STAGE: Deciphering Tissue Organization and Predicting Perturbation Responses through Transferable Spatial Representations

Anonymous authors
Paper under double-blind review

Abstract

Spatial transcriptomics measures gene expression while preserving tissue architecture, providing a basis for investigating tissue microenvironments and their functional responses. Learning shared representations that capture spatial organization and expression information across heterogeneous measurements is a key challenge for tissue analysis and downstream prediction. We present STAGE, a spatial transcriptomics foundation model for spatial domain analysis and expression recovery. STAGE jointly encodes gene expression and spatial geometry in cellular neighborhood graphs and uses hierarchical prototypes to organize representations at multiple resolutions. The model is pretrained on approximately 32.44 million cells spanning 18 tissue types and six spatial transcriptomics platforms. Across cross-platform, technical-replicate, and normal–tumor evaluations, both zero-shot and fine-tuned STAGE outperform six comparison methods in scenario-averaged spatial label consistency and embedding separation. Task-specific modules extend these representations to masked-cell expression recovery and spatial expression completion under known perturbation conditions. On 1,314 masked validation cells from colon Xenium, STAGE spatial reconstruction improves nonzero-expression Pearson correlation from 0.2305 for CIFM to 0.4139. On Perturb-map, conditioning CONCERT on STAGE expression representations reduces nonzero-gene MSE by 5.9% and energy distance by 5.6% relative to CONCERT. These results establish spatial representations as a shared basis for tissue organization analysis and expression reconstruction under known perturbation conditions.

1 Introduction

Spatial transcriptomics (ST) has fundamentally transformed biological research by enabling high-resolution localization of gene expression within intact tissue microenvironments (Bressan et al., 2023). By retaining spatial context, ST supports the delineation of tissue architecture, identification of spatial domains (or "niches"), and elucidation of microenvironmental organization underlying physiological functions and disease mechanisms (Rao et al., 2021). Recent work has further shown that cells of the same type can occupy distinct spatial niches and exhibit niche-specific cell–cell communication programs, underscoring the importance of explicitly modeling niche heterogeneity in spatial analysis (Qi et al., 2025).

Spatial transcriptomics platforms differ in gene coverage, spatial resolution, and measurement scale (McCombie et al., 2019; Janesick et al., 2023; Chen et al., 2015; He et al., 2022). These differences, together with batch effects, complicate the identification of shared tissue structure and the learning of comparable representations across samples (Figure 1a,c).

Spatial-domain identification is a core task in ST analysis (Zormpas et al., 2023). Louvain, Leiden, STAGATE, and GraphST provide complementary approaches to characterizing local spatial organization (Blondel et al., 2008; Traag et al., 2019; Dong & Zhang, 2022; Long et al., 2023b). A shared tissue representation should capture both local cellular neighborhoods and multi-scale organization while accommodating different gene panels.

The same spatial context can inform expression prediction as well as describe tissue organization. Perturbation-response prediction asks how altering a gene changes cellular expression. Predictions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

can prioritize targets and generate hypotheses for experimental validation when the number of possible perturbations exceeds screening capacity (Roohani et al., 2024). Responses can also depend on surrounding cells and local signals, making spatial context relevant to this task (Lin et al., 2025). CIFM connects cellular neighborhoods with expression recovery (You et al., 2025), while CONCERT models spatially dependent perturbation responses (Lin et al., 2025). These advances motivate using tissue representations to connect structural analysis with functional response prediction.

Recent foundation models such as scGPT (Cui et al., 2024) and Novae (Blampey et al., 2025) have expanded the scope of representation learning across heterogeneous transcriptomic measurements. scGPT learns from large-scale single-cell transcriptomes, while Novae introduces spatial embedding alignment across tissues. Building on these advances, STAGE couples the encoding of cellular neighborhoods with their coarse-to-fine organization during pretraining. Prototype-based objectives shape the neighborhood encoder, while the prototype vectors are jointly updated from the resulting representations, connecting local contextual learning with hierarchical organization.

We propose STAGE (Spatial Transcriptomics Analysis via Graph Embeddings with Hierarchical Prototypes), a foundation model for spatial representation learning and downstream tissue analysis (Figure 1). STAGE brings together three contributions: (1) **Joint expression–spatial encoding** uses shared gene embeddings and coordinate-aware graph encoding to represent expression and geometry within local and extended cellular neighborhoods; (2) **hierarchical prototype learning** combines balanced cross-view assignments, parent-level supervision, and confidence-weighted path consistency to organize neighborhood representations at multiple resolutions; and (3) **task-specific expression modeling** extends these representations to masked-cell recovery, spatial completion under known perturbation conditions, and population-response prediction. Across cross-platform, technical-replicate, and normal–tumor evaluations, both zero-shot and fine-tuned STAGE achieve lower scenario-averaged PAS and higher scenario-averaged ASW than six comparison methods under the specified evaluation protocols. Expression recovery and spatial completion provide additional evidence of the predictive usefulness of the learned representations.

2 Related work

2.1 Classical Spatial Clustering

Early spatial domain detection methods typically construct weighted graphs by combining gene expression and spatial proximity, followed by community detection. The Louvain algorithm, though efficient, often yields fragmented domains with poor spatial continuity. Leiden improves connectivity via a two-level node–edge refinement but remains sensitive to batch effects in multi-section integration. Graph neural networks (GNNs) effectively model gene expression and spatial structure for domain identification. STAGATE integrates expression similarity via graph attention to enable adaptive edge weighting and 3D alignment. GraphST applies contrastive learning to align local and global patterns across sections without explicit batch correction. Despite strong performance, GNN methods often rely on dataset-specific hyperparameters, limiting their cross-platform generalization.

2.2 Pretrained Foundation Models

To address the generalization limits of traditional methods, pretrained foundation models like scGPT and Novae have emerged to enable unified representation learning across platforms and tissues. scGPT, inspired by language modeling, is pretrained on large-scale single-cell data via masked-token prediction to learn transferable gene–gene and cell–gene relationships. Novae, a self-supervised graph model, introduces spatial embedding alignment loss for correcting platform discrepancies and supports zero-shot spatial domain prediction with integrated batch correction.

2.3 Expression Recovery and Perturbation Modeling

CIFM (Cellular Interaction Foundation Model) uses geometric graph neural networks and microenvironment-conditioned expression prediction to learn cellular interaction representations (You et al., 2025). CONCERT combines perturbation-attribute representations with a Gaussian-process variational autoencoder to model spatial dependencies and generate expression under specified conditions (Lin et al., 2025).

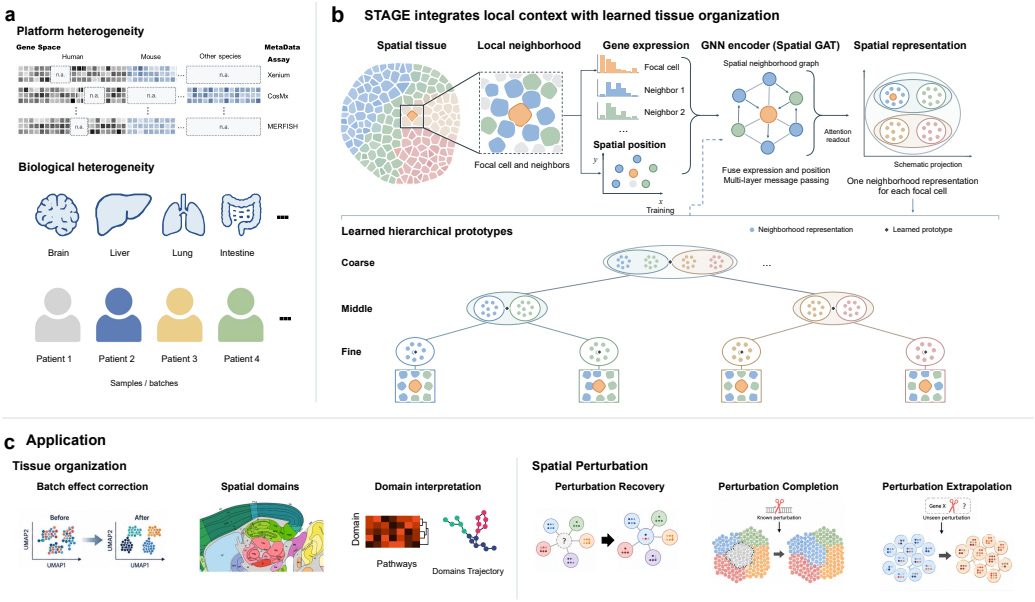

Figure 1: Overview of STAGE. **(a)** Heterogeneity in gene coverage, measurement platforms, tissues, and biological samples. **(b)** STAGE jointly encodes gene expression and spatial geometry within cellular neighborhood graphs using a graph attention network. Attention readout produces a representation of each focal cell's neighborhood, while learned hierarchical prototypes organize these representations at fine, intermediate, and coarse resolutions. **(c)** Applications include cross-batch integration, spatial-domain identification, and domain interpretation. Task-specific models further support masked-cell expression recovery, spatial expression completion under known perturbation conditions, and population-level response prediction for held-out perturbation targets.

For genetic perturbations, GEARS combines gene coexpression and Gene Ontology graphs to predict single- and multigene responses from single-cell perturbation screens (Roohani et al., 2024). STAGE learns transferable tissue representations through spatial graph encoding and hierarchical prototypes. We evaluate these representations in masked-cell recovery against CIFM, spatial completion by conditioning CONCERT on STAGE expression features, and population-response prediction for held-out targets. These tasks distinguish reconstruction under known conditions from generalization across perturbation targets.

3 Methodology

STAGE maps gene expression and cellular neighborhoods to spatial representations organized by hierarchical prototypes (Figure 2). Spatial graph encoding captures local tissue context, while prototype learning connects these representations across spatial resolutions. The learned encoder supports spatial-domain analysis and three expression-prediction tasks.

3.1 Spatial Graph Encoding

A tissue section provides normalized expression $X \in \mathbb{R}^{N \times |P|}$ for gene panel P and coordinates s_i for N cells. Each gene has a learnable identity embedding; expression-weighted aggregation over the measured panel produces a cell feature h_i^x. Sharing these gene embeddings across panels allows cells measured on different platforms to enter a common feature space.

For each center cell, we sample local and extended neighborhoods from the spatial graph. The local view captures nearby cellular interactions, whereas the extended view provides broader tissue context. Within each neighborhood, coordinates are centered and root-mean-square normalized to

obtain $\tilde{s}_i$. A spatial MLP f_s and a fusion projection f_f produce the encoder input:

$$h_i = f_f([h_i^x \,\|\, f_s(\tilde{s}_i)])\,. \tag{1}$$

A shared graph-attention encoder aggregates these features into graph representations. Squared-distance consistency between the two views ($\mathcal{L}_{\text{geo}}$) encourages stable neighborhood representations. Symmetric expression–coordinate alignment ($\mathcal{L}_{\text{align}}$) matches the two modalities from the same graph against other graphs in the batch. Graph construction, normalization, and both losses are detailed in Appendix B.

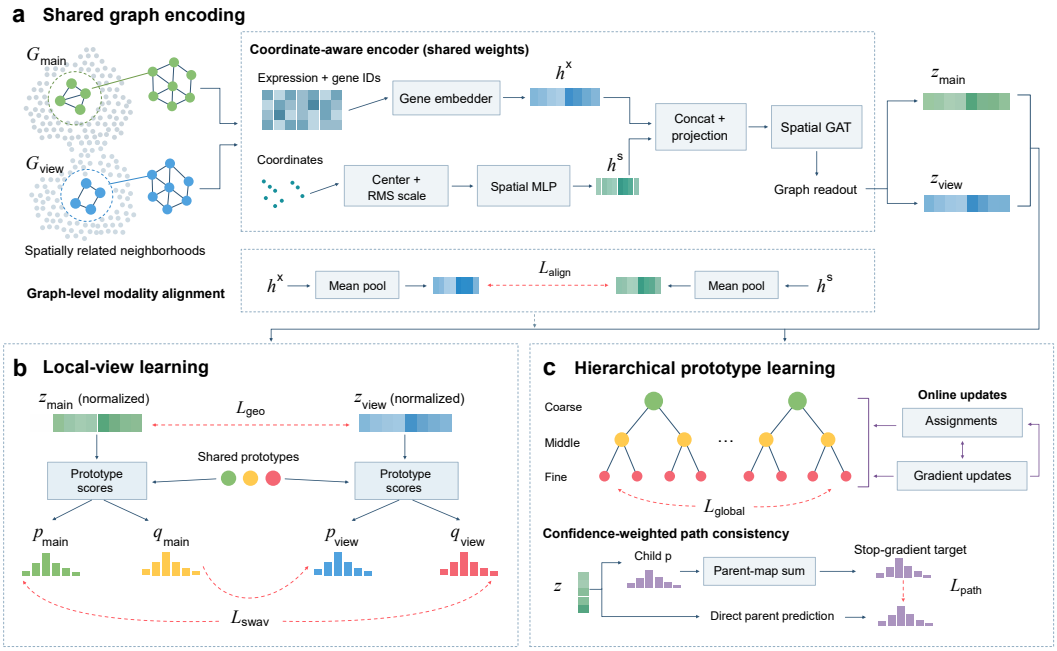

Figure 2: STAGE learning framework. **(a)** Shared graph encoding of expression and coordinates with graph-level modality alignment. **(b)** Local-view consistency and swapped prototype prediction using optimal transport. **(c)** Alternating prototype assignment and parameter updates, with confidence-weighted cross-level path consistency.

3.2 Local and Hierarchical Prototype Learning

Neighboring graph views learn compatible prototype assignments through swapped prediction (Caron et al., 2020). For embeddings z_i, prototype matrix C and temperature τ, let $p_i = \text{softmax}(Cz_i/\tau)$. Entropic optimal transport produces targets q_i with balanced assignment mass across the batch and prototypes, discouraging concentration on a few prototypes. Each view predicts its neighbor's target:

$$\mathcal{L}_{\text{swav}} = -\frac{1}{2N}\sum_{(i,j)}\sum_{k=1}^{K}\left(q_{ik}\log p_{jk} + q_{jk}\log p_{ik}\right). \tag{2}$$

Here N is the number of paired graph views in the mini-batch. This objective links local views through shared prototype identities without requiring annotated domains. The transport constraints and Sinkhorn updates are given in Appendix B.5.

To connect local patterns with broader tissue organization, prototypes form a tree from coarse roots to fine leaves, with one parent per non-root prototype. Training alternates balanced assignment and gradient-based parameter updates. The hierarchical loss $\mathcal{L}_{\text{global}}$ maps each highest-scoring child prototype to its parent and trains the corresponding parent-level prediction. Sinkhorn targets and hierarchy targets are computed separately; their precise update order is given in Appendix B.6.

We also require predictions at adjacent levels to agree. If $\pi_l(c)$ maps a child to its parent, the child probabilities induce the parent target $\widetilde{p}_i^l(u) = \sum_{c:\pi_l(c)=u} p_i^{l+1}(c)$. The path loss $\mathcal{L}_{\text{path}}$ penalizes disagreement between this target and the directly predicted parent distribution. Confidence weighting reduces the influence of uncertain assignments; gradients are stopped through the aggregated target and confidence weights. The full pretraining objective is

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{swav}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{global}}\mathcal{L}_{\text{global}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}}. \tag{3}$$

All terms apply from the first training epoch. Tree updates, confidence weighting, augmentation, and training settings are specified in Appendix B.

3.3 Expression Recovery and Perturbation-response Prediction

For **masked-cell recovery**, we hide the center-cell expression while retaining visible neighboring expression and spatial structure. A decoder reconstructs the center-cell expression from its STAGE neighborhood representation, $\widehat{x}_i = D_\phi(E_\theta(G_i^{\text{mask}}))$. The encoder and decoder are jointly optimized against the observed expression, teaching the representation to retain information needed for reconstruction.

For **spatial completion**, STAGE encodes expression in observed spatial neighborhoods. These expression representations condition the latent variables of the CONCERT expression model (Lin et al., 2025), allowing its predictions to incorporate the local tissue context. Following task-specific adaptation, the STAGE encoder remains fixed while CONCERT learns to reconstruct unobserved locations under known conditions. We denote this model as CONCERT + STAGE.

For **population response prediction**, we use a stratified mean-response predictor and add a residual conditioned on STAGE representations. Reference-cell embeddings and prototype features are pooled by batch b and cell type c into $h_{b,c}^{\text{STAGE}}$. Given reference expression $x_{b,c}^A$ and target embedding e_t, the head predicts

$$\widehat{y}_{b,c,t} = x_{b,c}^A + \mu_{b,c}^{\text{train}} + \gamma R_\phi\left(x_{b,c}^A, e_t, h_{b,c}^{\text{STAGE}}\right). \tag{4}$$

The term $\mu_{b,c}^{\text{train}}$ uses the same stratified estimator as the mean-response baseline, averaging training responses equally over targets within each batch–cell-type group. This empirical estimate is held fixed; R_ϕ learns deviations conditioned on the target, reference expression, and STAGE spatial features. The residual scale γ is selected on validation data; STAGE and target embeddings remain fixed. Task-specific losses and protocols are given in Appendix E.

4 Experiments

4.1 Experimental Setups

Data and tasks. STAGE is pretrained on 32.44 million cells spanning 18 tissues and six platform-labelled collections. All samples used for evaluation were excluded from the STAGE pretraining corpus.Spatial-domain evaluation covers 17 tissue sections in cross-platform, technical-replicate, and normal–tumor settings. We also evaluate masked-cell recovery on colon Xenium. Perturb-map profiles CRISPR-perturbed mouse lung tumors with Visium and tests spot-level expression completion under known conditions (Dhainaut et al., 2022; Lin et al., 2025). Perturb-Multi profiles spatial perturbations in mouse liver with MERFISH and tests hepatocyte population responses to held-out targets (Saunders et al., 2025). Dataset specifications are in Appendix A.

Comparisons and metrics. We compare zero-shot STAGE and four-epoch fine-tuning with Louvain, Leiden, STAGATE, GraphST, scGPT, and Novae. Domain evaluation measures distribution alignment (JSD↓), spatial continuity (FIDE↑), local label disagreement (PAS↓), and embedding separation (ASW↑). Downstream tasks use reconstruction error and expression correlations with task-specific baselines. Metric definitions and aggregation are in Appendix A; domain and downstream protocols are in Appendices C and E. Component ablations are reported in Section 4.5; additional settings and hyperparameter sensitivity analyses are provided in Appendix D.

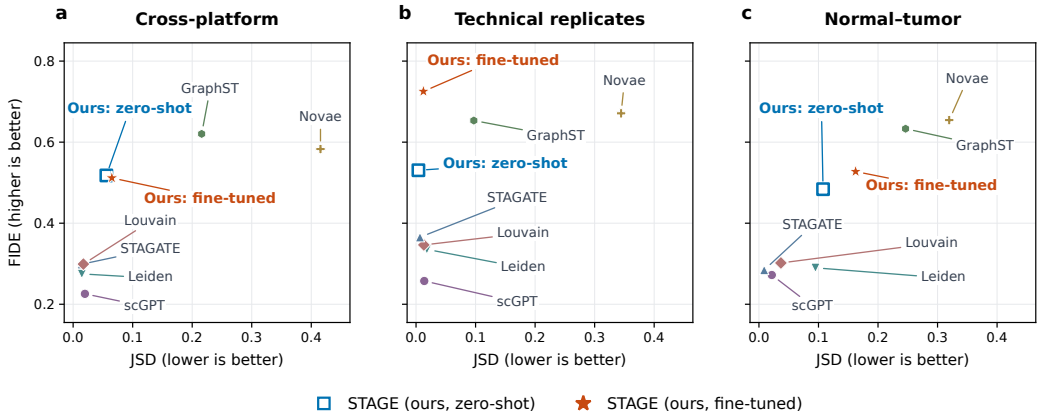

Figure 3: Spatial-domain distribution alignment (JSD) and continuity (FIDE) for (a) cross-platform pairs, (b) technical replicates, and (c) normal–tumor pairs. Points show means across 3, 3, and 4 pairs, respectively. Domain-assignment and statistical settings are defined in Appendix C.

4.2 High Cross-platform Consistency and Spatial Continuity

Across cross-platform pairs, technical replicates, and normal–tumor pairs, both zero-shot and fine-tuned STAGE achieve lower PAS and higher ASW than all six comparison methods (Table 1). These results demonstrate consistent advantages in local domain homogeneity and embedding separation under the evaluation protocol. Figure 3 complements this comparison with domain-distribution alignment and spatial continuity; complete results and protocols are in Appendix C.

For cross-platform pairs, both STAGE modes outperform all six baselines in PAS and ASW. Their shared reference prototypes provide a common domain vocabulary across the paired sections. FIDE further shows greater spatial continuity than Louvain, Leiden, STAGATE, and scGPT (Figure 3a). Representative domain maps are provided in Appendix F.

For same-tissue, same-platform technical replicates, both STAGE modes outperform all six baselines in PAS and ASW, while fine-tuned STAGE also achieves the highest FIDE among the compared methods (Figure 3b). The agreement across these measures supports spatial-domain identification across repeated measurements.

For normal and tumor tissues paired by organ, both STAGE modes again outperform all six baselines in PAS and ASW. Their FIDE is higher than Louvain, Leiden, STAGATE, and scGPT (Figure 3c). The three scenarios show that STAGE's advantages in local domain homogeneity and embedding separation extend across the evaluated measurement platforms, replicates, and tissue states. Molecular-marker interpretation is presented in Section 4.3; additional spatial-domain maps are provided in Appendix F.

Table 1: Spatial-domain homogeneity (PAS, lower is better) and embedding separation (ASW, higher is better). full scores, uncertainty estimates and assignment settings are in Appendix C.

Method	Cross-platform		Technical replicates		Normal–tumor	
	PAS↓	ASW↑	PAS↓	ASW↑	PAS↓	ASW↑
Louvain	0.7109	0.0530	0.6472	0.1267	0.7170	0.0440
Leiden	0.7176	0.0570	0.6507	0.0912	0.7343	0.0166
STAGATE	0.3947	0.0556	0.6287	0.1108	0.8163	0.0505
GraphST	0.4665	0.0666	0.2727	0.1057	0.4049	0.0592
Novae	0.3975	0.1420	0.2507	0.1127	0.3075	0.1399
scGPT	0.8491	0.0118	0.6319	0.1158	0.8390	-0.0205
STAGE (zero-shot)	0.2242	0.1945	0.1369	0.1677	0.1575	0.1942
STAGE (fine-tuned)	0.2625	0.1793	0.1515	0.1805	0.2212	0.1589

4.3 Molecular Characteristics and Spatial Organization of Tissue Domains

The spatial domains identified by STAGE provide interpretable tissue structure for downstream biological analyses. Taking the colorectal cancer sample as an example, Figure 4a shows the distribution of spatial domains within the tissue, and Figure 4b further shows their spatial adjacency.

In the gene expression analysis, we observed representative genes with similar spatial enrichment patterns, including *MZB1*, *POU2AF1*, *DERL3*, and *CD79A* (Figure 4c–f). Their co-enrichment in specific regions suggests the presence of a B cell-associated microenvironment. Combining spatial-domain identification with molecular expression patterns helps interpret tissue heterogeneity and its potential biological significance.

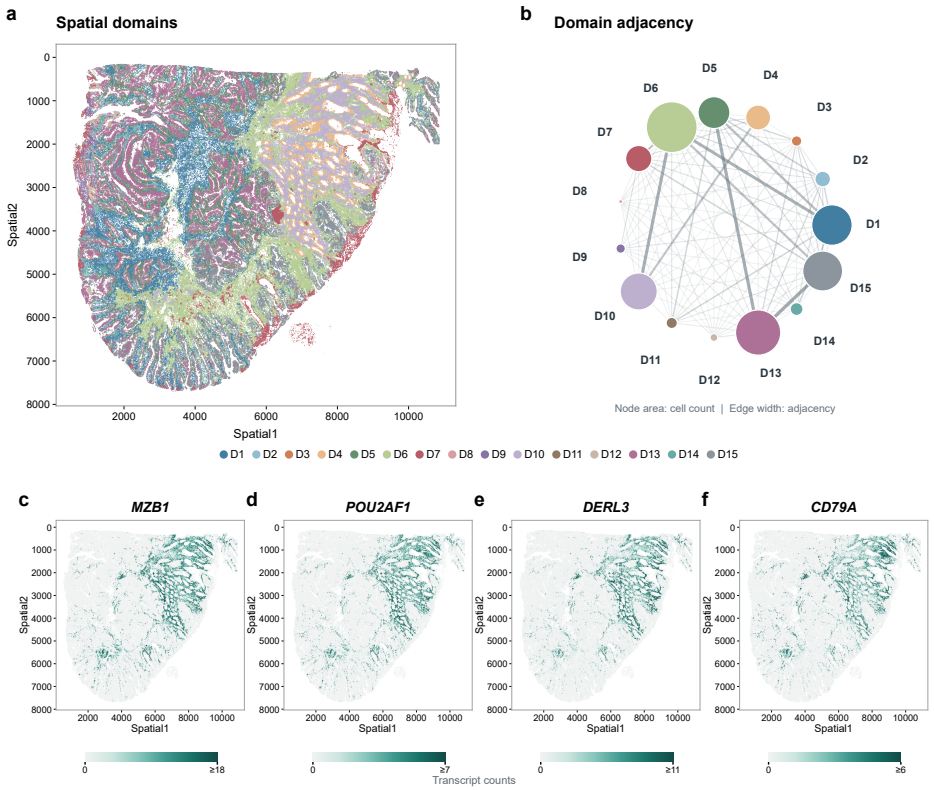

Figure 4: Spatial organization and molecular markers of colorectal cancer. (a) Fifteen spatial domains on the Xenium section. (b) Domain-adjacency graph: node area represents cell count and edge width represents spatial adjacency. (c–f) Spatial transcript counts for *MZB1*, *POU2AF1*, *DERL3*, and *CD79A*. Panels (a) and (b) share domain colors; each marker color scale saturates at its 99.5th percentile. Domain-assignment settings are given in Appendix C.1.

4.4 Spatial Expression Recovery and Perturbation-response Prediction

Spatial context improves reconstruction of masked-cell expression. On the same masked validation cells, STAGE spatial reconstruction increases nonzero-expression Pearson correlation from 0.2305 to 0.4139 and Spearman correlation from 0.4522 to 0.4903 relative to CIFM. Table 2 summarizes the downstream comparisons; detailed metrics and protocols are provided in Appendix E.

Table 2: Expression recovery and perturbation-response prediction. In (b), CONCERT + STAGE uses STAGE-encoded expression representations as latent context. In (c), STAGE denotes the complete population response head. NZ denotes observed nonzero expression; Δ denotes changes relative to matched controls. Comparisons are within each task; detailed protocols and remaining metrics are in Appendix E.

(a) Masked-cell recovery: colon Xenium validation			
Model	All MSE↓	NZ r ↑	NZ ρ ↑
CIFM	0.9829	0.2305	0.4522
STAGE (spatial)	0.8733	0.4139	0.4903
STAGE (cross-gene)	0.8304	0.2818	0.2748
(b) Perturb-map: spatial completion			
Model	NZ MSE↓	NZ r ↑	NZ ρ ↑
CONCERT	0.7506 ± 0.02640	0.8173 ± 0.0006000	0.7272 ± 0.0005000
Spatial kNN†	1.216 ± 0.1217	0.8365 ± 0.01020	0.7332 ± 0.01630
STAGE	0.7063 ± 0.003300	0.8221 ± 0.001200	0.7309 ± 0.001300
(c) Perturb-Multi: held-out targets			
Model	NZ MSE↓	r_Δ ↑	ρ_Δ ↑
Neighborhood expression	$0.003836 \pm 1.100\times10^{-5}$	0.5104 ± 0.001730	0.4641 ± 0.002690
Mean response†	0.004594 ± 0.0002200	0.4359 ± 0.01869	0.3814 ± 0.01497
STAGE	$0.003393 \pm 5.100\times10^{-7}$	$0.5843 \pm 7.100\times10^{-5}$	0.5357 ± 0.0002200

Panel (a) reports one model per method. Panels (b,c): mean $\pm$ training SD; † denotes between-section SD in (b) and between-fold SD in (c). All three sources of variation are defined in Appendix A.6.

On Perturb-map, conditioning CONCERT on STAGE expression representations reduces nonzero-gene MSE by 5.9% and energy distance by 5.6% relative to CONCERT. These gains show improved reconstruction of detected expression and better matching of condition-level expression distributions. In the additional Perturb-Multi held-target evaluation, the complete STAGE population response model reduces nonzero-gene MSE by 11.5% relative to the neighborhood expression model and improves delta Pearson correlation from 0.5104 to 0.5843. Complete comparisons are given in Appendix E.

4.5 Spatial Information and Hierarchical Consistency

We evaluate spatial information fusion and hierarchical consistency on three cross-platform tissue pairs: colorectal cancer, liver hepatocellular carcinoma, and ovarian cancer. The component definitions and evaluation settings are given in Appendix D.1.

Table 3: Spatial-information and hierarchical-consistency ablation of STAGE. Values are equal-weight means over three cross-platform tissue pairs after 60 training epochs (one trained model per configuration).

Spatial information fusion	Hierarchical consistency	JSD ↓	CNA ↑	PAS ↓	Embedding ASW ↑
No	No	0.0991	0.6220	0.1832	-0.0924
Yes	No	**0.0504**	0.8561	0.1187	0.1374
Yes	Yes	0.0517	**0.8660**	**0.1084**	**0.1557**

CNA denotes chance-corrected neighborhood agreement: $\mathrm{CNA} = \mathrm{clip}((a - \sum_k p_k^2)/(1 - \sum_k p_k^2), 0, 1)$, where a is same-domain agreement on undirected spatial six-neighbor edges and p_k is domain frequency. Embedding ASW uses the learned representations. The target number of domains is 15; at epoch 60, the base configuration has 12 domains for liver HCC and 15 for the other pairs; both augmented configurations have 15 throughout.

Table 3 shows complementary gains from the two components. Adding spatial information reduces PAS from 0.1832 to 0.1187 and increases embedding ASW from -0.0924 to 0.1374. Adding hierar-

chical consistency further reduces PAS to 0.1084 and increases embedding ASW to 0.1557; neighborhood agreement increases from 0.6220 to 0.8561 and then 0.8660. These results support the roles of spatial information and cross-level consistency in improving local label consistency and representation separation.

5 Conclusion

STAGE combines joint expression–geometry encoding on spatial graphs with hierarchical prototypes to learn transferable representations of tissue organization. Across the evaluated cross-platform, technical-replicate, and normal–tumor settings, both zero-shot and fine-tuned STAGE outperform six comparison methods in domain homogeneity and embedding separation. Task-specific modules extend these representations to masked-cell expression recovery, spatial completion under known perturbation conditions, and population-response prediction for held-out targets. STAGE connects tissue organization learning with expression prediction, providing shared representations for spatial-domain analysis, expression recovery, and perturbation-response modeling in heterogeneous spatial transcriptomic data.

References

Quentin Blampey, Hakim Benkirane, Nadège Bercovici, Kevin Mulder, Grégoire Gessain, Florent Ginhoux, Fabrice André, and Paul-Henry Cournède. Novae: a graph-based foundation model for spatial transcriptomics data. *Nature Methods*, 22(12):2539–2550, 2025. doi: 10.1038/s41592-025-02899-6.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. The dawn of spatial omics. *Science*, 381 (6657):eabq4964, 2023. doi: 10.1126/science.abq4964.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Jiang Chang, Junting Lu, Qingyi Liu, Tao Xiang, Shaosen Zhang, Yonglin Yi, Dongxu Li, Tianyuan Liu, Zeyuan Liu, Xinjie Chen, Zhenghao Dong, Cainan Li, HanZhang Yi, Siqi Yu, Luwei Huang, Fangfei Qu, Mengdi Wang, Dehe Wang, Hao Dong, Guoyu Cheng, Liang Zhu, Jiachen Li, Chenying Li, Pujie Wu, Xiaoting Xie, Andrew E. Teschendorff, Dongxin Lin, Xiaoqun Wang, and Chen Wu. Single-cell multi-stage spatial evolutional map of esophageal carcinogenesis. *Cancer Cell*, 43(3):380–397.e7, 2025. doi: 10.1016/j.ccell.2025.02.009.

Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015. doi: 10.1126/science.aaa6090.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Maxime Dhainaut, Samuel A. Rose, Guray Akturk, Aleksandra Wroblewska, Sebastian R. Nielsen, Eun Sook Park, Mark Buckup, Vladimir Roudko, Luisanna Pia, Robert Sweeney, Jessica Le Berichel, C. Matthias Wilk, Anela Bektesevic, Brian H. Lee, Nina Bhardwaj, Adeeb H. Rahman, Alessia Baccarini, Sacha Gnjatic, Dana Pe'er, Miriam Merad, and Brian D. Brown. Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell*, 185(7):1223–1239.e20, 2022. doi: 10.1016/j.cell.2022.02.015.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1):1739, 2022. doi: 10.1038/s41467-022-29439-6.

Bin Duan, Shaoqi Chen, Xiaojie Cheng, and Qi Liu. Multi-slice spatial transcriptome domain analysis with SpaDo. *Genome Biology*, 25(1):73, 2024. doi: 10.1186/s13059-024-03213-x.

Shanshan He, Ruchir Bhatt, Carl Brown, Emily A. Brown, Derek L. Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G. Garrison, Gary Geiss, Mark T. Gregory, Margaret L. Hoang, Rustem Khafizov, Emily E. Killingbeck, Dae Kim, Tae Kyung Kim, Youngmi Kim, Andrew Klock, Mithra Korukonda, Alecksandr Kutchma, Zachary R. Lewis, Yan Liang, Jeffrey S. Nelson, Giang T. Ong, Evan P. Perillo, Joseph C. Phan, Tien Phan-Everson, Erin Piazza, Tushar Rane, Zachary Reitz, Michael Rhodes, Alyssa Rosenbloom, David Ross, Hiromi Sato, Aster W. Wardhani, Corey A. Williams-Wietzikoski, Lidan Wu, and Joseph M. Beechem. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology*, 40(12):1794–1806, 2022. doi: 10.1038/s41587-022-01483-z.

Yunfei Hu, Manfei Xie, Yikang Li, Mingxing Rao, Wenjun Shen, Can Luo, Haoran Qin, Jihoon Baek, and Xin Maizie Zhou. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biology*, 25(1):212, 2024. doi: 10.1186/s13059-024-03361-0.

Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, 10x Development Teams, and Sarah E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1): 8353, 2023. doi: 10.1038/s41467-023-43458-x.

Zheng Li and Xiang Zhou. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biology*, 23(1): 168, 2022. doi: 10.1186/s13059-022-02734-7.

Xiang Lin, Zhenglun Kong, Soumya Ghosh, Manolis Kellis, and Marinka Zitnik. CONCERT predicts niche-aware perturbation responses in spatial transcriptomics. *bioRxiv*, 2025. doi: 10.1101/2025.11.08.686890. Preprint.

Brian Long, Jeremy Miller, and The SpaceTx Consortium. SpaceTx: A roadmap for benchmarking spatial transcriptomics exploration of the brain. *arXiv*, 2023a. doi: 10.48550/arXiv.2301.08436. Preprint.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1):1155, 2023b. doi: 10.1038/s41467-023-36796-3.

W. Richard McCombie, John D. McPherson, and Elaine R. Mardis. Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine*, 9(11):a036798, 2019. doi: 10.1101/cshperspect.a036798.

Jeffrey R. Moffitt, Dhananjay Bambah-Mukku, Stephen W. Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D. Perez, Nimrod D. Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018. doi: 10.1126/science.aau5324.

Juntian Qi, Zhengchao Luo, Chuan-Yun Li, Jinzhuo Wang, and Wanqiu Ding. Interpretable niche-based cell–cell communication inference using multi-view graph neural networks. *Nature Computational Science*, 5(6):444–455, 2025. doi: 10.1038/s43588-025-00809-6.

Jingyang Qian, Xin Shao, Hudong Bao, Yin Fang, Wenbo Guo, Chengyu Li, Anyao Li, Hua Hua, and Xiaohui Fan. Identification and characterization of cell niches in tissue from spatial omics data at single-cell resolution. *Nature Communications*, 16(1):1693, 2025. doi: 10.1038/s41467-025-57029-9.

Anjali Rao, Dalia Barkley, Gustavo S. França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021. doi: 10.1038/s41586-021-03634-9.

Pengfei Ren, Rui Zhang, Yunfeng Wang, Peng Zhang, Ce Luo, Suyan Wang, Xiaohong Li, Zongxu Zhang, Yanping Zhao, Yufeng He, Haorui Zhang, Yufeng Li, Zhidong Gao, Xiuping Zhang, Yahui Zhao, Zhihua Liu, Yuanguang Meng, Zhe Zhang, and Zexian Zeng. Systematic benchmarking of high-throughput subcellular spatial transcriptomics platforms across human tumors. *Nature Communications*, 16(1):9232, 2025. doi: 10.1038/s41467-025-64292-3.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Reuben A. Saunders, William E. Allen, Xingjie Pan, Jaspreet Sandhu, Jiaqi Lu, Thomas K. Lau, Karina Smolyar, Zuri A. Sullivan, Catherine Dulac, Jonathan S. Weissman, and Xiaowei Zhuang. Perturb-Multimodal: A platform for pooled genetic screens with imaging and sequencing in intact mammalian tissue. *Cell*, 188(17):4790–4809.e22, 2025. doi: 10.1016/j.cell.2025.05.022.

Qiang Shi, Yihan Chen, Yang Li, Shishang Qin, Yu Yang, Yang Gao, Linnan Zhu, Dongfang Wang, and Zemin Zhang. Cross-tissue multicellular coordination and its rewiring in cancer. *Nature*, 643 (8071):529–538, 2025. doi: 10.1038/s41586-025-09053-4.

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, Mojtaba Bahrami, and Fabian J. Theis. Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, 22(12):2525–2538, 2025. doi: 10.1038/s41592-025-02814-z.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

Marco Varrone, Daniele Tavernari, Albert Santamaria-Martínez, Logan A. Walsh, and Giovanni Ciriello. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics*, 56(1):74–84, 2024. doi: 10.1038/s41588-023-01588-4.

Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018. doi: 10.1126/science.aat5691.

Yuning You, Zitong Jerry Wang, Kevin Fleisher, Rex Liu, and Matt Thomson. Building foundation models to characterize cellular interactions via geometric self-supervised learning on spatial genomics. In *ICLR Workshop on Machine Learning for Genomics Explorations (MLGenX)*, 2025. URL `https://openreview.net/forum?id=hy12KXLT0K`.

Zhiyuan Yuan. MENDER: fast and scalable tissue structure identification in spatial omics data. *Nature Communications*, 15(1):207, 2024. doi: 10.1038/s41467-023-44367-9.

Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, 2024. doi: 10.1038/s41592-024-02215-8.

Eleftherios Zormpas, Rachel Queen, Alexis Comber, and Simon J. Cockell. Mapping the transcriptome: Realizing the full potential of spatial data analysis. *Cell*, 186(26):5677–5689, 2023. doi: 10.1016/j.cell.2023.11.003.

Appendix

Additional material is organized by topic: data and evaluation protocols (Appendix A), method and training details (Appendix B), spatial-domain comparisons (Appendix C), component ablation and hyperparameter sensitivity (Appendix D), expression and perturbation evaluation (Appendix E), qualitative analyses (Appendix F), and availability statements (Appendix G).

A Datasets and Evaluation Protocols

A.1 Pretraining Corpus

The pretraining corpus contains 109 input samples and 32,444,865 cells across 18 tissue categories. Table 4 gives exact counts by the six platform annotations. MERSCOPE, a commercial implementation of MERFISH, and MERFISH are listed as separate source collections. The tissue categories are bone, bone marrow, brain, breast, colon, femur, heart, kidney, liver, lung, lymph node, ovary, pancreas, prostate, skin, tonsil, uterus, and whole mouse.

Table 4: Pretraining inputs by platform annotation. Feature ranges describe stored expression matrices before model-specific gene filtering.

Platform	Samples	Cells	Stored features
Xenium	66	19,738,644	248–5,002
MERSCOPE	18	10,196,835	500
CosMx	14	2,472,622	980–21,569
MERFISH	5	28,317	155
BARISTAseq	3	5,257	54–62
STARmap	3	3,190	144–160
Total	109	32,444,865	

The principal source collections are the 10x Genomics Xenium releases, Vizgen FFPE human immuno-oncology collection, and CosMx public releases. The five MERFISH sections are mouse hypothalamic sections at bregma -0.04, -0.09, -0.14, -0.19, and -0.24 mm, originally measured by Moffitt et al. (2018), with processed data obtained from BASS (Li & Zhou, 2022). The small brain collections contain STARmap BZ5/BZ9/BZ14 controls (Wang et al., 2018) and BARISTAseq slices 1–3 from the SpaceTx BARISTASeq collection (Long et al., 2023a). The accompanying `pretraining_inventory.csv` lists every sample identifier, platform annotation, stored cell and feature counts, source resource, and spatial-domain evaluation membership. Collection-level URLs are distinguished from sample-specific releases.

A.2 Evaluation Dataset Specifications

The three spatial-domain scenarios comprise 17 distinct sections and 10 pairs. Cross-platform pairs use the colorectal, hepatocellular, and ovarian carcinoma samples measured by CosMx and Xenium (Ren et al., 2025). Technical replication uses the CosMx FFPE lung-cancer series Lung5 Rep1, Rep2, and Rep3, forming pairs 1–2, 1–3, and 2–3.

Normal–tumor pairs use the Xenium kidney preview, Xenium lung preview, and CosMx normal-liver/hepatocellular-carcinoma releases. The pancreatic pair combines the non-diseased pancreas section with the pancreatic adenocarcinoma section measured with multimodal cell segmentation. Pairing is defined by tissue type.

Table 5: Spatial-domain evaluation inputs. Counts describe cells and expression features before model-specific gene filtering. N/T denotes non-diseased/tumor. Full sample identifiers and pair membership are supplied in `evaluation_pairs.csv`.

Sample group	Platform/order	Cells	Features
Colorectal cancer	CosMx / Xenium	292,371 / 405,927	6,175 / 4,998
Hepatocellular carcinoma	CosMx / Xenium	237,030 / 271,687	6,175 / 4,699
Ovarian cancer	CosMx / Xenium	289,272 / 410,193	6,175 / 4,998
Lung5 Rep1	CosMx	100,292	980
Lung5 Rep2	CosMx	106,660	980
Lung5 Rep3	CosMx	100,264	980
Kidney N/T	Xenium	97,560 / 56,510	377 / 377
Lung N/T	Xenium	295,883 / 531,165	392 / 392
Liver N/T	CosMx	340,706 / 463,972	1,197 / 1,197
Pancreas N/T	Xenium	103,901 / 140,702	377 / 377

Evaluation design. Zero-shot STAGE uses pretrained weights without pair-specific gradient updates; fine-tuned STAGE receives four epochs of pair-specific updates. Both use the paired sections for unsupervised domain assignment. Masked-cell recovery is evaluated on 1,314 validation center cells. Perturb-map holds out spatial regions, and Perturb-Multi holds out targets during response-head fitting. Task-specific inputs, upstream adaptation, and target exposure are specified in Appendix E.

A.3 Expression and Perturbation Datasets

Perturb-map. This dataset measures the spatial consequences of CRISPR knockouts in a mouse model of lung adenocarcinoma (Dhainaut et al., 2022) (GEO: GSE193460). We use four 10x Visium tissue sections distributed with CONCERT (Lin et al., 2025), containing 6,363 spatial spots and an author-selected common panel of 1,053 genes. Each spot provides expression, spatial coordinates, and a perturbation or tissue-region annotation. The task is to reconstruct expression at held-out spatial locations under known conditions, using observed regions of the same section. The prediction unit is a Visium spot.

Perturb-Multi. Perturb-Multimodal (Perturb-Multi) combines pooled CRISPR perturbations with spatial RNA and protein imaging and single-cell RNA sequencing in mouse liver (Saunders et al., 2025). Our analysis uses the 209-gene spatial RNA measurements obtained by multiplexed error-robust fluorescence in situ hybridization (MERFISH), together with cell positions and perturbation identities. We evaluate directly perturbed hepatocytes (Hep), averaging expression within each batch–cell-type–target group to define one population response unit. The model predicts this population's post-perturbation expression for held-out targets. Thus, this experiment assesses population response generalization, whereas Perturb-map assesses spatial completion under known conditions.

A.4 Data Preprocessing

Genes are retained if detected in at least 50 cells, and cells without spatial coordinates are excluded. Spatial graphs use Delaunay triangulation with a radius of $100\,\mu\mathrm{m}$. Local neighborhoods include cells within $25\,\mu\mathrm{m}$; extended neighborhoods cover distances from 25 to $50\,\mu\mathrm{m}$.

A.5 Evaluation Metrics

JSD. Let $n_{s,k}$ count cells with output domain index k in section s, and $P_s(k) = n_{s,k}/\sum_j n_{s,j}$. The index set is the union of observed domains, with zero frequency for absent domains. STAGE uses shared reference prototypes; baselines retain their output cluster indices. The equally weighted Jensen–Shannon divergence (Duan et al., 2024) is

$$\mathrm{JSD}(P_1, P_2) = H\left(\frac{P_1 + P_2}{2}\right) - \frac{H(P_1) + H(P_2)}{2}, \quad H(P) = -\sum_k P(k)\log_2(P(k) + 10^{-8}). \tag{5}$$

The F1-score of inter-domain edges (FIDE score) is used to quantify the spatial domain continuity. Specifically, let $C = (C_i)_{1 \le i \le N}$ represent the categorical spatial domain predictions for N cells of a slide. The FIDE score is defined as:

$$\text{FIDE}(C, A) = \text{F1-score}\left((C_i, C_j)_{i,j \text{ s.t. } A_{ij}>0}\right)$$

where A_{ij} is positive when cells i and j are graph neighbors. We compute the macro-averaged F1 score between the domain labels at the two endpoints of every nonzero edge, averaging over observed domain labels. The adjacency A is the physical-neighbor graph shared by the evaluation methods. Slide scores are averaged within pairs before scenario-level aggregation. Intuitively, for an edge $i \leftrightarrow j$, the edge is considered an inter-domain edge if $C_i \neq C_j$.

Percentage of Abnormal Spots (PAS) (Yuan et al., 2024) quantifies local disagreement between spatial-domain labels. We select up to 100,000 cells per section without replacement using random seed 42, construct Euclidean spatial neighborhoods within this selected set, and exclude each center cell from its own neighbors. A cell is abnormal if at least six of its ten nearest neighbors have a different domain label. PAS is the fraction of abnormal cells, reported on the $[0, 1]$ scale; lower values indicate greater local spatial homogeneity.

ASW. For cell i, let a_i be its mean Euclidean embedding distance to other cells in the same domain, and b_i its mean distance to cells in the nearest other domain. For an evaluated cell set S, average silhouette width (Hu et al., 2024) is

$$\text{ASW}(S) = \frac{1}{|S|} \sum_{i \in S} \frac{b_i - a_i}{\max(a_i, b_i)}.$$

STAGE uses the domain-centroid transformation

$$\widetilde{z}_{s,i} = z_{s,i} - \mu_{s,k} + \mu_{r(k),k}, \qquad c_{s,i} = k, \quad r(k) = \arg\max_s n_{s,k}, \tag{6}$$

where $z_{s,i}$ is the cell embedding, $c_{s,i}$ is its fixed domain label, and $\mu_{s,k}$ is the mean of valid embeddings for domain k in section s. Centroids are computed before sampling. STAGE uses $\widetilde{z}$ and baselines use their respective embeddings to calculate a_i, b_i. Each section is sampled without replacement up to 5,000 cells for seeds 42–44; the three scores are averaged. Section and pair aggregation follows Appendix A.6.

NMI(Hu et al., 2024). Normalized Mutual Information (NMI) quantifies the similarity between two clusterings and ranges from 0 to 1. A higher NMI value, closer to 1, indicates better agreement between the two clustering assignments. NMI has been widely used to evaluate clustering accuracy in single-cell analyses and has also been applied to assess the performance of spatial domain identification algorithms. We use NMI to measure the agreement between the ground truth domain labels and the tissue structure identification results produced by each method.

Suppose P denotes the predicted spatial domain clustering and T denotes the ground truth clustering labels. Their entropies are $H(P)$ and $H(T)$, respectively, and the mutual information is $MI(P, T)$. NMI is then defined as:

$$\text{NMI} = \frac{MI(P, T)}{\sqrt{H(P)H(T)}}.$$

ARI(Hu et al., 2024). The Adjusted Rand Index (ARI) measures the similarity between two clustering assignments while accounting for chance agreement. ARI has an upper bound of 1 and can take negative values; larger values indicate greater agreement. To compute the ARI, a contingency table is constructed by comparing the true domain labels with the predicted tissue structure identification results for each spot.

The contingency table includes four entries: (1) TP: the number of spot pairs assigned to the same cluster in both the true and predicted clustering; (2) TN: the number of pairs assigned to different clusters in both; (3) FN: the number of pairs in the same true cluster but in different predicted clusters; and (4) FP: the number of pairs in different true clusters but in the same predicted cluster.

ARI is computed as:

$$\text{ARI} = \frac{TP + TN - E}{TP + TN + FP + FN - E},$$

where E is the expected index value under random clustering. The expected value is computed as:

$$E = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{TP + TN + FP + FN}.$$

A.6 Statistical Reporting

Spatial-domain scores are averaged over sections within each pair and then equally over three cross-platform, three technical-replicate, or four normal–tumor pairs. Fine-tuned STAGE means and sample standard deviations (SDs) in Table 7 use training seeds 0, 1, and 2; those in Table 8 describe between-pair variation for one trained model.

For Novae PAS and ASW, SDs describe clustering variation with seeds 0–2 on fixed representations. Cross-platform STAGATE PAS and ASW vary both training and clustering seeds together (42–44), whereas GraphST varies training seeds 42–44 with clustering fixed. Zero-shot STAGE and the other fixed baseline PAS/ASW predictions report between-pair SDs where available. Baseline FIDE point estimates match Figure 3; corresponding replicate-level uncertainty estimates are not available. Table symbols distinguish these sources of variation. ASW computation and cell subsampling are defined in Appendix A.5.

For Perturb-map and Perturb-Multi, neural-model means and sample SDs use independent training seeds 17, 29, and 43. Perturb-map scores are averaged within spatial blocks, conditions, and sections; deterministic spatial kNN reports between-section SD. Perturb-Multi scores give equal weight to each target; the deterministic mean-response baseline reports SD across held-out folds (five target folds). Each SD is interpreted according to its stated source of variation.

B Method and Training Details

B.1 Multi-scale Spatial Graph Construction

We define spatial neighborhoods and edge weights as follows:

Neighborhood Definition: For each cell v, we define multi-scale neighborhoods:

$$\text{Local: } \mathcal{N}_L(v) = \{u : d_{\text{spatial}}(u, v) \leq r_{\text{local}}\} \tag{7}$$

$$\text{Extended: } \mathcal{N}_V(v) = \{u : r_{\text{local}} < d_{\text{spatial}}(u, v) \leq r_{\text{view}}\} \tag{8}$$

where $d_{\text{spatial}}(u, v)$ represents Euclidean distance between spatial coordinates.

Edge Weight Computation: Spatial relationships are encoded through distance-based weights:

$$w_{uv} = \exp\left(-\frac{d_{\text{spatial}}(u, v)^2}{2\sigma^2}\right) \tag{9}$$

where σ is the spatial bandwidth parameter adjusted per platform resolution.

B.2 Gene Panel Alignment

Cross-platform gene panel differences are addressed through learnable gene embeddings $\{v_g\}_{g=1}^{G}$ where $v_g \in \mathbb{R}^E$. For a cell with expression x_i and panel P, the embedding is:

$$\text{embed}(x_i, P) = \frac{\sum_{j \in P} x_{ij} v_j}{\sqrt{\sum_{j \in P} \|v_j\|^2}} \tag{10}$$

The L2 normalization ensures consistent weighting across different panel sizes, analogous to PCA where components are trainable gene embeddings. This design allows the model to learn meaningful gene programs that generalize across platforms while adapting to panel-specific characteristics.

B.3 Spatial Coordinate Fusion and Consistency

Tissue geometry provides information about cellular organization that complements gene expression. STAGE jointly encodes spatial coordinates and expression within each neighborhood graph. For a

neighborhood graph g, we center and normalize coordinates by their root-mean-square distance from the center:

$$\tilde{s}_i = \frac{s_i - \mu_g}{\sqrt{|g|^{-1} \sum_{j \in g} \|s_j - \mu_g\|_2^2 + \varepsilon}}, \qquad i \in g. \tag{11}$$

An MLP f_s maps $\tilde{s}_i$ to a spatial embedding, which is fused with the expression embedding h_i^x by a learnable projection $f_f([h_i^x \parallel f_s(\tilde{s}_i)])$. The projection is initialized to preserve the expression branch, so spatial information is introduced without discarding the pretrained expression representation. During training, the spatial branch is independently dropped for each graph with probability p_{sp}.

The two sampled graph views provide the consistency signal. Let z_i^m and z_i^v be the normalized graph embeddings from the main and view neighborhoods. We impose a geometric consistency term:

$$\mathcal{L}_{\text{geo}} = \frac{1}{2N} \sum_{i=1}^{N} \|z_i^m - z_i^v\|_2^2 . \tag{12}$$

In parallel, we align the graph-level expression and spatial modalities. With e_g^x and e_g^s denoting mean-pooled expression and spatial embeddings for graph g, respectively, the symmetric alignment term is

$$\mathcal{L}_{\text{align}} = \frac{1}{2} \left[\text{CE}\big(\text{softmax}(E^x E^{s\top}/\tau), I\big) + \text{CE}\big(\text{softmax}(E^s E^{x\top}/\tau), I\big) \right], \tag{13}$$

where E^x and E^s stack the graph-level embeddings and I identifies corresponding graphs. Expression recovery uses task-specific fine-tuning, described in Section 3.3.

B.4 Network Architecture

Gene embeddings have dimension 512, with 60,697 vocabulary entries. The coordinate MLP is $2 \to 512 \to 512$ with ELU between its linear layers and L2-normalized output. Concatenated expression and coordinate embeddings are projected from 1,024 to 512 dimensions. The graph encoder contains 10 GATv2 layers (Brody et al., 2022), hidden width 128, 16 attention heads, ELU activations, and one-dimensional edge features; its output width is 64. The 128 hidden channels are distributed across the attention heads. Attention pooling applies a learned $64 \to 1$ gate and a $64 \to 64$ value projection, normalizes gates within each neighborhood graph, and sums the weighted values. Prototype banks and their active counts are specified in Appendix B.6.

B.5 Balanced Prototype Assignment

Given a batch of embeddings $Z \in \mathbb{R}^{B \times d}$ and prototype matrix $C \in \mathbb{R}^{K \times d}$, we solve the entropic optimal transport problem:

$$Q^* = \arg\min_{Q \in \mathcal{Q}} \langle Q, -ZC^T \rangle - \epsilon H(Q) \tag{14}$$

where $H(Q) = -\sum_{i,j} Q_{ij} \log Q_{ij}$ is the Shannon entropy, $\epsilon > 0$ is the regularization parameter, and $\mathcal{Q}$ is the transportation polytope, $\mathcal{Q} = \left\{ Q \in \mathbb{R}_+^{B \times K} : Q\mathbf{1}_K = \frac{1}{B}\mathbf{1}_B, Q^T\mathbf{1}_B = \frac{1}{K}\mathbf{1}_K \right\}$.

This formulation ensures that: (1) each sample contributes equal total assignment mass, (2) each prototype receives equal assignment mass, and (3) the assignment is smooth due to entropy regularization.

The normalized targets $q_i = BQ_i^*$ and predictions $p_i = \text{softmax}(Cz_i/\tau)$ enter the swapped loss in Equation 2.

B.5.1 Optimal Transport Solver

The optimal transport problem is efficiently solved using the Sinkhorn-Knopp algorithm, which alternates between row and column normalizations:

$$Q^{(t+1)} = \text{Diag}(u^{(t)}) \exp\left(\frac{ZC^T}{\epsilon}\right) \text{Diag}(v^{(t)}) \tag{15}$$

where $u^{(t)}$ and $v^{(t)}$ are renormalization vectors computed iteratively to enforce the marginal constraints. The marginal constraints are approximated by iterative normalization.

B.6 Hierarchical Assignments and Path Consistency

B.6.1 Prototype hierarchy.

We write the hierarchy in coarse-to-fine order as $\{\mathcal{C}^l\}_{l=1}^{L_p}$, with $L_p = 3$ and a unique parent $\pi_l(c)$ for each level-$(l+1)$ child. Each level has capacity for 512 trainable 64-dimensional prototype vectors. The number of active prototypes is data-dependent: during bottom-up initialization, active children are grouped into $\max(2, \lfloor K_{\text{child}}/2 \rfloor)$ parents. The active count is bounded by the capacity of each level. The 15 domains used for evaluation are a separate hierarchy cut.

B.6.2 Alternating assignment and parameter updates.

For each mini-batch, we encode both graph views and normalize prototype vectors. Similarities to the selected prototypes, concatenated across levels, define the transport problem in Appendix B.5. Three Sinkhorn column/row normalization iterations produce targets for each view; gradients are stopped through these targets. The symmetric swapped loss predicts the opposite view's targets.

The hierarchy supplies an additional parent-classification target. For a child-level distribution p_i^{l+1}, we take its highest-scoring active prototype and map it to its parent:

$$c_i^{l+1} = \arg\max_c p_i^{l+1}(c), \qquad u_i^l = \pi_l(c_i^{l+1}),$$
$$\mathcal{L}_{\text{global}} = -\frac{1}{2N(L_p-1)} \sum_{a\in\{m,v\}} \sum_i \sum_{l=1}^{L_p-1} \log p_{i,a}^l(u_{i,a}^l). \tag{16}$$

Here each p^l is a temperature-scaled softmax over the active prototypes at that level. The discrete child choice and parent map are treated as targets, while gradients update the predicted parent scores. Adam updates the encoder and prototype parameters using the combined objective. The procedure alternates assignment and gradient updates, with cross-level agreement encouraged by the parent loss and the soft path-consistency term below.

B.6.3 Cross-level Path Consistency.

The hierarchy associates each child prototype at level $l+1$ with a parent at level l, where level 1 is the coarsest level. To align assignments with this structure, let p_i^l be the soft assignment of graph embedding h_i to level-l prototypes and let $\pi_l(c)$ map a level-$(l+1)$ child to its parent. We aggregate child probabilities into the parent space:

$$\widetilde{p}_i^l(u) = \sum_{c:\pi_l(c)=u} p_i^{l+1}(c). \tag{17}$$

The independently predicted parent distribution is regularized toward this aggregated target for both graph views $a \in \{m, v\}$:

$$\mathcal{L}_{\text{path}} = \frac{1}{2N\sum_{l\in\mathcal{E}}\alpha_l} \sum_{a\in\{m,v\}} \sum_{i=1}^{N} \sum_{l\in\mathcal{E}} \frac{w_{i,a}^l \alpha_l}{K_l} \left\| p_{i,a}^l - \text{sg}(\widetilde{p}_{i,a}^l) \right\|_2^2, \tag{18}$$

where $w_{i,a}^l = [\max_u p_{i,a}^l(u) \max_u \widetilde{p}_{i,a}^l(u)]^\eta$ weights assignments by confidence with exponent $\eta > 0$, $K_l = 512$ is the parent-level capacity over which the squared differences are averaged (inactive probabilities are zero), and sg stops gradients through the aggregated target. Gradients are also stopped through the confidence weights. The set $\mathcal{E}$ contains the regularized parent–child transitions, with weights $\alpha_l = \rho^{L_p-l-1}$ with decay factor $\rho \in (0, 1]$ under our coarse-to-fine notation. This encourages the parent assignment to agree with the combined probabilities of its children.

B.6.4 Hierarchical Prototype Initialization

Prototype vectors are initialized with Kaiming uniform initialization and L2 normalization. The finest-level vectors are updated from nearest-prototype assignments using moving-average cell-group means. Active finest-level prototypes have accumulated at least two assignments; at least two vectors are retained if fewer satisfy this criterion. K-means then groups the active vectors to initialize the

next coarser level, and the procedure repeats upward. Cluster membership defines each child's parent. This procedure initializes the prototype tree; Sinkhorn supplies assignment targets during subsequent training. Downstream reference-based domain clustering is defined in Appendix C.1.

B.7 Biologically-motivated Data Augmentation

To enhance model robustness against technical variations commonly observed in spatial transcriptomics, we apply two biologically relevant augmentations:

Pseudo Batch Effect Simulation: We introduce artificial batch effects to reduce sensitivity to technical noise. For each cell i, we sample additive noise $a \sim \text{Exponential}(\lambda)^{|P|}$ and multiplicative factors $s \sim \mathcal{N}(0, \sigma^2 I_{|P|})$, then update expression as:

$$x_i^{(\text{noise})} = a + x_i \odot (1 + s) \tag{19}$$

This augmentation simulates the variability introduced by different experimental conditions, library preparation protocols, and sequencing depths.

Gene Panel Subsetting: We randomly subset the gene panel according to ratio $\gamma \in (0, 1)$, selecting $\lfloor \gamma |P| \rfloor$ genes to create $P' \subset P$. This augmentation simulates the effect of different gene panels across technologies, as if multiple platforms generated the data or the panel was updated during a longitudinal study.

These augmentations help the model generalize across varying experimental conditions and dataset characteristics, crucial for building a foundation model applicable to diverse spatial transcriptomics studies.

B.8 Pretraining Settings

The table below summarizes the spatial-domain pretraining settings.

Table 6: STAGE pretraining hyperparameters.

Parameter	**Value**
Learning rate	0.0001
Batch size	256
Hierarchical levels (L_p)	3
Prototype capacity per level	512
Temperature (τ)	0.1
Regularization (ϵ)	0.05
Global loss weight (λ_{global})	0.1
Geometry consistency weight (λ_{geo})	0.05
Expression–spatial alignment weight (λ_{align})	0.05
Path consistency weight (λ_{path})	0.0015
Spatial-branch dropout (p_{sp})	0.05
Path confidence exponent (η) / level decay (ρ)	3 / 0.7
Hierarchical learning starts	Epoch 1
Total training epochs	60
Pretraining epochs for domain-transfer experiments	30

Path consistency is applied to the transition between the intermediate and coarsest levels of the three-level hierarchy. Confidence weights use $\eta = 3$, and level weights use $\alpha_l = \rho^{L_p - l - 1}$ with $\rho = 0.7$. Hierarchical learning and path consistency are included from the first training epoch. The spatial-domain transfer experiments use the model pretrained for 30 epochs, followed by four epochs of pair-specific fine-tuning.

B.9 STAGE Training Algorithm

Algorithm 1 summarizes STAGE training.

Algorithm 1 STAGE Training Algorithm

Require: Spatial transcriptomics datasets $\{G_i\}$, hyperparameters $\{K, L_p, \lambda, \tau, \epsilon\}$
Ensure: Pre-trained model parameters θ, hierarchical prototypes $\mathcal{T}$

```
Initialize gene embeddings {v_g}, network parameters θ_0
Initialize normalized prototype vectors; construct parent maps bottom-up from active child prototypes
for epoch = 1 to E_total do
    for each mini-batch {G_i} do
        Generate augmented views {G'_i} via panel subsetting and noise
        Compute expression/spatial embeddings and fuse them with centered, RMS-scaled coordinates
        Compute graph embeddings from the main and view neighborhoods and stack them as rows of Z
        Compute stop-gradient targets for both views with three Sinkhorn iterations on selected prototype scores
        Compute the symmetric swapped-prediction loss in Equation 2
        Compute L_geo between the two graph views and symmetric L_align between expression and spatial graph embeddings
        Map the highest-scoring active child to its parent; compute parent-classification loss for both views
        Compute confidence-weighted cross-level path loss L_path by aggregating child assignments into parent space
        Update θ, T via gradient descent on L_swav + λ_global L_global + λ_geo L_geo + λ_align L_align + λ_path L_path
    end for
end for
```

C Spatial-Domain Evaluation

C.1 Domain Assignment and Adaptation

We evaluate two approaches to domain assignment, both starting from the same model pretrained for 30 epochs. Fine-tuning uses four epochs in 32-bit precision, and the largest section in each tissue pair serves as the reference. Zero-shot evaluation uses pretrained representations without updating the encoder. Domain-centroid correction is applied in both modes.

Prototype assignment. Prototypes and their hierarchy are initialized from the reference section and adapted with learning rate 10^{-3}. Cells are assigned to domains using this adapted hierarchy. The resulting partition contains 18 domains for the cross-platform liver pair, 16 for technical-replicate pairs 1–2 and 1–3, and 15 for the remaining pairs. The zero-shot control uses reference-initialized prototypes with the pretrained representations.

Reference-based clustering. Fine-tuning uses learning rate 10^{-4}. Slide-level assignments are balanced over active prototypes and averaged across hierarchy levels and both graph views. Domain identification fits 512 prototypes by spherical k-means to all valid reference cells, then merges them into 15 domains using cosine distance and average linkage, with clustering seed 0. The zero-shot control applies the same clustering procedure to pretrained representations. Cross-platform input panels are matched between zero-shot and fine-tuned evaluation; technical-replicate and normal–tumor evaluations use mode-specific panels.

Figure 3 uses prototype assignment. Table 7 and the spatial-domain maps use reference-based clustering. Table 8 compares both approaches with fine-tuning seed 0. Its reference-based clustering results use the same final-epoch representations and scoring procedure as seed 0 in Table 7; this seed also defines the spatial maps.

Evaluation controls. Comparisons use the same tissue sections and scenario definitions. FIDE evaluates labels on a coordinate-space physical-neighbor graph, independently of model-specific

representation-learning graphs. PAS uses a common ten-neighbor disagreement criterion, and ASW uses common Euclidean-distance and cell-subsampling procedures. Domain-assignment settings and uncertainty sources are specified separately to distinguish training, clustering, and tissue-pair variation.

C.2 Quantitative Comparisons

Table 7: Spatial continuity (FIDE), domain homogeneity (PAS), and embedding separation (ASW). Values are means with sample SD where available. Baseline FIDE point estimates match Figure 3. Symbols indicate between-pair (†), clustering (§), joint training and clustering (∗), or training variation with fixed clustering (⋄). Fine-tuned STAGE reports training SD. Values marked ‡ do not include independent training or clustering replicate estimates. Statistical definitions are given in Appendix A.6.

Methods	**FIDE↑**	**PAS↓**	**ASW↑**
Cross-Platform Paired Dataset			
Louvain	$0.2991^{\ddagger}$	$0.7109\ (\pm\ 0.1753)^{\dagger}$	$0.0530\ (\pm\ 0.0439)^{\dagger}$
Leiden	$0.2749^{\ddagger}$	$0.7176\ (\pm\ 0.1979)^{\dagger}$	$0.0570\ (\pm\ 0.0292)^{\dagger}$
STAGATE	$0.2988^{\ddagger}$	$0.3947\ (\pm\ 0.0061)^{*}$	$0.0556\ (\pm\ 0.0055)^{*}$
GraphST	$0.6204^{\ddagger}$	$0.4665\ (\pm\ 0.0054)^{\diamond}$	$0.0666\ (\pm\ 0.0025)^{\diamond}$
Novae	$0.5829^{\ddagger}$	$0.3975\ (\pm\ 0.0266)^{\S}$	$0.1420\ (\pm\ 0.0098)^{\S}$
scGPT	$0.2257^{\ddagger}$	$0.8491\ (\pm\ 0.2233)^{\dagger}$	$0.0118\ (\pm\ 0.1011)^{\dagger}$
STAGE (zero-shot)	$0.5239\ (\pm\ 0.0633)^{\dagger}$	$0.2242\ (\pm\ 0.0892)^{\dagger}$	$0.1945\ (\pm\ 0.0242)^{\dagger}$
STAGE (fine-tuned)	$0.5711\ (\pm\ 0.0014)$	$0.2625\ (\pm\ 0.0120)$	$0.1793\ (\pm\ 0.0150)$
Same-Tissue, Same-Platform Replicates			
Louvain	$0.3464^{\ddagger}$	$0.6472\ (\pm\ 0.0068)^{\dagger}$	$0.1267\ (\pm\ 0.0085)^{\dagger}$
Leiden	$0.3354^{\ddagger}$	$0.6507\ (\pm\ 0.0243)^{\ddagger}$	$0.0912\ (\pm\ 0.0360)^{\ddagger}$
STAGATE	$0.3654^{\ddagger}$	$0.6287\ (\pm\ 0.0231)^{\dagger}$	$0.1108\ (\pm\ 0.0159)^{\dagger}$
GraphST	$0.6533^{\ddagger}$	$0.2727\ (\pm\ 0.0100)^{\dagger}$	$0.1057\ (\pm\ 0.0132)^{\dagger}$
Novae	$0.6711^{\ddagger}$	$0.2507\ (\pm\ 0.0050)^{\S}$	$0.1127\ (\pm\ 0.0041)^{\S}$
scGPT	$0.2574^{\ddagger}$	$0.6319\ (\pm\ 0.2511)^{\dagger}$	$0.1158\ (\pm\ 0.1233)^{\dagger}$
STAGE (zero-shot)	$0.5246\ (\pm\ 0.0248)^{\dagger}$	$0.1369\ (\pm\ 0.0103)^{\dagger}$	$0.1677\ (\pm\ 0.0067)^{\dagger}$
STAGE (fine-tuned)	$0.6523\ (\pm\ 0.0081)$	$0.1515\ (\pm\ 0.0063)$	$0.1805\ (\pm\ 0.0121)$
Paired Normal and Tumor Tissues from the Same Organ			
Louvain	$0.3019^{\ddagger}$	$0.7170\ (\pm\ 0.1237)^{\dagger}$	$0.0440\ (\pm\ 0.0507)^{\dagger}$
Leiden	$0.2899^{\ddagger}$	$0.7343\ (\pm\ 0.2635)^{\ddagger}$	$0.0166\ (\pm\ 0.0648)^{\ddagger}$
STAGATE	$0.2845^{\ddagger}$	$0.8163\ (\pm\ 0.0970)^{\dagger}$	$0.0505\ (\pm\ 0.0245)^{\dagger}$
GraphST	$0.6331^{\ddagger}$	$0.4049\ (\pm\ 0.1745)^{\dagger}$	$0.0592\ (\pm\ 0.0321)^{\dagger}$
Novae	$0.6544^{\ddagger}$	$0.3075\ (\pm\ 0.0227)^{\S}$	$0.1399\ (\pm\ 0.0006)^{\S}$
scGPT	$0.2722^{\ddagger}$	$0.8390\ (\pm\ 0.2158)^{\dagger}$	$-0.0205\ (\pm\ 0.0307)^{\dagger}$
STAGE (zero-shot)	$0.4917\ (\pm\ 0.0804)^{\dagger}$	$0.1575\ (\pm\ 0.0504)^{\dagger}$	$0.1942\ (\pm\ 0.0357)^{\dagger}$
STAGE (fine-tuned)	$0.5681\ (\pm\ 0.0125)$	$0.2212\ (\pm\ 0.0137)$	$0.1589\ (\pm\ 0.0365)$

Table 8: Spatial-domain performance with prototype assignment and reference-based clustering. Values are means ± between-pair sample SD. Domain-assignment and statistical settings are defined in Appendix C.

Scenario / mode	JSD↓	FIDE↑	PAS↓	ASW↑
Prototype assignment				
Cross-platform / zero-shot	0.0558 (± 0.0271)	0.5178 (± 0.0689)	0.2158 (± 0.0879)	0.1916 (± 0.0315)
Cross-platform / fine-tuned	0.0650 (± 0.0526)	0.5117 (± 0.2279)	0.2493 (± 0.0575)	0.0635 (± 0.1690)
Technical replicates / zero-shot	0.0040 (± 0.0028)	0.5307 (± 0.0254)	0.1381 (± 0.0105)	0.1457 (± 0.0301)
Technical replicates / fine-tuned	0.0124 (± 0.0121)	0.7252 (± 0.0204)	0.1758 (± 0.0246)	0.1462 (± 0.0315)
Normal–tumor / zero-shot	0.1077 (± 0.0301)	0.4840 (± 0.0699)	0.1439 (± 0.0135)	0.1788 (± 0.0425)
Normal–tumor / fine-tuned	0.1624 (± 0.0606)	0.5275 (± 0.0603)	0.3427 (± 0.1014)	0.0755 (± 0.1208)
Reference-based clustering				
Cross-platform / zero-shot	0.0542 (± 0.0228)	0.5239 (± 0.0633)	0.2242 (± 0.0892)	0.1945 (± 0.0242)
Cross-platform / fine-tuned	0.0513 (± 0.0219)	0.5698 (± 0.0880)	0.2722 (± 0.0312)	0.1733 (± 0.0299)
Technical replicates / zero-shot	0.0049 (± 0.0027)	0.5246 (± 0.0248)	0.1369 (± 0.0103)	0.1677 (± 0.0067)
Technical replicates / fine-tuned	0.0059 (± 0.0036)	0.6591 (± 0.0310)	0.1552 (± 0.0146)	0.1942 (± 0.0539)
Normal–tumor / zero-shot	0.0955 (± 0.0389)	0.4917 (± 0.0804)	0.1575 (± 0.0504)	0.1942 (± 0.0357)
Normal–tumor / fine-tuned	0.1489 (± 0.0499)	0.5793 (± 0.0644)	0.2078 (± 0.0256)	0.1696 (± 0.0530)

Continuity and domain separation. The two domain-assignment approaches assess the effects of tissue-pair adaptation on domain continuity, local homogeneity, and embedding separation. Table 8 reports the complete comparison.

C.3 Baseline Training Settings

The following settings support the cross-platform comparisons in Table 7.

STAGATE. Training uses 1,000 epochs, Adam with learning rate 10^{-3} and weight decay 10^{-4}, and hidden dimensions 512 and 30. Cross-platform sections use a ten-neighbor spatial graph. Counts are normalized to 10^4 per cell and log-transformed before selecting up to 3,000 highly variable genes. Domains are identified by Louvain clustering on a 15-neighbor graph of the learned representations, with 15 target domains. Cross-platform training and clustering share seeds 42, 43, and 44.

GraphST. Training uses 600 epochs, Adam with learning rate 10^{-3} and no weight decay, 64 hidden dimensions, and a symmetric three-neighbor spatial graph with self-loops. Up to 3,000 highly variable genes are selected from counts before normalization to 10^4 per cell, log transformation, and scaling without centering, with values clipped at 10. Training combines expression reconstruction and contrastive learning with shuffled cell features. The final reconstruction is normalized to unit Euclidean norm. Domains are identified using 20 principal components (seed 42) and an EEE Gaussian mixture with 15 components (seed 2020). Training seeds are 42, 43, and 44 for cross-platform sections.

D Component Ablation and Hyperparameter Sensitivity

D.1 Spatial Information and Hierarchical Consistency

We evaluate spatial information fusion and hierarchical consistency on three cross-platform tissue pairs: colorectal cancer, liver hepatocellular carcinoma, and ovarian cancer. Spatial information fusion combines coordinate encoding with expression features and the constraints $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{align}}$; hierarchical consistency adds the cross-level path constraint $\mathcal{L}_{\text{path}}$.

The quantitative results are reported in Table 3 and discussed in Section 4.5.

D.2 Hyperparameter Sensitivity

Figure 5 compares hierarchy depth $L \in \{1, 2, 3, 4, 5\}$ and batch size $B \in \{64, 128, 256, 512, 1024\}$. The default settings $L = 3$ and $B = 256$ balance spatial continuity and distribution alignment.

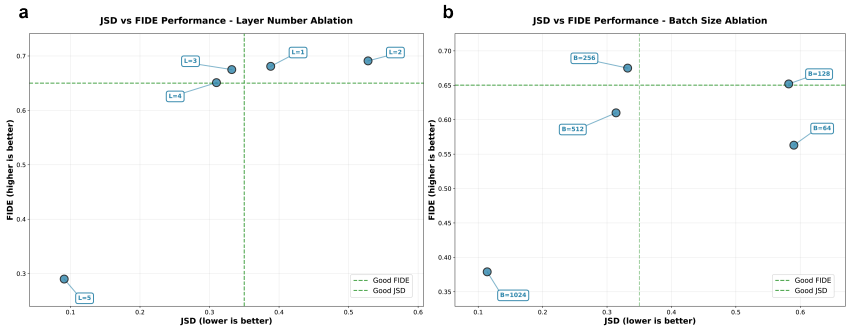

Figure 5: Sensitivity of spatial-domain identification to (a) hierarchy depth and (b) batch size.

Table 9 further evaluates the prototype-selection threshold θ, temperature τ, Sinkhorn regularization ϵ, global-loss weight λ, and prototype count K, with four values for each parameter. FIDE measures spatial continuity and JSD measures distribution alignment. The default settings yield FIDE of 0.675 and JSD of 0.332. Removing the hierarchical-loss constraint ($\lambda = 0$) increases JSD to 0.365. Increasing K to 1024 raises FIDE to 0.678 but also increases JSD to 0.345, illustrating the trade-off between the two metrics.

Table 9: Quantitative ablation study of STAGE hyperparameters. Each parameter is evaluated with four variations. Default configurations are shown in **bold**. Results demonstrate that the default settings achieve the best balance between spatial coherence (high FIDE) and batch correction (low JSD).

Hyperparameter	Value	FIDE (↑)	JSD (↓)	Remark
Threshold (θ)	0.90	0.648	0.341	Loose selection introduces noise
	0.95	0.662	0.338	
	0.99	**0.675**	**0.332**	**Optimal trade-off**
	0.995	0.668	0.334	Misses rare spatial niches
Temperature (τ)	0.05	0.665	0.336	Hard assignments, slightly unstable
	0.1	**0.675**	**0.332**	**Standard setting**
	0.2	0.655	0.320	Smooth distribution
	0.5	0.590	0.295	Over-smoothed; loss of detail
Sinkhorn (ϵ)	0.03	0.670	0.335	
	0.05	**0.675**	**0.332**	**Optimal transport balance**
	0.08	0.645	0.325	
	0.10	0.615	0.315	Trivial equipartition
Global Weight (λ)	0.0	0.662	0.365	Poor alignment without hierarchy
	0.05	0.670	0.348	
	0.1	**0.675**	**0.332**	**Best alignment–continuity balance**
	0.5	0.671	0.328	Strong constraint
Prototypes (K)	128	0.615	0.305	Under-clustering; low resolution
	256	0.652	0.318	
	512	**0.675**	**0.332**	**Sufficient capacity**
	1024	0.678	0.345	Overfits batch artifacts

E Expression Recovery and Perturbation Evaluation

E.1 Masked-cell expression recovery

To enable STAGE to recover missing expression, we mask the center-cell expression in a local spatial graph and encode the remaining visible expression together with the neighborhood structure. A decoder maps the resulting spatial representation to the center-cell expression:

$$\widehat{x}_i = D_\phi\big(E_\theta(G_i^{\text{mask}})\big)\,, \tag{20}$$

where G_i^{mask} denotes the masked graph, E_θ is the STAGE encoder, and D_ϕ is the expression decoder. The observed center-cell expression serves as the reconstruction target for optimizing the encoder and decoder, allowing STAGE to learn expression recovery from spatial context.

Table 2 compares CIFM (You et al., 2025) with two STAGE reconstruction methods on the same 1,314 validation center cells. Spatial reconstruction jointly optimizes STAGE spatial representation learning and center-cell reconstruction; cross-gene reconstruction uses shared expression patterns across genes in neighboring cells. Metrics use gene-wise standardized expression. NZ Pearson correlation r and Spearman correlation ρ assess reconstruction of detected expression and are computed over flattened cell–gene entries with nonzero original observed expression.

Spatial reconstruction increases nonzero-expression Pearson correlation from 0.2305 to 0.4139 and Spearman correlation from 0.4522 to 0.4903, improving both numerical agreement and rank correlation for detected expression.

Recovery decoder architecture. The spatial reconstruction model uses the 10-layer, 16-head encoder described in Appendix B.4. Masked centers are excluded from visible-expression encoding and represented by a learned 64-dimensional vector when restored to the decoding graph. A five-layer GATv2 decoder (Brody et al., 2022) has total hidden width 128, 16 heads, scalar edge features, and 64-dimensional outputs. A $64 \to 64 \to 512$ MLP with ELU between the linear layers maps each decoded node to gene-embedding space; inner products with the evaluated genes' 512-dimensional embeddings yield expression predictions. Coordinate embeddings provide spatial information to both encoding and decoding. All 599 evaluated output genes are retained for scoring; their standardized targets and nonzero-entry mask are shared across the compared recovery methods.

For cross-gene reconstruction, each one-hop neighbor contributes two 8-dimensional gene-program summaries (presence-weighted and abundance-weighted), a 32-dimensional projection of its 64-dimensional node embedding, nonzero support density, and spatial affinity. This 50-dimensional vector passes through a $50 \to 32 \to 32$ GELU network and attention pooling. A separate $8 \to 32$ gene projection is combined with the pooled context and their elementwise product; a $96 \to 32 \to 1$ GELU network supplies the gene-conditioned score. The accompanying scalar context calibration network is $12 \to 32 \to 32 \to 1$, including an 8-dimensional gene embedding in its input. These dimensions specify the cross-gene model separately from the spatial decoder above.

E.2 Spatial completion on Perturb-map

STAGE encodes the observed spatial neighborhoods to condition the latent variables of the CONCERT GP-VAE/LORD model (Lin et al., 2025), which reconstructs unobserved locations under known perturbation conditions.

Data sources are described in Appendix A.3. The comparison includes four tissue sections, 1,265 held-out Visium spots, and an author-selected common 1,053-gene output panel. Each neural model is trained for at most 300 epochs per section, with validation-based model selection. Seed and aggregation settings are defined in Appendix A.6. CONCERT + STAGE denotes CONCERT conditioned on STAGE-encoded expression representations of spatial neighborhoods, which are fixed after task-specific fine-tuning. CONCERT and within-condition spatial kNN serve as the expression-completion baselines. The common output panel contains 1,053 genes, including 947 covered by the STAGE input vocabulary. Validation and test expression is masked when constructing spatial context; neighborhood statistics are estimated from the training regions. Expression accuracy is evaluated over observed nonzero entries and all output genes, together with condition-level energy distance.

Table 10: Perturb-map expression completion evaluated over observed nonzero genes.

Model	MSE↓	MAE↓	$r \uparrow$	$\rho \uparrow$	$R^2 \uparrow$
CONCERT	0.7506 (±0.0264)	0.7033 (±0.0123)	0.8173 (±0.0006)	0.7272 (±0.0005)	0.2603 (±0.0258)
Within-condition spatial kNN†	1.2164 (±0.1217)	0.9158 (±0.0475)	0.8365 (±0.0102)	0.7332 (±0.0163)	−0.2279 (±0.1764)
CONCERT + STAGE	0.7063 (±0.0033)	0.6831 (±0.0014)	0.8221 (±0.0012)	0.7309 (±0.0013)	0.3010 (±0.0028)

Mean ± training SD; † denotes between-section SD.

Table 11: Perturb-map expression completion evaluated over all output genes.

Model	MSE↓	MAE↓	$r \uparrow$	$\rho \uparrow$	$R^2 \uparrow$
CONCERT	0.8912 (±0.0085)	0.7708 (±0.0062)	0.7675 (±0.0012)	0.6045 (±0.0009)	0.4929 (±0.0048)
Within-condition spatial kNN†	0.6883 (±0.0468)	0.5968 (±0.0211)	0.7885 (±0.0124)	0.5964 (±0.0167)	0.6100 (±0.0236)
CONCERT + STAGE	0.8832 (±0.0057)	0.7696 (±0.0031)	0.7728 (±0.0010)	0.6084 (±0.0004)	0.4983 (±0.0030)

Mean ± training SD; † denotes between-section SD.

Table 12: Condition-level energy distance on the complete Perturb-map output panel.

Model	Energy distance↓
CONCERT	21.3208 (±0.3928)
CONCERT + STAGE	20.1364 (±0.2341)

Mean ± training SD.

Nonzero-expression evaluation directly assesses reconstruction of detected signals. Within this scope, CONCERT + STAGE reduces MSE by 41.9% relative to spatial kNN and by 5.9% relative to CONCERT. It also reduces condition-level energy distance by 5.6% relative to CONCERT, indicating improved distribution matching. Evaluation over all output entries is reported separately in Table 11.

E.3 Population response prediction on Perturb-Multi

We use the spatial RNA measurements from the Perturb-Multimodal mouse liver study (Saunders et al., 2025). A response unit is the mean expression of directly perturbed hepatocytes sharing a batch, cell type, and target. The held-target protocol comprises five folds and 1,093 population response units spanning 202 targets. Evaluation uses the genes with nonzero post-perturbation population means among the 209 measured genes. Statistical aggregation follows Appendix A.6.

Across targets held out from response-head fitting, STAGE achieves MSE 0.0033934 and delta Pearson correlation 0.58426, compared with 0.0038360 and 0.51035 for the neighborhood expression model. This comparison evaluates the complete predictor combining reference expression, target embeddings, and STAGE context.

Reference populations and predictor inputs. Author-labelled control cells are partitioned into disjoint reference, training-contrast, and scoring subsets (A, B, and C) within each batch and cell type using seed 20260912. For hepatocytes, eligible controls have no called knockout or ambiguous source within two graph hops. Population A supplies reference expression and spatial context; B defines training contrasts, which are converted back to post-perturbation means for fitting; C supplies the matched control mean for evaluating predicted and observed expression changes. Context graphs exclude B, C, and response cells, while retaining eligible control and uncalled neighbours. Uncalled neighbours have unassigned perturbation status. Reference and response means are computed from separate cell populations.

The frozen STAGE encoder was previously adapted on eight TLS adaptation samples (PM1–PM8). It receives 195 of the 209 measured genes, while all 209 remain prediction outputs. Cell embeddings and prototype probabilities are separately reduced to 32 dimensions using SVD fitted on training-stratum reference-A centres, concatenated, and pooled within each batch–cell-type stratum. The target representation e_t is a frozen 128-dimensional Gene Ontology graph embedding pretrained on public perturbation data. Case-normalized gene-symbol matching provides representations for 183 of 202 targets; the remaining 19 receive zero vectors and a missingness indicator. Of these 202 targets, 90 occurred as training conditions and 18 as validation conditions in the public pretraining data. Held-target folds are defined at the spatial-response fitting stage.

Five target groups define held-target folds; each group is held out in turn, the next group is used for validation, and the others for fitting. Only the response head is optimized; the encoder and target representations remain fixed. Model selection uses validation target-averaged MSE, and scoring controls C are excluded from fitting.

The neighborhood expression model predicts responses from neighboring-cell expression. The mean-response baseline adds a training-derived, stratified mean response to the reference expression. STAGE uses the same mean-response estimator and adds the learned residual in Equation 4.

STAGE population response head. The mean-response baseline and STAGE share the following empirical estimator; STAGE additionally learns a conditional residual. Here $x^A_{b,c}$ is the mean expres-

sion of reference population A. The shared response is estimated within each batch–cell-type group by averaging equally over available training targets:

$$\mu_{b,c}^{\text{train}} = \frac{1}{|\mathcal{T}_{b,c}^{\text{train}}|} \sum_{u \in \mathcal{T}_{b,c}^{\text{train}}} \left(\overline{y}_{b,c,u}^{\text{train}} - x_{b,c}^{A} \right). \tag{21}$$

Both predictors use the same training-derived mean-response estimate, including target weighting and fallback for unseen groups. This training-derived statistic is fixed during residual optimization. For each training example, its own perturbation target is excluded when estimating the shared response. Validation and test predictions use training-derived statistics; a group absent from the training data uses the target-weighted global mean response. The residual predictor uses reference expression, the target representation, and pooled STAGE features to model target- and context-dependent deviations, minimizing target-averaged MSE for the resulting expression prediction. The residual scale γ is selected by validation performance from $\{0, 0.1, 0.25, 0.5, 0.75, 1\}$. Delta correlations compare predicted and observed expression changes relative to the same matched scoring-control population C, independently of the reference population A used as model input.

Conditional residual architecture. The response head has two branches. The first concatenates the 209-gene reference expression, the 128-dimensional target representation, and its availability indicator, then applies a $338 \to 64 \to 16$ MLP. The second combines the 64-dimensional pooled STAGE context with a missingness indicator and applies a $65 \to 64 \to 16$ MLP. Both use GELU after the hidden layer. For branch outputs $a, b \in \mathbb{R}^{16}$, the 209-gene residual is

$$\delta_\theta = d + Ua + V(a \odot b) + Wb, \tag{22}$$

where $d \in \mathbb{R}^{209}$ and $U, V, W \in \mathbb{R}^{209 \times 16}$ are learned. This residual is combined with the empirical mean response using the validation-selected scale γ. Response-head fitting optimizes these parameters with the STAGE encoder and target embedding fixed.

Table 13: Perturb-Multi population response prediction for held-out targets.

Model	MSE↓	$r \uparrow$	$\rho \uparrow$	$r_\Delta \uparrow$	$\rho_\Delta \uparrow$
Neighborhood expression model	0.0038360 ($\pm 1.1\times10^{-5}$)	0.98849 ($\pm 4.0\times10^{-5}$)	0.92843 ($\pm 6.7\times10^{-5}$)	0.51035 ($\pm$0.00173)	0.46408 ($\pm$0.00269)
Mean response†	0.0045941 ($\pm$0.00022)	0.98602 ($\pm$0.00063)	0.91360 ($\pm$0.00387)	0.43594 ($\pm$0.01869)	0.38141 ($\pm$0.01497)
STAGE	0.0033934 ($\pm 5.1\times10^{-7}$)	0.98979 ($\pm 1.0\times10^{-6}$)	0.93652 ($\pm 2.7\times10^{-5}$)	0.58426 ($\pm 7.1\times10^{-5}$)	0.53568 ($\pm$0.00022)

Mean $\pm$ training SD; † denotes between-fold SD.

F Qualitative Spatial and Molecular Analyses

F.1 Cross-Platform Domain Maps

Figure 6 compares spatial-domain maps in the matched colorectal cancer sections. Figure 13 also presents a comparison on expert-annotated MERFISH data, illustrating the recovery of anatomical spatial organization.

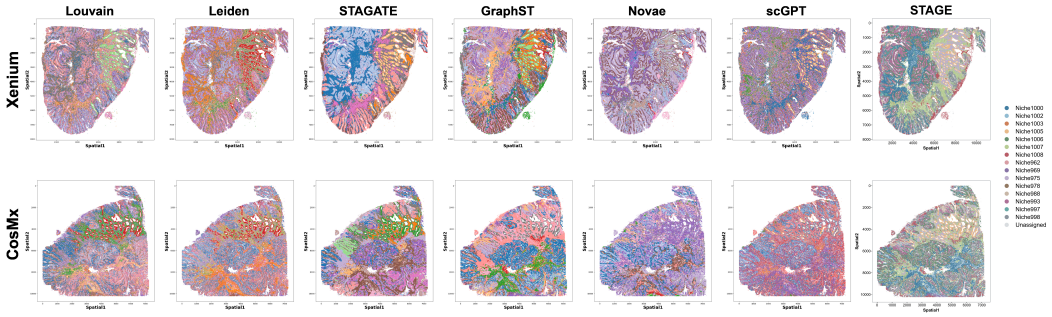

Figure 6: Spatial domains in matched colorectal cancer samples profiled by Xenium and CosMx. Colors distinguish inferred domains. STAGE settings are defined in Appendix C.1.

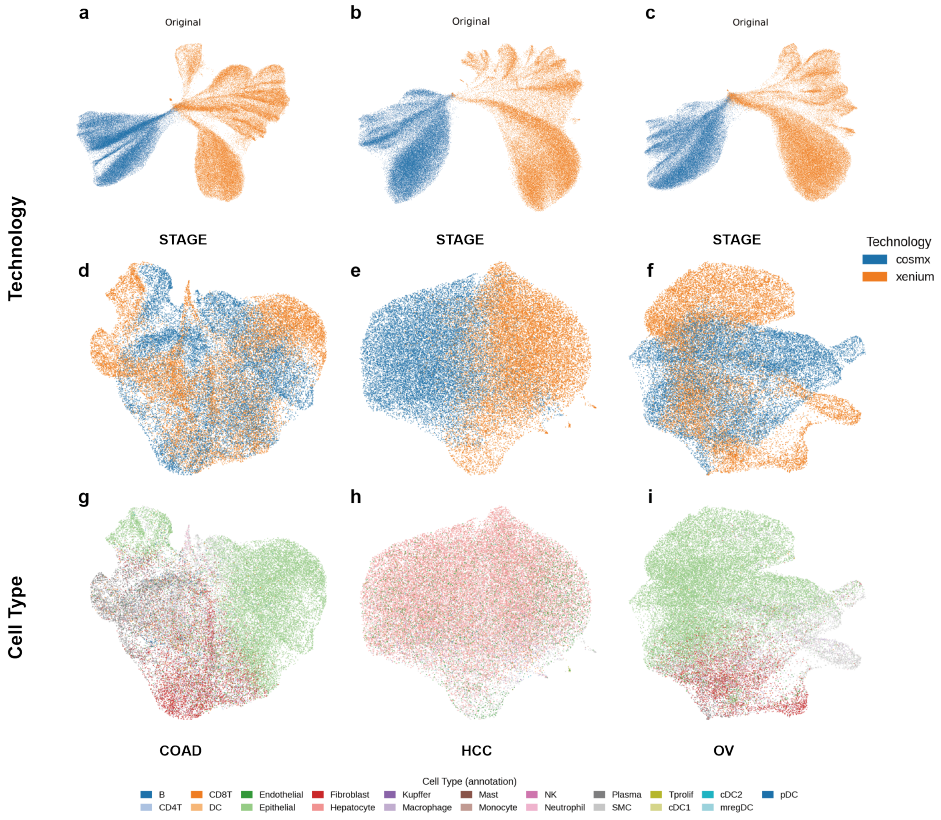

Figure 7: Cross-platform embedding alignment in paired colorectal, liver, and ovarian cancer samples measured by CosMx and Xenium. Each point represents a cell, colored by platform (blue, CosMx; orange, Xenium). Panels (a–c) show embeddings before correction, (d–f) after correction, and (g–i) the corresponding cell-type annotations.

F.2 Technical Replicates

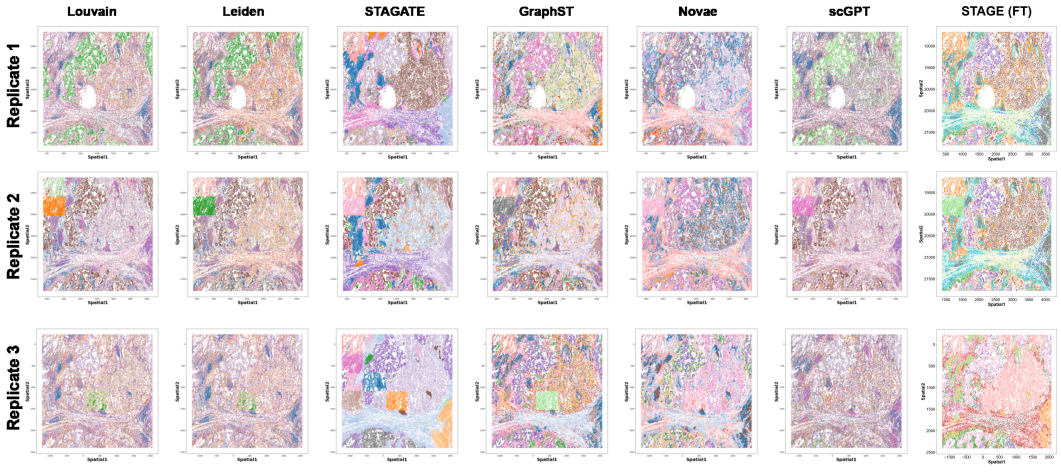

Figure 8: Spatial domains in three technical replicates of a lung tissue sample. Replicates 1 and 2 are from pair 1–2, and replicate 3 is from pair 1–3. Axes indicate tissue coordinates. STAGE settings are defined in Appendix C.1.

F.3 Shared and Disease-Specific Spatial Domains

Previous studies have shown that tumor and matched normal tissues often share spatial domains associated with conserved developmental structures, reflecting fundamental patterns of tissue compartmentalization (Shi et al., 2025). However, tumors can also develop disease-specific spatial domains during progression, driven by tumorigenesis and microenvironmental remodeling. For example, in esophageal cancer, it has been reported that tumor epithelial cells and fibroblasts form a distinct spatial niche absent in normal tissues, which is closely associated with the transition from precancerous lesions to invasive carcinoma (Chang et al., 2025).

The normal–tumor comparison examines spatial-domain composition across pathological states. Figure 9 shows the lung domains and their relative abundance identified by STAGE. Additional lung and pancreas comparisons are provided in Figures 10 and 11.

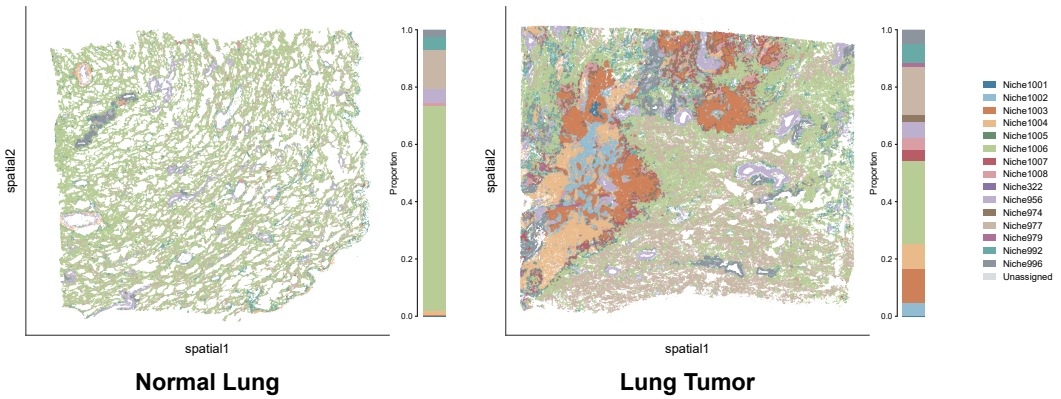

Figure 9: Spatial domains and their relative abundance in normal lung and lung tumor. Composition bars show the fraction of cells assigned to each domain, with a shared color legend across tissues. Domain-assignment settings are defined in Appendix C.1.

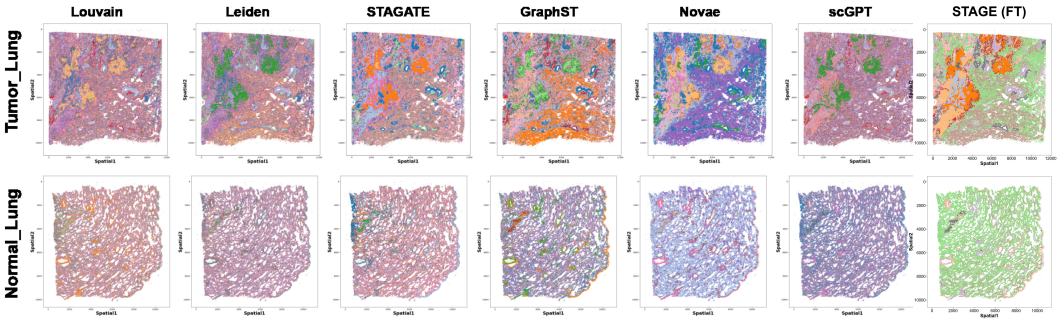

Figure 10: Spatial domains in lung cancer (top) and non-diseased lung tissue (bottom). Colors distinguish inferred domains. STAGE settings are defined in Appendix C.1.

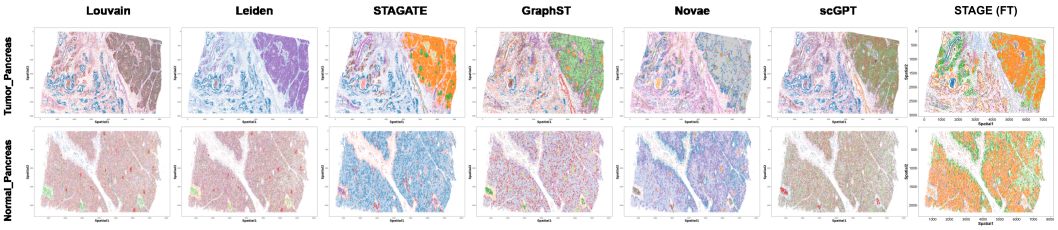

Figure 11: Spatial domains in pancreatic cancer (top) and non-diseased pancreatic tissue (bottom). Colors distinguish inferred domains. STAGE settings are defined in Appendix C.1.

F.4 Cross-tissue Domain Organization

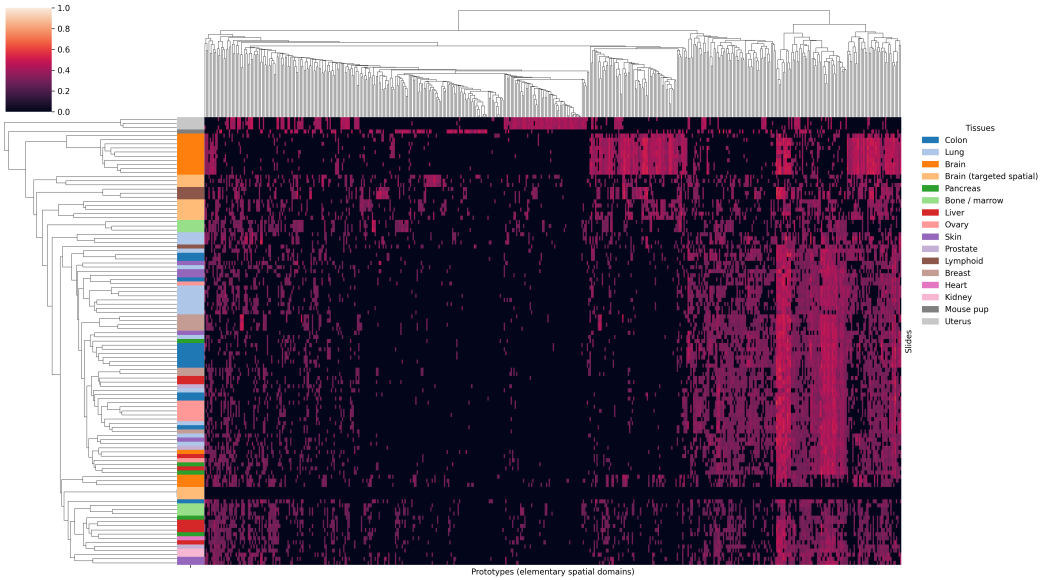

Figure 12: Cross-tissue spatial-domain composition and hierarchical prototype structure learned by STAGE.

From the perspective of tissue development and disease progression, spatial domains may partially overlap across different tissues or emerge in a condition-specific manner under particular physiological or pathological states.

As shown in Figure 12, STAGE generates spatial domain annotations across spatial transcriptomics slices from diverse tissue types, demonstrating its robustness in identifying both shared and tissue-specific domains. Lymph nodes and tonsils exhibit highly similar domain compositions, reflecting their common immune functions. Moreover, some domains identified in lymphoid tissues are also present in non-lymphoid tissues such as breast and lung, suggesting that certain spatial patterns may reflect underlying immune regulation or microenvironmental features rather than being strictly dictated by anatomical origin.

Hierarchical clustering based on domain composition (Figure 12) further shows that tissues with similar spatial characteristics cluster together and reveals organization complementary to conventional histological classifications. The spatial domains encode cell-type distributions together with contextual information such as disease states, local microenvironments, and functional programs. For example, immune activation, inflammation, or tumor progression can produce related spatial patterns across anatomically distinct tissues.

F.5 Comparison with Expert Annotations

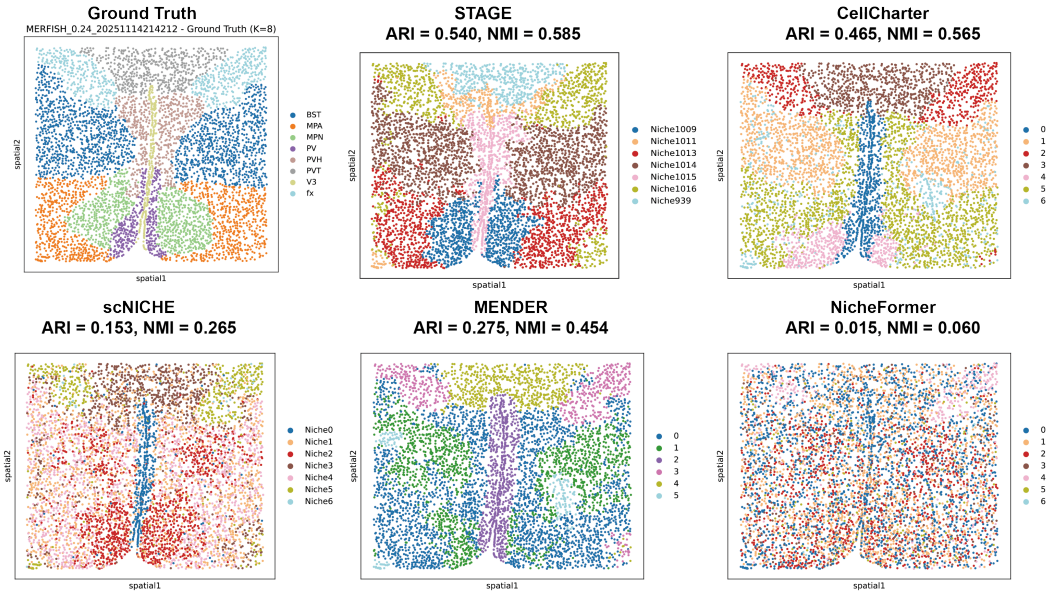

Figure 13: Spatial-domain identification in expert-annotated MERFISH data. Panels show the reference annotations and domains identified by STAGE, CellCharter (Varrone et al., 2024), scNiche (Qian et al., 2025), MENDER (Yuan, 2024), and Nicheformer (Tejada-Lapuerta et al., 2025). STAGE achieves ARI of 0.540 and NMI of 0.585 on this section.

G Availability and Disclosures

G.1 Code and Data Availability

The STAGE source code, pretrained model parameters, data-processing procedures, and experimental settings will be released upon acceptance. Evaluation code and usage examples will accompany the release to support reproduction of the reported results.

All evaluation datasets are publicly available through the sources cited in the main text. Training-data access information and preprocessing procedures will accompany the code release.

G.2 Use of Large Language Models

We used OpenAI language models to assist with manuscript drafting and revision, code development, and organization of experimental analyses. The authors are responsible for the scientific content, validation, and final approval of the manuscript.