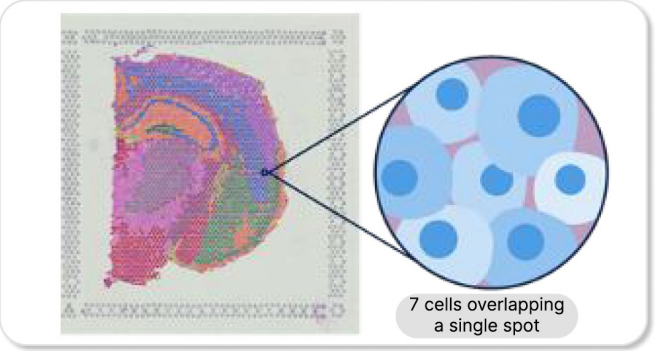
Beyond Local Neighborhoods: Global Similarity Learning for Spatial Transcriptomics

박찬영

Associate Professor, KAIST

Industrial and Systems Engineering / Graduate School of AI
cy.park@kaist.ac.kr

Background: Spatially Resolved Transcriptomics (SRT)



	Gene 1	Gene 2	Gene 3	Gene 4
Spot 1	18	1010	0	22
Spot 2	0	506	49	0
Spot 3	0	0	0	72

- SRT incorporate the spatial context and gene expression of cells
  - Allows for the identification of where specific genes are expressed within tissue
  - Enables the study of disease, such as cancer, by observing gene expression patterns in different tissue regions

Image credit: https://www.10xgenomics.com/analysis-guides/integrating-single-cell-and-visium-spatial-gene-expression-data

Background: Spatially Resolved Transcriptomics (SRT)

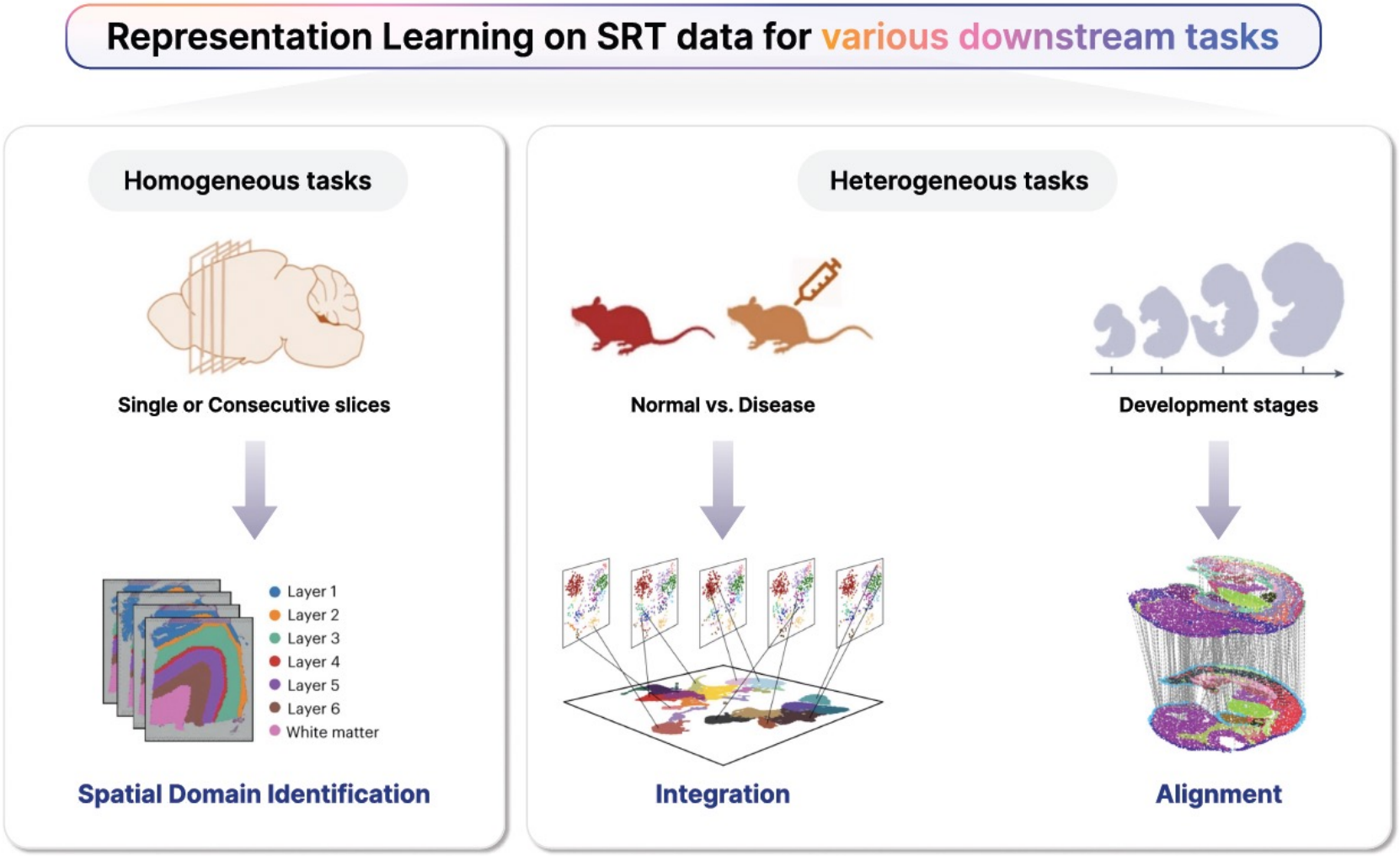
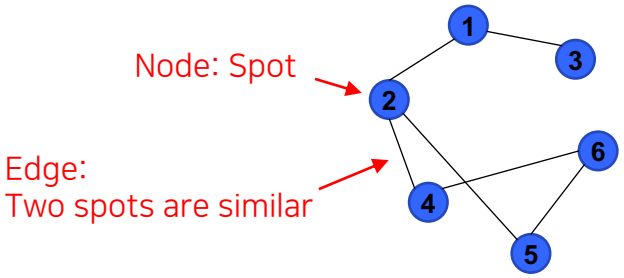


Image credit: [NatureComputationalScience2023] Integrating spatial transcriptomics data across different conditions, technologies and developmental stages
https://www.archrproject.com/bookdown/defining-cluster-identity-with-scrna-seq.html

Graph construction & Graph Neural Network

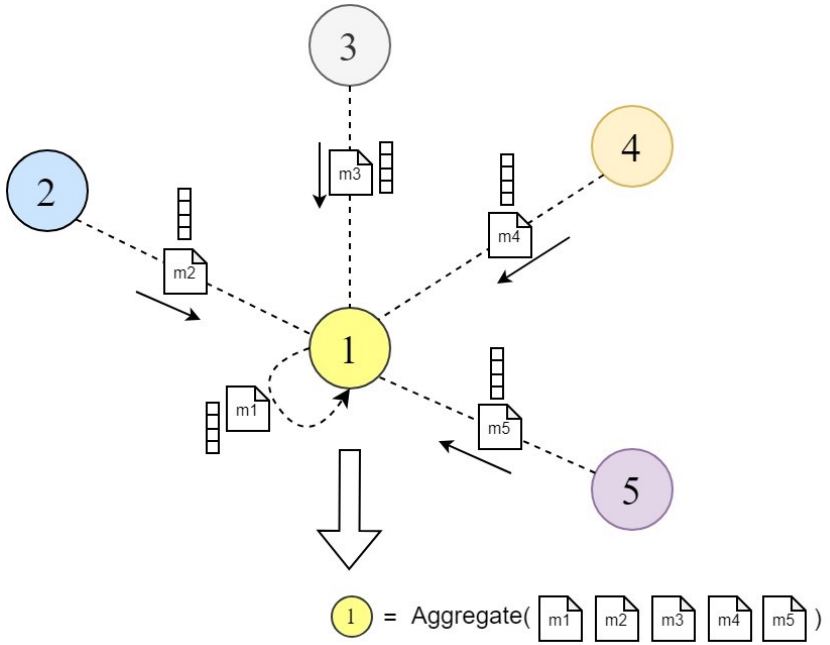
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Spot 1	5	0	18	0	0	0	0
Spot 2	10	29	0	5	0	0	0
Spot 3	0	40	5	0	10	0	0
Spot 4	0	0	0	19	0	37	0
Spot 5	9	0	0	0	0	0	3
Spot 6	0	0	0	97	7	0	8

① Calculate Spot-Spot Similarity



K (=2) Nearest Neighbor Graph Construction

② Representation Learning



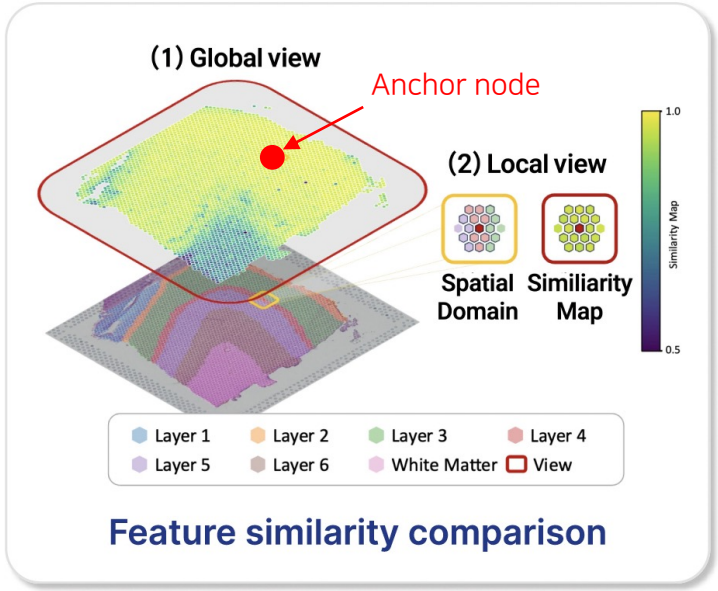
Graph Neural Network
(Focus on local Information)

STAGATE, SpaceFlow (2022, Nat. Comm.)
STAligner (2023, Nat. Comp. Sci)
SLAT (2023, Nat. Comm.)
GraphST (2024, Nat. Comm.)
...

Challenge 1: Continuous nature of SRT data

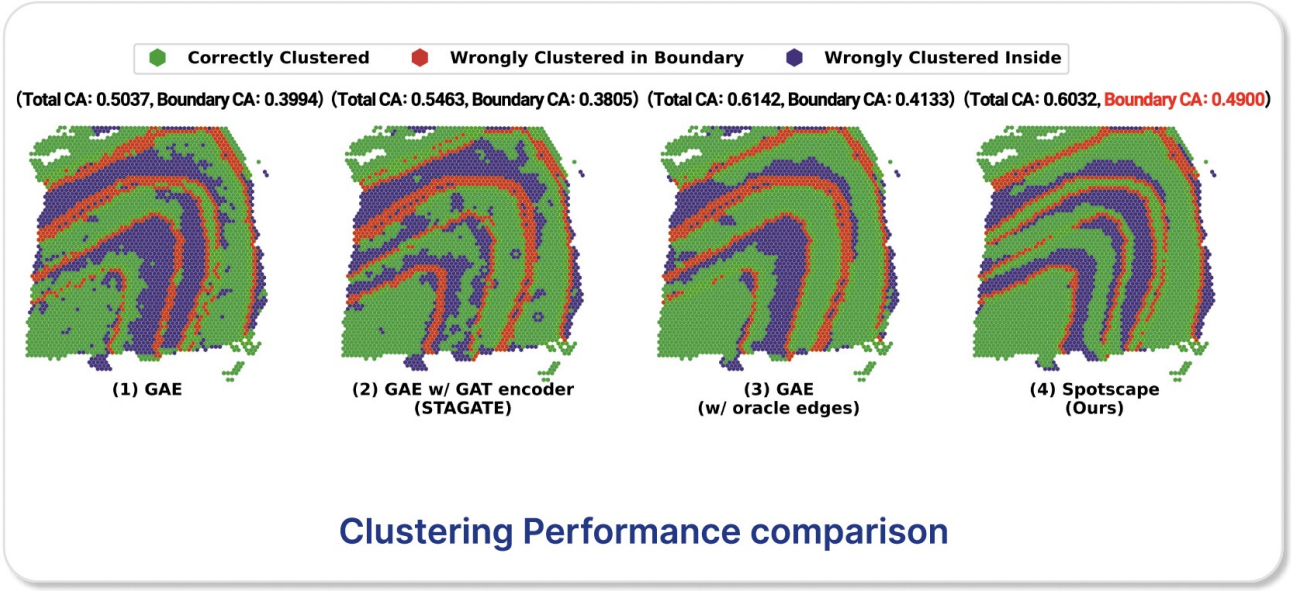
- Gene expression values vary smoothly along spatial coordinates

- TotalCA: 전체 spot 클러스터링 성능
- Boundary CA: Boundary spot 클러스터링 성능



Dorsolateral prefrontal cortex (DLPFC) data

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



(1) GNNs receive information from heterophilic nodes

(2) Hard to learn attention scores due to highly similar neighboring nodes

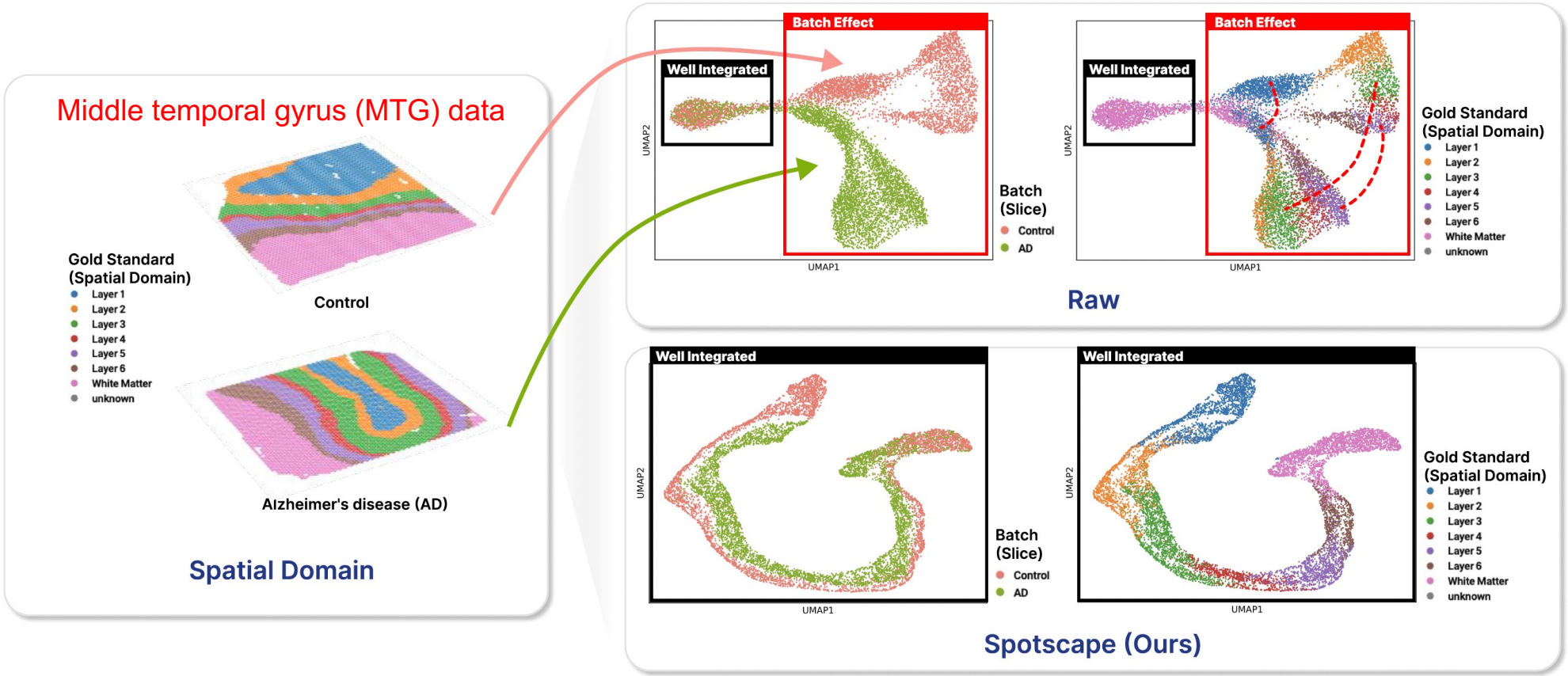
(3) Even oracle edges performs poorly on clustering near boundary

(4) Ours show good Boundary CA, while maintaining Total CA

Should capture relationship between spots in global context

Challenge 2: Batch effects in SRT data

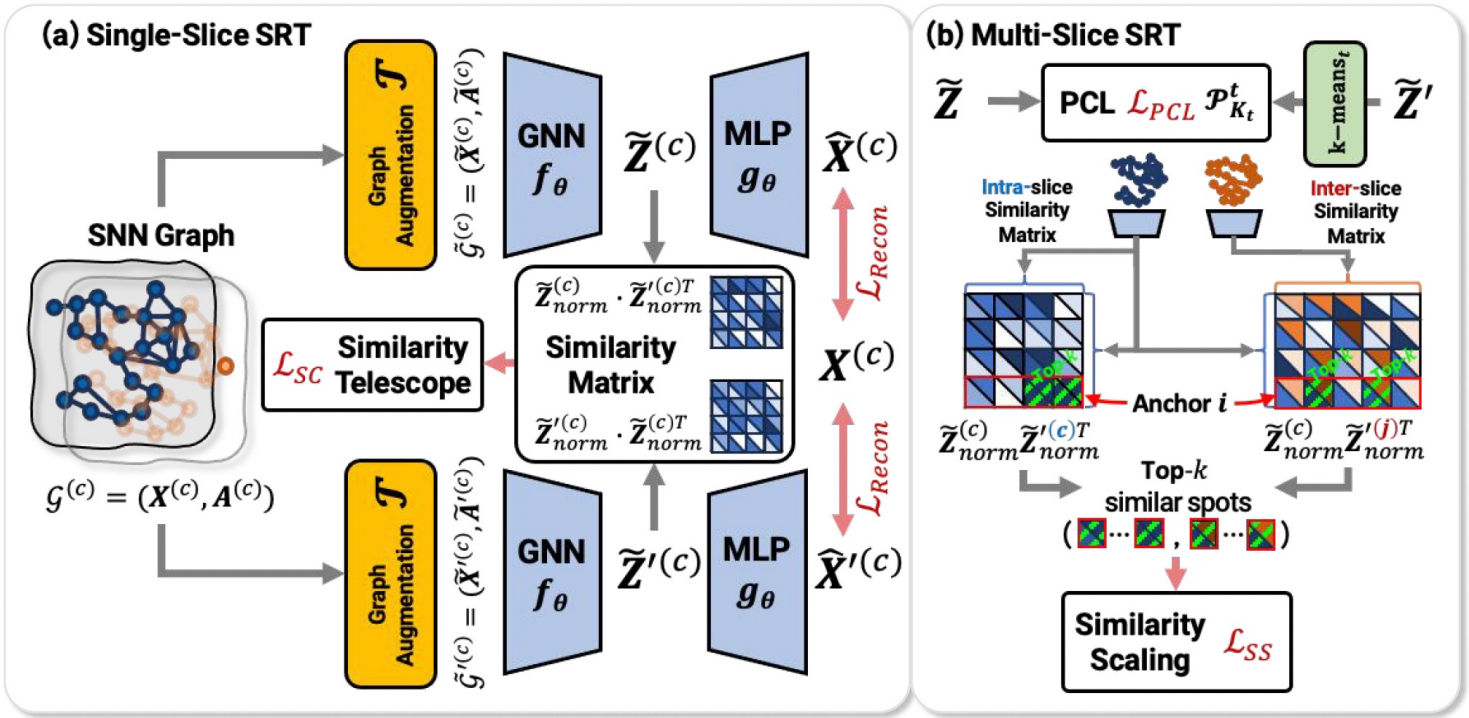
- Gene expression profiles from the same slice cluster together unexpectedly, regardless of their biological relevance



Should alleviate batch effects to extend multi-slices tasks

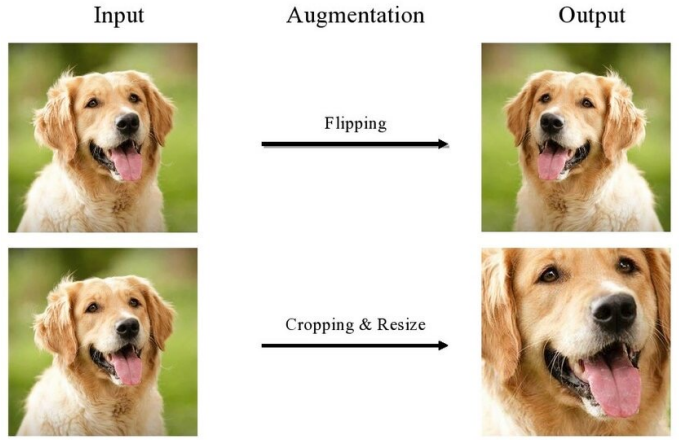
Proposed Method: SpotScape

- Key ideas
  - (1) Capture **global relationships between spots** by reflecting the global context among multiple spots
  - To mitigate batch effects,
    - (2) **balance the scales of intra-slice and inter-slice similarity** between spots
    - (3) **group spots from the same spatial domain** while distancing others

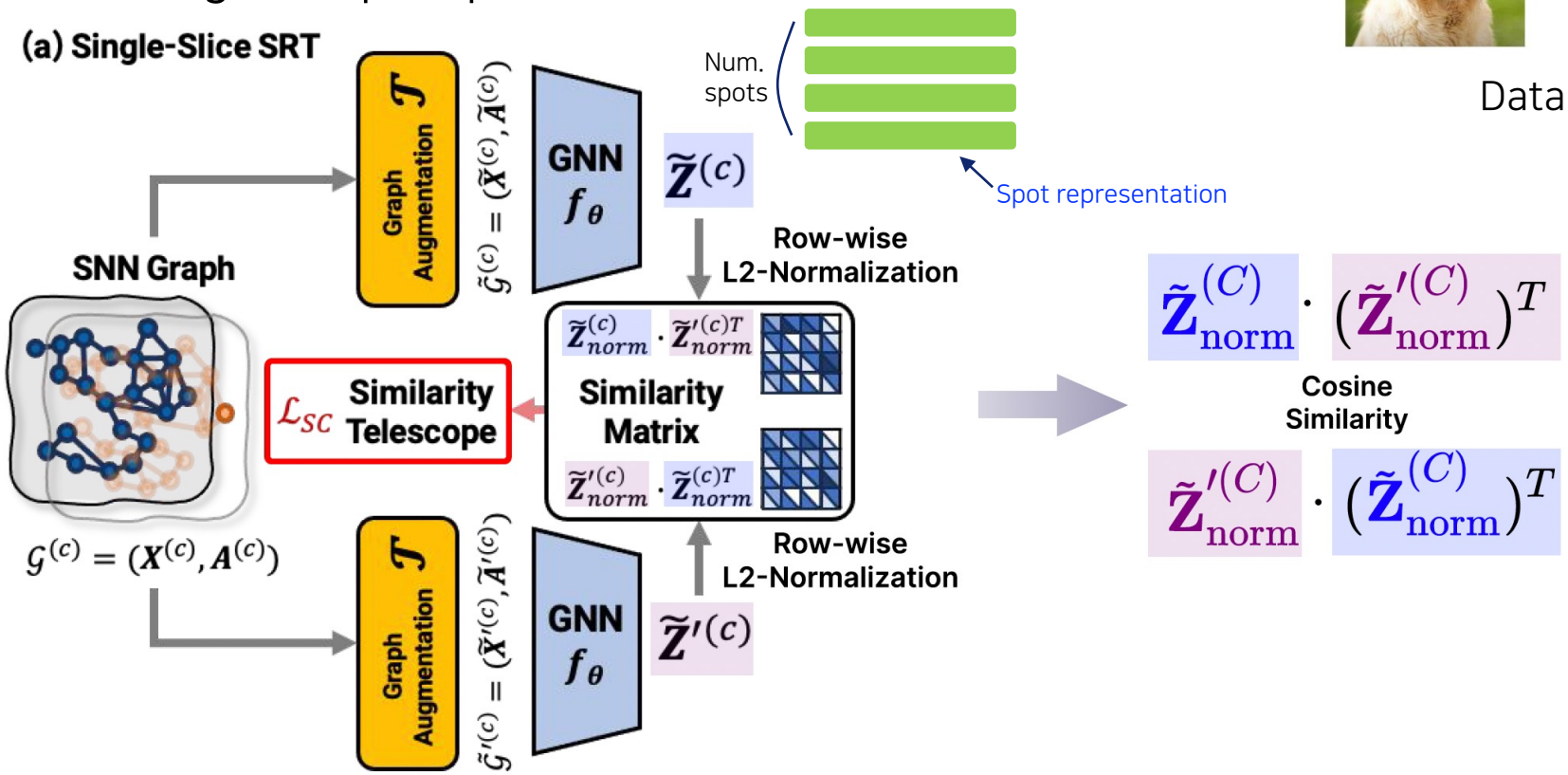


Similarity Consistency Loss

- (1) Capture **global relationships between spots** by reflecting the global context among multiple spots

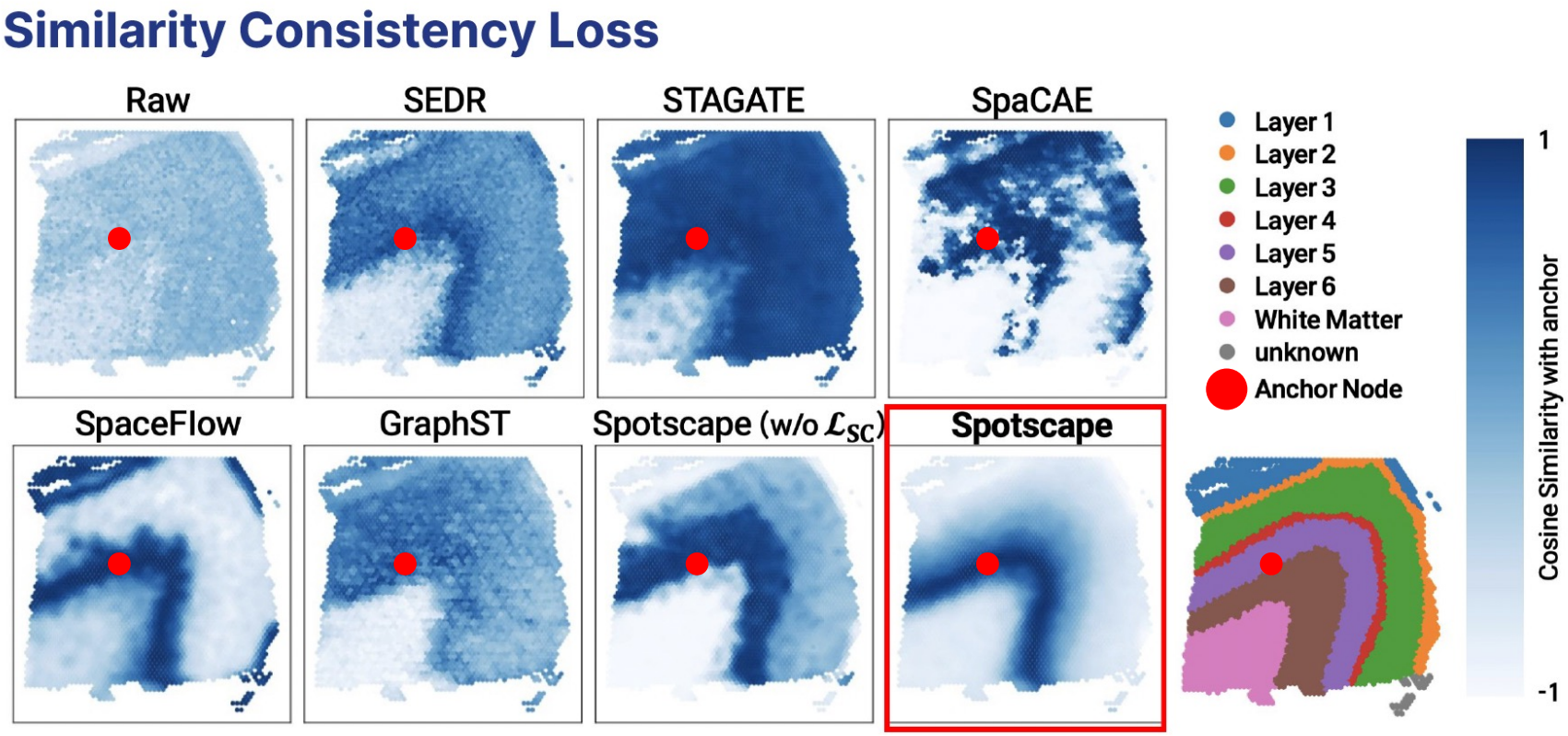


Data Augmentation



$$\mathcal{L}_{\text{SC}}(\tilde{Z}^{(C)}, \tilde{Z}^{\prime(C)}) = \text{MSE}(\tilde{Z}^{(C)}_{\text{norm}}(\tilde{Z}^{\prime(C)}_{\text{norm}})^T, \tilde{Z}^{\prime(C)}_{\text{norm}}(\tilde{Z}^{(C)}_{\text{norm}})^T)$$

Similarity Consistency Loss

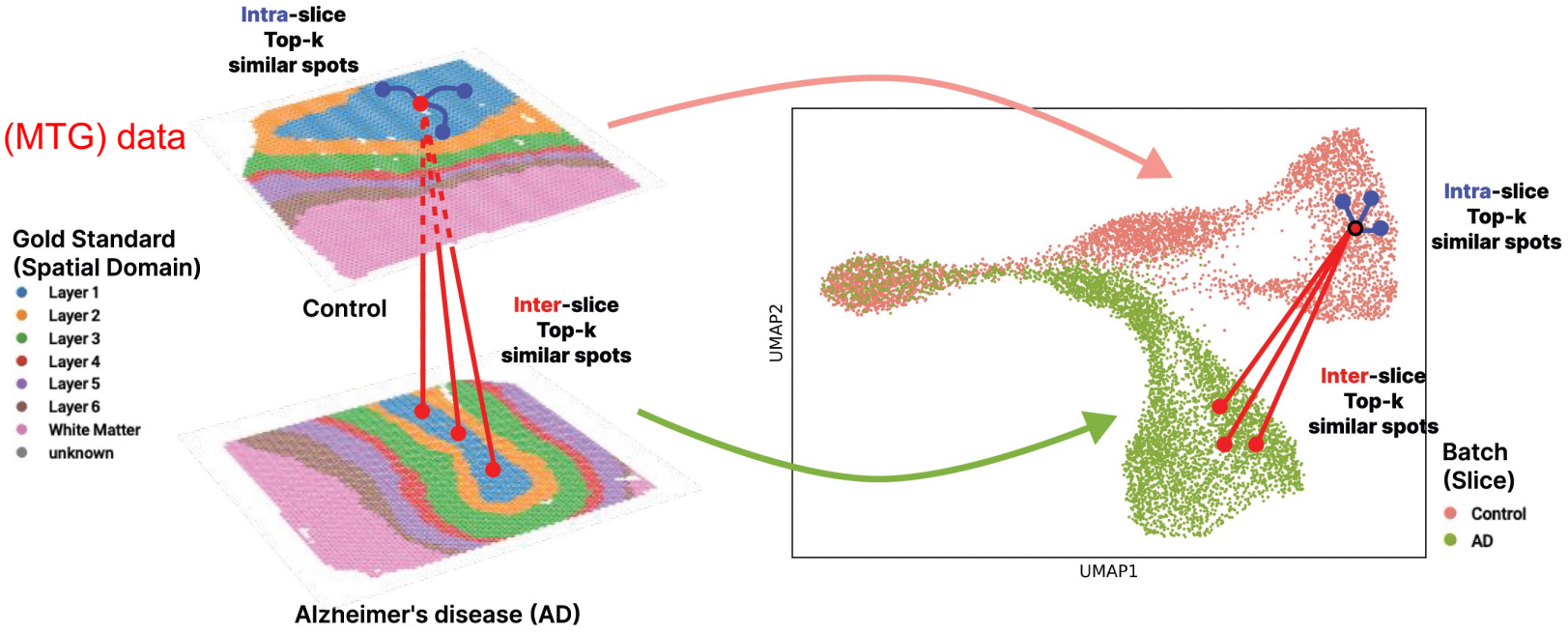


Spotscape captures relative similarity between spots regardless of the actual distance

Similarity Scaling Loss (Batch effect mitigation)

- (2) Balance the scales of intra-slice and inter-slice similarity between spots

Middle temporal gyrus (MTG) data



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

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

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

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

(i-th row, l-th row)

Prototypical Contrastive Loss (Batch effect mitigation)

- (3) **Group spots from the same spatial domain** while distancing others

Prototype set of t times k-means

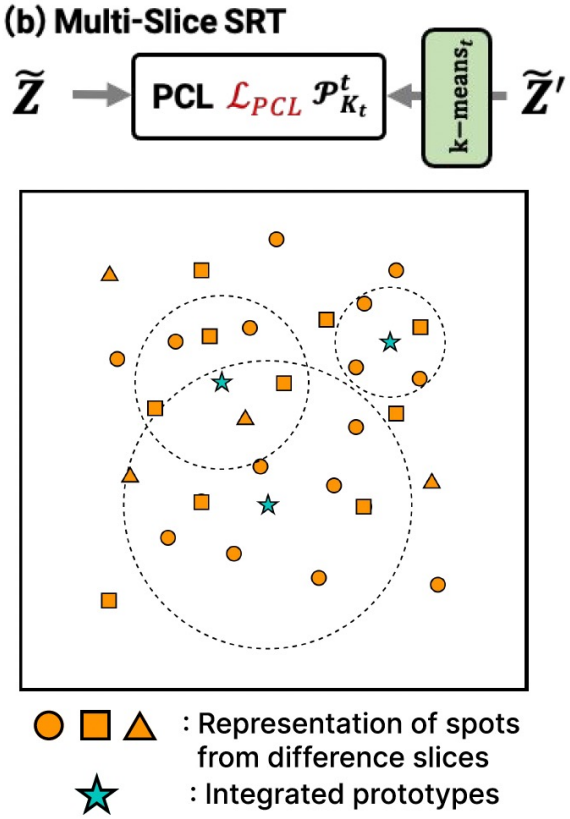
$\mathcal{P}^t_{K_t}$ ← k-means$_t$ ← $\tilde{Z}'$

Prototype

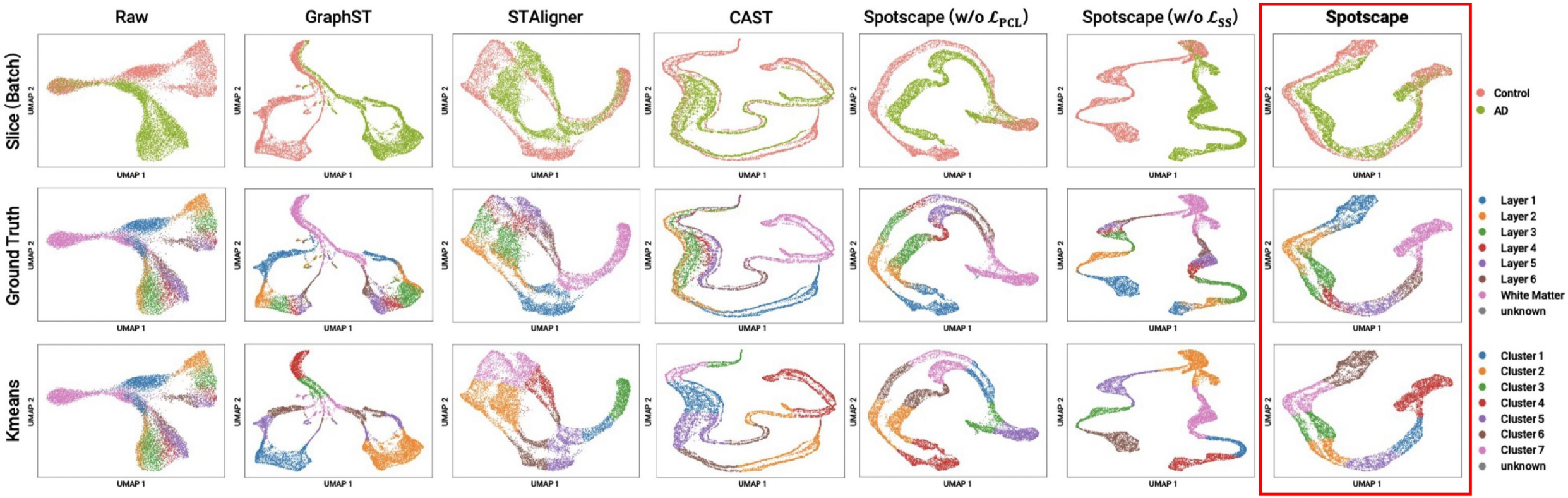
$$l_{\mathrm{PCL}}(\tilde{Z}_i, P_{\mathrm{set}}) = \frac{1}{T}\sum_{t=1}^{T} \log \frac{e^{(\mathrm{sim}(\tilde{Z}_i, p^t_{\mathrm{map}_t(i)})/\tau)}}{\sum_{j=1}^{K_t} e^{(\mathrm{sim}(\tilde{Z}_i, p^t_j)/\tau)}},$$

$$\mathcal{L}_{\mathrm{PCL}} = -\frac{1}{N_s}\sum_{i=1}^{N_s} l_{\mathrm{PCL}}(\tilde{Z}_i, P_{\mathrm{set}}).$$

- To avoid the risk of obtaining inaccurate prototypes, the PCL loss gets involved to the training procedure after a **warm-up period**



UMAP visualization result



Ours effectively integrates spot representations across slices while preserving biological meaning

Experiments: Datasets

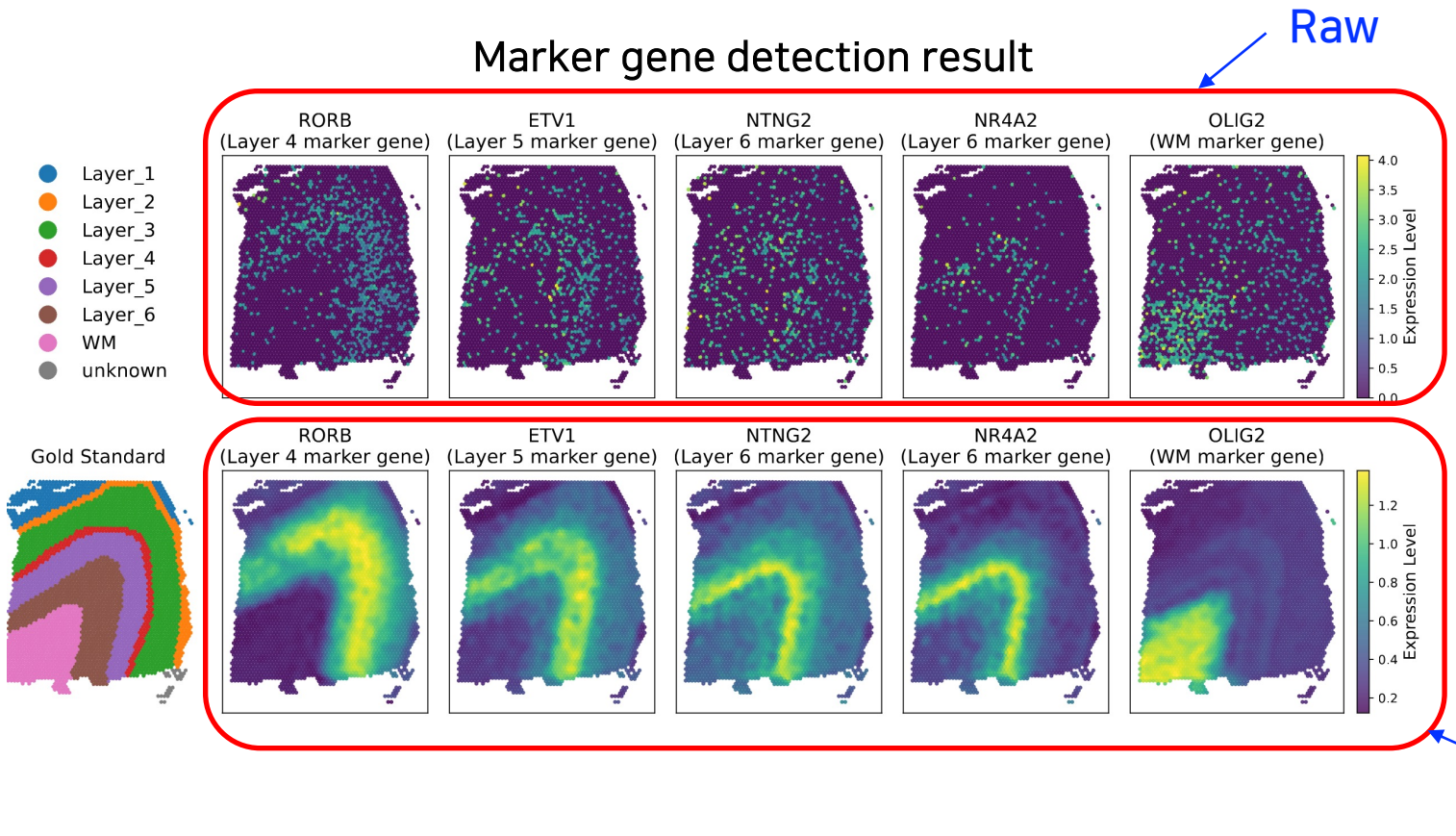
Dataset	Species	Tissue	Technology	Resolution	Spots	Genes	# Spatial Domains	Experimental Usage
DLPFC	Human	Brain (dorsolateral prefrontal cortex)	10x Visium	50 μm	3,460 ~ 4,789	33,538	5 ~ 7	- Single-slice SDI - Trajectory inference - Homogeneous multi-slice integration
MTG	Human	Brain (middle temporal gyrus)	10x Visium	50 μm	3,445 ~ 4,832	36,601	6 ~ 7	- Single-slice SDI - Heterogeneous integration (Control vs AD) - Batch-effect correction
Mouse Embryo	Mouse	Whole embryo	Stereo-seq	0.2 μm	30,756 ~ 55,295	25,485 ~ 27,330	18 ~ 19	- Single-slice SDI - Developmental trajectory inference - Heterogeneous slice alignment
NSCLC	Human	Non-small cell lung cancer	NanoString CosMX	Subcellular	960	11,756	4	- Single-slice SDI (High-resolution)
Breast Cancer	Human	Breast cancer	10x Visium	50 μm	4,992	18,085	11	Heterogeneous alignment (cross-sample / cell-type structure)
	Human	Breast cancer	10x Xenium	Subcellular	167,780	313	20	Cross-technology alignment (Visium ↔ Xenium)

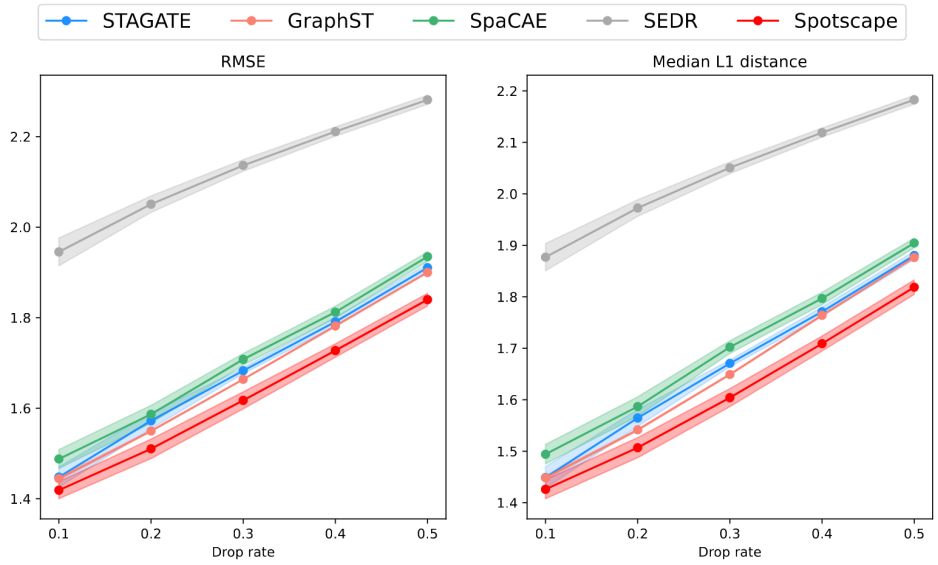
Experiments: Downstream Tasks

- Single-slice Tasks
  - Spatial Domain Identification
  - Trajectory Inference
  - Imputation

- Multi-slice Tasks
  - Integration
  - Alignment
  - Batch Effect Correction
  - Differentially Expressed Gene (DEG) Analysis

Single-slice Task: Imputation





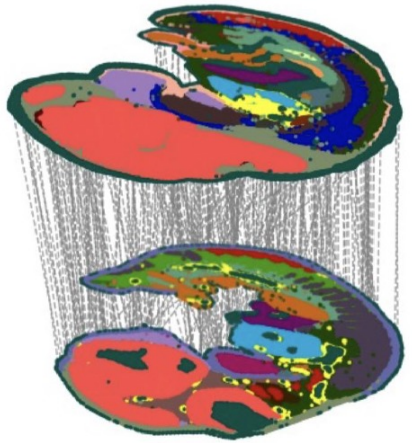
- Spotscape clarifies marker gene expression for easier identification in noisy raw data
- Spotscape achieves best imputation performance

Multi-slice Task: Integration / Alignment

Each patient has 4 slices

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)



Mouse Embryo

Control vs AD

Table 3 Heterogeneous integration performance on MTG data.

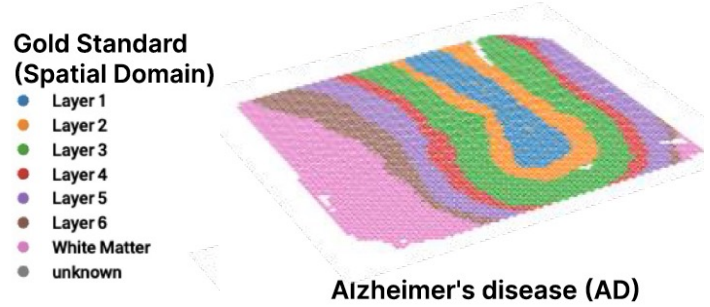
	Clustering Metric			Batch Effect Correction Metric			
	ARI	NMI	CA	Silhouette batch	kBET	Graph connectivity	PCR comparison
GraphST	0.23 (0.02)	0.42 (0.00)	0.39 (0.01)	0.56 (0.00)	0.02 (0.01)	0.65 (0.02)	0.00 (0.00)
STAligner	0.38 (0.03)	0.54 (0.03)	0.49 (0.02)	0.62 (0.04)	0.11 (0.08)	0.85 (0.04)	0.18 (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 $\mathcal{L}_{\mathrm{PCL}}$)	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 $\mathcal{L}_{\mathrm{SS}}$)	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)

Label Transfer ARI

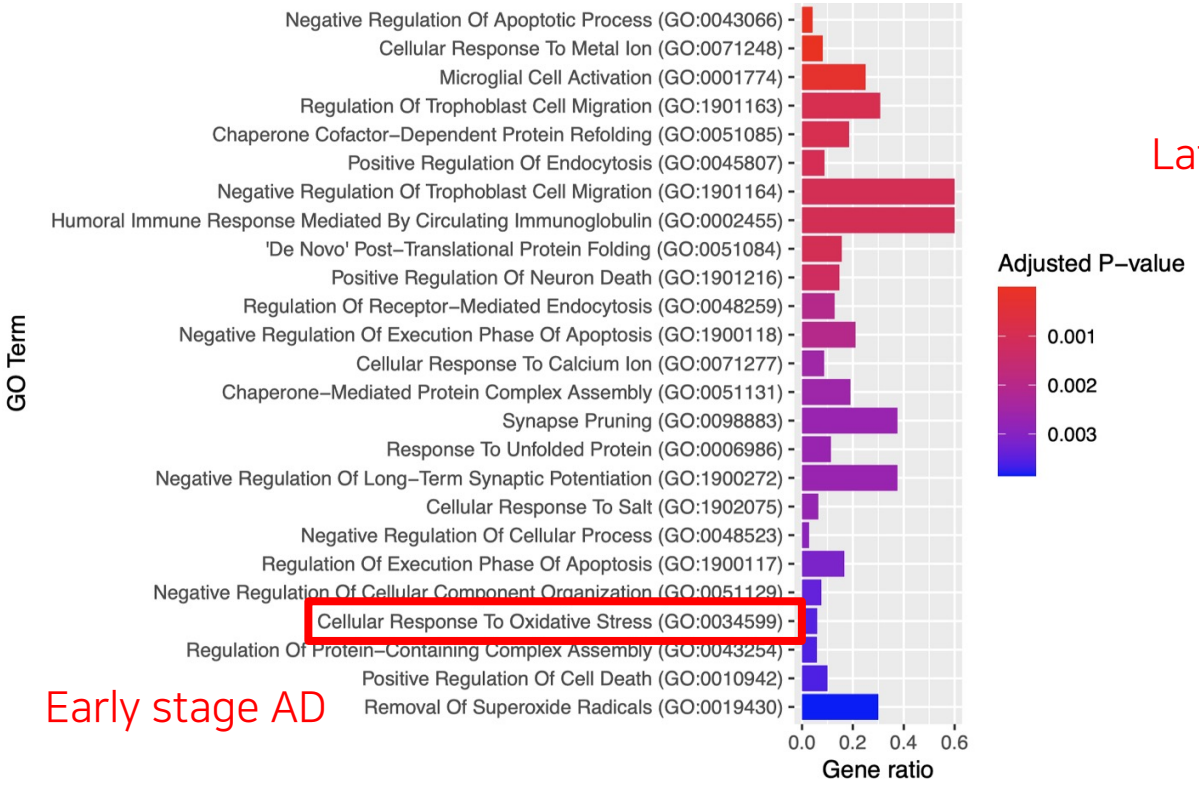
	LTARI
PASTE2	0.21 (0.02)
CAST	0.10 (0.00)
STAligner	0.46 (0.01)
SLAT	0.41 (0.11)
Spotscape	**0.51**** (0.01)

- **Homogeneous Integration:** Consistently outperforms all baseline methods when integrating multiple tissue slices from the same patient sample
- **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 and technologies

Multi-slice Task: DEG Analysis



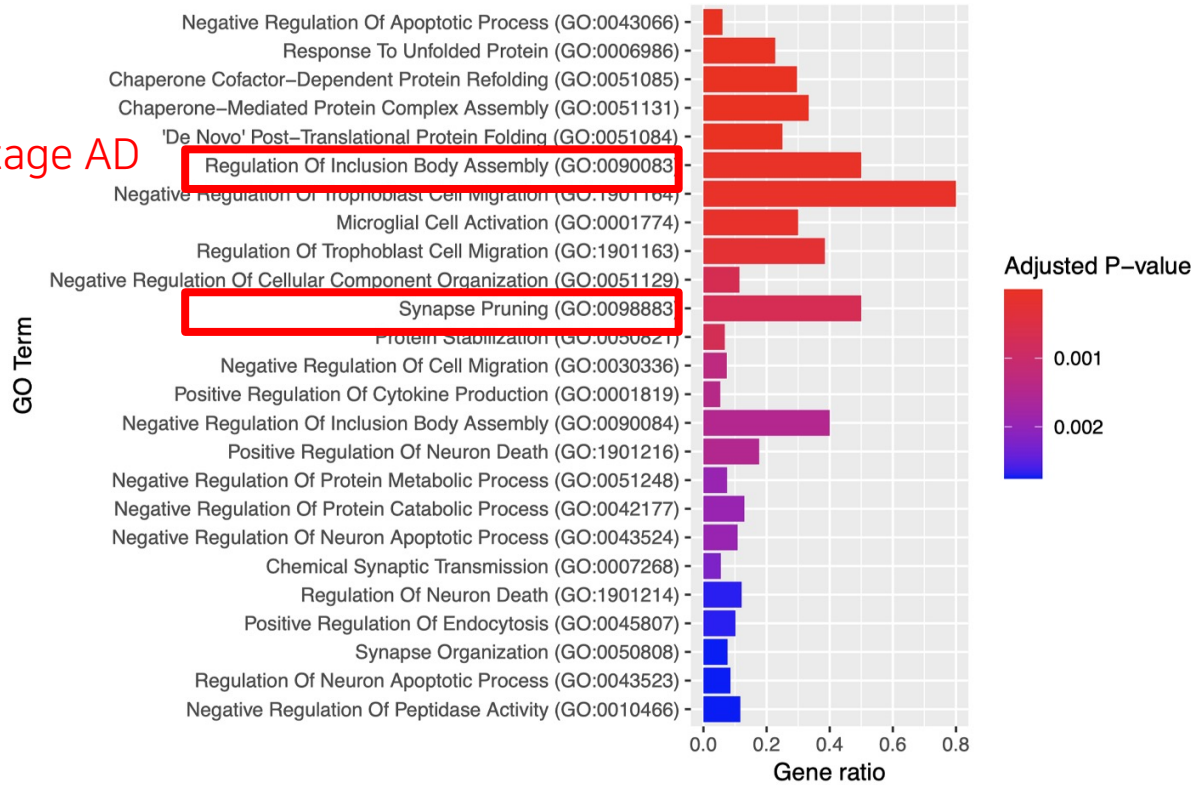
Layer 2



Early stage AD

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

Layer 5



Later stage AD

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 findings, as it correctly identified distinct early-stage (Layer 2) and late-stage (Layer 5) disease pathologies that align with known AD progression

Conclusion

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

- **Spotscape** learns **global similarity** for spatial transcriptomics beyond local neighborhoods
- **Similarity Telescope** improves representations, especially at **domain boundaries**
- **PCL + Similarity Scaling** enable robust **multi-slice integration** and **batch-effect mitigation**
- Consistent gains across **tasks**: SDI, trajectory, imputation, integration, alignment
- Validated on **diverse SRT platforms** (Visium, Stereo-seq, CosMX, Xenium)