SIGMMA: Hierarchical Graph-Based Multi-Scale Multi-modal Contrastive Alignment of Histopathology Image and Spatial Transcriptome

Dabin Jeong [1] **Amirhossein Vahidi** [1 2] **Ciro Ramírez-Suástegui** [1] **Marie Moullet** [1 2] **Kevin Ly** [1 2] **Mohammad Vali Sanian** [1] **Sebastian Birk** [1 3] **Yinshui Chang** [1] **Adam Boxall** [1] **Daniyal Jafree** [1] **Lloyd Steele** [1] **Vijaya Baskar MS** [1] **Muzlifah Haniffa*** [1 4 5] **Mohammad Lotfollahi*** [1 6 2]

Abstract

Recent advances in computational pathology have leveraged vision–language models to learn joint representations of Hematoxylin and Eosin (H&E) images with spatial transcriptomic (ST) profiles, but existing approaches typically align H&E tiles and ST profiles at a single scale, overlooking fine-grained cellular structures and their spatial organization. We propose SIGMMA, a multi-modal contrastive alignment framework for learning hierarchical H&E-ST representations. By enforcing multi-scale contrastive alignment, SIGMMA ensures coherent representations across modalities, while a graph-based modeling of cell interactions integrates both inter- and intra-subgraph relationships to capture cellular organization from fine to coarse scales. Across datasets, SIGMMA consistently improves gene-expression prediction and cross-modal retrieval performance, and its learned multi-scale embeddings recover tumor microenvironments and immune-exclusion programs in pancreatic cancer.

1. Introduction

Tissue architecture is hierarchically organized across spatial scales: from the *microenvironment* of interacting cells, to the *mesoenvironment* of cellular neighborhoods, up to the *macroenvironment* of tissue-level structures, e.g. tertiary lymphoid structures (Teillaud et al., 2024). Capturing this hierarchy requires both morphological and molecular views: H&E imaging characterizes cellular morphology, while single-cell spatial transcriptomics (ST) profiles individual cells in 2D (Zhang et al., 2025). Jointly modeling the two modalities reveals molecular heterogeneity not visible from H&E alone (Fig. 1c).

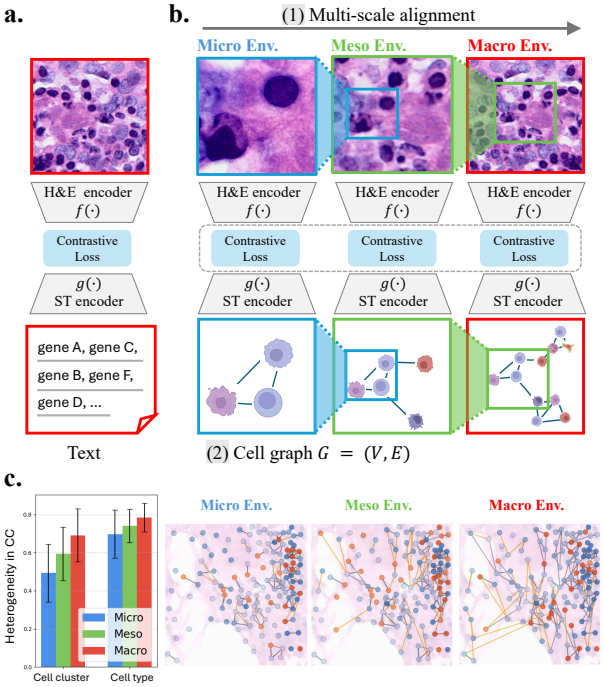

Figure 1. **Motivation. (a)** Limitations of prior VL-based H&E-ST alignment. **(b)** SIGMMA addresses these via (1) multi-scale alignment and (2) cell-graph representation that preserves 2D coordinates and cell-cell relationships. **(c)** SIGMMA captures multi-scale information, with ST representations becoming more heterogeneous at larger scales. CC, connected component.

Why graph structure for ST? Recent contrastive approaches learn joint H&E-ST embeddings for cross-modal retrieval and image-to-expression prediction, often adapting

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK [2]Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK [3]Institute of AI for Health, Helmholtz Center Munich, Neuherberg, Germany [4]Cambridge University Hospitals Foundation Trust, Addenbrooke's Hospital, Cambridge, UK [5]Cambridge Institute for Therapeutic Immunology and Infectious Diseases (CIITID), University of Cambridge, UK [6]Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. Correspondence to: Muzlifah Haniffa <mh32@sanger.ac.uk>, Mohammad Lotfollahi <ml19@sanger.ac.uk>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026. Copyright 2026 by the author(s).

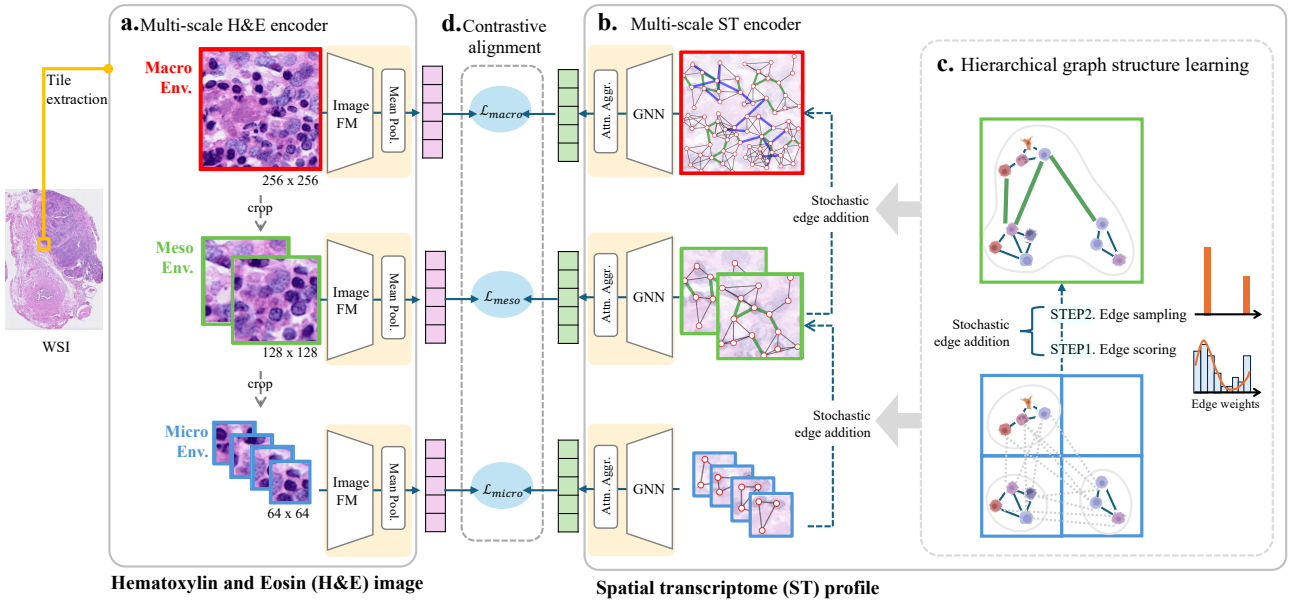

Figure 2. **SIGMMA schematic**. Given paired H&E-ST tiles, SIGMMA aligns them at multiple scales. **(a)** Multi-crop H&E side (Sec. 3.1); **(b)** hierarchical graph structure learning on the ST side (Sec. 3.2); **(c)** stochastic edge addition with neighbor-patch constraint; **(d)** multi-modal multi-scale contrastive alignment (Sec. 3.3). FM, foundation model; GNN, graph neural network.

vision–language (VL) models from H&E-biomedical text (Chen et al., 2025a; Glettig et al., 2025; Zou et al., 2025). However, these methods represent ST as a 1D gene sequence aggregated across cells, discarding 2D spatial organization and cell-cell interactions (Fig. 1a). Graph representations instead preserve spatial topology, explicitly modeling cell-cell interactions and tissue context (Fig. 1b).

Why hierarchical multi-scale alignment? Multi-scale alignment requires correspondence across ROI granularities, but graph-structured ST complicates this: message passing expands the receptive field via graph connectivity rather than image ROI scales, causing cross-modal scope mismatch. While multi-scale contrastive alignment captures coarse-to-fine tissue features (Fig. 1b), reconciling graph receptive field with image ROI remains, to our knowledge, unaddressed (Suppl. Fig. 4).

Thus, we propose SIGMMA for multi-scale H&E-ST alignment, with the following contributions:

- **Graph-structured ST representation.** A cell graph with a hierarchical module capturing local and long-range dependencies via intra- and inter-subgraph relationships.
- **Multi-scale cross-modal alignment.** A multi-scale contrastive objective that aligns micro, meso, and macro embeddings across modalities.
- **ROI scale reconciliation.** Progressive expansion of the graph receptive field to match the image ROI size.

2. Related work

ST at single-cell resolution. ST has emerged as a powerful approach to map gene expression within the spatial context of tissues. Specifically, it measures gene expression together with 2D spatial coordinates, indicating the location and level of expression of specific genes. There are two main techniques for measuring ST: Visium (Williams et al., 2022) and Xenium (Janesick et al., 2023). Visium is a sequencing-based platform which captures transcriptomic signals at the spot level, where each spot typically aggregates the expression profiles of multiple neighboring cells. In contrast, the Xenium platform utilizes high-resolution *in situ* hybridization and imaging to measure gene expression at the cellular/subcellular levels, offering deeper insights into cell–cell interactions. In this work, we use Xenium rather than Visium because Xenium provides cell-level spatial transcriptomics, enabling alignment with H&E images while explicitly modeling each cell's 2D spatial context.

Tiling of H&E WSI image. Whole-slide images (WSIs) are gigapixel-scale, making direct application of vision models computationally prohibitive and forcing heavy downsampling that removes critical cellular-level signals (Hou et al., 2016). Since discriminative patterns are small, sparse, and spatially scattered, tile-level modeling enables vision models to learn high-resolution local features by training on small image tiles, and leads to WSI-level tasks by aggregating tile-level embeddings (Azizi et al., 2023; Saillard et al., 2024; Chen et al., 2024; Jaume et al., 2024b;a). As molecular phenotypes and cellular contexts vary across lo-

calized regions, tile-level alignment can provide a more fine-grained correspondence between image features and transcriptomic signals than slide-level alignment. Motivated by this, our work focuses on tile-level alignment between H&E and ST features, enabling cross-modal learning at a spatially-resolved and fine-grained level.

Foundation models for H&E and ST. Foundation models have recently emerged in computational histopathology for both H&E images and ST. For H&E image, DINO (Oquab et al., 2023)-based vision foundation models enable scalable learning of morphology-rich representations that generalize across slides (Chen et al., 2024; Saillard et al., 2024; Vorontsov et al., 2024). Extending this line of work, hierarchical transformers leverage the intrinsic multi-scale structure of WSIs and learn representations across cellular, tissue, and slide levels (Chen et al., 2022). In parallel, ST foundation models, inspired by large language models, capture cell-level representations by modeling each cell's gene expression profile as a sequence using transformer architectures (Wang et al., 2025; Birk et al., 2025), or by encoding the spatial relationships among cells through graph-based representations and graph neural networks (Birk, 2025; Blampey et al., 2024). These uni-modal foundation models provide generalizable representations for H&E and ST, serving as building blocks for downstream multi-modal alignment. In this work, we build upon these foundation models to learn a unified cross-modal representation between H&E and ST.

H&E-ST contrastive alignment. Early ST-prediction methods regressed spot-level expression directly from H&E using CNN or transformer backbones (Chung et al., 2024; Ganguly et al., 2025), later recasting the task as node-level regression over spatial graphs whose nodes are proximity-connected spots (Ganguly et al., 2025). High-resolution ST then shifted the paradigm from spot to cellular/subcellular modeling, motivating cell-graph approaches (Ge et al., 2025) and diffusion-based image-to-expression generation at subcellular resolution (Xu et al., 2025). In parallel, contrastive methods align H&E and ST rather than predicting one from the other, enriching cross-modal representations and improving downstream prediction (Xie et al., 2023; Redekop et al., 2025); VL frameworks extend this by pairing H&E tiles with biomedical text or gene-token sequences (Huang et al., 2023; Lu et al., 2024; Albastaki et al., 2025). More recent work performs spatially resolved alignment between image regions and Visium spots (Chen et al., 2025a; Gindra et al., 2025; Guo et al., 2025), with subsequent studies extending it to cell-level alignment on Xenium (Glettig et al., 2025). In contrast, our framework introduces graph-based multi-scale alignment between H&E and ST: we represent each ST tile as a cell graph built from cell coordinates and align it with the H&E tile across multiple spatial scales, maintaining spatial consistency and enabling fine-grained cell-level correspondence.

Discussion on the positioning of SIGMMA relative to prior H&E-to-ST prediction methods is provided in Suppl. Sec. B

3. Method: SIGMMA

The problem is formulated as follows. We consider paired H&E images and ST profiles from the same tissue section, $(\mathcal{I}, \mathcal{S})$, tessellated into $m \times m$ tiles $\{(I_i, S_i)\}_{i=1}^{n}$. We train an H&E encoder f and an ST encoder g producing $z_i^I = f(I_i)$ and $z_i^S = g(S_i)$, jointly optimized so that paired embeddings align in a shared latent space. SIGMMA (Fig. 2) consists of three components: (i) a multi-scale H&E encoder, (ii) a hierarchical-graph ST encoder, and (iii) cross-modal multi-scale contrastive alignment.

3.1. Multi-scale H&E encoder

Each tile $I \in \mathbb{R}^{m \times m \times 3}$ is partitioned via a multi-crop strategy (Guo et al., 2024; Liu et al., 2024) into 4×4 micro patches $\{I_{\text{micro},j}\}_{j=1}^{16}$, 2×2 meso patches $\{I_{\text{meso},j}\}_{j=1}^{4}$, and one macro patch (Fig. 2a). Each patch is resized to the H&E foundation model's input resolution (Chen et al., 2024), encoded by f, and grid-wise mean-pooled:

$$
\begin{aligned}
z^I_{\text{micro}} &= \text{Pool}_I\big(f(I_{\text{micro}})\big),\\
z^I_{\text{meso}} &= \text{Pool}_I\big(f(I_{\text{meso}})\big),\\
z^I_{\text{macro}} &= \text{Pool}_I\big(f(I_{\text{macro}})\big).
\end{aligned}
$$

3.2. Multi-scale ST encoder

For each tile S, we build a geometric graph $G = (V, E)$, where V are cells and E encodes proximity-based cell-cell interactions (Team, 2025). Node embeddings are initialized by an ST foundation model (Birk, 2025).

Hierarchical graph structure learning. To reconcile graph receptive field with image ROI, we hierarchically expand subgraphs by linking neighbors via stochastic edge addition (cf. (Piao et al., 2022; 2024)) followed by GNN layers. Given GNN-derived embeddings h_u, h_v, an edge score $s_{uv} = \sigma(\text{MLP}([h_u, h_v]))$ parametrises a Bernoulli distribution from which a binary p_{uv} is drawn; the Gumbel-softmax (Jang et al., 2016) provides differentiable sampling at temperature τ. We start from a micro subgraph $G_{\text{micro}} = (V_{\text{micro}}, E_{\text{micro}})$ where cells in I_{micro} are connected by spatial proximity, and recursively grow neighborhoods:

$$
\begin{aligned}
\mathcal{N}^{(l-1)}_{\text{meso}}(u) &= \mathcal{N}^{(l-2)}_{\text{micro}}(u) \cup \{v : p^{(l-1)}_{uv} = 1\},\\
\mathcal{N}^{(l)}_{\text{macro}}(u) &= \mathcal{N}^{(l-1)}_{\text{meso}}(u) \cup \{v : p^{(l)}_{uv} = 1\}.
\end{aligned}
$$

Table 1. Cross-modal retrieval across HEST1k and in-house datasets. R, recall.

	H&E → ST														
Dataset	*HEST1k-LUAD*			*HEST1k-PAAD*			*HEST1k-SKCM*			*HEST1k-IDC*			*in-house skin*		
Model	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%
CLIP	0.278	0.452	0.566	0.195	0.338	0.471	0.290	0.495	0.586	0.342	0.532	0.668	0.347	0.503	0.617
PLIP	0.367	0.526	0.621	0.187	0.336	0.469	0.253	0.414	0.527	0.356	0.536	0.665	0.370	0.539	0.650
BLEEP	0.419	0.554	0.630	0.152	0.182	0.212	0.318	0.500	0.614	**0.443**	**0.603**	**0.704**	0.426	0.550	0.623
OmiCLIP	0.281	0.453	0.596	0.177	0.320	0.485	0.231	0.382	0.532	0.342	0.520	0.636	0.329	0.502	0.605
SIGMMA	**0.590**	**0.728**	**0.826**	**0.402**	**0.630**	**0.813**	**0.333**	**0.559**	**0.731**	0.394	0.570	0.687	**0.472**	**0.591**	**0.687**

	ST → H&E														
Dataset	*HEST1k-LUAD*			*HEST1k-PAAD*			*HEST1k-SKCM*			*HEST1k-IDC*			*in-house skin*		
Model	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%
CLIP	0.297	0.413	0.526	0.141	0.284	0.390	0.285	0.473	0.591	0.445	**0.675**	**0.798**	0.354	0.513	0.619
PLIP	0.330	0.483	0.618	0.213	0.358	0.475	0.274	0.435	0.543	0.440	0.639	0.767	0.371	0.552	0.665
BLEEP	0.415	0.568	0.654	0.030	0.121	0.212	0.330	0.494	0.580	**0.502**	0.655	0.754	0.419	0.561	0.634
OmiCLIP	0.281	0.501	0.599	0.165	0.318	0.435	0.242	0.403	0.495	0.412	0.612	0.742	0.335	0.514	0.632
SIGMMA	**0.602**	**0.768**	**0.813**	**0.304**	**0.505**	**0.652**	**0.333**	**0.500**	**0.602**	0.399	0.611	0.750	**0.459**	**0.620**	**0.708**

This yields scale-specific topologies $G_{\text{micro}}, G_{\text{meso}}, G_{\text{macro}}$. Graph-level representations follow:

$$z_*^S = \text{Pool}_S(g(G_*)), \quad * \in \{\text{micro}, \text{meso}, \text{macro}\},$$

with global attention pooling (Li et al., 2015).

Neighbor-patch constraint. To restrict edge addition spatially, we partition the 2D coordinate grid into blocks of size $b \times b$ and allow (p, q) only if $\lfloor x_p/b \rfloor = \lfloor x_q/b \rfloor$ and $\lfloor y_p/b \rfloor = \lfloor y_q/b \rfloor$, with b=4, 2, 1 for micro, meso, macro respectively. This prevents cross-block edges and aligns graph growth with the ROI hierarchy.

3.3. Multi-modal multi-scale contrastive alignment

We use bidirectional InfoNCE (Oord et al., 2018):

$$\mathcal{L}_{I\to S} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i^I, z_i^S)/\tau)}{\sum_j \exp(\text{sim}(z_i^I, z_j^S)/\tau)},$$

$$\mathcal{L}_{S\to I} \text{ defined symmetrically,}$$

$$\mathcal{L}_{\text{ALIGN}}(z^I, z^S) = \tfrac{1}{2}(\mathcal{L}_{I\to S} + \mathcal{L}_{S\to I}).$$

The total objective sums per-scale losses (Fig. 2d):

$$\mathcal{L} = \mathcal{L}_{\text{MICRO}} + \mathcal{L}_{\text{MESO}} + \mathcal{L}_{\text{MACRO}},$$

where $\mathcal{L}_* = \mathcal{L}_{\text{ALIGN}}(z_*^I, z_*^S)$.

4. Experiments

Datasets. We benchmark on HEST-1k (Jaume et al., 2024a), the largest public paired H&E-ST resource, using its four Xenium subsets (IDC, PAAD, SKCM, LUAD) plus an in-house skin dataset. Each WSI is tessellated into 256×256 tiles at $20\times$ ($0.5\,\mu$m/px), following (Chen et al., 2024; Vorontsov et al., 2024; Saillard et al., 2024). Preprocessing and splits: Suppl. Sec. C.

Baselines and metrics. We compare against: (1) H&E foundation model UNI (Chen et al., 2024); (2) VL models CLIP (Radford et al., 2021) and PLIP (Huang et al., 2023); (3) H&E-ST contrastive baselines OmiCLIP (Chen et al., 2025a) and BLEEP (Xie et al., 2023). All baselines are fine-tuned with original or grid-searched hyperparameters (Suppl. Sec. D). For gene-expression prediction, we follow HEST-1k linear probing (PCA-256 + ridge regression on top 50 HVGs), reporting tile-level PCC and MSE (mean±std). For cross-modal retrieval, we report Recall@p% (p=5, 10, 15) on held-out tiles.

4.1. Quantitative: benchmarking on gene expression prediction and cross-modal retrieval tasks.

We evaluate alignment via gene expression prediction and cross-modal retrieval. Multi-modal alignment consistently enriches H&E embeddings across backbones (ResNet50, H-Optimus-0, UNI), with up to 67% lower MSE and 56% higher PCC on UNI (Suppl. Tab. 6); we adopt UNI as backbone. SIGMMA achieves the highest PCC on gene expression prediction (Suppl. Tab. 4) and delivers consistent gains on cross-modal retrieval in both H&E ↔ ST directions (Tab. 1)—best on most datasets, second on IDC. See Suppl. Sec. F for further analysis.

4.2. Qualitative: cell-aware attention

We probed attention maps of the H&E encoder fine-tuned with SIGMMA (Suppl. Fig. 6). Unlike UNI, which attends to tissue boundaries and cell-sparse adipose regions (lipid-washed white intra-tissue areas, row 1), SIGMMA sharply localizes on individual nuclei in cell-dense regions (heads 5-6) and suppresses out-of-tissue activations (rows 3-4, heads 2/4/6), shifting attention toward cell-dense morphology.

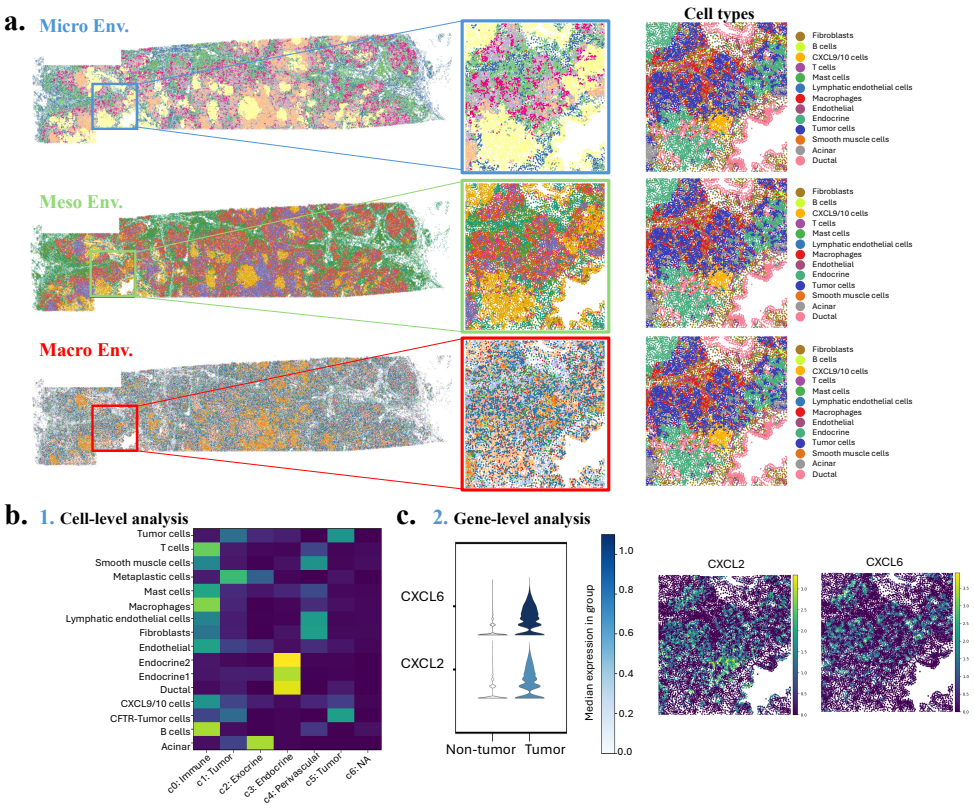

Figure 3. **Biological interpretation of SIGMMA cell-level embeddings (PAAD). (a)** Microenvironment clusters across scales (left), close-up of tumor/non-tumor boundary (middle), reference cell-type annotation (right). **(b)** Cell-type proportions per micro-scale microenvironment. **(c)** *CXCL2*/*CXCL6* expression in tumor vs. non-tumor microenvironments (violin) and spatial projection (right).

4.3. Ablation

We ablate three core components (Suppl. Sec. E): (1) cell graph, (2) multi-scale loss, (3) graph sparsification via stochastic edge addition. Multi-scale loss and graph sparsification contribute most, with the full combination performing best. Adopting a cell graph alone degrades performance (Suppl. Tab. 5 row 2) because the GNN receptive field on sparsely distributed Xenium cells misaligns with its H&E crop —confirming the graph and multi-scale components are interdependent by design.

4.4. Biological application

We tested whether SIGMMA's multi-scale embeddings capture meaningful biology in a public PAAD section (Suppl. Sec. C). Clustering and projecting onto the section (Fig. 3a) yields spatially coherent domains; at micro/meso scales, SIGMMA cleanly delineates tumor nests from surrounding tissue without cell-type supervision. Per-microenvironment cell-type composition (Fig. 3b) recovers six structures: two tumor-dominant, one immune infiltrate (T/B cells excluded from tumor), and three pancreatic compartments (perivascular, endocrine, exocrine), consistent with known pancreatic tumor organization. Differential expression between tumor-associated and immune-excluded microenvironments identifies *CXCL2*/*CXCL6* (Hu et al., 2021b)—chemokines driving immunosuppressive myeloid recruitment and T/B exclusion —with tumor-localized spatial expression (Fig. 3c). SIGMMA thus recovers hallmarks of immune exclusion and tumor-specific molecular programs.

5. Conclusion

We presented SIGMMA, a hierarchical framework aligning H&E-ST representations across multiple scales. To our knowledge, we are the first to formulate and address the graph receptive field-ROI mismatch arising in cell-resolution ST. Limitations and future work are discussed in Suppl. Sec. F.2.

Impact Statement

This paper aims to advance machine learning in the biomedical domain by integrating whole-slide imaging with gene expression profiling. We do not identify specific consequences requiring special emphasis at this time.

References

Albastaki, S., Sohail, A., Ganapathi, I. I., Alawode, B., Khan, A., Javed, S., Werghi, N., Bennamoun, M., and Mahmood, A. Multi-resolution pathology-language pre-training model with text-guided visual representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25907–25919, 2025. 3

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023. 2

Birk, S. Stemo. https://github.com/Lotfollahi-lab/stemo, 2025. 3, 11, 14

Birk, S., Bonafonte-Pardàs, I., Feriz, A. M., Boxall, A., Agirre, E., Memi, F., Maguza, A., Yadav, A., Armingol, E., Fan, R., et al. Quantitative characterization of cell niches in spatially resolved omics data. *Nature Genetics*, pp. 1–13, 2025. 3

Blampey, Q., Benkirane, H., Bercovici, N., Andre, F., and Cournede, P.-H. Novae: A graph-based foundation model for spatial transcriptomics data. pp. 2024.09.09.612009, 2024. doi: 10.1101/2024.09.09.612009. 3, 14

Chadoutaud, L., Lerousseau, M., Herrero-Saboya, D., Ostermaier, J., Fontugne, J., Barillot, E., and Walter, T. scellst predicts single-cell gene expression from h& e images. *Nature Communications*, 17(1):1194, 2026. 9

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16144–16155, 2022. 3

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862, 2024. 2, 3, 4, 11, 14, 15

Chen, W., Zhang, P., Tran, T. N., Xiao, Y., Li, S., Shah, V. V., Cheng, H., Brannan, K. W., Youker, K., Lai, L., et al. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, pp. 1–15, 2025a. 2, 3, 4, 9, 11

Chen, Y.-A., Watson, C., and Beštak, K. PALOM - Piecewise alignment for layers of mosaics, July 2025b. URL https://github.com/labsyspharm/palom. 11

Chung, Y., Ha, J. H., Im, K. C., and Lee, J. S. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11591–11600, 2024. 3, 9, 10

Dong, K. and Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1): 1739, 2022. 9

Fu, X., Cao, Y., Bian, B., Wang, C., Graham, D., Pathmanathan, N., Patrick, E., Kim, J., and Yang, J. Y. H. Spatial gene expression at single-cell resolution from histology using deep learning with ghist. *Nature methods*, 22(9):1900–1910, 2025. 9

Ganguly, A., Chatterjee, D., Huang, W., Zhang, J., Yurovsky, A., Johnson, T. S., and Chen, C. Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15611–15620, 2025. 3, 9, 10

Ge, Y., Leng, J., Tang, Z., Wang, K., U, K., Zhang, S. M., Han, S., Zhang, Y., Xiang, J., Yang, S., et al. Deep learning-enabled integration of histology and transcriptomics for tissue spatial profile analysis. *Research*, 8: 0568, 2025. 3

Gindra, R. H., Palla, G., Nguyen, M., Wagner, S. J., Tran, M., Theis, F. J., Saur, D., Crawford, L., and Peng, T. A large-scale benchmark of cross-modal learning for histology and gene expression in spatial transcriptomics. *arXiv preprint arXiv:2508.01490*, 2025. 3

Glettig, M., Ehrensperger, T., Yates, J., and Boeva, V. H&enium, applying foundation models to computational pathology and spatial transcriptomics to learn an aligned latent space. *bioRxiv*, pp. 2025–07, 2025. 2, 3

Guo, Y., Liu, J. S., Cheng, H., and Ma, Y. Jade: Joint alignment and deep embedding for multi-slice spatial transcriptomics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

Guo, Z., Xu, R., Yao, Y., Cui, J., Ni, Z., Ge, C., Chua, T.-S., Liu, Z., and Huang, G. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pp. 390–406. Springer, 2024. 3

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017. 11

He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., and Zou, J. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020. 9, 10

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 13, 14

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2424–2433, 2016. 2

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., and Li, M. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11): 1342–1351, 2021a. 9

Hu, J., Zhao, Q., Kong, L.-Y., Wang, J., Yan, J., Xia, X., Jia, Z., Heimberger, A. B., and Li, S. Regulation of tumor immune suppression and cancer cell survival by cxcl1/2 elevation in glioblastoma multiforme. *Science advances*, 7(5):eabc2511, 2021b. 5

Huang, T., Liu, T., Babadi, M., Jin, W., and Ying, R. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *arXiv preprint arXiv:2506.05361*, 2025. 10

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., and Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 3, 4

Janesick, A., Shelansky, R., Gottscho, A. D., Wagner, F., Williams, S. R., Rouault, M., Beliakoff, G., Morrison, C. A., Oliveira, M. F., Sicherman, J. T., et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications*, 14(1):8353, 2023. 2, 11

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

Jaume, G., Doucet, P., Song, A., Lu, M. Y., Almagro Pérez, C., Wagner, S., Vaidya, A., Chen, R., Williamson, D., Kim, A., et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024a. 2, 4

Jaume, G., Oldenburg, L., Vaidya, A., Chen, R. J., Williamson, D. F., Peeters, T., Song, A. H., and Mahmood, F. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9632–9644, 2024b. 2, 10

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12): 1289–1296, 2019. 14

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015. 4

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024. 3

Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij, K., Chen, A., et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature communications*, 14(1):1155, 2023. 9

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024. 3

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

Piao, Y., Lee, S., Lee, D., and Kim, S. Sparse structure learning via graph neural networks for inductive document classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 11165–11173, 2022. 3, 11

Piao, Y., Lee, S., Lu, Y., and Kim, S. Improving out-of-distribution generalization in graphs via hierarchical semantic environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27631–27640, 2024. 3

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 4

Redekop, E., Pleasure, M., Wang, Z., Flores, K., Sisk, A., Speier, W., and Arnold, C. W. Spade: Spatial transcriptomics and pathology alignment using a mixture of data experts for an expressive latent space. *arXiv preprint arXiv:2506.21857*, 2025. 3

Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., and Vert, J.-P. H-optimus-0, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0. 2, 3, 4, 14, 15

Team, S. D. squidpy.gr.spatial_neighbors: Spatial neighbor graph construction in squidpy. https://squidpy.readthedocs.io/en/stable/api/squidpy.gr.spatial_neighbors.html, 2025. Accessed: 2025-10-23. 3, 11

Teillaud, J.-L., Houel, A., Panouillot, M., Riffard, C., and Dieu-Nosjean, M.-C. Tertiary lymphoid structures in anticancer immunity. *Nature Reviews Cancer*, 24(9):629–646, 2024. 1

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10):2924–2935, 2024. 3, 4, 15

Wang, C., Cui, H., Zhang, A., Xie, R., Goodarzi, H., and Wang, B. scgpt-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. *bioRxiv*, pp. 2025–02, 2025. 3

Weng, Z., Fang, Y., Qian, J., Wang, X., Cooper, L. A., Cai, W., and Zhou, B. Hifusion: Hierarchical intra-spot alignment and regional context fusion for spatial gene expression prediction from histopathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10630–10637, 2026. 9, 10

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., and Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome medicine*, 14(1):68, 2022. 2

Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., and Bader, G. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023. 3, 4, 9

Xu, M., Gupta, S., Hu, X., Li, C., Abousamra, S., Samaras, D., Prasanna, P., and Chen, C. Topocellgen: Generating histopathology cell topology with a diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20979–20989, 2025. 3

Xue, S., Zhu, F., Chen, J., and Min, W. Inferring single-cell resolution spatial gene expression via fusing spot-based spatial transcriptomics, location, and histology using gcn. *Briefings in Bioinformatics*, 26(1):bbae630, 2025. 9

Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y., and Yang, Y. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022. 9

Zhang, P., Chen, W., Tran, T. N., Zhou, M., Carter, K. N., Kandel, I., Li, S., Hoi, X. P., Sun, Y., Lai, L., et al. Thor: a platform for cell-level investigation of spatial transcriptomics and histology. *Nature Communications*, 16(1): 7178, 2025. 1

Zou, J., Xiao, K., Chen, Z., Pei, J., Xu, J., Chen, T., Hou, L., Wu, C., She, Y., Yuan, Z., et al. Predicting spatial transcriptomics from h&e image by pretrained contrastive alignment learning. *bioRxiv*, pp. 2025–06, 2025. 2

Supplementary Material

We begin in Sec. A by elaborating on the motivation behind SIGMMA's design, focusing on the graph receptive field (RF)-ROI mismatch that arises uniquely in cell-resolution spatial transcriptomics and which the main text could not fully expand upon. Sec. 2 then situates SIGMMA against prior work, detailing how it differs from existing H&E-ST alignment methods. Building on this, details on the dataset are provided in Sec. C, and methodological descriptions are included in Sec. D. Additional results that could not be presented in the main text due to space constraints are shown in Sec. E. Finally, Sec. F discusses challenging cases, limitations, and future research directions.

A. Motivation

In Visium, spots lie on a fixed hexagonal grid (∼100 µm spacing, ∼55 µm diameter) determined by array design, so GNN message passing expands the receptive field (RF) uniformly and naturally coincides with image ROI crops at each scale. In Xenium, nodes are individual cells at their true 2D coordinates, and because cell density reflects tissue biology rather than a regular lattice, graph edges cross ROI boundaries unpredictably (Suppl. Fig. 4). The resulting RF–ROI mismatch is negligible for Visium but systematic for Xenium, and to our knowledge has not been identified in prior work. The mismatch is structural: it requires the simultaneous presence of (i) an ST-side graph, (ii) irregularly distributed cells as nodes, and (iii) multi-scale image ROIs. Because no prior work satisfies all three, the mismatch is structurally absent rather than implicitly addressed:

- **(A) No ST-side graph.** TRIPLEX (Chung et al., 2024), HiFusion (Weng et al., 2026), BLEEP (Xie et al., 2023), OmiCLIP (Chen et al., 2025a), ST-Net (He et al., 2020), sCellST (Chadoutaud et al., 2026), and GHIST (Fu et al., 2025) have no GNN RF to misalign.

- **(B) Graph on a regular grid.** Hist2ST (Zeng et al., 2022) and MERGE (Ganguly et al., 2025) build graphs over Visium spots, so the RF coincides with image ROIs by construction.

- **(C) Cell graph without multi-scale alignment.** SpaGCN (Hu et al., 2021a), STAGATE (Dong & Zhang, 2022), and GraphST (Long et al., 2023) use ST-only clustering with no H&E coupling; scstGCN (Xue et al., 2025) performs single-scale super-resolution without an alignment objective.

The regime of *cell-level Xenium graphs with multi-scale cross-modal alignment* is therefore genuinely under-explored, and it is precisely here that the RF–ROI mismatch first becomes substantive. SIGMMA's neighbor-patch constraint (Sec. 3.2) addresses it explicitly by restricting edges to cells within the same spatial block, aligning the GNN RF with the pixel-space ROI hierarchy.

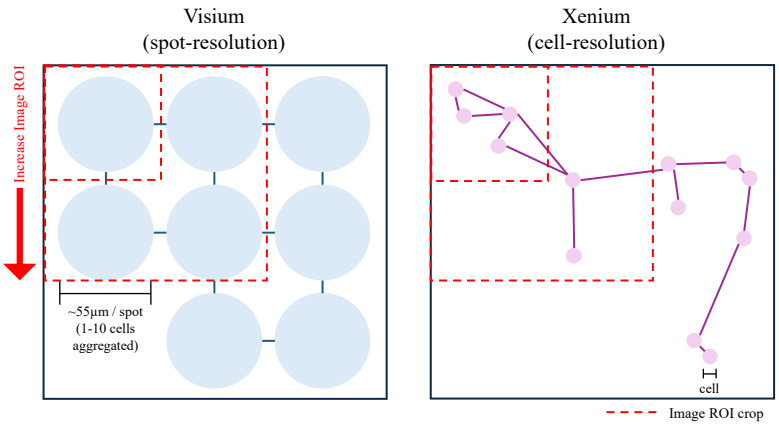

Figure 4. **The graph RF–ROI mismatch is unique to cell-resolution spatial transcriptomics.** In Visium, the fixed hexagonal grid ensures that GNN receptive field expansion corresponds naturally to image ROI crops. In Xenium, the irregular spatial distribution of cells in a spatial graph causes the GNN RF to cross ROI boundaries unpredictably, creating a systematic mismatch at every scale.

B. Related work

H&E to ST prediction. Prior work on H&E to ST prediction formulates the histology-transcriptomics relationship as a supervised mapping from H&E to gene expression. ST-Net (He et al., 2020) regresses per-spot expression from a single-scale image encoder, with no spatial hierarchy on either modality. TRIPLEX (Chung et al., 2024) introduces multi-scale image context through parallel encoders at the spot, neighborhood, and whole-slide levels, but fuses them into a single per-spot prediction; ST remains a flat per-spot vector. HiFusion (Weng et al., 2026) decomposes each spot image into multi-resolution sub-patches to capture intra-spot heterogeneity, yet the hierarchy is again confined to the image side. MERGE (Ganguly et al., 2025) builds a patch graph with shortcut edges between cluster centroids, but all nodes share a single patch resolution and ST serves only as a per-node target. STFlow (Huang et al., 2025) models joint expression across spots via flow matching, but the H&E-ST coupling itself is a unidirectional supervised mapping at Visium spot resolution. Across these methods, all of which operate on Visium spot data, image encoders and nonlinear decoders are jointly optimized under regression or generative objectives and evaluated by end-to-end prediction accuracy.

SIGMMA addresses a different problem formulation. Rather than decoding histology into expression, it models the H&E–ST relationship as symmetric cross-modal representation alignment via contrastive learning between modality-specific encoders. The two paradigms thus target distinct objectives—*representation quality* versus *prediction accuracy*.

This shift in objective is reflected in how the multi-scale hierarchy is constructed: rather than confining hierarchy to the image side, SIGMMA builds it on *both* modalities and explicitly aligns them across scales. On the ST side, this takes the form of a hierarchy over a cell-level spatial graph in which micro-, meso-, and macro-scale representations correspond to progressively expanded GNN receptive fields, regulated by a neighbor-patch constraint that preserves spatial correspondence with image-side ROIs. Such a hierarchy is meaningful only when ST itself exhibits internal spatial structure over which scale can be expanded—the regime enabled by cell-resolution platforms such as Xenium, and not the spot-grid setting for which earlier methods were designed. The two formulations are therefore complementary rather than competing: SIGMMA embeddings could be coupled with prediction-oriented decoders, and advances in such decoders may inform future extensions of SIGMMA.

C. Data

C.1. Data acquisition

HEST1k dataset. The HEST1k dataset (Jaume et al., 2024b) is a publicly available benchmark comprising paired H&E image and Xenium-ST data. Table 2 provides a comprehensive summary of all Xenium sections within HEST1k, including tissue types, cell counts, and the presence of cell-type annotations. All sections in the table have paired post-Xenium H&E morphology images. Cell-type annotations were derived from the *TENX116* section, which serves as the reference dataset for the biological analyses presented in Sec. 4.4.

Table 2. Summary of HEST1k H&E-Xenium ST sections.

Tissue type	Section ID	# of cells	Cell-type annotations	Source
LUAD	TENX118	162,254		Public
	TENX141	161,000		Public
PAAD	TENX116	190,965	O	Public
	TENX140	235,099		Public
	TENX126	140,702		Public
SKCM	TENX115	106,980		Public
	TENX117	87,499		Public
IDC	TENX95	574,852		Public
	TENX99	892,966		Public
	NCBI783	142,272		Public
	NCBI785	167,780		Public

In-house dataset. The in-house dataset consists of 21 Xenium ST sections derived from 13 patients with either eczema or skin warts. Eczema samples were profiled using the Xenium Prime 5K Human Pan-Tissue & Pathways Panel, whereas skin wart samples were processed with the Xenium Immuno-Oncology Panel. Tissue sections were placed into the active capture area on Xenium slides and stored accordingly until being processed for in situ gene expression according to the

manufacturer's protocol. All Xenium slides were processed using the Xenium Prime 5K Human Pan Tissue & Pathways Panel, including the standard cell segmentation antibody staining. Due to confidentiality, detailed section-level metadata cannot be shared; however, we report an average cell count of approximately 16,704±6,693 cells per section.

C.2. Data preprocessing

H&E image. Since the H&E image and Xenium ST profile are acquired on different imaging systems, their spatial coordinates are not directly comparable. To place morphological information from the H&E image into the same spatial coordinate system, whole-slide H&E image and immunofluorescence (IF) were registered using Palom (Chen et al., 2025b), a piecewise registration framework for layers of mosaics. Image registration was performed using the DAPI (IF) and green (H&E) channels. Coarse affine alignment was initialized using 4,000 keypoints, followed by local shift refinement and constraint optimization. The final H&E image was reconstructed using a blockwise affine transformation and rescaled to 0.5μm/pixel at 20x magnification level. The Public HEST1k dataset was provided with registration already completed. All sections are tessellated into 256×256 pixel-sized tiles.

ST Xenium profile. All Xenium ST data is processed with the 10X platform (Janesick et al., 2023), which provides built-in cell segmentation and outputs cell-resolved features by default. We extracted ST tiles corresponding to the same spatial region as each H&E tile and constructed tile-level representations by aggregating the ST profiles of all cells within each tile. For graph representation of each ST tile, we constructed cell graphs using `squidpy.gr.spatial_neighbors` with default parameters (Team, 2025), which constructs a 6-nearest-neighbor graph from Euclidean coordinates, encoding local spatial proximity among neighboring cells. For sequence representation of ST for the VL model, following Loki pipeline (Chen et al., 2025a), we selected the top 50 highly variable genes by tile-level mean expression and ranked their gene names in descending mean expression.

C.3. Data splitting strategy

HEST1k Dataset. Due to the limited number of paired H&E–Xenium ST sections in public datasets (Tab. 2), section-level split is not feasible, as it would not provide enough data for stable model training. Additionally, assigning spatially adjacent tiles to different splits can lead to spatial leakage across data subsets, since such tiles often share morphological and molecular characteristics. To address these issues, we employ a *spatially-stratified tile split* for the HEST1k dataset, assigning validation and test tiles to non-overlapping spatial regions (Fig. 5). All splits use an 8:1:1 ratio for training, validation, and testing.

In-house dataset. Given the larger number of available sections, we employ a section-level split for the in-house dataset. Each section is assigned entirely to one subset to avoid information leakage across data splits. The dataset is randomly partitioned into training, validation, and test sets with an 8:1:1 ratio.

D. Method

D.1. Implementation detail

We use UNI (Chen et al., 2024) as the H&E encoder backbone, which is a ViT-L/16 architecture. For each H&E tile, we generate multi-scale crops and resize each patch to the ViT input resolution, 224×224 pixels. Each patch is encoded using the UNI model, and the CLS token output is used as its patch-level embedding. Patch embeddings within each scale are then mean-pooled and L2-normalized to produce a single representation per scale. This procedure is applied from the micro to macro scale, yielding a set of multi-scale H&E embeddings.

For the ST encoder, node embeddings are initialized with STEMO features (Birk, 2025), followed by SAGEConv (Hamilton et al., 2017)-based message passing and stochastic edge addition layer (Piao et al., 2022). All GNN components are implemented using the `dgl` library. Graph-level embeddings are then obtained via attention pooling. This procedure is applied hierarchically from micro to macro scales to produce multi-scale ST representations.

At each scale, H&E and ST embeddings are passed through simple MLP projection heads to obtain vectors of the same dimensionality, and a symmetric InfoNCE loss is then applied to align the modalities. The hidden dimensions for H&E and ST embeddings are set to 256 or 512. All projection and encoder layers use LeakyReLU activations.

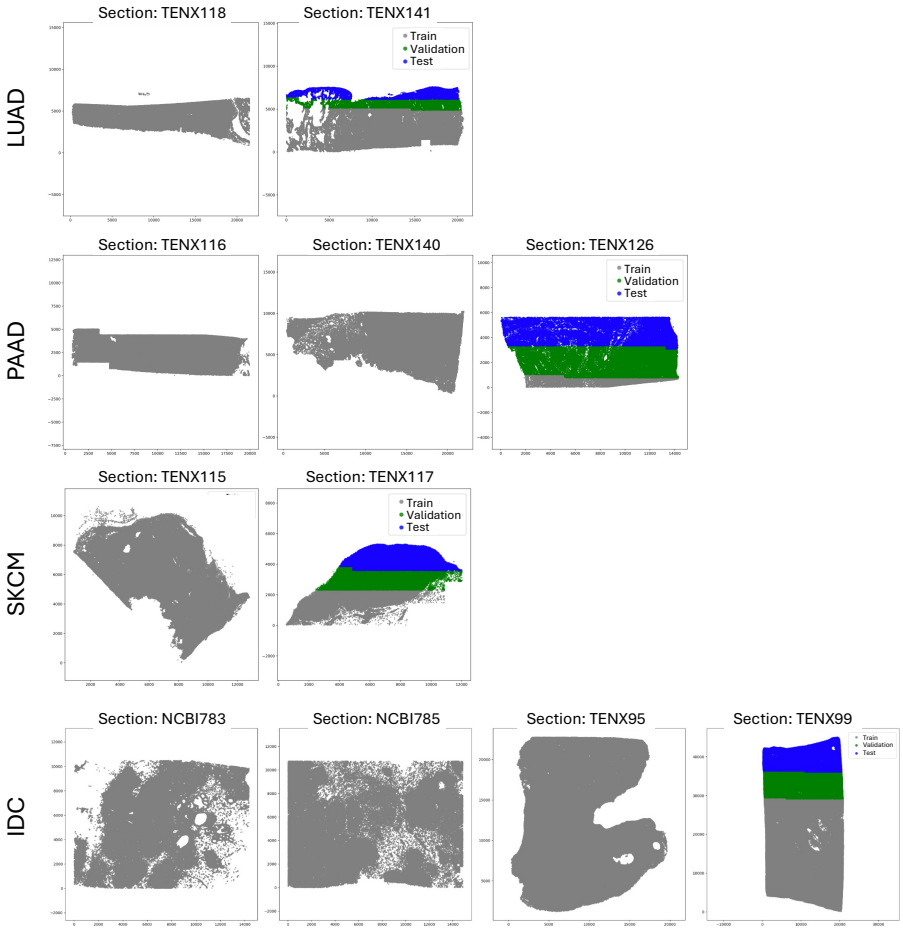

Figure 5. Train-validation-test split of HEST1k dataset.

We optimize the model using AdamW with a StepLR scheduler and weight decay. To obtain a larger effective batch size, which is essential for contrastive learning, we employ 10 steps of gradient accumulation with a batch size of 128 and incorporate cross-batch negatives. All experiments are conducted on a single A100 SXM4 GPU with 80GB RAM.

The hyperparameters were tuned by grid search, as a combination of learning rate [0.001, 0.0001, 0.00001, 0.000001], weight decay value [0.001, 0.0001, 0.00001], dropout rate [0.1, 0.2, 0.3, 0.4, 0.5], and feature dimension of H&E/ST embeddings [256, 512]. Each hyperparameter combination was fed into the model, and the validation loss was calculated. The model checkpoint with the smallest evaluation loss is saved for testing. You can find the optimized hyperparameter settings in Tab. 3 and in the code under *config/{dataset}_{model_name}.yaml*.

D.2. Evaluation metrics.

Gene expression prediction. For each tile, we assess how accurately the model predicts the expression vector over the top G (G=50) highly variable genes. Let $y_{i,g}$ and $\hat{y}_{i,g}$ denote the ground-truth and predicted expression values of gene g in tile i. Tile-level MSE and PCC are computed across genes as follows:

$$\text{MSE}_i = \frac{1}{G}\sum_{g=1}^{G}(y_{i,g} - \hat{y}_{i,g})^2 \tag{1}$$

$$\text{PCC}_i = \frac{(\mathbf{y}_i - \bar{\mathbf{y}}_i)^\top(\hat{\mathbf{y}}_i - \bar{\hat{\mathbf{y}}}_i)}{\|\mathbf{y}_i - \bar{\mathbf{y}}_i\|_2 \, \|\hat{\mathbf{y}}_i - \bar{\hat{\mathbf{y}}}_i\|_2} \tag{2}$$

We report dataset-level performance as the mean $\pm$ standard deviation of MSE_i and PCC_i across all tiles.

Table 3. Hyperparameters for SIGMMA training.

Hyperparameters	Value
Image encoder input dimension	1024
Image encoder output dimension	512
Spatial encoder input dimension	384
Spatial encoder output dimension	512
Number of GNN layers	2
Dropout rate	0.1
Temperature τ	0.01
Batch size	128
Gradient accumulation steps	10
Effective batch size	1280
Number of epochs	100
Weight decay	1×10^{-3}
Learning rate (LR)	1×10^{-4}
LR scheduler	Step decay
LR step size	100
LR gamma	0.1
Optimizer	Adam

Cross-modal retrieval. For a query tile q from one modality (HE or ST), we rank all tiles in the other modality by embedding similarity. Retrieval accuracy is measured by whether the paired tile appears within the top-$p\%$. Recall@$p\%$ is defined as follows:

$$\text{Recall@}p\% = \frac{1}{N}\sum_{q=1}^{N} \mathbb{1}(y_q \in \text{TopK}(q)), \tag{3}$$

where N denotes the number of query tiles, $K = \lfloor p\% \times N \rfloor$ and $\mathbb{1}(\cdot)$ is the indicator function.

D.3. Fine-tuning of baselines.

All baselines were either fine-tuned on our standardized data splits or evaluated using their official checkpoints. Each model was fine-tuned using the loss function specified in the original implementation. For BLEEP, no pretrained BLEEP checkpoint is provided, and only the ResNet50 (He et al., 2016) backbone is publicly available; therefore, we initialized BLEEP with this backbone and fine-tuned the model on our dataset to ensure a consistent and fair comparison.

E. Result

E.1. Performance on gene expression prediction task

Here, our focus is to evaluate the quality of the learned image representation for gene expression prediction (Tab. 4). Therefore, to avoid introducing biases from different methods' gene expression decoders, we use the image embedding output by each method and apply a ridge regression for gene expression prediction for each method.

Table 4. Gene expression prediction across HEST1k and in-house datasets.

Dataset	*HEST1k-LUAD*		*HEST1k-PAAD*		*HEST1k-SKCM*		*HEST1k-IDC*		*in-house skin*	
Model	MSE (↓)	PCC (↑)	MSE (↓)	PCC (↑)	MSE (↓)	PCC (↑)	MSE (↓)	PCC (↑)	MSE (↓)	PCC (↑)
UNI	0.046±0.041	0.476±0.064	0.008±0.008	0.470±0.064	0.073±0.080	0.666±0.032	0.046±0.041	0.476±0.064	0.094±0.072	0.418±0.014
CLIP	0.052±0.052	0.467±0.088	0.009±0.010	0.245±0.081	0.080±0.066	0.541±0.018	0.052±0.052	0.467±0.088	0.103±0.084	0.330±0.022
PLIP	0.027±0.016	0.561±0.059	0.011±0.012	0.432±0.032	0.060±0.055	0.612±0.058	0.053±0.050	0.465±0.089	0.107±0.084	0.331±0.015
BLEEP	**0.011±0.011**	0.252±0.082	**0.004±0.008**	0.124±0.137	**0.012±0.006**	0.594±0.232	**0.004±0.003**	0.443±0.159	**0.035±0.008**	0.292±0.034
OmiCLIP	0.022±0.013	0.613±0.034	0.018±0.016	0.480±0.026	0.083±0.057	0.481±0.061	0.053±0.044	0.472±0.055	0.118±0.093	0.230±0.025
SIGMMA	0.015±0.007	**0.741±0.023**	0.015±0.015	**0.485±0.036**	0.051±0.048	**0.744±0.052**	0.051±0.043	**0.510±0.072**	0.060±0.032	**0.452±0.025**

E.2. Ablation study

Here, we present ablation results for the core components of the gene expression prediction task and cross-modal retrieval task (Tab. 5) on the HEST1k-LUAD dataset. Each component contributes to performance gains in both tasks: adding the multi-scale loss and graph sparsification progressively improves prediction/retrieval accuracy, and integrating all components yields the best overall performance.

Table 5. Ablation of core components of SIGMMA on HEST1k-LUAD for task 1 and task 2.

Task 1. Gene expression prediction.

Components			Task 1.	
Cell graph	Multi-scale	Graph sparsif.	MSE (↓)	PCC (↑)
			0.032±0.018	0.345±0.035
✓			0.039±0.018	0.268±0.032
✓	✓		0.020±0.014	0.645±0.046
✓	✓	✓	**0.015±0.007**	**0.741±0.023**

Task 2. Cross-modal retrieval.

Components			H&E → ST			ST → H&E		
Cell graph	Multi-scale	Graph sparsif.	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%
			0.517	0.694	0.768	0.529	0.673	0.780
✓			0.480	0.639	0.737	0.459	0.621	0.722
✓	✓		0.550	0.667	0.786	0.514	0.685	0.761
✓	✓	✓	**0.590**	**0.728**	**0.826**	**0.602**	**0.768**	**0.813**

We additionally explored (1) different H&E image backbones, ResNet50 (He et al., 2016), H-Optimus-0 (Saillard et al., 2024), and UNI (Chen et al., 2024), and (2) different ST backbones for cell embedding initialization, Harmony (Korsunsky et al., 2019), Novae (Blampey et al., 2024), and STEMO (Birk, 2025). These experiments assess the impact of backbone choices and support the architectural decisions made in our final model. SIGMMA, built upon the UNI vision backbone, consistently achieves the best performance across both downstream tasks (Tab. 6).

On the other hand, among the ST backbones, STEMO tended to perform well overall across both downstream tasks (Tab. 7). Accordingly, these results justify selecting UNI and STEMO as the vision and ST backbones in our framework.

Table 6. Ablation of vision backbones on HEST1k-LUAD for task 1 and 2.

Task 1. Gene expression prediction.

Model	MSE (↓)	PCC (↑)
ResNet50	0.052±0.047	0.365±0.079
SIGMMA (ResNet50)	0.031± 0.035	0.389±0.064
H-Optimus-0	0.035±0.034	0.512±0.078
SIGMMA (H-Optimus-0)	0.020±0.018	0.673±0.030
UNI	0.046±0.041	0.476±0.064
SIGMMA (UNI)	**0.015±0.007**	**0.741 ±0.023**

Task 2. Cross-modal retrieval.

Model	H&E → ST			ST →H&E		
	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%
SIGMMA (ResNet50)	0.086	0.199	0.260	0.135	0.229	0.333
SIGMMA (H-Optimus-0)	0.563	0.676	0.777	0.554	0.688	0.774
SIGMMA (UNI)	**0.590**	**0.728**	**0.826**	**0.602**	**0.768**	**0.813**

Table 7. Ablation of ST backbone in HEST1k-LUAD dataset for task 1 and 2.

Task 1. Gene expression prediction.

Model	MSE (↓)	PCC (↑)
SIGMMA (Harmony)	**0.005±0.005**	0.498±0.076
SIGMMA (Novae)	0.011±0.003	0.606±0.048
SIGMMA (STEMO)	0.015±0.007	**0.741±0.023**

Task 2. Cross-modal retrieval.

Model	H&E → ST			ST →H&E		
	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%
SIGMMA (Harmony)	0.529	0.723	0.820	0.526	0.730	**0.830**
SIGMMA (Novae)	0.039	0.086	0.148	0.049	0.099	0.148
SIGMMA (STEMO)	**0.590**	**0.728**	**0.826**	**0.602**	**0.768**	0.813

E.3. Attention map analysis

To probe what each model attends to in the input H&E tile, we extracted self-attention weights from the final transformer block ($L = 24$) of the UNI ViT-L/16 image encoder, comparing the pretrained UNI backbone (baseline) to the same backbone after SIGMMA fine-tuning. For each input tile, we visualize the class-token–to-patch attention distribution from each of the six attention heads in this last layer, reshaping the per-patch attention scores into a 2D map aligned to the input image. As an anatomical reference for nuclei-rich regions, we overlay the built-in Xenium cell segmentation as a blue contour on each input tile, allowing direct visual comparison of attended regions against cell-dense areas (Fig. 6).

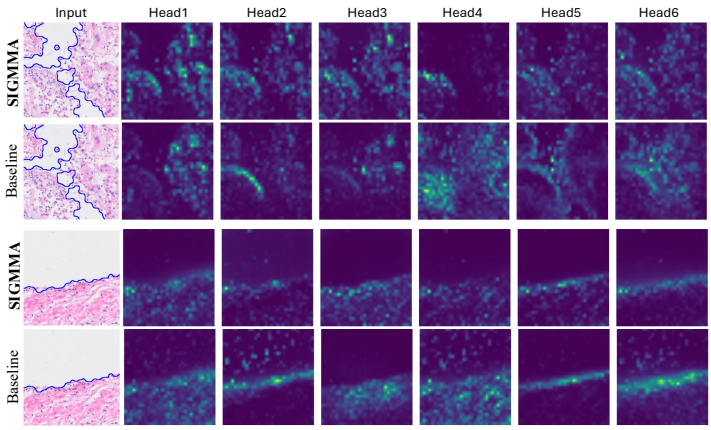

Figure 6. Attention maps from six attention heads in the last encoder layer ($L = 24$) of the UNI image encoder backbone fine-tuned with SIGMMA, illustrating class-token–to-patch attention distributions. Blue contour overlaid on the input images indicates the cell-segmentation mask, marking the boundaries of cell-rich regions.

E.4. Sensitivity analysis on tile size

We further conduct a sensitivity analysis on tile size, as tile resolution is a common source of variability in histopathology pipelines. Tile size 224 is a common choice in general vision models, while 256 is the standard image size used in many histopathology foundation models (Chen et al., 2024; Vorontsov et al., 2024; Saillard et al., 2024). As shown in Tab. 8, performance varies moderately across 224, 256, and 512 pixel tiles, but SIGMMA shows no collapse or strong dependence on any specific configuration. Mid-sized tiles (224/256 pixel) perform well across both tasks, while 512 pixel tiles show reduced performance, potentially due to increased heterogeneity within larger regions. Overall, this sensitivity analysis shows that while performance varies with tile resolution, SIGMMA remains generally robust, and medium tile sizes tend to provide favorable performance across tasks.

Table 8. Sensitivity analysis on tile size for task 1 and 2.

Task 1. Gene expression prediction.

Tile size	MSE (↓)	PCC (↑)
224 × 224	0.012±0.005	0.677±0.020
256 × 256	0.015±0.007	0.741±0.023
512 × 512	0.011±0.004	0.438±0.065

Task 2. Cross-modal retrieval.

	H&E → ST			ST → H&E		
Tile size	R@5%	R@10%	R@15%	R@5%	R@10%	R@15%
224 × 224	0.616	0.755	0.817	0.598	0.744	0.794
256 × 256	0.590	0.728	0.826	0.602	0.768	0.813
512 × 512	0.453	0.605	0.721	0.523	0.640	0.733

F. Discussions

F.1. Analysis of challenging cases

Understanding the PCC–MSE discrepancy. Although SIGMMA achieves high PCC across genes, indicating that it accurately captures the relative variation of expression across tiles, inspection of calibration plots reveals a consistent miscalibration in absolute prediction values. As shown in Fig. 7, the regression line has slopes < 1 and positive intercepts, indicating that the model underestimates variation while introducing a systematic bias. This global calibration mismatch increases MSE despite preserving rank-order consistency, explaining the discrepancy between PCC and MSE observed in Tab. 4.

Embedding similarity challenges IDC retrieval. As shown in Fig. 8, H&E embeddings in IDC from SIGMMA are highly homogeneous, making many tiles nearly indistinguishable and inherently limiting ST→HE retrieval. By contrast, CLIP produces more dispersed H&E embeddings for IDC, indicating greater apparent variability. This difference in feature distribution explains why ST→HE retrieval drops for SIGMMA specifically on IDC.

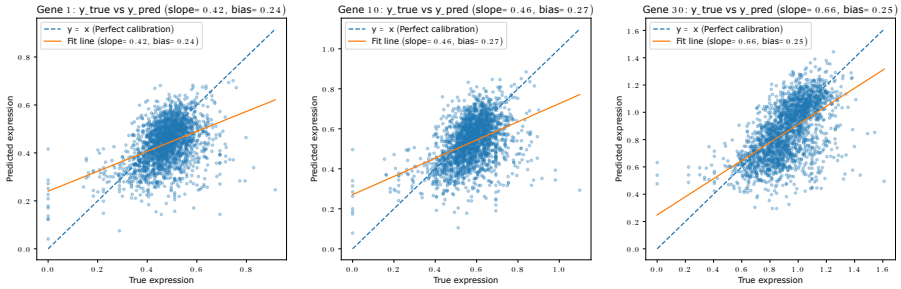

Figure 7. Calibration analysis of gene expression predictions.

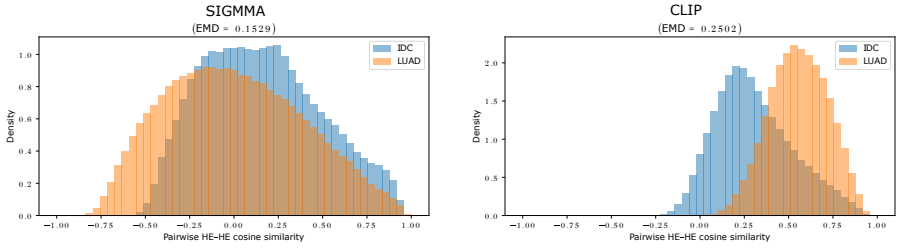

Figure 8. Cosine similarity of H&E embeddings.

F.2. Limitations and future work

Limitations. Although SIGMMA learns meaningful multi-modal representation, several limitations remain. The model's generalizability is constrained by the limited range and diversity of available paired H&E–Xenium ST datasets, hindering robust performance across heterogeneous tissue types. Additionally, because the approach relies on hierarchical spatial graphs constructed from cell segmentation, its effectiveness is inherently dependent on segmentation quality, which may be variable in complex tissue contexts.

Future work. Future extensions of SIGMMA include evaluating the framework on other single-cell–resolution spatial transcriptomics platforms (e.g., CosMx, MERSCOPE) to assess cross-platform robustness. In addition, SIGMMA's hierarchical design naturally scales to the WSI-level task, which we plan to explore as a next step. Finally, extending retrieval evaluation beyond within-tissue settings to cross-tissue scenarios may reveal how well the learned representations generalize across distinct morphological and molecular contexts.