



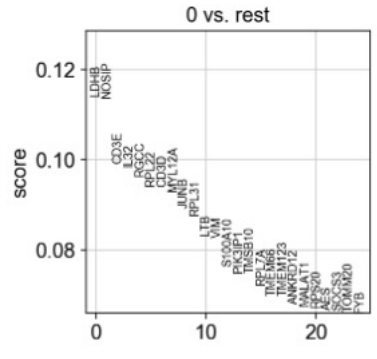
1) Generate Raw data using sequencing machine
2) Pre-processing
3) Downstream Analysis
   - Cell Annotation (Clustering)
   - Differentially expression gene analysis
   - Trajectory inference
   - Compositional analysis

- Single-cell RNA Sequencing Data Imputation Using Bi-level Feature Propagation, **Briefings in Bioinformatics 2024**
- Single-cell RNA-seq data imputation using Feature Propagation, **ICML 2023 CompBio Workshop (Best Paper Award)**

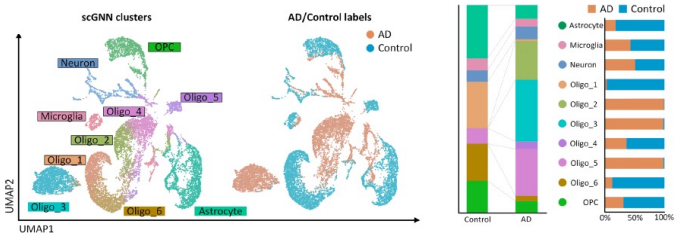
- Deep Single-cell RNA-seq Data Clustering with Graph Prototypical Contrastive Learning, **Bioinformatics 2023**



Cell Annotation (Clustering)



Differentially expression gene analysis



Compositional analysis

[Molecular Systems Biology 2019] Current best practices in single-cell RNA-seq analysis: a tutorial

Briefings in Bioinformatics, 2024, **25(3)**, bbae209

https://doi.org/10.1093/bib/bbae209

Problem Solving Protocol

Single-cell RNA sequencing data imputation using bi-level feature propagation

Junseok Lee[1,‡], Sukwon Yun[2,‡], Yeongmin Kim[3], Tianlong Chen[2,4,5], Manolis Kellis[4,5], Chanyoung Park[1,*]

[1]Department of Industrial and Systems Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
[2]Department of Computer Science, 201 S. Columbia St. CB 3175, UNC-Chapel Hill, Chapel Hill, NC 27599, United States
[3]School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
[4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, United States
[5]Broad Institute of MIT and Harvard, Merkin Building, 415 Main St., Cambridge, MA 02142, United States

*Corresponding author. Department of Industrial and Systems Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. E-mail: cy.park@kaist.ac.kr

‡Junseok Lee and Sukwon Yun contributed equally to this work.

Introduction

Target task **Imputation (Denoising)** **on scRNA-seq Data to enhance the performance of downstream tasks**

	Gene1	Gene2	Gene3	Gene 4
Cell 1	18	1010	0	22
Cell 2	0	506	49	2
Cell 3	0	49	0	0

Ground Truth

Sequencing Machine

	Gene1	Gene2	Gene3	Gene 4
Cell 1	18	1010	0	22
Cell 2	0	506	49	0
Cell 3	0	0	0	0

Observed (with dropout)

Imputation Model

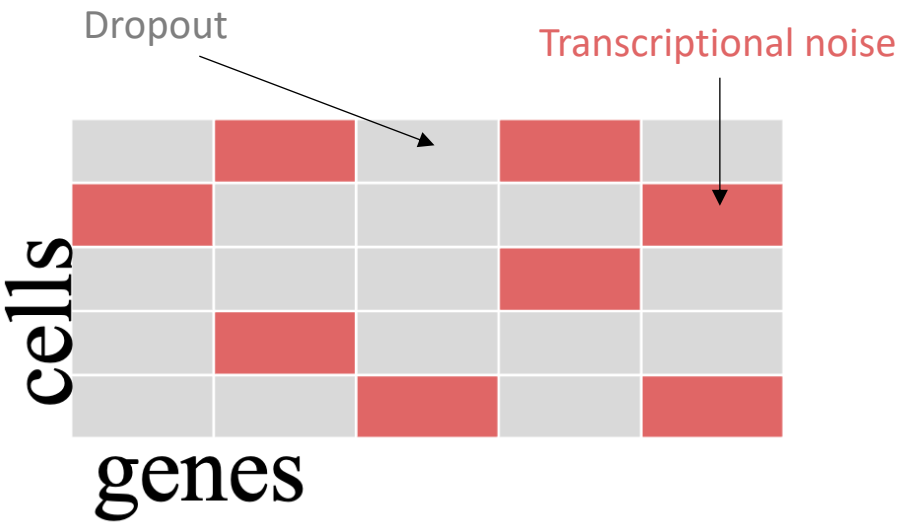
	Gene1	Gene2	Gene3	Gene 4
Cell 1	18	1010	0	22
Cell 2	0	506	49	2
Cell 3	0	49	0	0

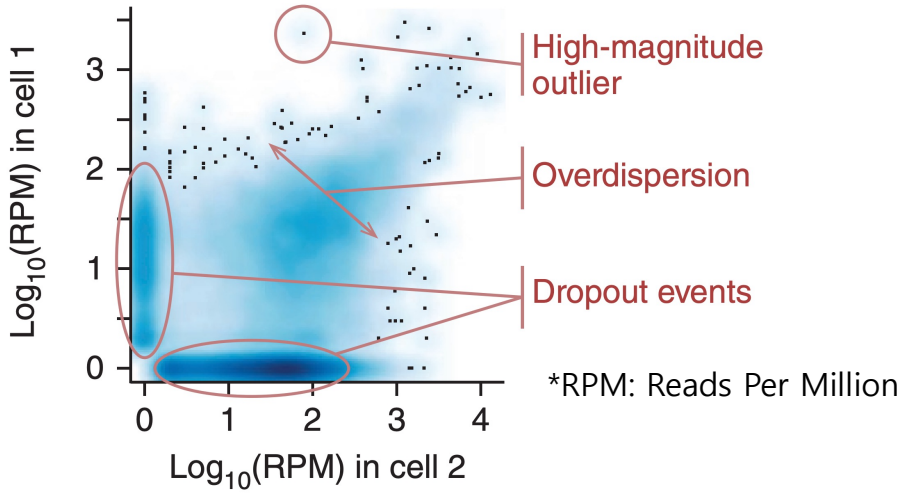
Imputed

Challenges

Challenge 1 Noise in both zero- and non-zero values

- **Zero-values (True zero? or False zero?):** Often regarded as a dropout (e.g., false-zeros), but there can be true zeros
- **Non-zero values (Noise):** Might capture biologically irrelevant signals (e.g., batch effects, transcriptional noise)





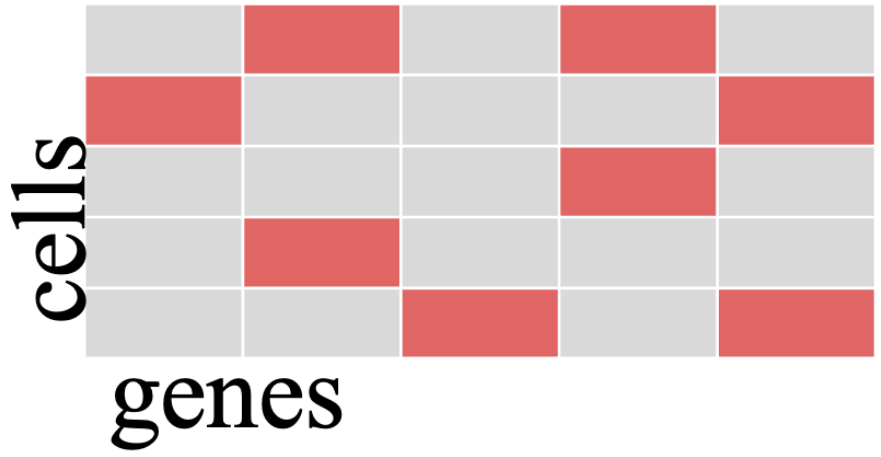
Careful handling of both zero-values and non-zero values is crucial

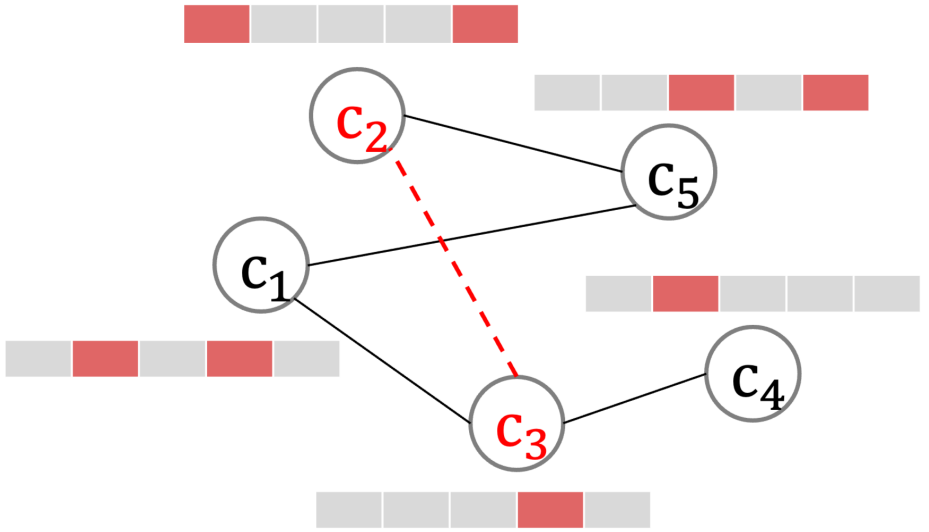
[Nature Methods 2014] Bayesian approach to single-cell differential expression analysis

Challenges

Challenge 2 **Biologically relevant graph structure is not provided**

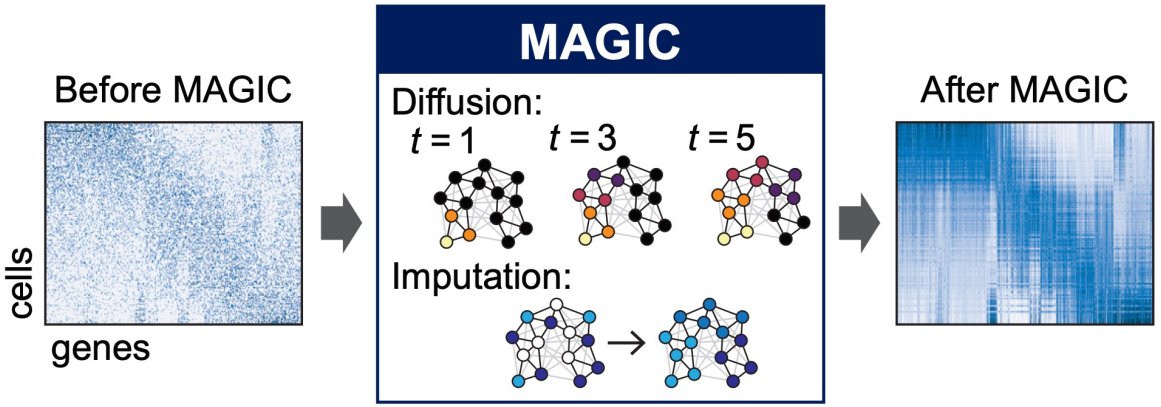
- kNN Graph based on initial sparse matrix may not be optimal





< kNN cell-cell Graph on sparse cell-gene matrix >

RELATED WORKS MAGIC



1) Calculate the Exponentiate Markov matrix M (KNN Graph)

2) Iterate diffusion process until convergence

$$D_{imputed} = M^t * D$$

Initial cell-gene matrix

Idea

Graph structure can capture relationships among cells
→ leveraging message-passing schemes for information propagation

Limitation

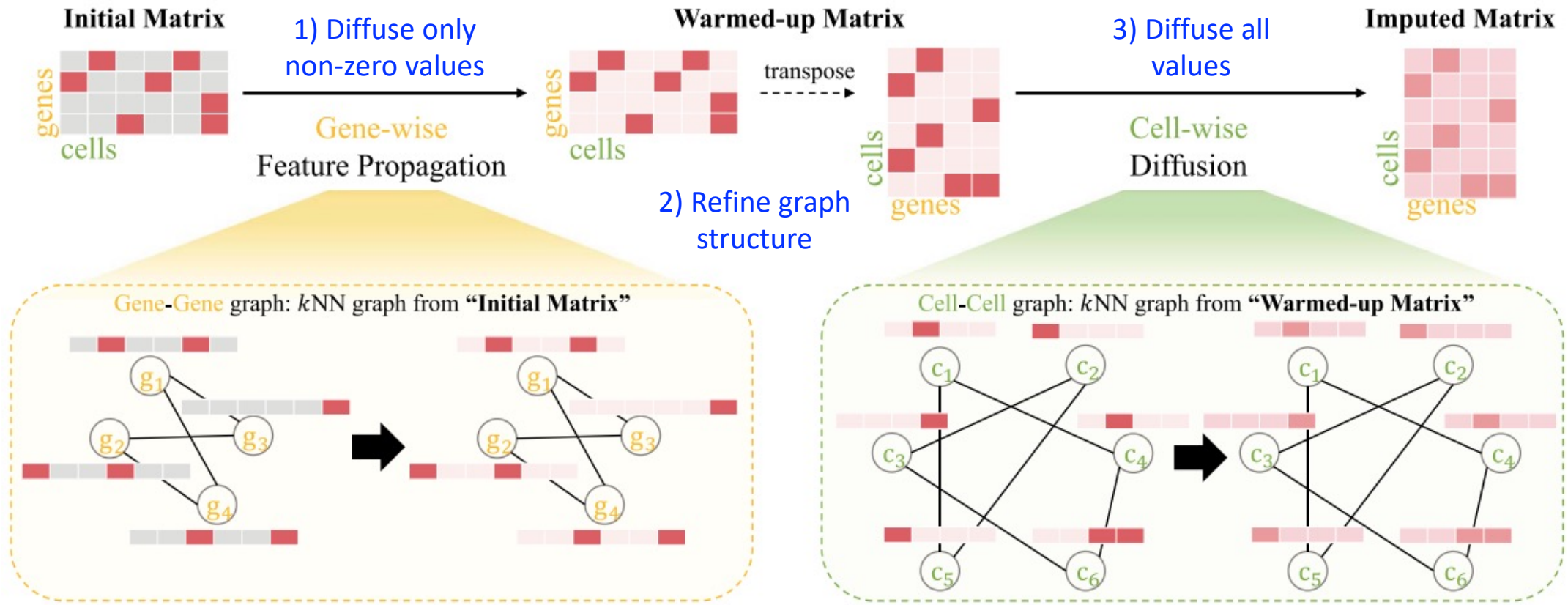
- Depends on the quality of constructed cell-cell graph (scRNA-seq data is inherently sparse and noisy)
- Diffusing zero values involve the risk of propagating noise, when diffused zeros are false zero (dropout)
- Do not leverage relationship information between genes

[Cell 2018] Recovering Gene Interactions from Single-Cell Data Using Data Diffusion

Single-cell Bi-level Feature Propagation (scBFP)

Key Idea

- Step 1: Diffuse only non-zero values to mitigate the risk of propagating false-zeros to non-zero values
- Step 2: Refine the kNN graph after warm-up phase to construct a cell-cell graph
- Step 3: Leverage cell-cell relationships by conducting diffusion on the refined cell-cell graphs



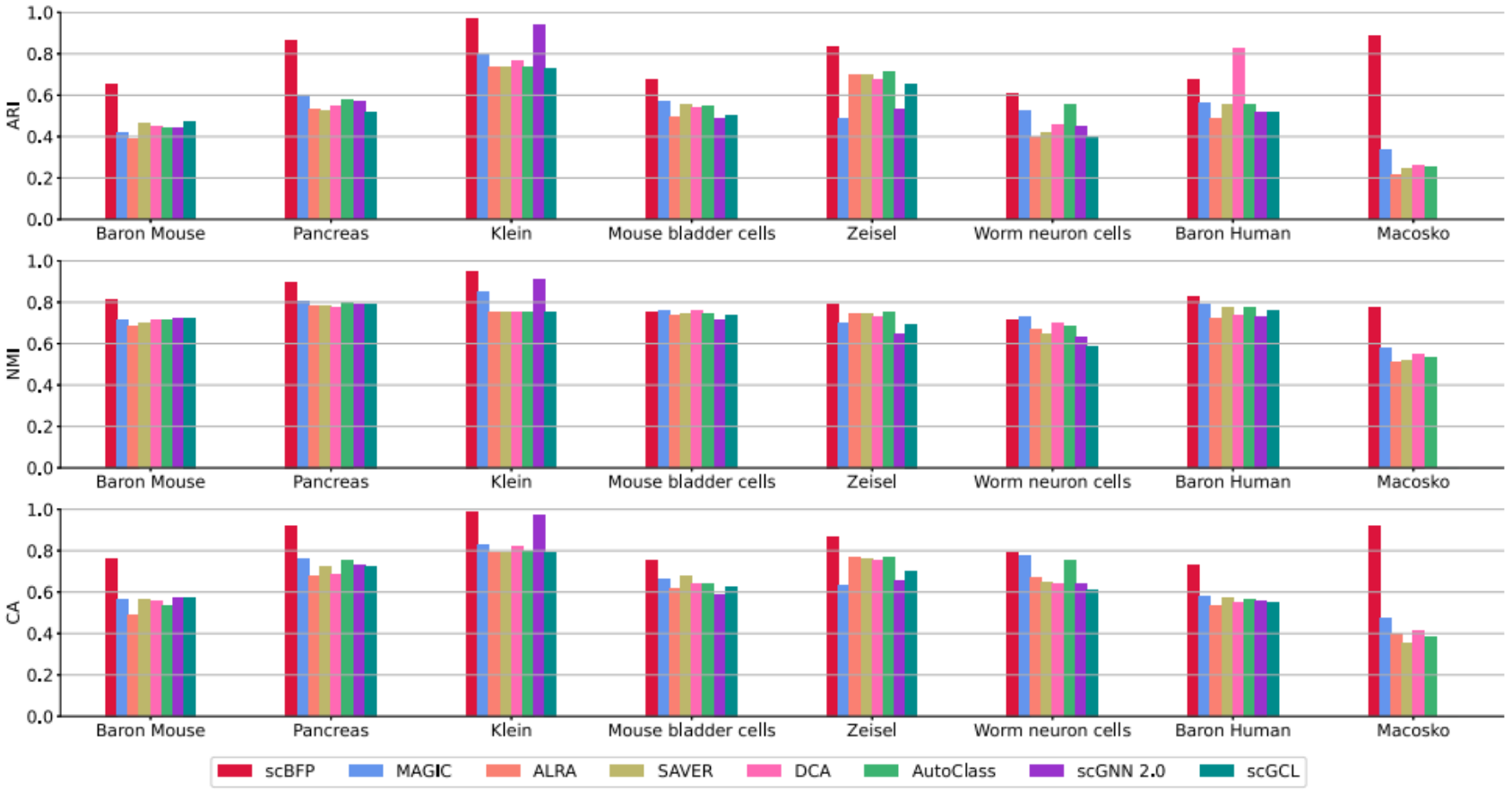
EXPERIMENTS

Extensive experiments on various scRNA-seq downstream tasks

- **Cell Clustering** : scBFP helps to conduct improved cell clustering
- **DEG Detection** : scBFP helps to detect differentially expressed genes
- **Dropout value recovery** : scBFP effectively recover dropout values
- **Enrichment analysis** : scBFP enriches relevant genes in lung cancer data

EXPERIMENTS

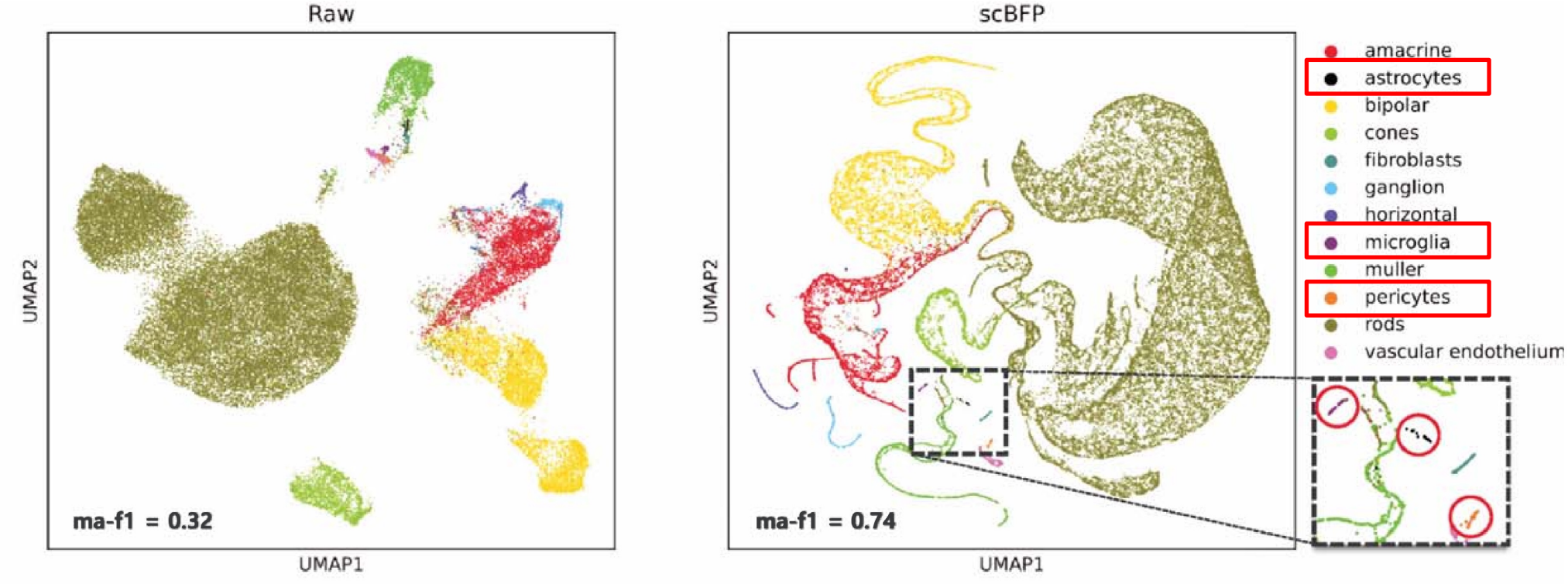
Cell Clustering



- scBFP consistently outperforms the state-of-the art baselines on eight datasets

EXPERIMENTS

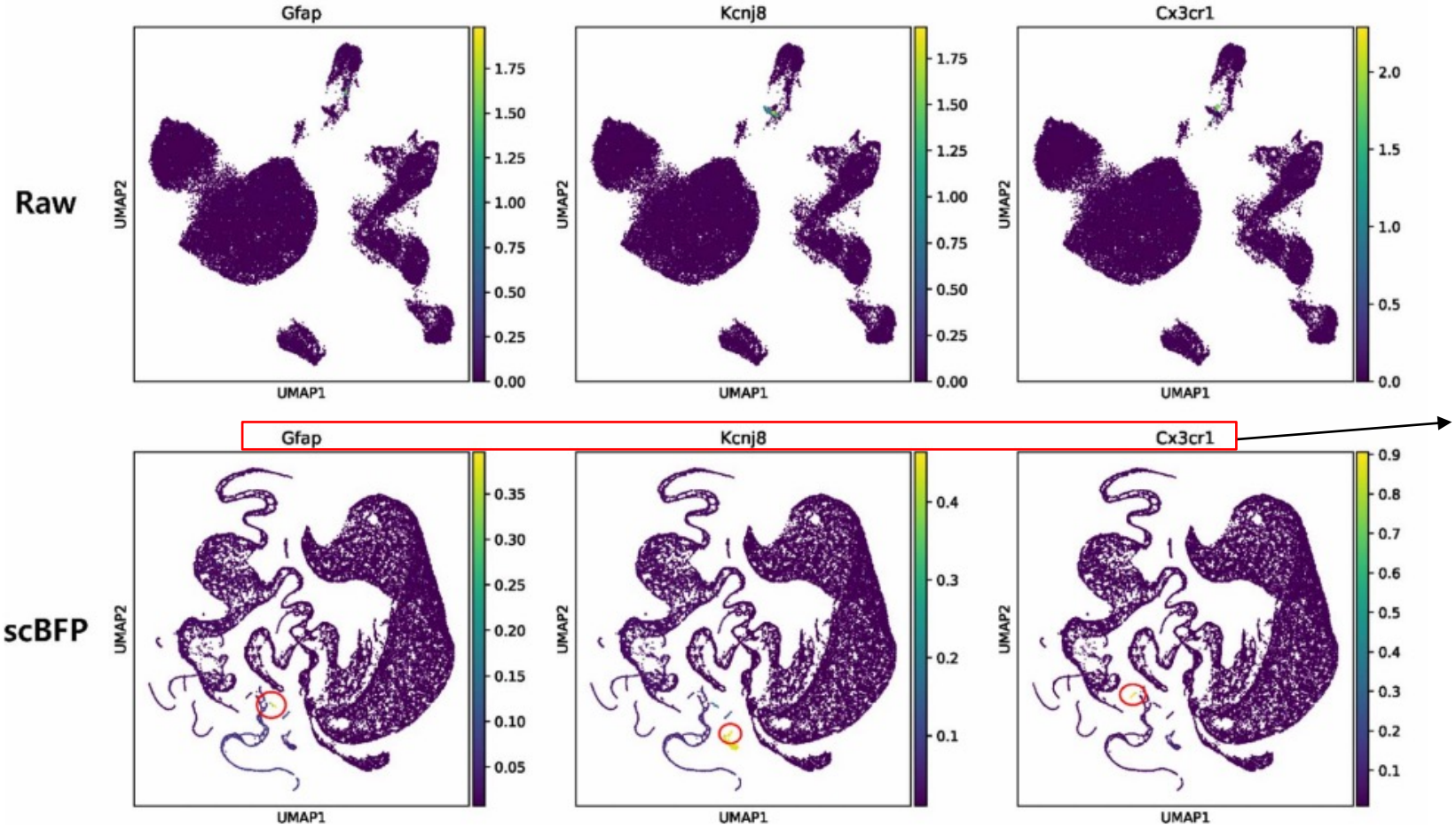
Cell Clustering



- scBFP also captures the rare cell types in the imbalance dataset

EXPERIMENTS

DEG Detection: Visualize the expression levels of well known marker genes

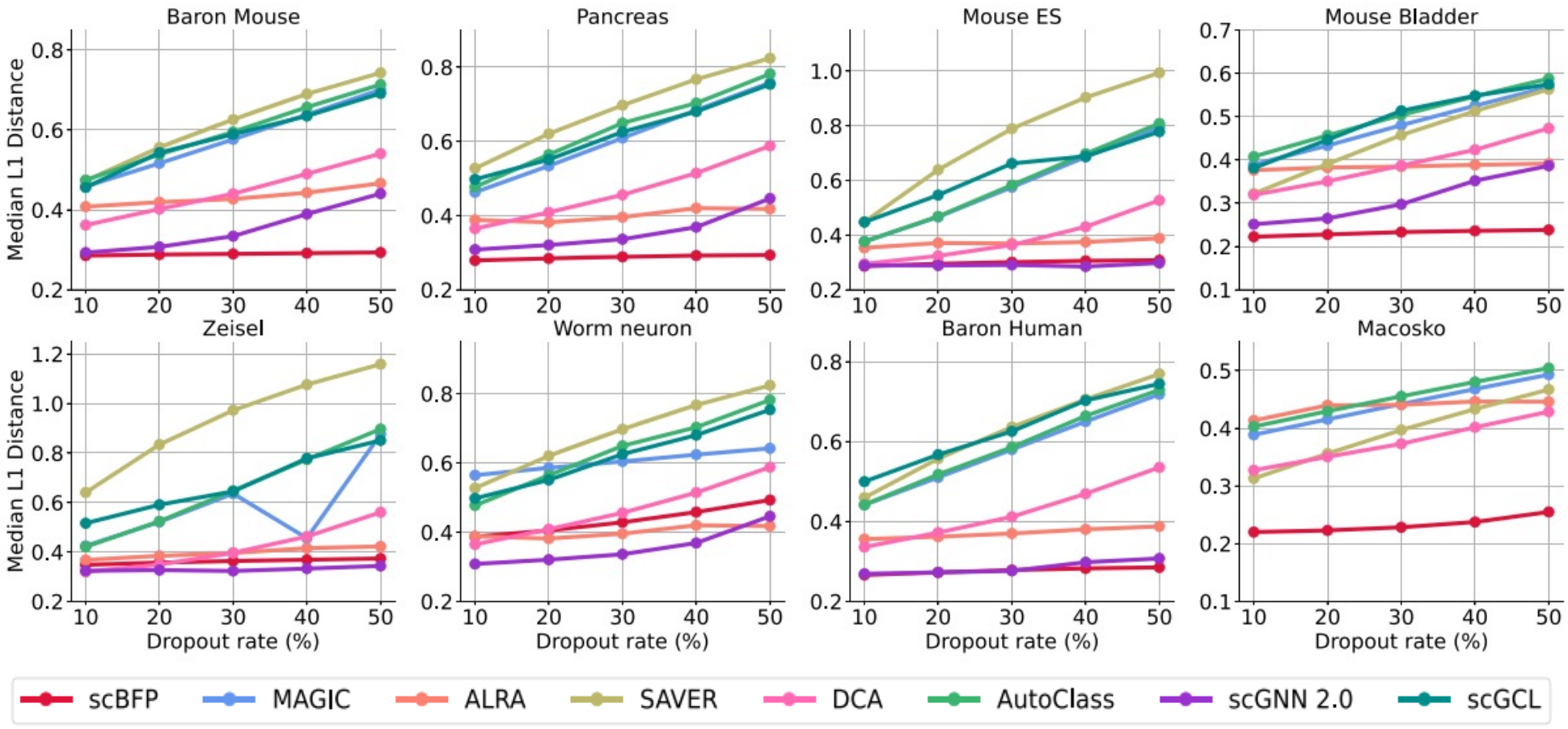


Marker genes for
astrocytes, pericytes, microglia

- scBFP also identifies Differentially Expressed Genes (i.e., marker genes) associated with rare cell types

EXPERIMENTS

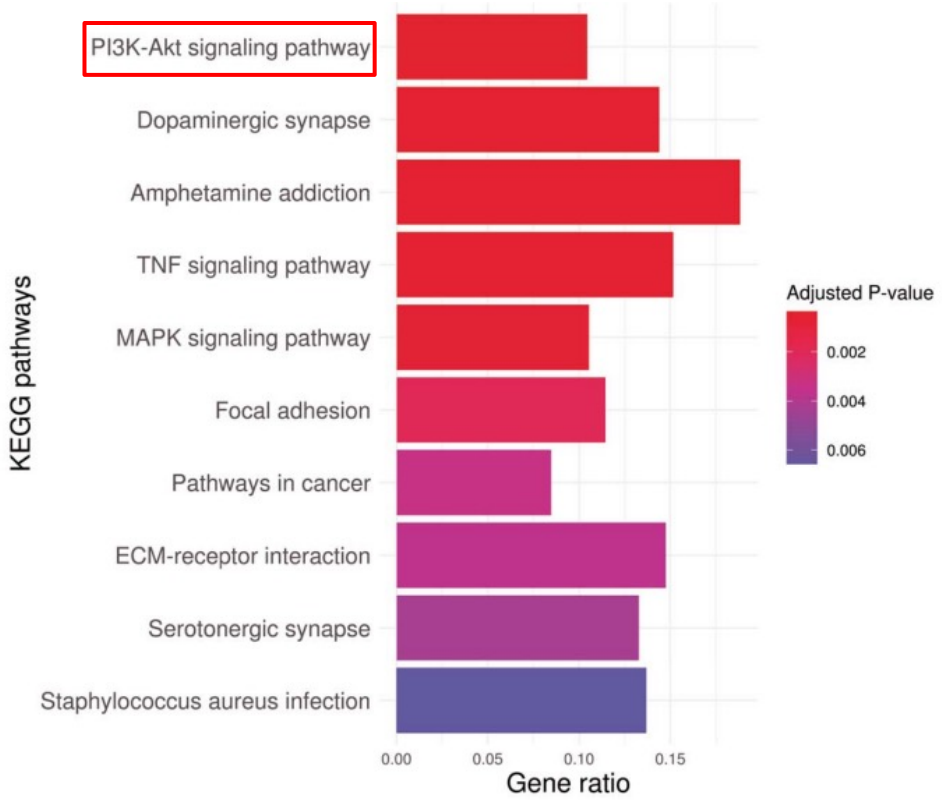
Dropout value recovery



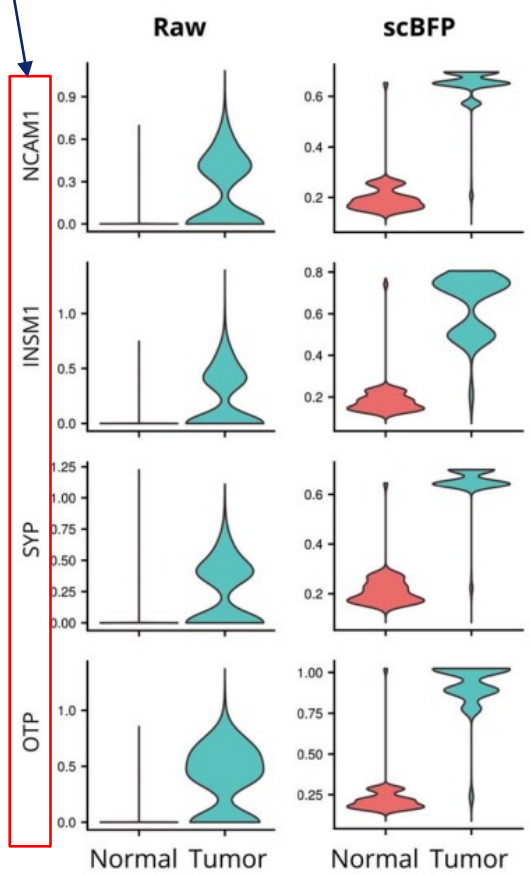
- scBFP also successes to robustly recover the original values even on the high dropout rate

EXPERIMENTS

Enrichment Analysis (Lung Tumor Cancer)



4 marker genes for lung carcinoid



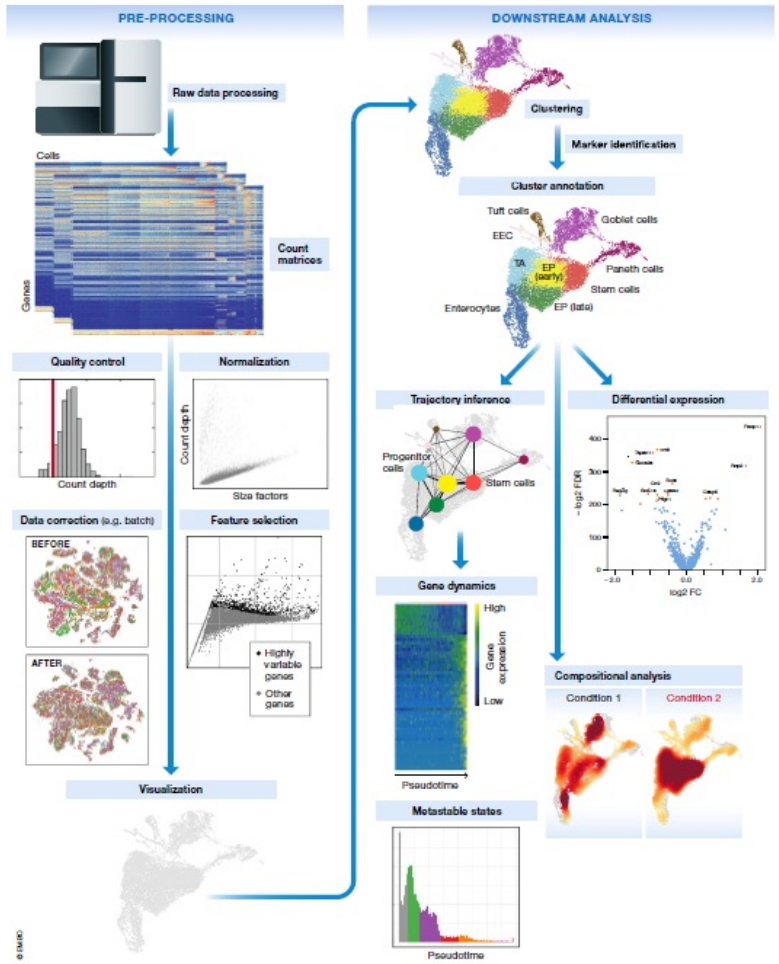
Normal tissue의 발현값은 대부분 0이었으나, 이를 제대로 impute해서 normal과 Tumor 간의 구분이 명확해짐

- DEGs derived from scBFP show a clear enrichment in cancer-relevant KEGG pathways

  → For example, **'PI3K/Akt signaling pathway'** plays a role in promoting tumor cell growth

BACKGROUND Single Cell RNA-sequencing (scRNA-seq)

Pipeline of scRNA-seq Data analysis



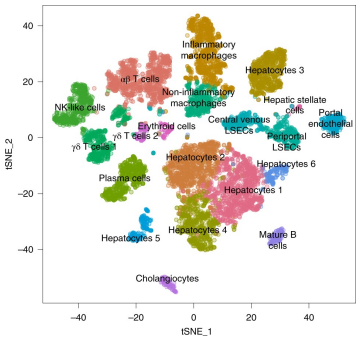
1) Generate Raw data using sequencing machine

2) Pre-processing

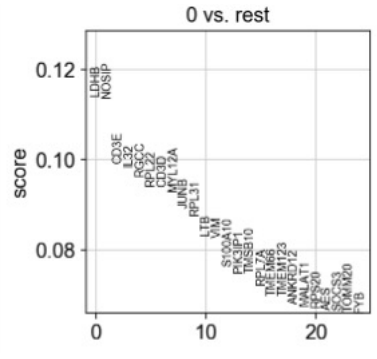
- Single-cell RNA Sequencing Data Imputation Using Bi-level Feature Propagation, **Briefings in Bioinformatics 2024**
- Single-cell RNA-seq data imputation using Feature Propagation, **ICML 2023 CompBio Workshop (Best Paper Award)**

3) Downstream Analysis

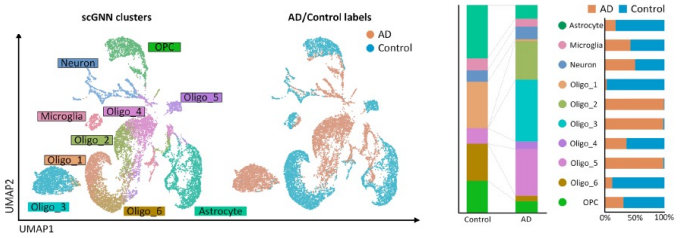
- Cell Annotation (Clustering)
  - Deep Single-cell RNA-seq Data Clustering with Graph Prototypical Contrastive Learning, **Bioinformatics 2023**
- Differentially expression gene analysis
- Trajectory inference
- Compositional analysis



Cell Annotation (Clustering)



Differentially expression gene analysis



Compositional analysis

[Molecular Systems Biology 2019] Current best practices in single-cell RNA-seq analysis: a tutorial

ICML
International Conference
On Machine Learning
2025

Global Context-aware Representation Learning for Spatially Resolved Transcriptomics

Yunhak Oh [*1] **Junseok Lee** [*2] **Yeongmin Kim** [3] **Sangwoo Seo** [2] **Namkyeong Lee** [2] **Chanyoung Park** [1 2]

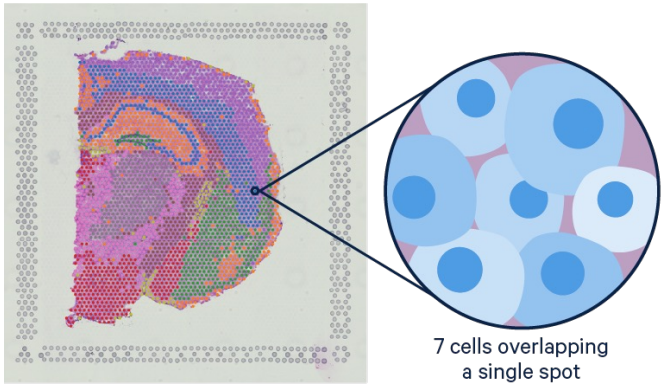
BACKGROUND Spatial Single-cell RNA-seq data

- Limitation of scRNA-seq data: Loss of the **spatial context** of cells due to tissue dissociation

- Spatial Information in tissue is key to understanding biological functions and interactions
  - e.g., Cells of the same type tend to gather closely

- Spatially Resolved Transcriptomics (SRT)

- Incorporate the **spatial context of the cells**
- Enable to comprehend complex transcriptional structure and mechanism of disease



Spatial Coordinates

	Gene1	Gene2	Gene3	Gene 4
Spot 1	18	1010	0	22
Spot 2	0	506	49	0
Spot 3	0	0	0	0

Gene Expression

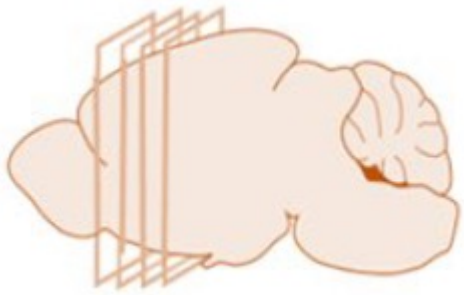
BACKGROUND Spatial Single-cell RNA-seq data Integration

Analyzing individual slices has limited ability to detect spatially varying, lowly expressed transcripts

Spatial transcriptomics (ST) studies **generate data from multiple tissue slices**

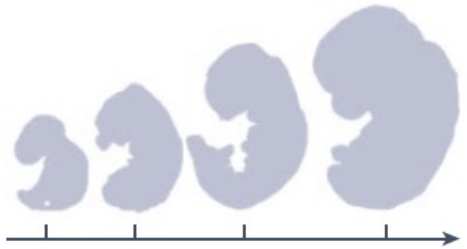
1) Homogeneous Slices

- Consecutive slices



2) Heterogeneous Slices

- Development stages



- Normal versus disease

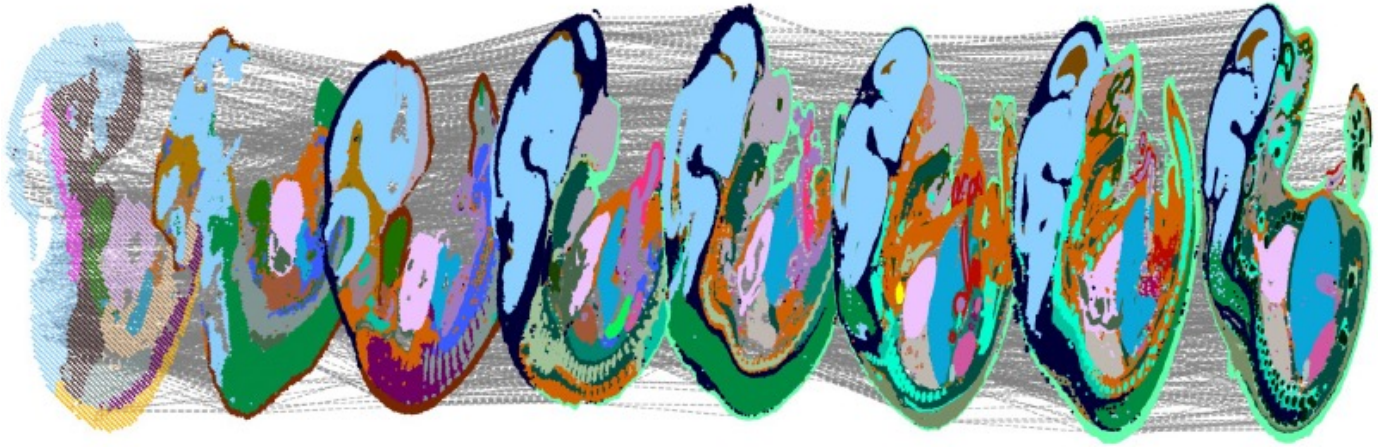
DOWNSTREAM TASK Spatial Single-cell RNA-seq data Integration

Target data **Spatial Single-cell RNA sequencing (scRNA-seq) Data**

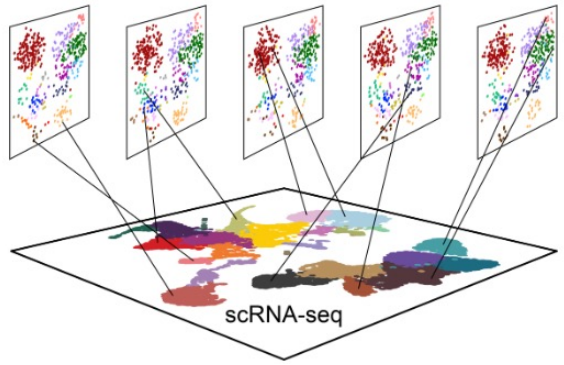
Target task

- Identify spatial domain (**Clustering**)
- **Align** spatial data across multiple slices
- **Integrate** the representations of multiple slices

- Alignment across development stages



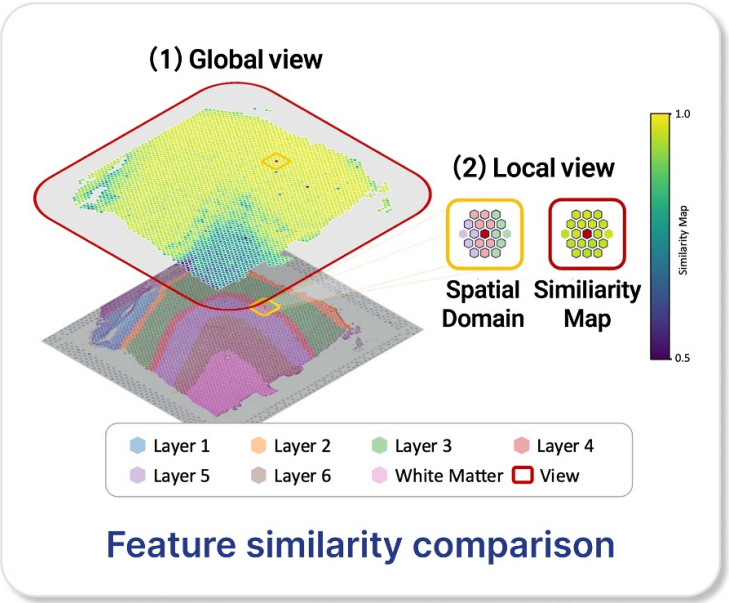
- Integration of multiple slices

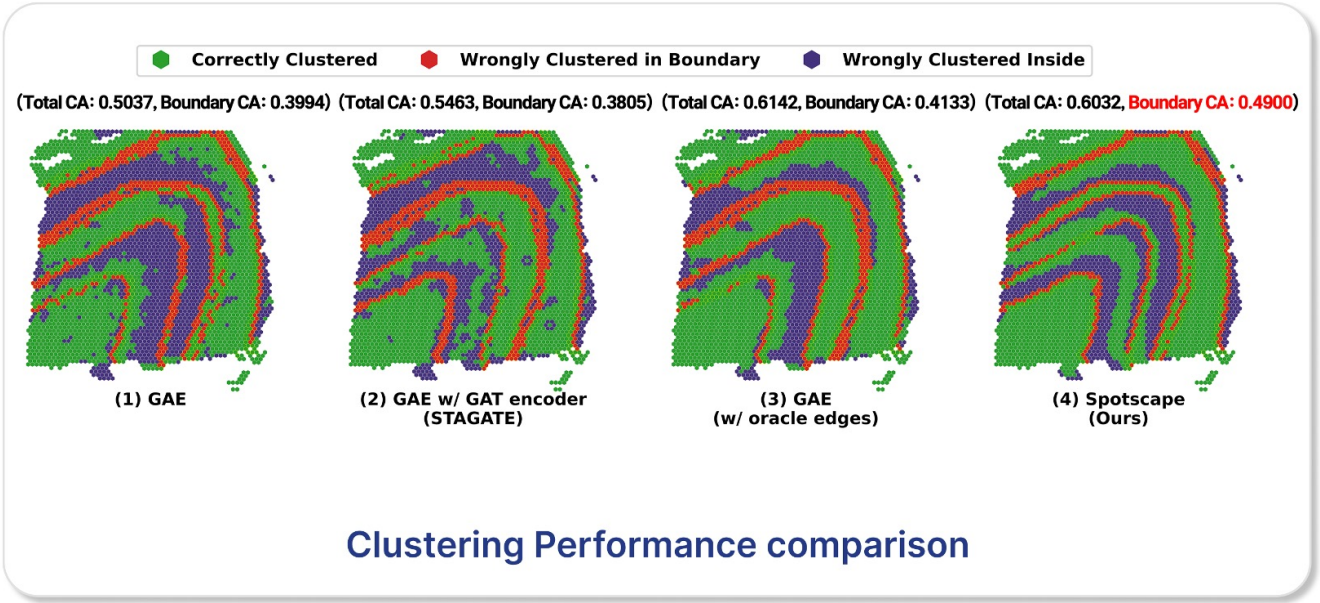


CHALLENGES

Continuous Nature of biological systems: Gene expression values vary smoothly along spatial coordinates

Local (spatially close) spots have high similarity, regardless of the spatial domain





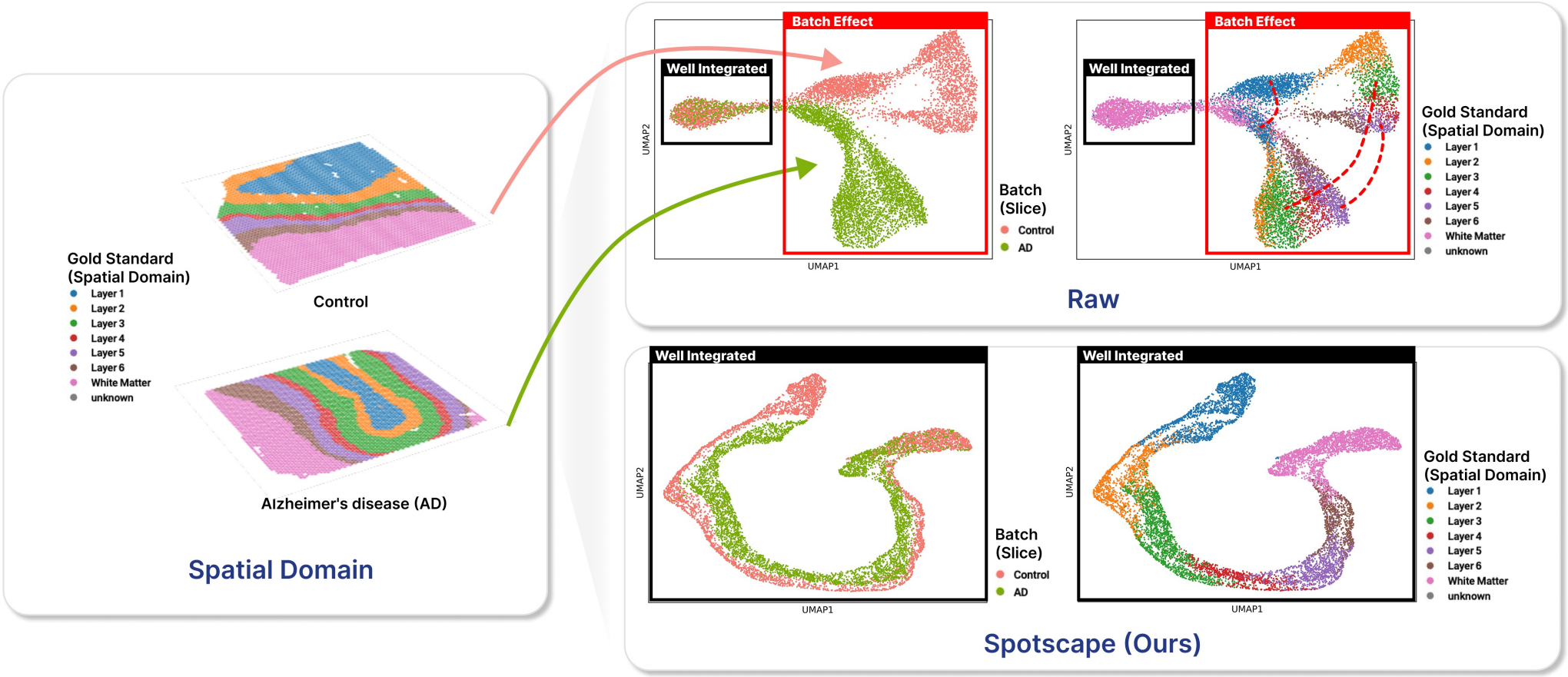
- Results in a negative effects on GNNs by acquiring information from **heterophilic nodes**
- **Difficult to learn the appropriate attention scores** that gives high scores to homophilic nodes due to their low feature difference
- Even if it can learn appropriate edges (attentions), **local view has insufficient information**

Learn the similarities and differences between cells from a global perspective

CHALLENGES

Batch Effect in SRT data

- Gene expression profiles from the same slice cluster together unexpectedly, regardless of their biological relevance

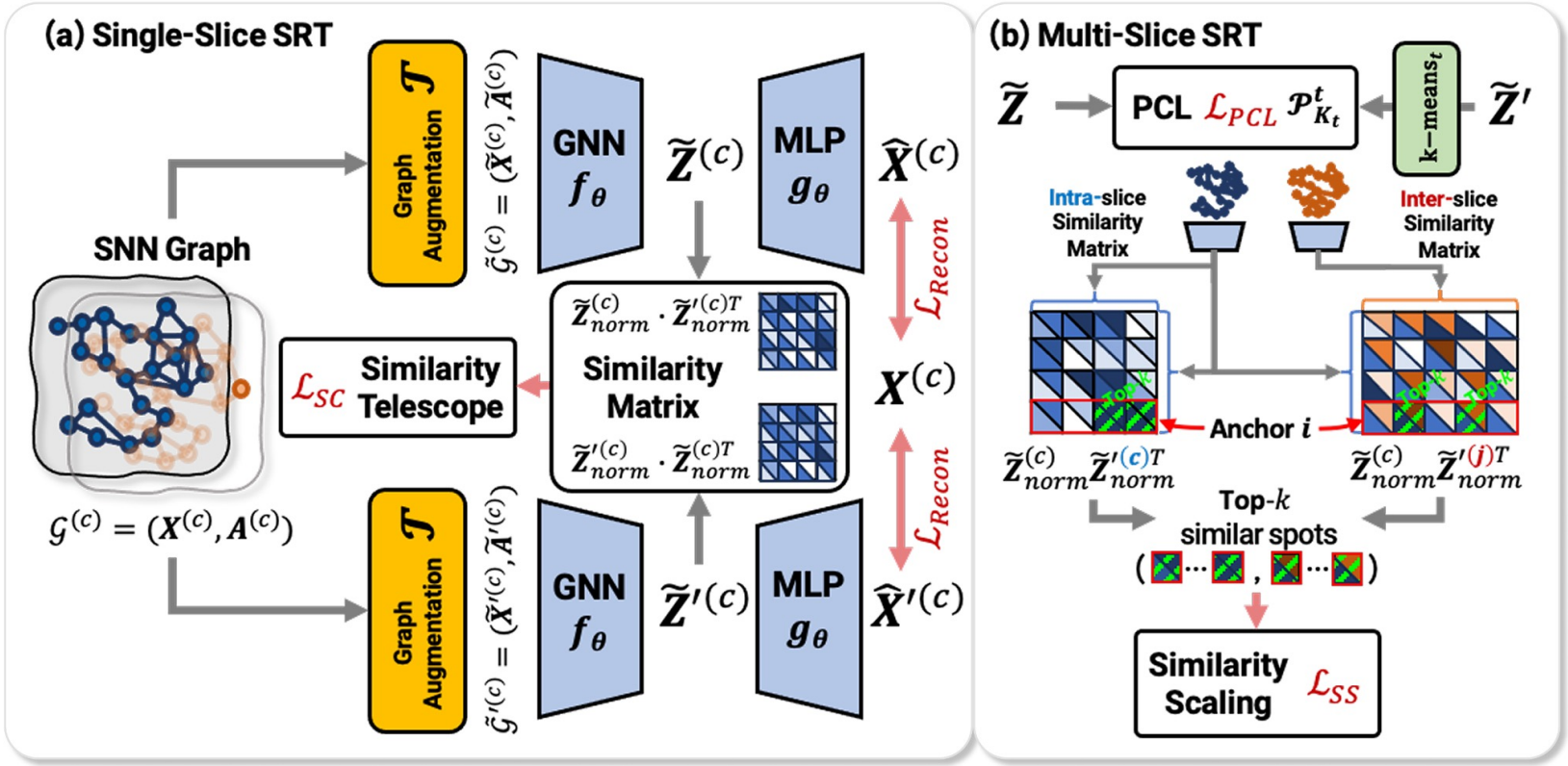


Should alleviate batch effects to extend multi-slice tasks

Spotscape

Key Ideas

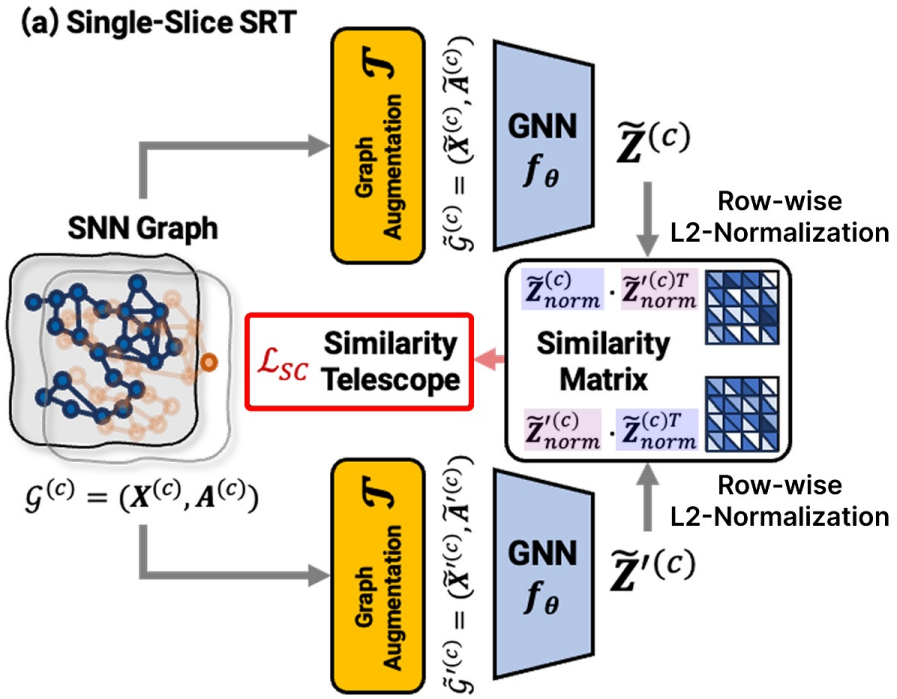
- Capturing **global relationships** between cells by learning **robust similarities** with respect to different augmentation
- Explicitly **balance the similarity scales of inter- and intra-relationships** to mitigate batch effects
- **Grouping spots from the same spatial domain while distancing others** in latent space using PCL to mitigate batch effect

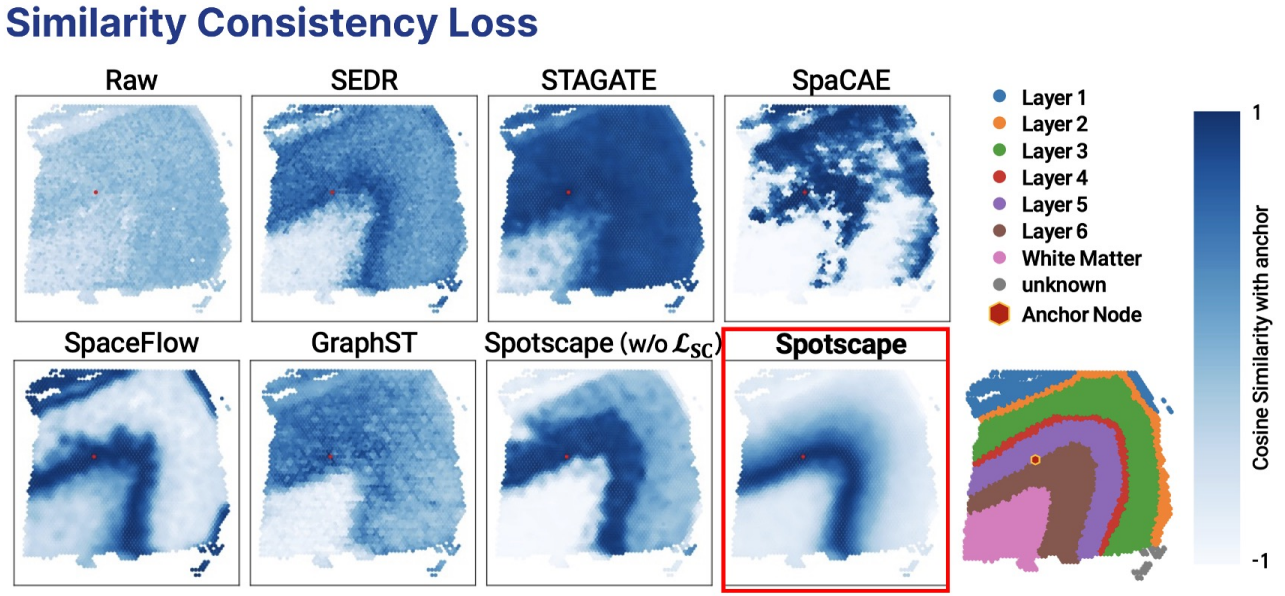


Spotscape

Effectiveness of Similarity Telescope (Similarity Consistency loss)

- Capturing **global relationships** between cells by learning **robust similarities** with respect to different augmentation





- **Spotscape** captures relative similarities between spots, aligning with the spatial dynamics of SRT data, unlike baselines

- **Spotscape** exhibits varying similarity level based on the true spatial domains, accurately reflecting spatial distance relationship

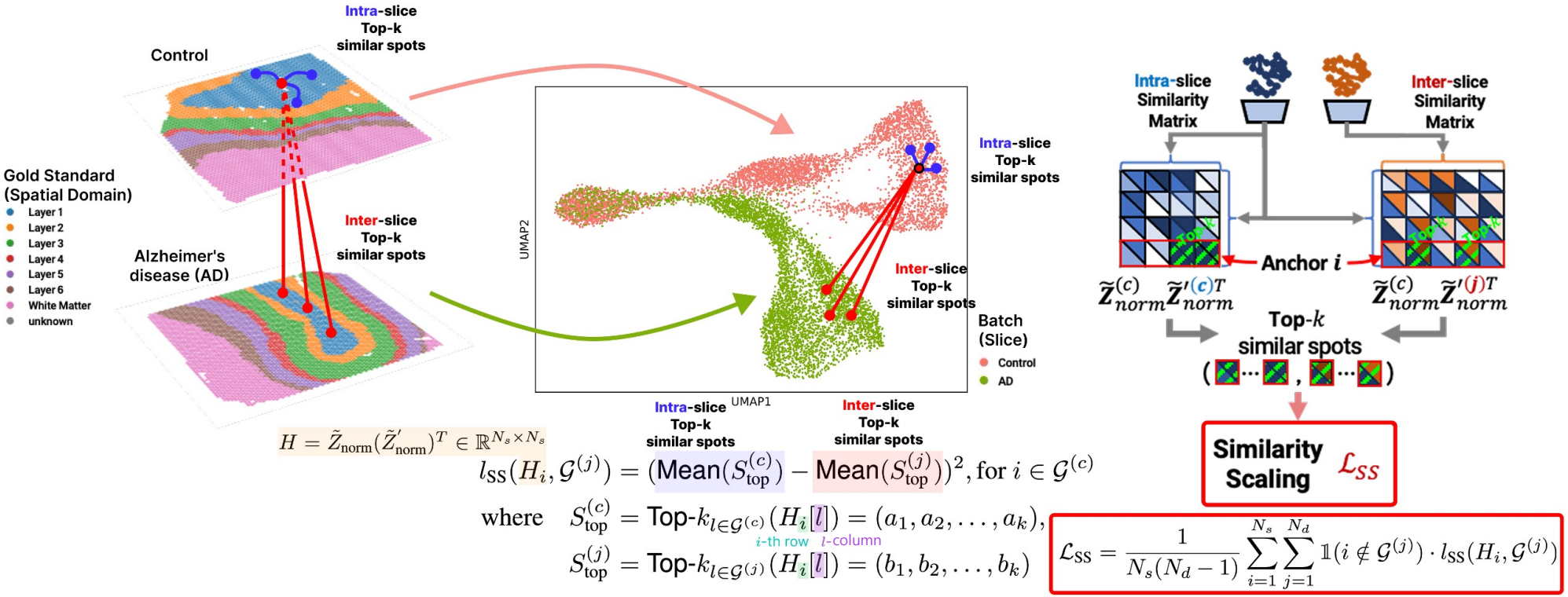
Spotscape

Concept of Similarity Scaling Loss

- Explicitly **balance the similarity scales of inter- and intra-relationships** to mitigate batch effects

Batch Effects: Experimental condition or noise ≫ Biological relevance

(b) Multi-Slice SRT



$$H = \tilde{Z}_{\text{norm}}(\tilde{Z}'_{\text{norm}})^T \in \mathbb{R}^{N_s \times N_s}$$

$$l_{\text{SS}}(H_i, \mathcal{G}^{(j)}) = (\text{Mean}(S_{\text{top}}^{(c)}) - \text{Mean}(S_{\text{top}}^{(j)}))^2, \text{for } i \in \mathcal{G}^{(c)}$$

$$\text{where} \quad S_{\text{top}}^{(c)} = \text{Top-}k_{l \in \mathcal{G}^{(c)}}(H_i[l]) = (a_1, a_2, \ldots, a_k),$$

(i-th row, l-column)

$$S_{\text{top}}^{(j)} = \text{Top-}k_{l \in \mathcal{G}^{(j)}}(H_i[l]) = (b_1, b_2, \ldots, b_k)$$

$$\mathcal{L}_{\text{SS}} = \frac{1}{N_s(N_d - 1)} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} \mathbb{1}(i \notin \mathcal{G}^{(j)}) \cdot l_{\text{SS}}(H_i, \mathcal{G}^{(j)})$$

EXPERIMENTS

Various downstream tasks on both single- and multi-slice tasks

- **Single-slice Tasks**
  - Spatial domain identification
  - Trajectory inference
  - Denoising & Imputation
- **Multi-slice Tasks**
  - Integration
  - Alignment
  - Differentially expressed gene (DEG) analysis

EXPERIMENTS Single-slice Tasks

Spatial Domain Identification

(a) DLPFC (Patient 1)

	Slice 151673			Slice 151674			Slice 151675			Slice 151676		
	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA
SEDR	0.36 (0.08)	0.49 (0.08)	0.55 (0.06)	0.37 (0.08)	0.48 (0.07)	0.51 (0.07)	0.33 (0.06)	0.45 (0.05)	0.51 (0.03)	0.29 (0.03)	0.41 (0.04)	0.47 (0.02)
STAGATE	0.37 (0.04)	0.55 (0.03)	0.52 (0.04)	0.34 (0.03)	0.50 (0.02)	0.51 (0.03)	0.33 (0.03)	0.50 (0.03)	0.48 (0.03)	0.33 (0.00)	0.47 (0.01)	0.52 (0.01)
SpaCAE	0.21 (0.01)	0.37 (0.01)	0.43 (0.01)	0.25 (0.03)	0.38 (0.01)	0.44 (0.03)	0.23 (0.03)	0.41 (0.03)	0.42 (0.04)	0.23 (0.02)	0.34 (0.02)	0.43 (0.03)
SpaceFlow	0.42 (0.06)	0.57 (0.05)	0.57 (0.03)	0.37 (0.04)	0.51 (0.03)	0.53 (0.03)	0.38 (0.07)	0.55 (0.06)	0.53 (0.05)	0.38 (0.05)	0.51 (0.05)	0.53 (0.04)
GraphST	0.20 (0.02)	0.34 (0.03)	0.41 (0.02)	0.27 (0.02)	0.41 (0.01)	0.46 (0.01)	0.22 (0.02)	0.34 (0.01)	0.40 (0.02)	0.26 (0.05)	0.40 (0.05)	0.45 (0.04)
Spotscape	**0.48**** (0.02)	**0.64**** (0.01)	**0.61**** (0.02)	**0.47**** (0.04)	**0.60**** (0.02)	**0.60**** (0.03)	**0.45**** (0.02)	**0.60*** (0.01)	**0.59**** (0.02)	**0.42*** (0.05)	**0.58**** (0.04)	**0.57*** (0.03)

(a) DLPFC (Patient 2)

	Slice 151507			Slice 151508			Slice 151509			Slice 151510		
	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA
SEDR	0.29 (0.06)	0.39 (0.07)	0.45 (0.06)	0.21 (0.02)	0.31 (0.02)	0.39 (0.02)	0.37 (0.04)	0.47 (0.04)	0.51 (0.05)	0.31 (0.05)	0.44 (0.04)	0.47 (0.04)
STAGATE	0.41 (0.01)	0.53 (0.01)	0.59 (0.00)	0.32 (0.01)	0.49 (0.00)	0.54 (0.01)	0.41 (0.02)	0.57 (0.02)	0.61 (0.04)	0.32 (0.03)	0.50 (0.02)	0.50 (0.02)
SpaCAE	0.28 (0.06)	0.41 (0.06)	0.46 (0.06)	0.20 (0.04)	0.31 (0.05)	0.40 (0.04)	0.31 (0.01)	0.44 (0.02)	0.50 (0.04)	0.27 (0.02)	0.42 (0.03)	0.45 (0.02)
SpaceFlow	0.55 (0.03)	0.68 (0.02)	0.71 (0.05)	0.44 (0.04)	0.57 (0.03)	0.58 (0.04)	0.53 (0.05)	0.66 (0.02)	0.65 (0.04)	0.50 (0.03)	0.64 (0.01)	0.61 (0.02)
GraphST	0.31 (0.01)	0.45 (0.01)	0.50 (0.01)	0.34 (0.01)	0.45 (0.02)	0.53 (0.02)	0.35 (0.01)	0.51 (0.01)	0.55 (0.02)	0.30 (0.02)	0.47 (0.01)	0.49 (0.03)
Spotscape	**0.60**** (0.03)	**0.72**** (0.01)	**0.76**** (0.03)	**0.48*** (0.05)	**0.64**** (0.03)	**0.63**** (0.02)	**0.59**** (0.01)	**0.71**** (0.01)	**0.70**** (0.02)	**0.53*** (0.04)	**0.67**** (0.02)	**0.64** (0.04)

(a) DLPFC (Patient 3)

	Slice 151669			Slice 151670			Slice 151671			Slice 151672		
	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA
SEDR	0.24 (0.07)	0.40 (0.07)	0.48 (0.06)	0.24 (0.06)	0.39 (0.05)	0.48 (0.05)	0.37 (0.10)	0.50 (0.09)	0.59 (0.07)	0.49 (0.09)	0.58 (0.06)	0.66 (0.07)
STAGATE	0.29 (0.05)	0.45 (0.07)	0.52 (0.04)	0.20 (0.01)	0.38 (0.01)	0.44 (0.01)	0.40 (0.07)	0.49 (0.03)	0.63 (0.06)	0.38 (0.02)	0.51 (0.04)	0.54 (0.01)
SpaCAE	0.21 (0.02)	0.28 (0.03)	0.43 (0.02)	0.21 (0.03)	0.28 (0.02)	0.43 (0.04)	0.38 (0.16)	0.29 (0.01)	0.49 (0.05)	0.25 (0.04)	0.35 (0.05)	0.50 (0.01)
SpaceFlow	0.30 (0.07)	0.48 (0.03)	0.51 (0.05)	0.34 (0.05)	0.50 (0.03)	0.56 (0.05)	0.54 (0.04)	0.67 (0.02)	0.67 (0.04)	0.60 (0.06)	0.70 (0.02)	0.73 (0.06)
GraphST	0.17 (0.04)	0.26 (0.04)	0.43 (0.02)	0.14 (0.01)	0.23 (0.00)	0.37 (0.01)	0.30 (0.05)	0.38 (0.03)	0.54 (0.03)	0.23 (0.01)	0.32 (0.02)	0.49 (0.01)
Spotscape	**0.46**** (0.02)	**0.58**** (0.01)	**0.65**** (0.02)	**0.45**** (0.04)	**0.56**** (0.03)	**0.66**** (0.03)	**0.68**** (0.10)	**0.74**** (0.04)	**0.79**** (0.08)	**0.75**** (0.04)	**0.74**** (0.02)	**0.84**** (0.05)

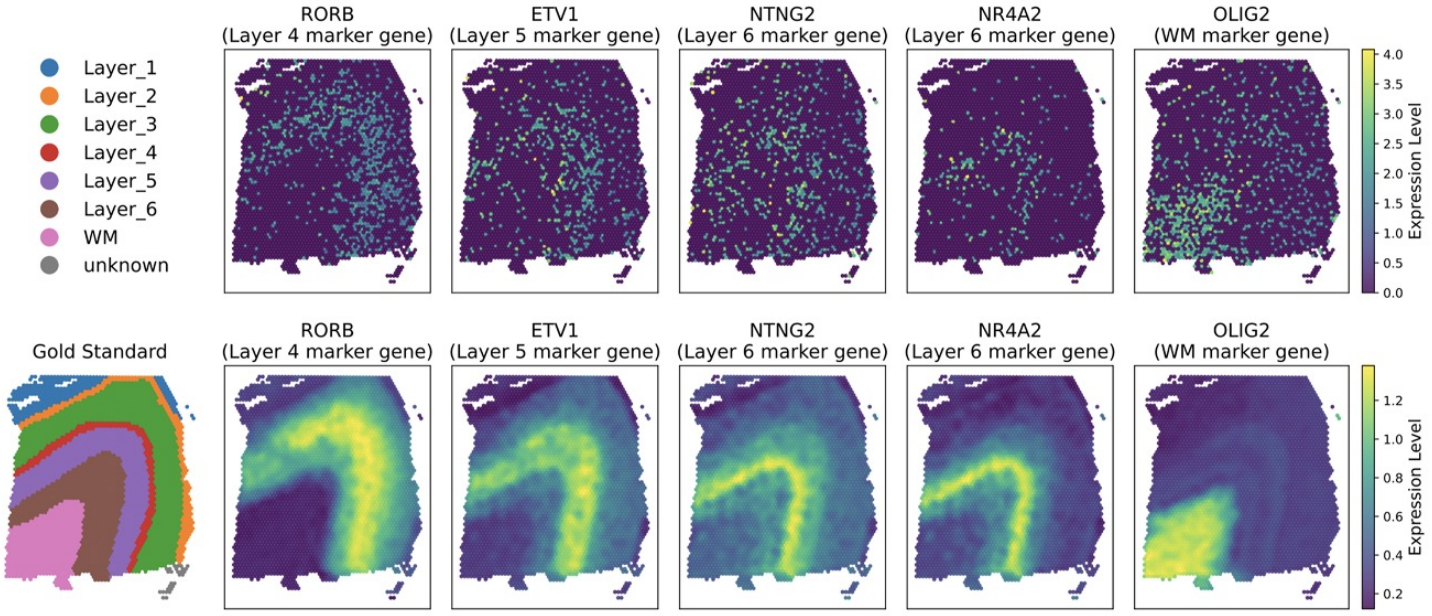
	(b) MTG - Control Group			(b) MTG - AD Group			(c) Mouse Embryo			(d) NSCLC		
	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA	ARI	NMI	CA
SEDR	0.41 (0.02)	0.59 (0.02)	0.52 (0.02)	0.43 (0.08)	0.59 (0.07)	0.57 (0.07)	0.32 (0.02)	0.56 (0.01)	0.42 (0.02)	0.44 (0.08)	0.46 (0.06)	0.70 (0.08)
STAGATE	0.54 (0.00)	0.65 (0.00)	0.59 (0.00)	0.51 (0.01)	0.61 (0.01)	0.59 (0.01)	0.36 (0.01)	0.60 (0.01)	0.47 (0.01)	0.35 (0.05)	0.41 (0.04)	0.64 (0.02)
SpaCAE	0.37 (0.03)	0.52 (0.00)	0.44 (0.03)	0.22 (0.01)	0.40 (0.01)	0.40 (0.01)	0.34 (0.01)	0.60 (0.01)	0.48 (0.02)	0.32 (0.05)	0.38 (0.03)	0.62 (0.02)
SpaceFlow	0.66 (0.03)	0.74 (0.01)	0.70 (0.03)	0.54 (0.01)	0.71 (0.00)	0.65 (0.01)	0.42 (0.03)	0.60 (0.02)	0.49 (0.03)	0.53 (0.03)	0.52 (0.02)	**0.75** (0.02)
GraphST	0.38 (0.00)	0.51 (0.00)	0.48 (0.00)	0.43 (0.06)	0.55 (0.05)	0.55 (0.04)	0.34 (0.01)	0.59 (0.02)	0.45 (0.01)	0.30 (0.00)	0.38 (0.00)	0.65 (0.00)
Spotscape	**0.73**** (0.02)	**0.78**** (0.01)	**0.75**** (0.03)	**0.68**** (0.02)	**0.75**** (0.01)	**0.77**** (0.03)	**0.44** (0.01)	**0.63**** (0.01)	**0.54**** (0.01)	**0.57**** (0.02)	**0.57**** (0.01)	0.74 (0.01)

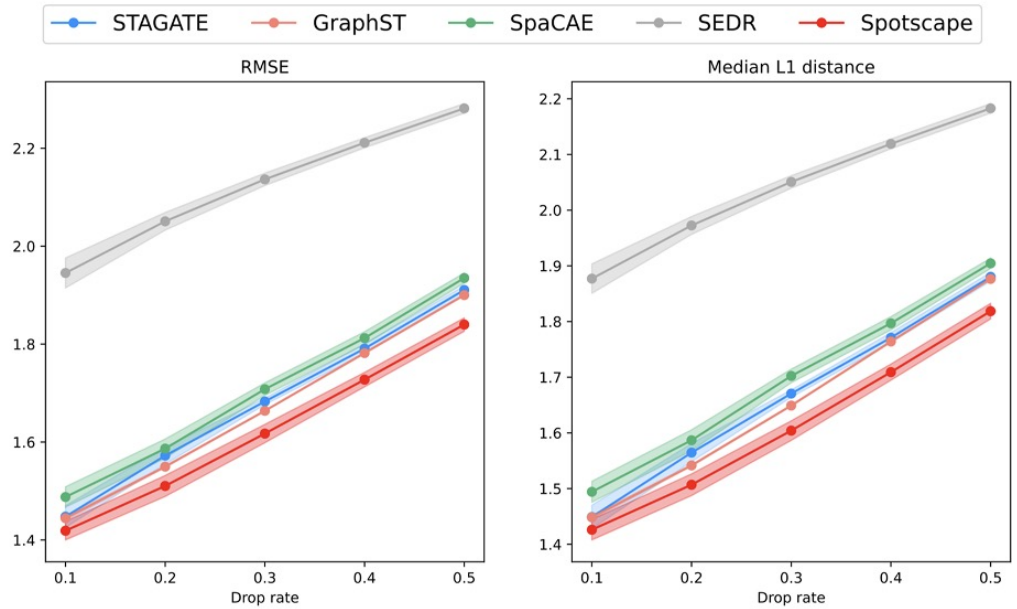
All experiments are repeated over 10 runs with different random seeds, and we report the mean and standard deviation of the results. For all experimental results, **Bold** indicates the best performance, underlining denotes the second-best, and an asterisk (*) marks statistically significant improvements of Spotscape over the top-performing baseline based on a paired t-test (**: $p < 0.01$, *: $p < 0.05$), with the numbers in parentheses representing the standard deviation.

- **Spotscape consistently outperforms** all baselines across 4 datasets and 16 slices
- **Move beyond the limited insights** of local neighbor analysis by **capturing global contextual information**

EXPERIMENTS Single-slice Tasks

Denoising & Imputation





- **Spotscape** clarifies marker genes expression for easier identification in noisy raw data in denoising task
- **Spotscape** achieves **best performance in imputation**, leading baselines on RMSE and L1-distance metrics

EXPERIMENTS Multi-slice Tasks

Integration

Table 2. Homogeneous integration performance on DLPFC data.

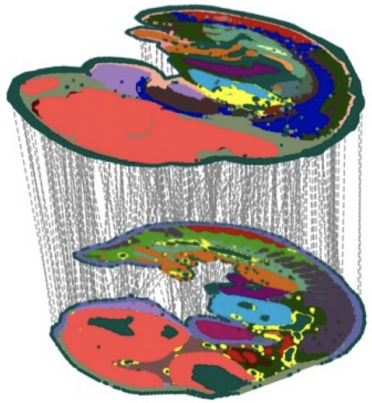
	Patient 1			Patient 2			Patient 3		
	ARI	**NMI**	**CA**	**ARI**	**NMI**	**CA**	**ARI**	**NMI**	**CA**
SEDR	0.38 (0.06)	0.49 (0.06)	0.56 (0.06)	0.32 (0.05)	0.44 (0.07)	0.48 (0.07)	0.43 (0.02)	0.51 (0.01)	0.56 (0.03)
STAGATE	0.31 (0.03)	0.46 (0.03)	0.49 (0.03)	0.30 (0.02)	0.46 (0.01)	0.48 (0.02)	0.31 (0.09)	0.43 (0.06)	0.54 (0.08)
SpaCAE	0.21 (0.03)	0.36 (0.02)	0.40 (0.02)	0.12 (0.06)	0.19 (0.07)	0.32 (0.05)	0.13 (0.05)	0.14 (0.05)	0.43 (0.06)
SpaceFlow	0.48 (0.03)	0.60 (0.02)	0.60 (0.02)	0.44 (0.05)	0.59 (0.02)	0.58 (0.04)	0.51 (0.02)	0.60 (0.01)	0.69 (0.05)
GraphST	0.18 (0.01)	0.32 (0.01)	0.38 (0.02)	0.25 (0.01)	0.39 (0.01)	0.42 (0.02)	0.25 (0.04)	0.30 (0.04)	0.50 (0.01)
PASTE	0.34 (0.00)	0.45 (0.00)	0.54 (0.00)	0.17 (0.00)	0.28 (0.00)	0.40 (0.00)	0.29 (0.00)	0.43 (0.00)	0.54 (0.00)
STAligner	0.38 (0.04)	0.52 (0.04)	0.55 (0.04)	0.29 (0.02)	0.45 (0.02)	0.48 (0.03)	0.37 (0.06)	0.47 (0.05)	0.59 (0.06)
CAST	0.26 (0.02)	0.37 (0.03)	0.42 (0.03)	0.30 (0.04)	0.43 (0.05)	0.47 (0.03)	0.38 (0.06)	0.40 (0.04)	0.56 (0.05)
Spotscape	**0.57**** (0.03)	**0.70**** (0.02)	**0.67**** (0.03)	**0.53**** (0.02)	**0.67**** (0.01)	**0.63**** (0.02)	**0.63**** (0.09)	**0.68**** (0.03)	**0.75**** (0.09)

Table 3. Heterogeneous integration performance on MTG data.

	Clustering Metric			Batch Effect Correction Metric			
	ARI	**NMI**	**CA**	**Silhouette batch**	**kBET**	**Graph connectivity**	**PCR comparison**
GraphST	0.23 (0.02)	0.42 (0.00)	0.39 (0.01)	0.56 (0.00)	0.02 (0.01)	0.65 (0.02)	0.00 (0.00)
STAligner	0.38 (0.03)	0.54 (0.03)	0.49 (0.02)	0.62 (0.04)	0.11 (0.08)	0.85 (0.04)	0.18 (0.10)
CAST	0.48 (0.07)	0.52 (0.06)	0.59 (0.06)	0.45 (0.02)	**0.11** (0.02)	0.81 (0.06)	**0.97** (0.03)
Spotscape (w/o $\mathcal{L}_{PCL}$)	0.61 (0.03)	0.71 (0.01)	0.70 (0.02)	0.67 (0.01)	0.03 (0.00)	0.79 (0.03)	0.50 (0.04)
Spotscape (w/o $\mathcal{L}_{SS}$)	0.47 (0.09)	0.60 (0.04)	0.59 (0.06)	0.24 (0.01)	0.00 (0.00)	0.63 (0.00)	0.00 (0.00)
Spotscape	**0.72**** (0.04)	**0.76**** (0.01)	**0.81**** (0.05)	**0.69**** (0.01)	0.08 (0.02)	**0.86** (0.03)	0.60 (0.08)

All experiments are repeated over 10 runs with different random seeds, and we report the mean and standard deviation of the results. For all experimental results, **Bold** indicates the best performance, underlining denotes the second-best, and an asterisk (*) marks statistically significant improvements of Spotscape over the top-performing baseline based on a paired t-test (**: $p < 0.01$, *: $p < 0.05$), with the numbers in parentheses representing the standard deviation.

Alignment



	LTARI
PASTE2	0.21 (0.02)
CAST	0.10 (0.00)
STAligner	0.46 (0.01)
SLAT	0.41 (0.11)
Spotscape	**0.51**** (0.01)

- **Homogeneous integration:** Consistently outperforms all baselines when integrating multi-slices from the same patient
- **Heterogeneous integration:** Integrates diverse samples (e.g., Control vs. AD) by correcting batch effects, significantly outperforming competitors
- **Multi-slice Alignment:** Outperforms even specialized tools, successfully aligning slices across different developmental stages

EXPERIMENTS Multi-slice Tasks

DEG Analysis

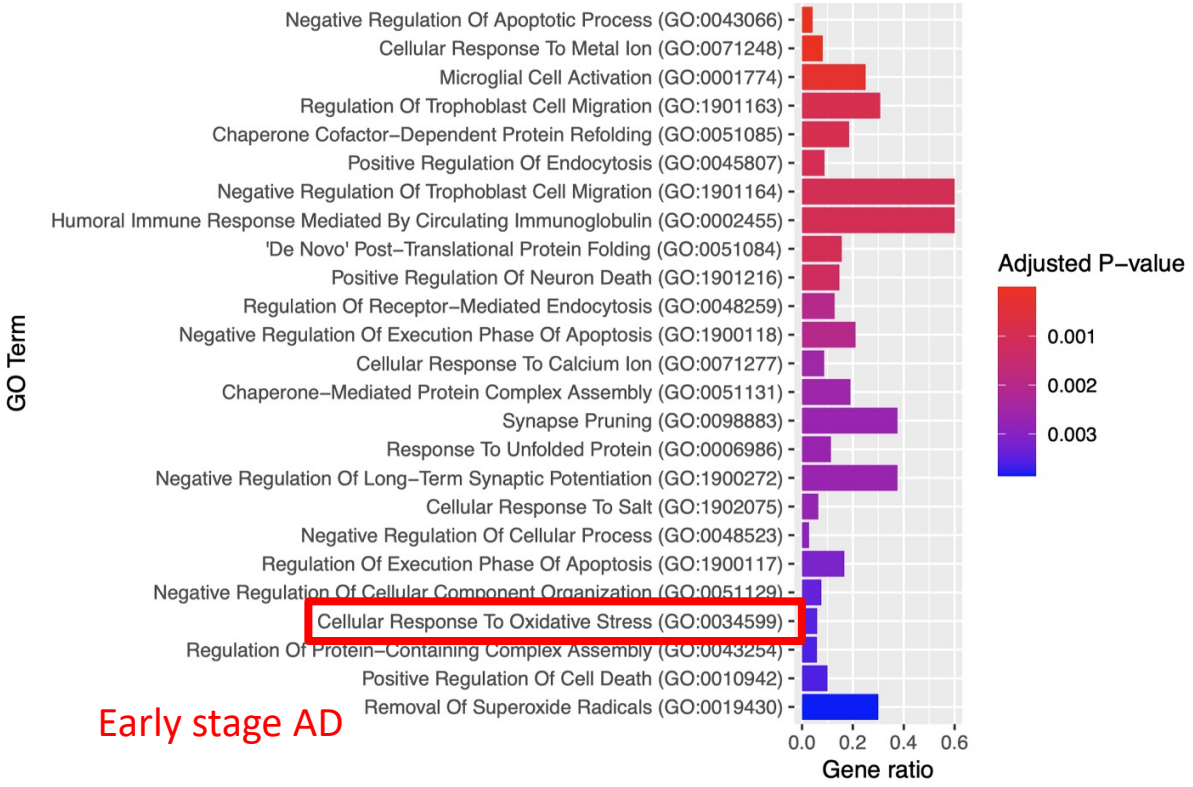


Figure 29. Top 25 biological process that DEGs between AD and Control enriched in a cluster assigned to layer 2.

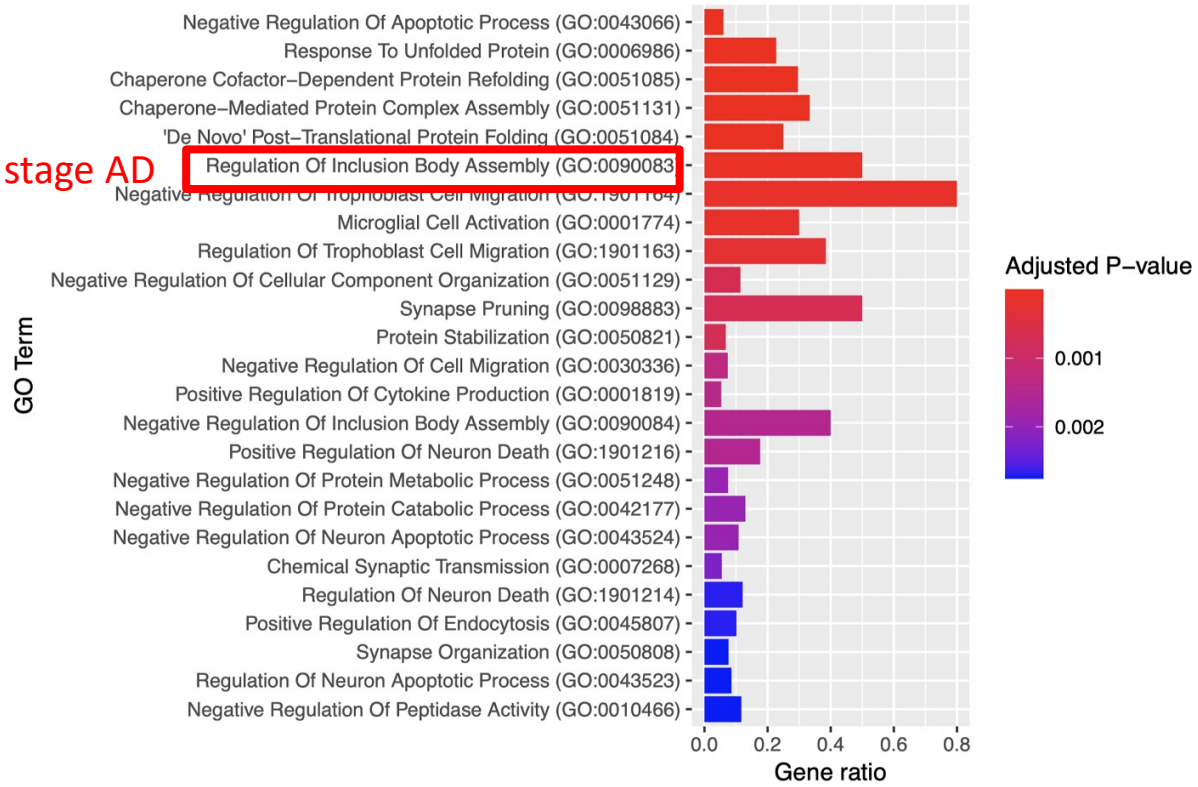
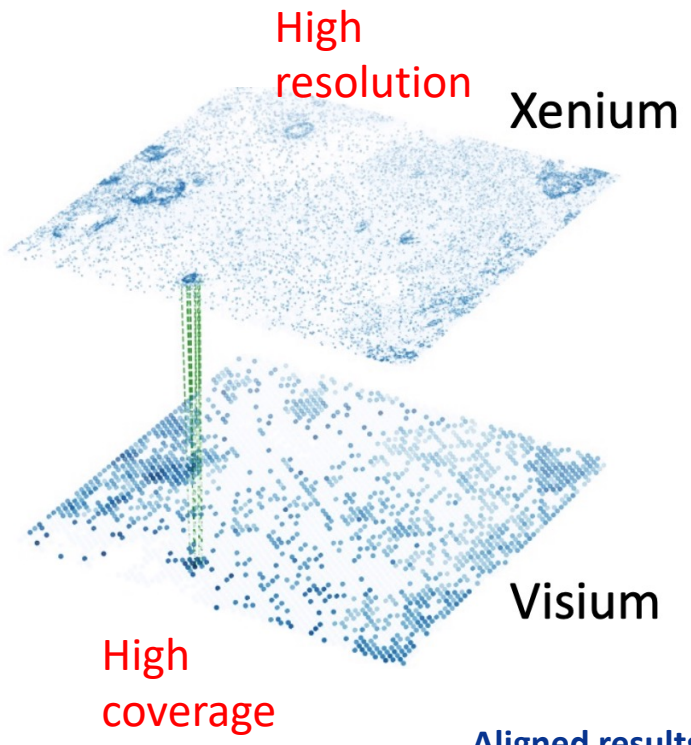
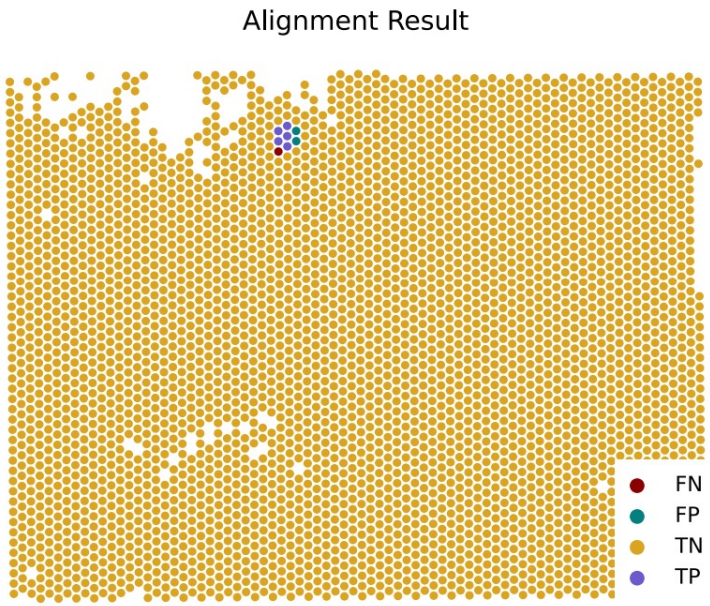


Figure 30. Top 25 biological process that DEGs between AD and Control enriched in a cluster assigned to layer 5.

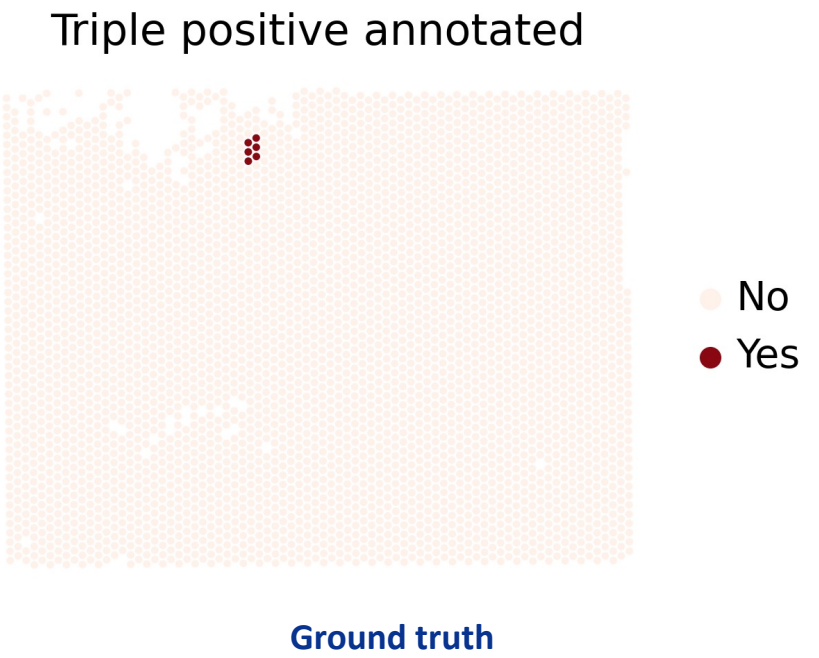
- **Biological Validation:** Gene analysis of Control vs. Alzheimer's samples validates **Spotscape**'s finding, as it correctly identified distinct early-stage (Layer 2) and late-stage (Layer 5) disease pathologies that align with known AD progression

Cross Technology Alignment





Aligned results of Spotscape



Ground truth

• Spotscape leverages cross-technology alignment to successfully identify and map even **extremely rare cell types** (e.g., triple-positive cancer cells)

SUMMARY

- Single-cell RNA seq 데이터의 noise 문제 연구
  - Noise in Zero & Non-zero
- Spatial transcriptomics 데이터
  - Boundary의 cell type clustering 문제
  - Multi-slice alignmnet의 Batch effect 문제