SELF-SUPERVISED GRAPH-BASED REFINEMENT OF SPATIAL TRANSCRIPTOMICS AT SUBSPOT RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics enables the study of gene expression within tissue context, but current sequencing-based technologies operate at a spatial resolution where each measurement aggregates signals from multiple cells. While deconvolution methods aim to estimate each spot's cell-type composition, they do not provide fine-grained spatial structure. We propose *Square*, a self-supervised framework based on graph attention networks to infer plausible subspot-level gene expression patterns from spatial transcriptomics data. The method leverages spatial neighborhood structure through message passing and introduces virtual nodes that represent latent subspot locations, enabling the redistribution of observed transcriptomic signals into finer spatial units. Importantly, our approach operates without requiring external references such as single-cell RNA-seq data or histological images. Rather than recovering true cellular-resolution measurements, *Square* aims to infer spatially consistent fine-grained structure that is compatible with the observed data and underlying spatial dependencies. Across synthetic, ST, Visium and Visium HD samples, we show that the inferred subspots exhibit improved spatial coherence, reduced mixing of cell-type signals, and enhanced alignment with known biological organization, while mitigating sparsity. These results suggest that graph-based modeling of spatial context can effectively refine coarse spatial transcriptomics data, providing a principled approach to studying fine-scale tissue organization when direct cellular-resolution measurements are unavailable.

1 INTRODUCTION

The advent of spatial transcriptomics has enabled the precise mapping of gene expression profiles within tissue architecture (Moses and Pachter, 2022), providing key insights into cellular composition and disease processes (Piwecka et al., 2023; Arora et al., 2023). While imaging-based spatial transcriptomic technologies can yield subcellular transcriptional resolution, they are constrained to a few thousand genes (Rao et al., 2021). In contrast, sequencing-based technologies provide an unbiased view of the transcriptional architecture of the tissue, but at the cost of cellular resolution, as their minimal spatial unit can contain several cells. To address this, several computational methods, termed deconvolution methods, have been developed for deciphering the spot composition of different cell types (Ma and Zhou, 2022; Andersson et al., 2020; Zhou et al., 2023; Lopez et al., 2022). However, recent studies have reported a lack of consistency among them (Yan and Sun, 2023).

The recent release of Visium HD (Oliveira et al., 2025) has partially alleviated the resolution limitations of sequencing-based technologies by moving from Visium's $55\mu m$-diameter spots to $2 \times 2\mu m$ bins. However, this increased spatial granularity comes at the expense of significantly sparser gene expression profiles, as fewer transcripts are captured per location (Kamel et al., 2025).

Consequently, a central challenge in spatial transcriptomics is not only to characterize the composition of coarse spatial units, but also to infer finer-grained spatial structure from noisy aggregated measurements and the observed spatial context.

Reference-free spatial enhancement. Many current methods for spatial refinement rely on side information such as (paired) single-cell RNA-seq data (scRNA-seq), histology images, or an assumed number of cell types, which can improve performance when available but can restrict applicability and introduce biases (Cang and Nie, 2020; Hu et al., 2023; Zhang et al., 2024). In practice, such

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.

auxiliary data are often either unavailable, imperfectly matched to the tissue under study, or weakly related to local gene expression patterns. For example, the extent to which image intensities directly predict transcriptional variation remains uncertain (Li et al., 2024). This motivates a need for spatial refinement methods that rely only on the measured transcriptomic signal and spatial coordinates, enabling broader applicability across tissues and platforms.

Contributions. To address this challenge, we propose *Square*, a self-supervised framework based on graph attention networks that leverages spatial dependencies to infer fine-grained spatial gene expression patterns without external references such as scRNA-seq or histology. Key components include virtual nodes representing latent subspots within each spot and the use of a variational autoencoder to account for sparsity and distribution shift across resolutions. *Square* extends the idea of deconvolution by redistributing each spot's observed transcriptomic signal across spatially coherent latent subspots, modeling sparsity and local spatial context while remaining consistent with the observed data and neighboring regions. Across synthetic and real datasets spanning ST, Visium, and Visium HD, *Square* produces: i) subspot-level transcriptional signals with improved spatial organization within original spots, ii) reduced mixing of cell-type signals, and iii) biologically meaningful patterns that were missed at the original resolution. Thus, our results supports self-supervised graph-based spatial refinement as a principled strategy for studying fine-scale tissue organization when direct cellular-resolution measurements are unavailable.

2 Towards Fine-Grained Spatial Resolution with *Square*

Tissue organization is not stochastic and generally follows structured biological patterns, such as immune infiltration or cell proliferation, as cells of the same cell type typically assemble together (Shah et al., 2016; Stoltzfus et al., 2020; Russ et al., 2021). Evidence of spatial autocorrelation of marker genes (Ma and Zhou, 2022) further substantiates this hypothesis. *Square* builds on this phenomenon to infer fine-grained spatial structure by redistributing transcriptomic signal across latent subspots within each original spot, leveraging information from the spot itself, and its neighbors.

Problem statement. A sequencing-based spatial transcriptomic experiment generates a gene expression matrix $\mathbf{X} \in \mathbb{R}^{G \times S}$, containing the expression of G genes across S spots, and a coordinate matrix $\mathbf{Z} \in \mathbb{R}^{2 \times S}$, containing the coordinates (x, y) of each spot. Our goal is to infer a finer-grained spatial representation, corresponding to an increased effective spatial resolution, by decomposing each spot into k latent subspots, yielding an expanded expression matrix $\mathbf{X}_e \in \mathbb{R}^{G \times kS}$ and the corresponding coordinates $\mathbf{Z}_e \in \mathbb{R}^{2 \times kS}$. These subspots are confined to the spatial extent of their originating spot and represent a partition of the original measurement, rather than newly observed locations. In what follows, k is set to 9 to guarantee an isotropic expansion of both x and y axes.

2.1 Overview of *Square*

Pooling-based self-supervision. To generate the training set, $k = 9$ initial spots are grouped into *super-spots* through a *pooling mask* and summing their transcriptional signals (App. Fig. 5A). Thus, each training instance consists of a super-spot and its m neighboring spots ($m = 16$ by default), which provide (self-)supervision by using the original pooled spots as ground truth. During training, virtual nodes are introduced within each super-spot as proxies for the subspots that we aim to predict (App. Fig. 5B). An attention mechanism then leverages neighboring information to predict the spatially structured subspot-level gene expression signal, enabling the decomposition of aggregated signals into finer-grained components.

Robust subspot inference under distribution shift. At inference time, each original spot is decomposed into a set of latent subspots using its m surrounding neighbors ($m = 8$ by default; App. Fig. 5C). This, however, leads to a distribution shift between training and inference, as the central node used during training (the *super-spot*) has a significantly denser profile due to the pooling. To address this, *Square* incorporates a variational autoencoder (VAE) (Kingma and Welling, 2013) that learns a latent representation of the spots, capturing the underlying biological signal while mitigating sparsity effects. Probabilistic frameworks are widely used in (single-cell) omics data to model uncertainty, sparsity, and technical variability (Lopez et al., 2018; 2022; Bergenstråhle et al., 2022). Additionally, the difference in the number of neighbors between training and inference is

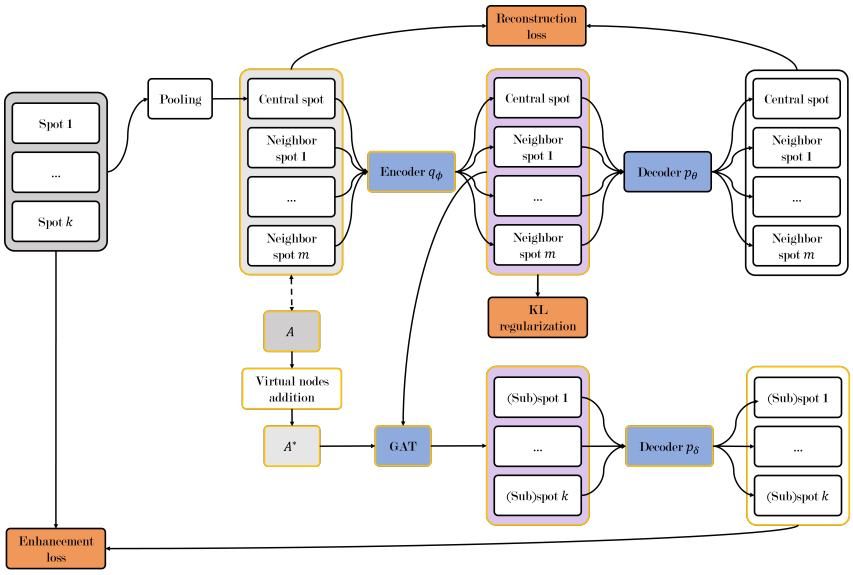


Figure 1: Overview of *Square*: Gray boxes represent inputs (dark and light gray represent strict inputs or generated by deterministic operations, respectively), white boxes the generated values, blue boxes the trainable modules and orange boxes the terms of the loss function. Yellow boxes show the procedure to generate subspots during inference. Purple boxes represent the resolution-agnostic low-dimensional space where the generated data inside them are located.

handled via the inductive learning provided by graph attention networks (GAT), which can operate on a variable number of nodes at inference time (Fig. 1).

Next, we describe the different components of *Square* in detail.

2.2 Graph-based Spatial Representation

Each training data instance is represented as a graph $\mathcal{G}$ comprising a central spot and its m neighboring spots. Following recent works on learning on graphs (Qian et al., 2024; Pham et al., 2017), we introduce k virtual nodes representing latent subspots within each spot. These nodes are connected to the central node, to the original neighbors, and are also interconnected among themselves (App. Fig. 5B-C). The spatial structure is thus represented by the adjacency matrix A of $\mathcal{G}$ as:

$$A^*_{ij} = \begin{cases} d(s_i, s_j)^{-1} & \text{if } s_i \text{ and } s_j \text{ are both original or both virtual nodes} \\ t & \text{if } s_i \text{ and } s_j \text{ correspond to the central spot and a virtual node} \\ 1 & \text{if } s_i \text{ and } s_j \text{ correspond to a neighbouring spot and a virtual node} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $d(s_i, s_j)$ denotes the Euclidean distance between the coordinates of (super)spots s_i and s_j. The coordinates of the virtual nodes are located between the central and neighboring nodes to mirror the original topology defined by $\mathbf{Z} \in \mathbb{R}^{2\times S}$, and $t \geq 1$ is a model hyperparameter.

2.3 Mitigating the Training-Inference Distribution Shift

To address the sparsity differences between super-spots and the original spots, we use a standard VAE architecture where both the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$ are parametrized as neural networks with non-linear activations, and the latent space is regularized towards a standard normal distribution. This formulation enables the model to learn a smooth, low-dimensional latent representation of the (super)spot data, capturing the dominant biological signal while reducing the impact of sparsity, noise, and batch effects (Lopez et al., 2018) (see App. A.4.4).

In our set-up, nodes in the graph $\mathcal{G}$ are represented by the VAE-encoded latent variables $\mathbf{z} \in \mathbb{R}^d$ as node features (with d set to 50% of the number of genes). The encoding process is independent of

the graph structure, so each spot is embedded based solely on its own transcriptomic profile. The graph $\mathcal{G}_{\mathbf{z}_\phi}$ is populated as follows:

$$\mathbf{z}_\phi = \{\mathbf{z}^{(i)}\}_{i=1}^{1+k+m} \sim q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}), \tag{2}$$

where k and m denote the number of virtual nodes and neighboring spots, respectively.

Virtual nodes' feature initialization. The transcriptomic profile of a virtual node is initialized by averaging the transcriptomic profiles of its connected spots in the graph (approach selected experimentally, App. A.4.2). These profiles are then mapped to the latent space using the VAE encoder $q_\phi(\mathbf{z}|\mathbf{x})$, enabling their integration into the graph-based inference process.

2.4 Graph Attention-Based Subspot Inference

Square builds on the capacity of Graph Neural Networks (GNNs) to infer novel embeddings for each node by leveraging both the topology of the graph and the features of neighboring nodes. Specifically, *Square* relies on a Graph Attention Network (GAT) (Brody et al., 2021), which assigns attention coefficients to neighboring nodes in order to learn their relative importance. This allows the model to compute a weighted aggregation of the representations of the neighbors $\mathcal{N}(i)$ of a node i, rather than assuming uniform contributions. This mechanism is defined through an edge scoring function $e : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that captures the importance of each neighbor:

$$e(\boldsymbol{h}_i, \boldsymbol{h}_j) = \text{LeakyReLU}\left(\boldsymbol{a}^{\mathrm{T}} \cdot [\boldsymbol{W}_\delta \boldsymbol{h}_i \parallel \boldsymbol{W}_\delta \boldsymbol{h}_j]\right), \tag{3}$$

where $\boldsymbol{a} \in \mathbb{R}^{2d'}$ is a learnable attention vector, $\boldsymbol{W}_\delta \in \mathbb{R}^{d' \times d}$ stands for the standard GNN (learnable) parametric function, and $\parallel$ denotes concatenation. The resulting scores are then normalized across all neighbors $j \in \mathcal{N}(i)$ and used to weight their contributions. The updated node representations after $l+1$ layers are then computed as:

$$\boldsymbol{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \text{softmax}_j\left(e\left(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_j^{(l)}\right)\right) \cdot \boldsymbol{W}_\delta \boldsymbol{h}_j\right), \tag{4}$$

where σ denotes a non-linear activation function. Finally, the updated embeddings of the virtual nodes are passed through a multi-layer perceptron (MLP) with parameters θ (the VAE decoder) to predict subspot-level gene expression in the original feature space.

2.5 Training *Square*

The VAE produces latent representations for all nodes (i.e., original spots, superspots, and virtual nodes), which are then used by the GAT to learn spatial dependencies across the graph. Since the VAE (encoder and decoder with parameters ϕ and θ) is trained jointly with the GAT (with parameters δ), the objective function combines two log-likelihoods: one corresponding to the VAE decoder $p_\theta(\mathbf{x}|\mathbf{z})$, and another corresponding to the graph-based prediction $p_\delta(\mathbf{x}|\mathbf{z}, \mathbf{A})$. In addition, a Kullback–Leibler (KL) divergence term regularizes the latent space by encouraging the encoded distribution $q_\phi(\mathbf{z}|\mathbf{x})$ to match the prior $p(\mathbf{z}) \sim \mathcal{N}(0, I)$. The resulting objective we maximize is:

$$\mathcal{L}_{\phi,\theta,\delta} = \mathcal{L}_{\phi,\theta}^{\text{recons}} + \mathcal{L}_{\phi,\delta}^{\text{enhance}} - \mathcal{L}_{\phi}^{\text{KL}}, \tag{5}$$

with

$$\mathcal{L}_{\phi,\theta}^{\text{recons}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]; \mathcal{L}_{\phi,\delta}^{\text{enhance}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\delta(\mathbf{x}|\mathbf{z}, \mathbf{A})]; \mathcal{L}_{\phi}^{\text{KL}} = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})).$$

The term $p_\delta(\mathbf{x}|\mathbf{z}, \mathbf{A})$ reconstructs the transcriptomic signal from the original spots used to form each superspot, enabling supervision through the decomposition of known aggregated measurements. In contrast, $p_\theta(\mathbf{x}|\mathbf{z})$ reconstructs gene expression in a resolution-agnostic manner.

Training data is generated by applying the *pooling mask* across the entire tissue slide and subsequently split into training and validation sets (80%/20%). The split is performed in a balanced manner to ensure comparable mean spatial autocorrelation between the two sets. Superspots with less than 2 neighbors are omitted during training. Model optimization is carried out using the *Adam* optimizer (Kingma and Ba, 2017), with a learning rate scheduler (patience of 10 epochs) and a batch size of 32.

Final configuration. Model architecture optimization was performed with Optuna (Akiba et al., 2019) for a total of 500 trials. The list of tested hyperparameters and the selected configuration are detailed in App. A.1.3. *Square*'s pseudo-code is provided in App. A.1.2.

3 Related work

Computational approaches for modeling spatial transcriptomics data can be broadly categorized into deconvolution-based methods and spatial refinement methods.

Deconvolution methods. Methods such as CARD (Ma and Zhou, 2022) and DestVI (Lopez et al., 2022) aim to estimate the cell-type composition of each spot by leveraging reference datasets or probabilistic models. While these approaches provide insight into the mixture of cell types within each spatial unit, they do not explicitly model fine-grained spatial structure or decompose spots into subspot-level representations. Some methods, such as CARD, incorporate spatial priors and can be used to impute values at unobserved locations, but their primary focus remains on estimating cell-type proportions rather than modeling intra-spot spatial organization.

Spatial refinement methods. Several approaches have been proposed to enhance the spatial granularity of transcriptomics data without relying on external references. BayesSpace (Zhao et al., 2021) performs spatial clustering followed by refinement in a low-dimensional space, while STAGE (Li et al., 2024) learns a mapping between gene expression and spatial coordinates to generate intermediate locations. Although these methods attempt to recover finer spatial patterns, they often rely on clustering assumptions, fixed-resolution schemes, or lack robustness across datasets. Note that resolution-enhancement methods relying on H&E images (Zhang et al., 2024; 2025; Hua et al., 2025; Zhao et al., 2024) are not benchmarked herein.

Taken together, existing approaches either estimate cell-type composition without modeling intra-spot spatial structure, or perform spatial refinement without explicitly decomposing aggregated measurements, limiting their ability to capture fine-grained spatial organization.

Our approach. To address this, *Square* formulates subspot inference as a graph-based learning problem, leveraging spatial dependencies through attention mechanisms. The method operates in a self-supervised manner and explicitly models the decomposition of aggregated measurements into latent subspot representations.

4 Experiments

4.1 Datasets and evaluation metrics

Simulated spot-level profiles. In this dataset, spot-level counts (dimension 100) are sampled from one of 5 negative binomial distributions, without enforcing a specific biological structure (App. Fig. 8A,B, pseudo-code in App. A.2.1). We generate a dense 90×90 square grid, mimicking that of ST. To generate ground truth, groups of 3×3 spots (without overlap) are aggregated into low-resolution spots. This dataset provides a controlled, albeit simplified, setting for quantitative evaluation.

Model-based Synthetic Mouse Olfactory Bulb (MOB). We use the simulator *scDesign3* (Song et al., 2024), based on a 10X Chromium scRNA-seq dataset (Tepe et al., 2018) and an ST dataset (Ståhl et al., 2016) from a MOB sample. The dataset contains 278 spots with gene expression vectors of dimension 182, comprising mixtures of up to four cell types. While no subspot-level ground truth is available, we expect inferred representations to preserve the cell-type proportions of the original spots. See App. A.2.2 for details of the dataset generation process.

Human Glioblastoma (GBM). We included a 10x Visium whole-transcriptome spatial transcriptomics dataset of a fresh-frozen human glioblastoma multiforme tissue section from a single donor (10x Genomics, 2020), comprising 3,468 spots with a median of 4,326 genes, generated using the Visium Spatial Gene Expression platform and sequenced on an Illumina NovaSeq 6000.

Mouse Small Intestine (MSI). To evaluate whether inferred subspots capture fine-grained biological structure, we use a Visium HD dataset from mouse small intestine (10x Genomics, 2024), which

provides measurements at near-cellular scale. We restrict analysis to a region of interest containing three intestinal villi (App. Fig. 10A), enabling tractable computation while preserving biological structure. An aligned H&E image is used for preprocessing. See App. A.2.3 for details.

Data preprocessing: Raw counts are: (i) normalized to account for sequencing depth, (ii) log-transformed to stabilize the variance, and (iii) highly variable genes (500 by default) as selected to reduce noise and dimensionality. For Visium HD, we perform cell segmentation on the H&E image. Bins whose centers fall within each cell polygon are pooled to obtain cell-level profiles, and cells containing multiple bins are discarded. *Square* is applied exclusively to bin-level data (without leveraging any information from the image). This aggregation step is performed only for evaluation.

Evaluation metrics. We evaluate *Square* using quantitative metrics across three scenarios: (i) simulated spot-level data with full ground truth, (ii) MOB data with known cell-type proportions, and (iii) real datasets without ground truth. In scenario (i), we use spot-wise distance and similarity metrics, and the adjusted rand index (ARI). In scenario (ii), we apply k-means clustering to inferred subspots and compare the entropy of cluster-frequency vectors with true cell-type proportions, also reporting RMSE between them. In scenario (iii), we assess biological consistency and changes in cell-type purity. Results are averaged over five runs. See App. A.3 for details.

4.2 Ablation study

To assess the contribution of three key components of *Square*, we consider the following variants: *i)* without virtual nodes, where only original nodes are updated and subspot representations are obtained by averaging neighboring node outputs; *ii)* without the VAE, using raw gene expression as node features; and *iii)* without attention, replacing the GAT with a *vanilla* GNN. Experiments are conducted on the *simulated* dataset.

All ablated variants underperform *Square* across most metrics (Table 10). Although the model without VAE achieves the lowest MSE, it performs worse on all other metrics and exhibits the highest variance, indicating that ignoring the distribution shift negatively impacts overall performance (see App. A.4.4). This highlights that MSE alone does not fully capture the quality of inferred spatial structure. Qualitative differences further support these findings. Incorporating attention enables heterogeneous subspot representations within each original spot, leading to more faithful preservation of spatial boundaries. Models without attention exhibit uniform message propagation, leading to a lack of smoothness in the reconstructed expression (App. Fig. 14A, left panels), whereas attention-based models produce more consistent boundaries that better reflect the underlying spatial structure (App. Fig. 14A, middle panels). Further validation of the attention coefficients computed by *Square* is included in App. A.4.5. Besides, not only virtual nodes contribute to more accurate reconstruction, but provide also more flexibility for settings with $k \neq 9$, where aggregating neighbors' updated features is not straightforward. Notably, similar results hold if the *simulated* dataset is generated under varying sparsity levels, controlled by the zero-inflation parameter π (see App. A.4.1).

Overall, these results highlight the complementary roles of the VAE (robustness to sparsity), attention (modeling spatial heterogeneity), and virtual nodes (enabling subspot-level decomposition).

Table 1: **Ablation Study:** Performance comparison across model variants.

Model	MSE ↓	SCD ↓	$JD_{>\mu}$ ↓	JD_{30} ↓	ARI ↑
w/o virtual nodes	$\mathbf{0.29 \pm 0.10}$	0.423 ± 0.005	0.408 ± 0.006	0.5918 ± 0.0024	0.80 ± 0.03
w/o VAE	2.31 ± 0.69	0.511 ± 0.042	0.443 ± 0.027	0.6028 ± 0.0111	0.59 ± 0.24
w/o attention	1.18 ± 0.27	0.462 ± 0.031	0.402 ± 0.014	0.5953 ± 0.0093	0.65 ± 0.37
Square	0.39 ± 0.08	$\mathbf{0.412 \pm 0.012}$	$\mathbf{0.396 \pm 0.005}$	$\mathbf{0.5848 \pm 0.0018}$	$\mathbf{0.88 \pm 0.08}$

4.3 Benchmarking on simulated spot-level profiles

We next compare *Square* with BayesSpace, STAGE and CARD (implementation details in App. A.6). Since STAGE generates four subspots per each initial spot but the ground truth consists of nine subspots, spot-wise metrics have been tailored by averaging the closest neighbors in the ground truth.

Nevertheless, we do not compute ARI metrics since averaged profiles may not capture the spatial distribution of cells faithfully (e.g., oversmoothed boundaries between cell-type-specific regions).

Square outperforms competing methods in most evaluation metrics (Table 2), achieving a substantially lower MSE and the best performance in JD metrics, while maintaining competitive SCD values. Notably, *Square* is the only method that achieves meaningful agreement with the ground-truth cell-type assignments (ARI), whereas competing methods perform at near-random levels.

Among the compared methods, BayesSpace exhibits the closest performance, but remains consistently outperformed. CARD produces expression values outside the original data range, leading to high MSE despite preserving relative biological trends. STAGE shows poor robustness, with high variance across runs, likely due to the mismatch between the number of generated and ground-truth subspots.

Overall, these results highlight the ability of *Square* to infer subspot-level representations that better reflect both the transcriptional profiles and the underlying spatial organization of the data.

Table 2: **Benchmarking on the simulated dataset:** Performance comparison.

Method	MSE ↓	SCD ↓	$JD_{>\mu}$ ↓	JD_{30} ↓	ARI ↑
BayesSpace	2.33 ± 0.00	$\mathbf{0.410 \pm 0.000}$	0.408 ± 0.000	0.5867 ± 0.0002	0.67 ± 0.00
STAGE	40.18 ± 67.57	0.607 ± 0.325	0.484 ± 0.129	0.6451 ± 0.0074	-
CARD	9.66 ± 0.00	0.430 ± 0.002	0.448 ± 0.001	0.5988 ± 0.0005	0.00 ± 0.00
Square	$\mathbf{0.39 \pm 0.08}$	0.412 ± 0.012	$\mathbf{0.396 \pm 0.005}$	$\mathbf{0.5848 \pm 0.0018}$	$\mathbf{0.88 \pm 0.08}$

4.4 *Square* CAPTURES FINE-GRAINED SPATIAL ORGANIZATION BETTER THAN COMPETING METHODS

The *synthetic MOB* dataset provides a more realistic setting, with two interior regions dominated by specific cell types, yielding low-entropy spots; and a more heterogeneous exterior region with higher-entropy spots (Fig. 2A, App. Fig. 9). To evaluate fine-grained structure, we cluster the inferred subspots (k-means, $k = 6$), and compute the entropy on cluster-frequency vectors (Fig. 2B).

Spearman correlations between spot entropies across coarseness levels (k) show a maximum of 0.35 ($k = 6$; p-value $< 10^{-5}$; Fig. 2D). While the correlation is bounded by the differing nature of the variables (continuous cell-type entropy versus discrete cluster-based entropy), it remains consistently higher for *Square* compared to STAGE, BayesSpace, and CARD (Fig. 2E). We further annotate inferred subspots (Fig. 2C; App. Fig. 11-12) to construct cell-type frequency vectors per spot, which are compared to the ground truth. The resulting RMSE values for *Square* are concentrated around 0.2, whereas competing methods show higher error distributions (Fig. 2F).

Overall, these results indicate that *Square* effectively captures fine-grained spatial organization, producing subspot-level representations that better reflect both transcriptional heterogeneity and spatial structure compared to existing methods.

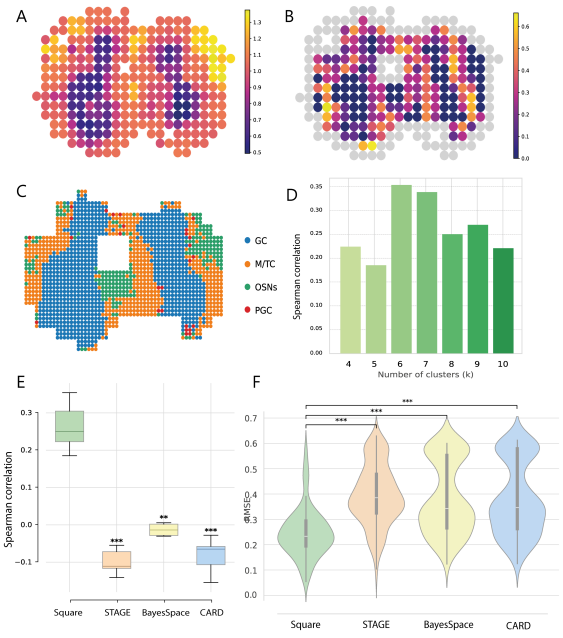


Figure 2: **A.** Spatial entropy of true cell-type proportions. **B.** Cluster-frequency entropy ($k = 6$). **C.** Enhanced sub-spots. **D-E.** Spearman correlation between entropies of *cell-type proportion* and *cluster-frequency vectors* for $k \in \{4, 10\}$ for *Square* (**D**), and against competing methods (**E**) (Mann-Whitney U test). **F.** RMSE distribution of true cell-type and inferred proportions across spots for *Square* and competing methods (t-test).

5 *Square* UNCOVERS BIOLOGICALLY MEANINGFUL INFORMATION

5.1 RECOVERING HYPOXIA-DRIVEN SPATIAL ORGANIZATION IN GBM

We build upon the findings of Greenwald et al. (2024), which propose a hypoxia-driven organizational model of GBM comprising five distinct layers. Their analysis demonstrated that the heightened cellular organization associated with hypoxia is not limited to hypoxic cell populations, but also extends to neighboring non-hypoxic cells (referred to as *MES-Hyp*). To quantify organization, the authors introduced the notion of spatial coherence for metaprogrammes (MPs), defined based on the number of neighboring spots sharing the same MP. By constructing subsamples as sets of concentric rings centered on each spot, this metric enabled the classification of spots into *structured* or *disorganized* states, further stratified by malignancy level (computed from the CNA signal) into four classes: *struct-malig*, *disorg-malig*, *struct-norm* and *disorg-norm*.

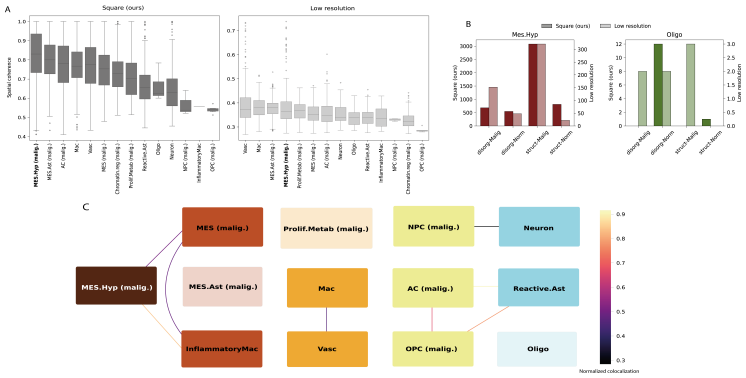


Figure 3: **A.** Spatial coherence for all metaprogrammes (in descending order (original data and *Square*-derived). **C.** Proportions of organization and malignancy division for *MES-Hyp* and *Oligo*. **D.** Organizational model inferred from the consensus set of interactions using *Square*-derived data.

We applied *Square* to the GBM dataset using the 562 genes defining the 14 proposed MPs, and their corresponding signatures to annotate the spots. The resulting representations exhibit clearer separation between MPs in terms of spatial coherence (Fig. 3A), aligning with the expected differences in their organizational behavior. In particular, the increased organization associated with hypoxia becomes more apparent when analyzing *Square*'s inferred subspot-level structure. *Square*-derived data also maintains the proportions of organizational and malignancy classes consistently observed on other GBM samples (Fig. 3B). These include the enrichment of *Oligo* in *disorg-norm* regions and the concentration of *MES-Hyp* in structured tumor cores. Notably, the higher frequency of *Oligo* is not clearly observable at the original spot level but becomes evident in the inferred representations.

Finally, we evaluate spatial relationships between cell states using the coupling measures (regional composition, adjacency, and colocalization) proposed in Greenwald et al. (2024) (see App. A.3.3). The consensus interaction network derived from *Square* preserves the layered structure reported in the original study (Fig. 3C), with no strong interactions between non-consecutive layers. This provides further evidence that the inferred subspot representations capture meaningful spatial organization consistent with known biological structure.

5.2 *Square* PRESERVES BIOLOGICALLY MEANINGFUL ORGANIZATION IN A HIGH-RESOLUTION SETTING

By aggregating 2-μm bins into 8-μm bins on the Visium HD MSI dataset, we assess *Square*'s ability to recover fine-grained spatial structure from coarser measurements ($k = 16, m = 8$). We further examine whether the inferred representations preserve biologically meaningful organization along the crypt–villus axis, a canonical differentiation trajectory in the intestinal epithelium. Along this axis, proliferative progenitor cells in the crypt progressively differentiate into mature enterocytes toward the villus tip (Bahar Halpern et al., 2020). We leverage prior studies, notably Mariadason et al. (2005), which characterized transcriptional programs underlying epithelial maturation. These results show that genes involved in proliferation and *Wnt* signaling are enriched in crypt regions, whereas differentiation-associated genes increase toward the villus. Using this framework, we evaluate whether *Square* preserves the expected spatial gradients. We define a continuous spatial coordinate by computing geodesic distances from manually annotated crypt seeds, yielding a crypt-villus axis aligned with tissue morphology (App. A.3.3).

Overall, the cellular profiles obtained from *Square*-derived bins exhibited higher count levels than the original 2 μm bins. Analysis of the highly variable genes and selected differentially expressed genes along the crypt–villus axis revealed enhanced expression patterns, particularly for: (i) *Max*, a transcription factor with peak expression in the villus region; and (ii) *Cldn2*, which encodes the claudin-2 transmembrane protein involved in epithelial tight junctions and is a known *Wnt* target, leading to reduced expression toward the villus tip (Mariadason et al., 2005) (Fig. 4A). Importantly, these expression profiles recapitulate the expected spatial gradient along the axis, a pattern that is less evident in the original data (Fig. 4B). Leveraging fold-change (FC) measurements across ten stratified regions along the axis, we further evaluate the relationship between each gene's FC and the correlation between its expression and the distance to the crypt. As expected, a positive association is observed in both the raw and *Square*-derived data, showcasing *Square*'s ability to preserve correlation structure (Fig. 4C).

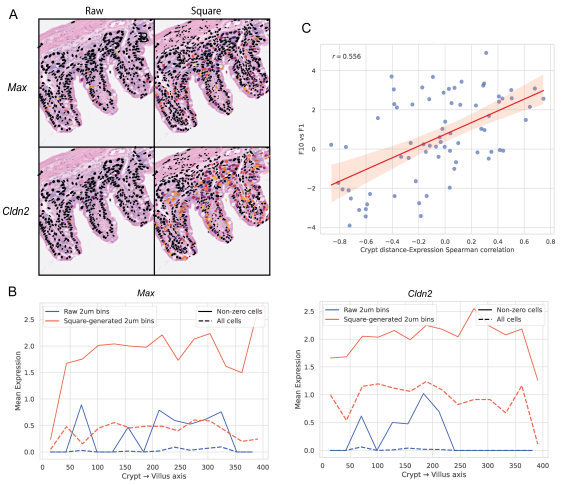


Figure 4: **A.** Spatial maps of segmented cells generated from raw 2 μm (left) and *Square*-derived bins (right). **B.** Mean expression profiles of *Max* (left) and *Cldn2* (right) along the crypt–villus axis. x-axis: geodesic distance to the crypt seeds. **C.** Scatter plot with each point representing a gene. x-axis: Spearman correlation between gene expression and geodesic distance to the crypt; y-axis: FC in expression between regions 1 (villus) and 10 (crypt) (Mariadason et al. (2005)).

By anchoring our analysis to established crypt–villus gene expression programs, these results indicate that *Square* preserves spatial organization and reveals known differentiation-associated transcriptional gradients that are less discernible at the original resolution. This supports its ability to retain biologically meaningful signal while improving the interpretability of spatial gene expression patterns.

These results highlight the applicability of *Square* to Visium HD data, where high spatial granularity is accompanied by substantial sparsity, demonstrating its ability to recover coherent spatial patterns and improve the interpretability of gene expression signals.

6 Conclusion

We introduced *Square*, a self-supervised framework that leverages graph attention networks to infer fine-grained spatial gene expression patterns from aggregated spatial transcriptomics data. By modeling spatial dependencies and decomposing each spot into latent subspots, the method enables a refined representation of spatial structure without requiring external references.

Our results show that this approach effectively captures spatial heterogeneity, reduces signal mixing, and reveals biologically meaningful patterns that are less apparent at the original resolution. Overall, *Square* provides a principled framework for studying fine-scale tissue organization from coarse spatial measurements, facilitating improved interpretation of spatial transcriptomics data.

7 Limitations

While *Square* infers fine-grained spatial structure, subspots remain model-based estimates rather than true cellular-resolution measurements. Performance may decrease in heterogeneous tissues or where long-range interactions dominate, though robustness analyses showed only weak dependence on spatial autocorrelation ($R^2 = 0.08$; App. A.4.6). See further details in App. A.5.

REFERENCES

10x Genomics. Human glioblastoma multiforme visium spatial gene expression dataset. Spatial Gene Expression dataset generated using Visium technology, Oct. 2020.

10x Genomics. Visium hd spatial gene expression library, mouse small intestine (ffpe). https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-intestine, 2024. Analyzed with Space Ranger v3.0; Visium CytAssist platform; accessed 2026-05-04.

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

A. Andersson, J. Bergenstråhle, M. Asp, L. Bergenstråhle, A. Jurek, J. Fernández Navarro, and J. Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):565, 2020.

R. Arora, C. Cao, M. Kumar, S. Sinha, A. Chanda, R. McNeil, D. Samuel, R. K. Arora, T. W. Matthews, S. Chandarana, et al. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature communications*, 14(1):5029, 2023.

K. Bahar Halpern, H. Massalha, R. K. Zwick, A. E. Moor, D. Castillo-Azofeifa, M. Rozenberg, L. Farack, A. Egozi, D. R. Miller, I. Averbukh, et al. Lgr5+ telocytes are a signaling source at the intestinal villus tip. *Nature communications*, 11(1):1936, 2020.

L. Bergenstråhle, B. He, J. Bergenstråhle, X. Abalo, R. Mirzazadeh, K. Thrane, A. L. Ji, A. Andersson, L. Larsson, N. Stakenborg, et al. Super-resolved spatial transcriptomics by deep data fusion. *Nature biotechnology*, 40(4):476–479, 2022.

S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):2084, 2020.

A. C. Greenwald, N. G. Darnell, R. Hoefflin, D. Simkin, C. W. Mount, L. N. G. Castro, Y. Harnik, S. Dumont, D. Hirsch, M. Nomura, et al. Integrative spatial analysis reveals a multi-layered organization of glioblastoma. *Cell*, 187(10):2485–2501, 2024.

J. Hu, K. Coleman, D. Zhang, E. B. Lee, H. Kadara, L. Wang, and M. Li. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with tesla. *Cell systems*, 14(5):404–417, 2023.

Y. Hua, Y. Zhang, Z. Guo, S. Bian, and Y. Zhang. Imspire: image feature-aided spatial resolution enhancement method. *Science China Life Sciences*, 68(1):272–283, 2025.

M. Kamel, Y. Song, A. Solbas, S. Villordo, A. Sarangi, P. Senin, M. Sunaal, L. C. Ayestas, C. Levin, S. Wang, et al. Enact: End-to-end analysis of visium high definition (hd) data. *Bioinformatics*, 41 (3):btaf094, 2025.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

S. Li, K. Gai, K. Dong, Y. Zhang, and S. Zhang. High-density generation of spatial transcriptomics with stage. *Nucleic Acids Research*, 52(9):4843–4856, 2024.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

R. Lopez, B. Li, H. Keren-Shaul, P. Boyeau, M. Kedmi, D. Pilzer, A. Jelinski, I. Yofe, E. David, A. Wagner, et al. Destvi identifies continuums of cell types in spatial transcriptomics data. *Nature biotechnology*, 40(9):1360–1369, 2022.

Y. Ma and X. Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.

J. M. Mariadason, C. Nicholas, E. Kaitlin, M. Zhuang, H. J. Smartt, B. G. Heerdt, W. Yang, G. A. Corner, A. J. Wilson, L. Klampfer, et al. Gene expression profiling of intestinal epithelial cell maturation along the crypt-villus axis. *Gastroenterology*, 128(4):1081–1088, 2005.

L. Moses and L. Pachter. Museum of spatial transcriptomics. *Nature methods*, 19(5):534–546, 2022.

M. F. d. Oliveira, J. P. Romero, M. Chung, S. R. Williams, A. D. Gottscho, A. Gupta, S. E. Pilipauskas, S. Mohabbat, N. Raman, D. J. Sukovich, et al. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics*, pages 1–12, 2025.

T. Pham, T. Tran, H. Dam, and S. Venkatesh. Graph classification via deep learning with virtual nodes. *arXiv preprint arXiv:1708.04357*, 2017.

M. Piwecka, N. Rajewsky, and A. Rybak-Wolf. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nature Reviews Neurology*, 19(6):346–362, 2023.

C. Qian, A. Manolache, C. Morris, and M. Niepert. Probabilistic graph rewiring via virtual nodes. *Advances in Neural Information Processing Systems*, 37:28359–28392, 2024.

A. Rao, D. Barkley, G. S. França, and I. Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

D. E. Russ, R. B. P. Cross, L. Li, S. C. Koch, K. J. Matson, A. Yadav, M. R. Alkaslasi, D. I. Lee, C. E. Le Pichon, V. Menon, et al. A harmonized atlas of mouse spinal cord cell types and their spatial organization. *Nature communications*, 12(1):5722, 2021.

S. Shah, E. Lubeck, W. Zhou, and L. Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 2016.

D. Song, Q. Wang, G. Yan, T. Liu, T. Sun, and J. J. Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, 42(2):247–252, 2024.

P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

C. R. Stoltzfus, J. Filipek, B. H. Gern, B. E. Olin, J. M. Leal, Y. Wu, M. R. Lyons-Cohen, J. Y. Huang, C. L. Paz-Stoltzfus, C. R. Plumlee, et al. Cytomap: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell reports*, 31(3), 2020.

B. Tepe, M. C. Hill, B. T. Pekarek, P. J. Hunt, T. J. Martin, J. F. Martin, and B. R. Arenkiel. Single-cell rna-seq of mouse olfactory bulb reveals cellular heterogeneity and activity-dependent molecular census of adult-born neurons. *Cell reports*, 25(10):2689–2703, 2018.

L. Yan and X. Sun. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. *Bioinformatics*, 39(1):btac805, 2023.

D. Zhang, A. Schroeder, H. Yan, H. Yang, J. Hu, M. Y. Lee, K. S. Cho, K. Susztak, G. X. Xu, M. D. Feldman, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature biotechnology*, 42(9):1372–1377, 2024.

P. Zhang, C. Gao, Z. Zhang, Z. Yuan, Q. Zhang, P. Zhang, S. Du, W. Zhou, Y. Li, and S. Li. Systematic inference of super-resolution cell spatial profiles from histology images. *Nature communications*, 16(1):1838, 2025.

C. Zhao, Z. Xu, X. Wang, S. Tao, W. A. MacDonald, K. He, A. C. Poholek, K. Chen, H. Huang, and W. Chen. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in bioinformatics*, 25(2):bbae052, 2024.

E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. Taylor, P. Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.

Z. Zhou, Y. Zhong, Z. Zhang, and X. Ren. Spatial transcriptomics deconvolution at single-cell resolution using redeconve. *Nature Communications*, 14(1):7930, 2023.

A Appendix

A.1 *Square*'s additional details

A.1.1 Training and inference graphs

Fig. 5 depicts the way in which superspots are generated from the original data, as well as the considered graphs at training and inference time.

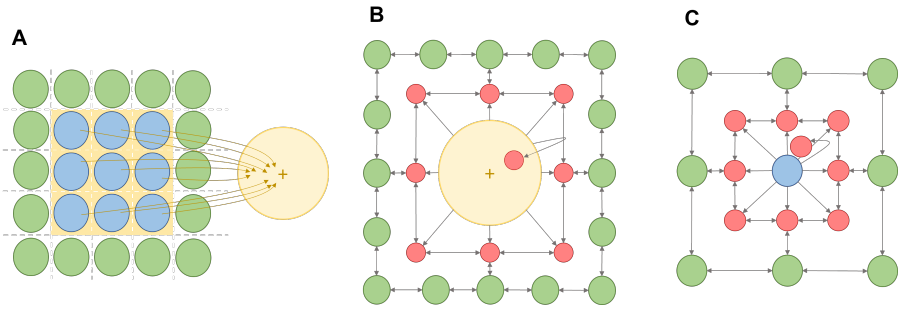


Figure 5: Overview of the data generation and graph construction process of *Square*. **A.** Training data generation via a pooling mask, where groups of neighboring spots are aggregated into super-spots while retaining the original spots as supervision. **B.** Training graph construction: a super-spot (yellow) and its neighbors (green) are augmented with virtual nodes (red) representing latent subspots within each spot. These virtual nodes serve as proxies for a finer-grained decomposition of the original measurements. **C.** Inference graph construction: each original spot (blue) is decomposed into a set of latent subspots (red) using its spatial neighbors (green). The inferred subspots are confined to the spatial extent of the original spot and represent a partition of its aggregated signal, rather than newly observed spatial locations.

A.1.2 *Square*'S PSEUDO-CODE

Algorithm 1 summarizes the steps followed by the proposed method *Square*.

Algorithm 1 Training and Inference Procedure for *Square*

Input: Spatial transcriptomics counts data $\mathbf{X}$, spots' coordinates array $\mathbf{Z}$, resolution factor $k = 9$, training neighbors $m_{\text{train}} = 16$, inference neighbors $m_{\text{inf}} = 8$
Output: Predicted high-resolution spot profiles ($\mathbf{X}_e$ and $\mathbf{Z}_e$)
// Data preprocessing
Normalize and logarithmize data
Select most highly variable genes
// Training Data Generation
for each non-edge region in X **do**
    Group k initial spots into a super-spot by summing their transcriptional signals
    Keep the k initial spots to serve as ground truth
    Identify m_{train} neighboring spots for each super-spot
    Generate training instance with paired matrices containing the super-spot and neighbors, and the ground-truth initial spots
end for
// Graph Construction for Each Training Instance
for each training instance **do**
    Initialize graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with central super-spot and neighbors
    Add k virtual nodes representing enhanced subspots
    Connect virtual nodes to central node, neighbors, and among themselves
    Set edge weights:
        $w = d(s_i, s_j)^{-1}$ for original neighbors or virtual neighbors
        $w = t$ for edges connecting central node to virtual nodes
        $w = 1$ for edges connecting original nodes to virtual nodes
    Initialize virtual nodes with the average of their closest neighbors
end for
// Model Training
for each training batch **do**
    Encode (with parameters ϕ) the super-spot, virtual nodes' and neighbors' profiles ($\mathbf{x}$) as Gaussian distributions ($\mathbf{z}$)
    $\mathcal{L}_{\phi}^{\text{KL}} = D_{KL}(q_\phi(\mathbf{z}||\mathbf{x}))$
    Reparametrization trick: $\mathbf{z} = \mu + \sigma \bigodot \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
    Decode (with parameters θ) latent representations back to gene space
    $\mathcal{L}_{\phi,\theta}^{\text{recons}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$
    Populate the training graph with latent representations
    Encode graph embeddings using a Graph Attention Network (GAT)
    Decode (with parameters δ) GAT output back to gene space
    $\mathcal{L}_{\phi,\delta}^{\text{enhance}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\delta(\mathbf{x}|\mathbf{z}, \mathbf{A})]$
    Compute total loss: $\mathcal{L}_{\phi,\theta,\delta} = \mathcal{L}_{\phi,\theta}^{\text{recons}} + \mathcal{L}_{\phi,\delta}^{\text{enhance}} - \mathcal{L}_{\phi}^{\text{KL}}$
    Update parameters $\{\phi, \theta, \delta\}$ using stochastic gradient descent (SGD):
        $\{\phi, \theta, \delta\} \leftarrow \{\phi, \theta, \delta\} - \eta \nabla_{\{\phi,\theta,\delta\}} \mathcal{L}_{\phi,\theta,\delta}$
end for
// Inference
for each spot in X **do**
    Identify m_{inf} neighboring spots
    Construct graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with central spot, neighbors, and k virtual nodes as in training
    Encode observed profiles into latent space using trained encoder (ϕ)
    Populate graph with latent embeddings
    Perform message passing using the trained GAT
    Decode (with parameters δ) the embeddings of virtual nodes into gene expression space ($\mathbf{X}_e$)
    Assign each predicted virtual node its corresponding coordinate within the central spot to form the enhanced high-resolution map ($\mathbf{Z}_e$)
end for

A.1.3 HYPERPARAMETER TUNING

We performed an optimization of model architecture and hyperparameters using Optuna[1], a framework for efficient hyperparameter search. We used its Tree-structured Parzen Estimator (TPE) sampler to suggest new hyperparameter configurations, and conducted a total of 500 trials. The list of considered hyperparameters and values is listed in Table 3.

Model optimization was conducted using the simulated dataset with available ground truth. Tables 4-7 show the top 10 trials based on MSE, cosine distance, correlation distance, and ARI, respectively. A visual summary of these results is also provided in Fig. 6 and 7. Based on the obtained results, we set the hyperparameters of the final *Square* model to those specified in Table 8.

Hyperparameter	**Tried values**
Adjacency matrix weight between central node and virtual nodes (t)	$\{1.0, 1.2, \ldots, 3.0\}$
Embedded space dimension / Gene space dimension ($\propto d$)	$\{0.3, 0.4, \ldots, 0.8\}$
GAT layers	1, 2, 3
GAT dropout	$\{0.0, 0.1, \ldots, 0.5\}$
GAT aggregation function	*sum*, *mean*, *mul*, *min*, *max*
GAT learnable skip-connection	*True*, *False*
GAT LeakyReLU angle of the negative slope	log-uniform(0.1, 1.0)
$p_\delta(\mathbf{x}\|\mathbf{z}, \mathbf{A})$ decoder layers	1, 2, 3

Table 3: Hyperparameters tested in *Square*, together with the considered values. Hyperparameter tuning was conducted with Optuna.

Table 4: Top 10 trials optimized for MSE.

Trial	Value ↓	t	Dim. ratio	GAT layers	GAT dropout	Aggr. function	Skip-connections	LReLU neg. slope	$p_\delta(\mathbf{x}\|\mathbf{z}, \mathbf{A})$ layers
113	0.208	2.4	0.55	1	0.5	*sum*	*True*	0.118	2
407	0.209	2.6	0.55	1	0.5	*sum*	*False*	0.109	2
215	0.209	2.0	0.55	1	0.5	*sum*	*True*	0.116	2
65	0.209	1.8	0.55	1	0.5	*sum*	*False*	0.813	2
200	0.21	2.0	0.55	1	0.5	*sum*	*True*	0.125	2
336	0.21	2.2	0.55	1	0.5	*sum*	*False*	0.12	2
370	0.21	2.0	0.55	1	0.5	*sum*	*True*	0.124	1
483	0.21	2.4	0.55	1	0.5	*sum*	*False*	0.116	2
401	0.21	2.6	0.55	1	0.5	*sum*	*False*	0.101	2
204	0.21	2.0	0.55	1	0.5	*sum*	*True*	0.114	2

[1]Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

Table 5: Top 10 trials optimized for Cosine distance.

Trial	Value ↓	t	Dim. ratio	GAT layers	GAT dropout	Aggr. function	Skip-connections	LReLU neg. slope	$p_\delta(\mathbf{x}\|\mathbf{z},\mathbf{A})$ layers
65	0.005	1.4	0.55	1	0.5	*sum*	*False*	0.813	2
249	0.005	2.8	0.55	1	0.4	*sum*	*True*	0.222	2
485	0.005	2.8	0.55	1	0.4	*sum*	*False*	0.78	2
458	0.005	2.6	0.55	1	0.5	*sum*	*False*	0.633	2
152	0.005	1.2	0.55	1	0.4	*sum*	*True*	0.748	2
247	0.005	2.8	0.55	1	0.4	*sum*	*True*	0.22	2
187	0.005	2.6	0.55	1	0.4	*sum*	*True*	0.654	2
206	0.005	2.6	0.55	1	0.4	*sum*	*True*	0.671	2
372	0.005	2.4	0.55	1	0.4	*sum*	*True*	0.353	2
203	0.005	2.4	0.55	1	0.4	*sum*	*True*	0.66	2

Table 6: Top 10 trials optimized for Correlation distance.

Trial	Value ↓	t	Dim. ratio	GAT layers	GAT dropout	Aggr. function	Skip-connections	LReLU neg. slope	$p_\delta(\mathbf{x}\|\mathbf{z},\mathbf{A})$ layers
468	0.169	2.4	0.5	3	0.2	*mean*	*True*	0.648	2
232	0.17	1.0	0.5	3	0.2	*mean*	*True*	0.49	1
401	0.17	1.0	0.5	3	0.2	*mean*	*True*	0.534	1
399	0.173	2.0	0.5	3	0.2	*mean*	*True*	0.234	1
474	0.174	2.6	0.5	3	0.2	*mean*	*True*	0.62	1
229	0.174	1.0	0.5	3	0.2	*mean*	*True*	0.48	1
170	0.175	2.0	0.5	3	0.2	*mean*	*True*	0.46	1
451	0.175	2.0	0.5	3	0.2	*mean*	*True*	0.558	1
21	0.176	2.6	0.75	2	0.1	*min*	*False*	0.726	2
315	0.176	2.2	0.5	3	0.2	*min*	*True*	0.494	1

Table 7: Top 10 trials optimized for ARI.

Trial	Value ↑	t	Dim. ratio	GAT layers	GAT dropout	Aggr. function	Skip-connections	LReLU neg. slope	$p_\delta(\mathbf{x}\|\mathbf{z},\mathbf{A})$ layers
266	0.733	2.4	0.6	1	0.1	*sum*	*False*	0.239	3
420	0.733	1.8	0.6	1	0.1	*sum*	*False*	0.156	3
119	0.733	1.8	0.6	1	0.1	*sum*	*False*	0.197	3
280	0.732	2.2	0.6	1	0.1	*sum*	*False*	0.24	2
190	0.732	2.2	0.6	1	0.1	*sum*	*False*	0.2	3
398	0.732	1.8	0.6	1	0.1	*sum*	*False*	0.176	3
294	0.731	1.6	0.6	1	0.1	*sum*	*False*	0.261	3
218	0.731	2.0	0.6	1	0.1	*sum*	*False*	0.172	3
174	0.731	2.0	0.6	1	0.1	*sum*	*False*	0.179	3
390	0.731	2.0	0.6	1	0.1	*sum*	*False*	0.198	3

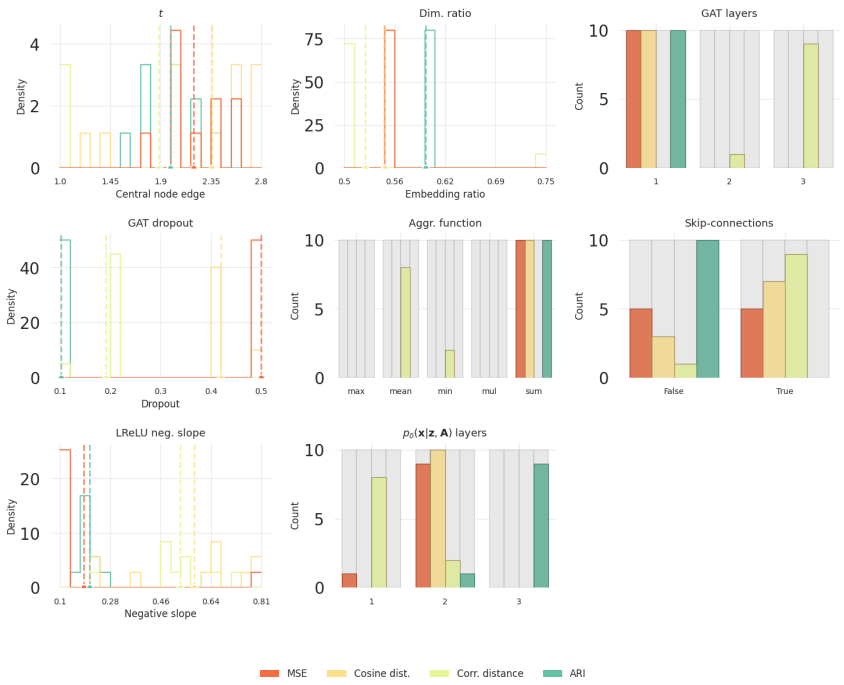


Figure 6: Results of the optimization process: density of the hyperparameters *central node edge, embedding ratio, dropout rate* and *negative slope* for the top 10 trials (dashed lines represent the mean value); and histograms of *GAT layers, Aggregation function, skip-connections* and $p_\delta(\mathbf{x}|\mathbf{z}, \mathbf{A})$ *decoder layers* also for the top 10 trials. Results are shown for MSE, cosine distance, correlation distance and ARI metrics.

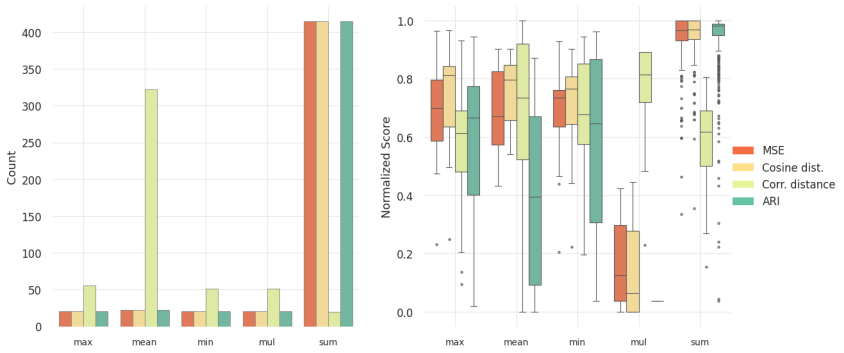


Figure 7: Results of the optimization process for hyperparameter *Aggregation function* for the 500 trials. Results are shown for MSE, cosine distance, correlation distance and ARI metrics. For the normalized score, the MSE is shown as $1 - MSE$ so that higher values are better.

Hyperparameter	**Optimal value**
Adjacency matrix weight between central node and virtual nodes (t)	2.0
Embedded space dimension / Gene space dimension	0.55
GAT layers	1
GAT dropout	0.2
GAT aggregation function	*sum*
GAT learnable skip-connection	*True*
GAT LeakyReLU angle of the negative slope	0.2
$p_\delta(\mathbf{x}\|\mathbf{z}, \mathbf{A})$ decoder layers	2

Table 8: Final hyperparameters' values for *Square* after Optuna optimization.

A.2 DATASETS: ADDITIONAL DETAILS

A.2.1 SIMULATED DATASET

In this dataset, the spots' transcriptional profiels are generated from Zero-Inflated Negative Binomial (ZINB) distributions with different parameters: the expected number of events k occurring with a given success probability p, and the zero-inflated probability π, which controls the desired sparsity. This is possible given that the number of successes k is drawn from a Poisson distribution with rate $\lambda = 10$, making values near zero highly unlikely. Consequently, the negative binomial component contributes minimally to the overall sparsity, which is instead primarily governed by the zero-inflation parameter π.

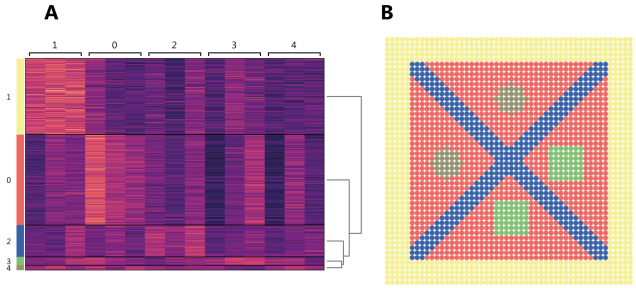


Figure 8: Simulated dataset, with $\pi = 0$, $N = 50$, $r = 3$ and $w = 4$, where N denotes grid-width, r the radius of the generated circles as well as half of the side of the generated squares, and w is the width of the generated diagonals. **A.** Heatmap of marker genes across the five spot transcriptional profiles. **B.** Spatial structure of the generated spatial data displaying different geometries and patterns.

Algorithm 2 *Simulated* spot-level data generation

Require: Sparsity parameter π, grid size N, shape radius r, and band width w

Initialize grid $G \in \{1, \ldots, 5\}^{N \times N}$ with label 1
Define inner-*Square* region $\mathcal{R}_2 = \{(i,j) : |i - c_1| \leq r,\ |j - c_2| \leq r\}$ and set $G[\mathcal{R}_2] = 2$
Define diagonal-band region $\mathcal{R}_3 = \{(i,j) \in \mathcal{R}_2 : |i - j| \leq w \ \vee \ |i + j - C| \leq w\}$ and set $G[\mathcal{R}_3] = 3$
$t_1, \ldots, t_4 \overset{i.i.d.}{\sim} \text{Uniform}(\{4, 5\})$
Let $\{p_1, \ldots, p_4\}$ be four positions symmetrically arranged around the grid center
for $i = 1, \ldots, 4$ **do**
  if $t_i = 4$ **then**
    $\mathcal{S}_i = \{(u,v) : (u - p_i^{(1)})^2 + (v - p_i^{(2)})^2 \leq r^2\}$
  else if $t_i = 5$ **then**
    $\mathcal{S}_i = \{(u,v) : |u - p_i^{(1)}| \leq r,\ |v - p_i^{(2)}| \leq r\}$
  end if
  $G[\mathcal{S}_i] = t_i$
end for

Flatten grid: $c \leftarrow \text{vec}(G)$
Initialize expression matrix X
for each cell type $k \in \{1, \ldots, 6\}$ **do**
  $M \leftarrow (c = k)$
  Sample gene-wise modifiers $k_g \sim \text{Poisson}(10)$
  for each gene g **do**
    $z_g \sim \text{Bernoulli}(\pi)$
    $y_g \sim \text{NegBin}(\max(k_g, 1), 0.3)$
    $X[M, g] \leftarrow (1 - z_g) \cdot y_g$
  end for
end for

Construct data object with X, labels c, and spatial coordinates

A.2.2 MOUSE OLFACTORY BULB (MOB) DATASET

The mouse olfactory bulb (MOB) dataset was generated using the statistical simulator *scDesign3*, using a 10x Chromium scRNA-seq reference and an ST MOB sample. The dataset contains 278 spots with gene expression vectors of dimension 182, comprising mixtures of up to four cell types (Fig. 9): granule cell (GC), periglomerular cell (PGC), mitral and tufted cell (M/TC), and olfactory sensory neuron (OSNs).

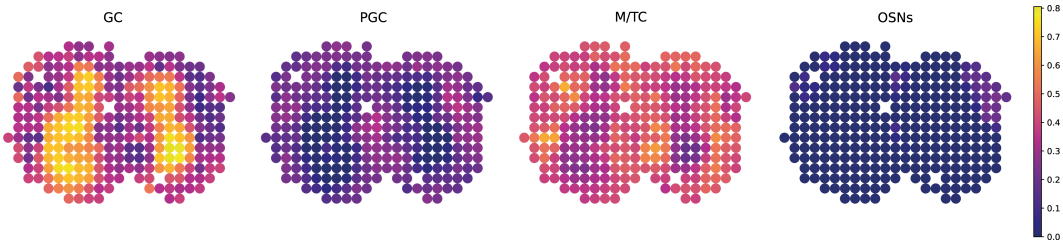


Figure 9: Mouse Olfactory Bulb dataset's ground truth proportions.

To model the single-cell data, gene expression distributions were learned while accounting for cell-type-specific effects. For each gene, expression counts were modeled using a negative binomial distribution in which both the mean and dispersion parameters depended on cell type. This allowed the model to capture differences in expression levels and variability across cell populations. In addition to marginal distributions, dependencies between genes were modeled using a Gaussian copula, enabling the preservation of correlation structures observed in the original data. Based on the fitted models, a new synthetic single-cell dataset was generated that retained the statistical characteristics of the reference dataset.

In parallel, the spatial transcriptomics data were modeled to capture spatial variation in gene expression. For each gene, the mean expression was represented as a smooth function of spatial coordinates, allowing flexible spatial patterns across the tissue. A Gaussian process-based smoother was used to model these spatial effects, while dispersion was assumed to be constant. As in the single-cell setting, gene–gene dependencies were incorporated through a Gaussian copula, ensuring realistic multivariate structure in the simulated data.

To integrate both modalities, a cell-type signature matrix was constructed from the modeled single-cell data by averaging gene expression across cells belonging to the same cell type. This signature matrix was then combined with the spatial expression estimates to infer cell-type proportions at each spatial location using a deconvolution approach based on `CIBERSORT`[2]. The resulting proportions describe the relative contribution of each cell type within every spatial spot.

Synthetic spatial transcriptomics data were subsequently generated by aggregating simulated single-cell profiles according to the inferred proportions. For each spatial location, a fixed number of cells was sampled across cell types, and their gene expression profiles were combined proportionally to the estimated fractions. The aggregated counts were rescaled to control library size and rounded to integer values. This procedure yields spatial observations representing mixtures of multiple cell types, thereby mimicking the compositional nature of spatial transcriptomics measurements.

The final output consists of a simulated single-cell dataset and a spatial dataset in which each location reflects a controlled mixture of cell types with realistic gene expression distributions and spatial structure.

A.2.3 MOUSE SMALL INTESTINE (MSI) DATASET

We used the publicly available Visium HD spatial transcriptomics dataset of mouse small intestine (MSI) generated with the CytAssist platform by 10x Genomics. The dataset consists of a formalin-fixed paraffin-embedded (FFPE) tissue section obtained from a healthy 8-week-old male C57BL/6

[2]Newman, A. M. et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature methods, 12(5), 453-457.

mouse. Tissue was processed using the Swiss roll technique and sectioned at a thickness of 5 µm prior to hematoxylin and eosin (H&E) staining and imaging. Imaging was performed using an Olympus VS200 slide scanner at 20× magnification.

Spatial gene expression profiling was conducted using the Visium HD Spatial Gene Expression assay with the CytAssist instrument and the mouse transcriptome probe set (v2.0). Library preparation followed the manufacturer's protocol, and sequencing was carried out on an Illumina NovaSeq 6000 platform using a paired-end configuration (43 bp read 1, 50 bp read 2, and dual 10 bp indices), yielding approximately 321.6 million reads.

Raw sequencing data were processed using Space Ranger (v3.0), which performs alignment to the reference transcriptome, tissue detection, and image registration, producing feature-barcode matrices for downstream analysis. The dataset represents a healthy intestinal tissue sample and is distributed under a Creative Commons Attribution 4.0 International license.

To characterize the crypt–villus axis, we manually defined "seed" points within the crypt region by drawing a line connecting villus tips and translating it perpendicularly toward the crypt (Fig. 10B). Each cell was assigned a position based on its distance to the nearest seed. To account for irregular tissue geometry, distances were computed using a geodesic metric—the shortest path length on an eight-nearest-neighbor graph obtained via the Dijkstra algorithm (Fig. 10C). This framework was then used to assess whether established marker genes exhibit the expected spatial gradients along the axis.

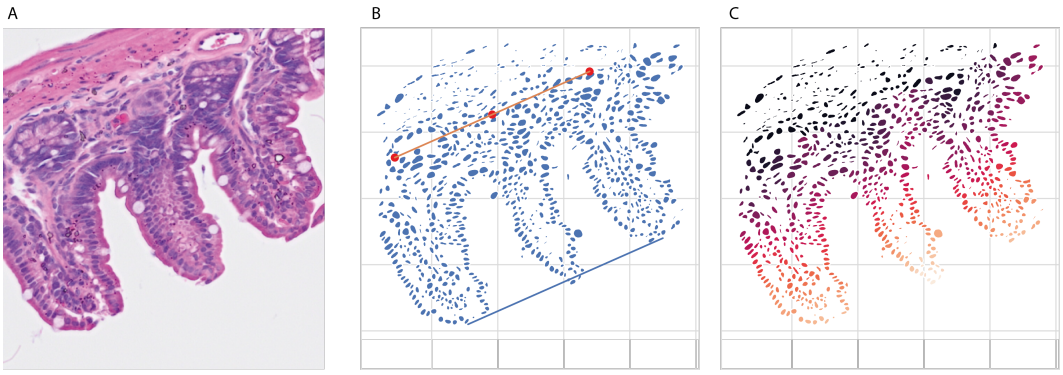


Figure 10: Visium HD mouse small intestine dataset. **A.** Selected region of interest (ROI) from the hematoxylin and eosin (H&E) image, including the three analyzed villi. **B.** Spatial map of segmented cells, with manually defined seeds (red dots) placed in the crypt regions. **C.** Spatial heatmap of the geodesic distance, used as a proxy for the crypt–villus axis.

A.3 EVALUATION METRICS

We devise several quantitative metrics that help in evaluating the effectiveness of *Square* in different scenarios. Next, we describe them in more detail, based on the particularities of each of the considered datasets.

A.3.1 SIMULATED DATASET

We compute the following spot-wise metrics that directly compare the transcriptional profiles of the inferred subspots with those of the original data: the Mean Error (MSE) and the rank-based Spearman correlation distance (SCD), the latter to measure the preservation of the relative ranking of highly expressed genes. To further evaluate whether the most highly expressed genes are retained after resolution enhancement, we compute two metrics based on the Jaccard distance (JD) between the sets of top-expressed genes in the ground truth and the generated data: for the genes above the mean ($JD_{>\mu}$) and for the top 30 most expressed genes (JD_{30}).

Additionally, we consider a metric to capture whether the generated spots maintain the cell-type distribution of the original spots. For each cell-type, we generate a representative by computing the mean of the corresponding single-cell transcriptional profiles. Then, the enhanced subspots are assigned the cell-type whose representative is closer in Euclidean distance to their inferred transcriptional profile. Using these predicted and true cell-type assignments, we compute the Adjusted Rand Index (ARI) to quantify the agreement between them. This metric evaluates how well the enhanced data preserves both the spatial distribution and cell-type identity compared to the ground truth.

A.3.2 MODEL-BASED SYNTHETIC MOUSE OLFACTORY BULB (MOB) DATASET

Since only the cell-type proportions of the original spots are available, we consider a clustering-based assessment to determine whether the generated subpots accurately reflect the underlying cell-type distribution. We first apply k-means to the inferred subspots with varying values of k, allowing for less and more coarse clustering. For a given k, we generate a *cluster-frequency* vector for every original spot, with the ith entry reflecting the proportion of subspots assigned to cluster i. Since cluster assignments may not reflect cell-types, we cannot directly compare this vector with the cell-type proportion vector. Instead, we compute the entropy of each of them, and then evaluate the Spearman correlation between them across spots. We expect the entropies between the vectors to be correlated, as the original cell-type distribution should be reflected in the clustering assignment.

To further evaluate the inferred subspots, we annotate each of them by first clustering them, and then assigning to each cluster the cell-type with more overlap between its marker genes and the cluster significative genes. A cell-type frequency vector is then constructed for each spot and compared to the ground truth cell-type proportion vector via the RMSE.

A.3.3 REAL DATASETS

For the considered 10x Visium dataset of a human **GBM tissue**, we recreate the main findings of Greenwald et al. [3], which analyzed 13 GBM spatial samples (note that the considered GBM sample does not belong to these 13). We adopted a metric introduced by the authors that quantifies the extent to which each state is enriched in neighboring versus dispersed spots, referred to as "spatial coherence". For each transcriptional state or metaprogram (MP), we define the set

$$S_{i,k} = \{j \in \mathcal{N}_i | m(j) = k\}, \tag{6}$$

where $\mathcal{N}_i$ denotes the set of neighbors of node i and $m(j)$ is the MP of node j. Then, for each spot $j \in S_{i,k}$, we define

$$A(j,k) = \frac{1}{|\mathcal{N}_j|} \sum_{l \in \mathcal{N}_j} \mathbf{1}(m(l) = k) \tag{7}$$

[3] Greenwald, A. C. et al. (2024). Integrative spatial analysis reveals a multi-layered organization of glioblastoma. Cell, 187(10), 2485-2501.

Then, the "spatial coherence", hereby denoted as C, for spot i and MP k is computed as follows:

$$C(i,k) = \frac{1}{|S_{i,k}|} \sum_{j \in S_{i,j}} A(j,k) \tag{8}$$

Furthermore, in order to quantify the spatial relationships between cell states, we reproduced three complementary measures of cell type coupling at distinct spatial resolutions: *i)* regional composition, defined within a neighborhood of spots determined by the number of concentric rings around a given spot; *ii)* adjacency, assessed using a single spot and its immediate neighbors; and *iii)* colocalization, inferred from spot-level deconvolution using MP scores. A consensus set of state–state interactions have been defined by selecting strong interactions supported by multiple coupling measures.

The Visum HD **MSI dataset** includes measurements at 2 and 8 μm resolution, allowing us to *in silico* derive 2 μm bins from 8 μm bins ($k = 16$), and subsequently evaluate whether *Square* can recover fine-grained spatial structure from coarser measurements ($k = 16, m = 8$). Further, to determine whether it preserves biologically meaningful spatial organization, we focused on the crypt–villus axis, a canonical differentiation trajectory in the intestinal epithelium. To establish a framework for systematically characterizing this axis, we manually defined "seeds" within the crypt region: to ensure accurate positioning, we first drew a line connecting the villus tips and then translated this line perpendicularly toward the crypt (Fig. 10B). We then computed, for each cell, its distance to the nearest seed, thereby defining a proxy for the crypt–villus axis. To mitigate biases introduced by Euclidean distance in irregular geometries such as intestinal tissue, we instead used a geodesic distance defined as the number of nodes along the shortest path between two cells (computed via the Dijkstra algorithm) on a graph constructed using the eight nearest neighbors (Fig. 10C). Using this metric, we assessed whether well-established marker genes along the crypt–villus axis exhibit the expected spatial gradients.

A.4 ADDITIONAL RESULTS

A.4.1 ABLATION STUDY: PERFORMANCE ACROSS DIFFERENT LEVELS OF SPARSITY

Table 9: **Ablation Study:** Performance comparison across model variants.

Model	Sparsity (π)	MSE ↓	SCD ↓	JD$_{>\mu}$↓	JD$_{30}$↓	ARI ↑
w/o virtual nodes	0%	**0.29 ± 0.10**	0.423 ± 0.005	0.408 ± 0.006	0.5918 ± 0.0024	0.80 ± 0.03
w/o VAE		2.31 ± 0.69	0.511 ± 0.042	0.443 ± 0.027	0.6028 ± 0.0111	0.59 ± 0.24
w/o attention		1.18 ± 0.27	0.462 ± 0.031	0.402 ± 0.014	0.5953 ± 0.0093	0.65 ± 0.37
Square		0.39 ± 0.08	**0.412 ± 0.012**	**0.396 ± 0.005**	**0.5848 ± 0.0018**	**0.88 ± 0.08**
w/o virtual nodes	10%	**2.37 ± 0.14**	0.531 ± 0.008	0.414 ± 0.028	0.6101 ± 0.0026	0.78 ± 0.05
w/o VAE		3.85 ± 0.26	0.588 ± 0.018	**0.307 ± 0.007**	0.6159 ± 0.0045	0.53 ± 0.27
w/o attention		3.10 ± 0.34	0.562 ± 0.034	0.324 ± 0.026	0.6127 ± 0.0120	0.54 ± 0.43
Square		2.39 ± 0.13	**0.523 ± 0.008**	0.406 ± 0.025	**0.6054 ± 0.0027**	**0.82 ± 0.07**
w/o virtual nodes	20%	**4.43 ± 0.12**	0.626 ± 0.007	0.476 ± 0.030	0.6312 ± 0.0024	0.53 ± 0.08
w/o VAE		5.92 ± 0.46	0.685 ± 0.021	**0.331 ± 0.009**	0.6394 ± 0.0059	0.33 ± 0.27
w/o attention		5.00 ± 0.18	0.650 ± 0.014	0.370 ± 0.035	0.6322 ± 0.0056	0.41 ± 0.23
Square		4.47 ± 0.09	**0.621 ± 0.005**	0.465 ± 0.025	**0.6275 ± 0.0024**	**0.56 ± 0.08**

A.4.2 VIRTUAL NODES' INITIALIZATION

We evaluated different approaches to initialize the virtual nodes' transcriptional profiles. Specifically, we tested: i) averaging the transcriptional profiles of the virtual node's neighboring spots; ii) imputation with a GNN (see below); iii) an all zero-vector; and iv) a random initialization.

The considered GNN is designed to predict dataset-specific spots in between others. A prior training is performed by generating 100 different graphs with the entire dataset grid, randomly removing 20% of the spots, which are used to train the network with MSE as loss. The GNN contains 2 layers and the number of neurons at the final layer equals the number of genes. Training is conducted using SGD and a learning rate scheduler with a patience parameter of 5 epochs. This network is then used to populate the transcriptional profile of virtual nodes, which are also embedded into the latent space using the encoder $q_\phi(\mathbf{z}|\mathbf{x})$.

The experiment is conducted with the *simulated* dataset, to have access to ground truth. The results are provided in Table 10. Overall, we observed the *mean* obtains the best results, and hence this approach is selected to initialize the virtual nodes in *Square*.

Table 10: **Virtual nodes' initialization:** Performance comparison across different approaches.

Approach	MSE ↓	SCD ↓	JD$_{>\mu}$ ↓	JD$_{30}$ ↓	ARI ↑
mean	1.72 ± 0.03	**0.438 ± 0.030**	**0.469 ± 0.023**	**0.8135 ± 0.0012**	**0.71 ± 0.11**
GNN	**1.01 ± 0.19**	1.049 ± 0.093	0.658 ± 0.027	0.8217 ± 0.0020	0.00 ± 0.00
zero	1.73 ± 0.04	0.468 ± 0.023	0.477 ± 0.026	0.8149 ± 0.0013	0.57 ± 0.12
random	1.49 ± 0.07	0.668 ± 0.087	0.541 ± 0.029	0.8172 ± 0.0014	0.23 ± 0.16

A.4.3 SYNTHETIC MOUSE OLFACTORY BULB (MOB) DATA

In order to annotate the generated spots after resolution enhancement, we leverage the 50 most significative marker genes of each one of the cell types. Fig. 11 shows the overlap between them, which is relatively large. Exclusive marker genes for each cell type are: *CNTNAP4* for granule cells; *CPE* for mitral and tufted cells; and *OMP*, *SPON1*, and *AI593442* for olfactory sensory neurons.

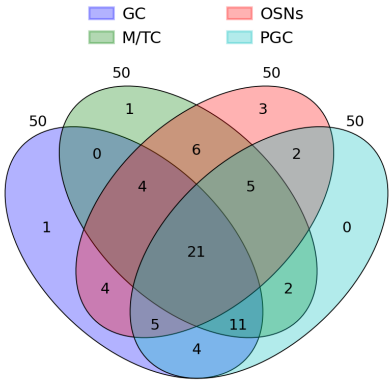


Figure 11: Venn diagram of the marker genes of MOB dataset cell types.

Figure 12 shows the annotation of the datasets generated by BayesSpace, CARD and STAGE. All three methods impose an excessively smooth, continuous spatial distribution of cell types across the tissue. Note that they overlook the presence of periglomerular cells, with a discrete appearance over the tissue, fact which makes its detection clearly challenging. In contrast, *Square* is able to recover the presence of these cells (Fig. 2D), even when they are blurred by the rest.

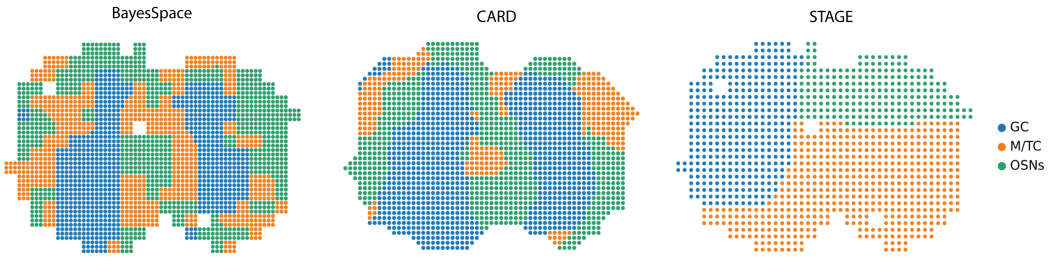


Figure 12: Annotation of datasets generated by BayesSpace, CARD and STAGE.

A.4.4 GENE AND LATENT SPACES

The distribution shift that arises when considering graph nodes at different resolutions must be empirically validated. To this end, we examined how node features cluster in low-dimensional representations, both before and after encoding them with the trained encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$. Visualizing the first two principal components (Fig. 13) reveals that original spots (blue) and superspots (red) are clearly separated in gene expression space, whereas they largely overlap in the latent space, suggesting that the encoder effectively mitigates the distribution shift.

Additionally, the silhouette scores obtained from a k-means clustering with an arbitrary $k = 5$ are 0.517 in the gene expression space and 0.385 in the latent space, further indicating reduced cluster separation in the latent representation.

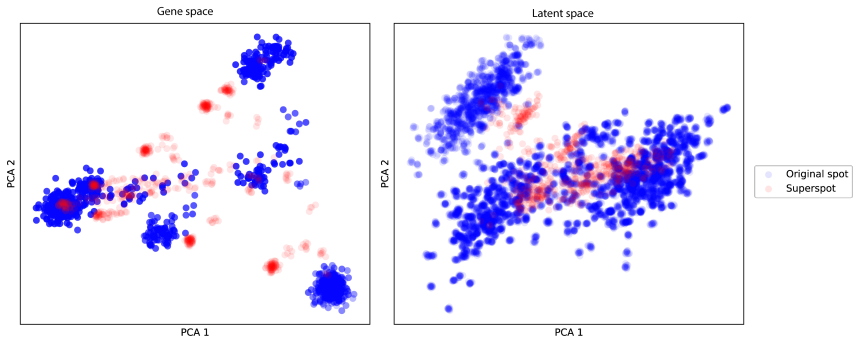


Figure 13: Projection on the two first principal components of PCA of the original transcriptional profile (left panel) and the generating embeddings after passing them through the VAE encoder q_ϕ (right panel) of original spots (blue) and superspots (red) from the *simulated* dataset.

A.4.5 ATTENTION

To validate the attention paradigm implemented in the graph neural network, we further analyzed the impact of attention mechanisms on the reconstruction of homogeneous shapes and examined the learned attention coefficients.

First, we assessed how incorporating attention affects the fidelity of spatial boundary reconstruction. Models without attention exhibit uniform message propagation across all neighboring nodes, resulting in blurred or inconsistent cluster boundaries in the reconstructed expression (Fig. 14A, left panels). In contrast, attention-based models produce heterogeneous subspot representations within each original spot, yielding sharper and more coherent spatial boundaries that better reflect the underlying tissue structure (Fig. 14A, middle panels). To further validate this, we performed an ablation study in which edges associated with higher attention coefficients, defined as $\alpha = \frac{1}{H}\sum_{i=1}^{H} \mathbf{a}_i$ (where H denotes the number of attention heads), were removed from the graph. This ablation resulted in markedly less well-defined clustering in both the circular and diagonal spatial regions (Fig. 14A, right panels), confirming that high-attention edges are critical for preserving spatial structure.

Regarding the distribution of attention coefficients α, we investigated whether virtual nodes assigned greater attention weight to the central node — which carries the primary information that virtual nodes are intended to represent. As shown in Fig. 14B, the attention coefficients assigned to central nodes are systematically higher than those assigned to peripheral neighbors, as reflected in the box plot distributions. Additionally, we verified that attention coefficients decay with increasing distance between connected nodes (Fig. 14C), consistent with the expectation that spatially proximal nodes should have greater influence during message aggregation. Together, these results confirm that the attention mechanism learns a geometrically meaningful weighting scheme that prioritizes relevant structural information.

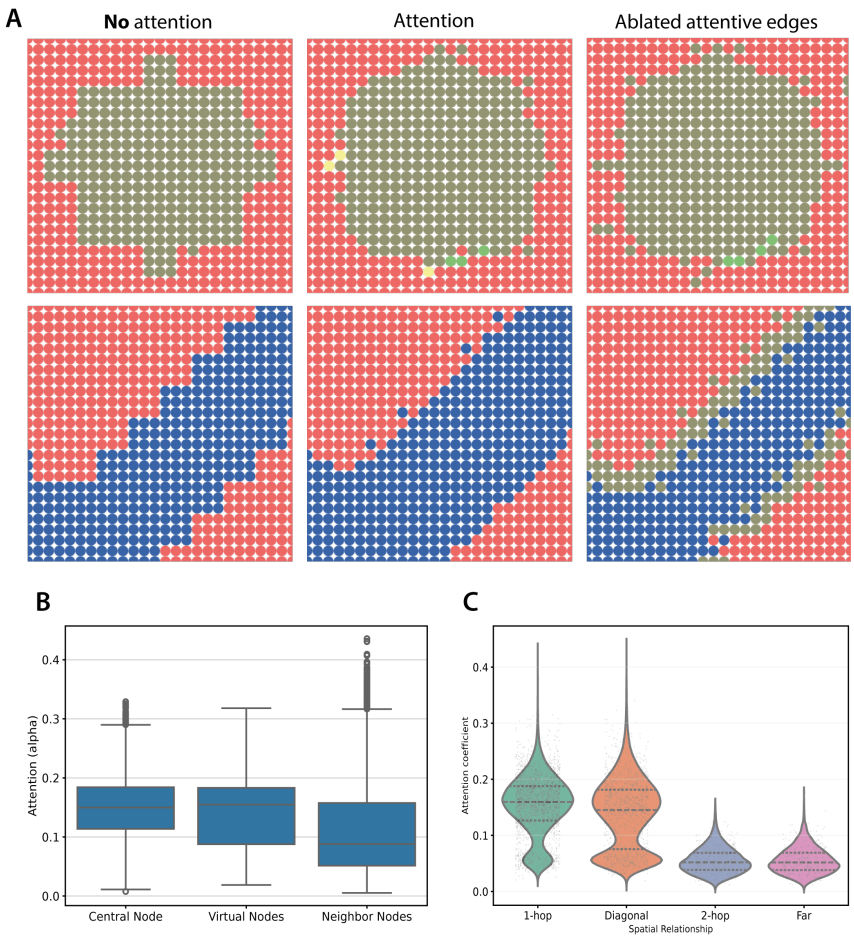


Figure 14: Validation of the attention mechanism in the graph neural network. **A.** Spatial reconstruction of two synthetic tissue regions (circular boundary, top row; diagonal boundary, bottom row) under three conditions: model without attention (left panels), attention-based model (middle panels), and attention-ablated model in which edges with the highest attention coefficients are removed (right panels). Dots represent individual subspots colored by cluster identity. Attention-based models produce sharper and more coherent spatial boundaries compared to the uniform propagation of attention-free models, while ablation of high-attention edges results in degraded boundary definition. **B.** Box plots of attention coefficients for nodes targeted by virtual nodes across relative positions, showing that central nodes consistently receive higher weights. **C.** Violin plots illustrating the decay of attention coefficients as a function of distance between connected nodes, confirming that the attention mechanism learns a geometrically meaningful weighting scheme that prioritizes spatially proximate neighbors.

A.4.6 ROBUSTNESS

We evaluated the performance of *Square* in challenging biological settings characterized by weak spatial autocorrelation. Specifically, using the simulated spot-level dataset, we analyzed *Square*'s robustness to varying levels of Moran's index. The Moran's index is computed for each graph in the validation set. Intuitively, high positive Moran's index values correspond to graphs with more pronounced boundaries separating distinct regions. This situation commonly arises at interfaces between different transcriptomic profiles (from the considered five), where accurate reconstruction is inherently more difficult. To evaluate the method's performance, for each graph in the validation set we compute the loss $\mathcal{L}_{\phi,\delta}^{\text{enhance}}$ (see Eq. equation 2.5) between the generated and ground-truth spots.

Results show a slight decrease in performance as Moran's index increases (Fig. 15). Nevertheless, the observed performance degradation is minimal, with an R^2 of 0.08, indicating a weak linear relationship between spatial autocorrelation and model performance.

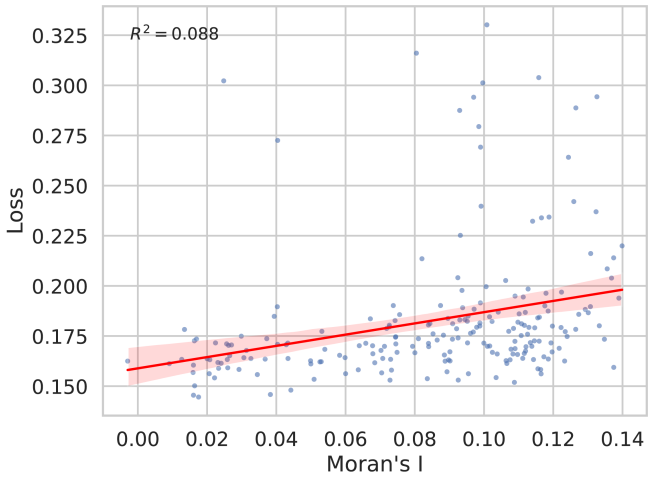


Figure 15: Scatter plot where every point is a graph instance from the validation set. The *x*-axis represents the Morans's index of the graph; and the *y*-axis the obtained loss.

A.5 LIMITATIONS

While *Square* provides a framework for inferring fine-grained spatial structure, several limitations should be noted. First, the inferred subspots do not correspond to directly observed cellular-resolution measurements, but rather represent model-based estimates consistent with the observed data and spatial context. As a result, they should be interpreted as plausible spatial decompositions rather than exact reconstructions of underlying cellular states.

Second, the method relies on spatial dependencies encoded in the graph structure. In regions with weak spatial organization or highly heterogeneous tissue architecture, the inferred subspot representations may be less reliable. Nevertheless, we evaluated this in a dedicated experiment (App. A.4.6) and observed only a slight performance decrease when increasing Moran's index ($R^2 = 0.08$), indicating a weak relationship between spatial autocorrelation and model performance.

Third, the model assumes that local neighborhoods provide sufficient information to infer finer spatial patterns. This assumption may not hold in cases with strong long-range interactions or when relevant biological signals are not captured in the observed gene expression profiles.

Finally, while *Square* does not require external reference data, incorporating complementary modalities (e.g., histology or scRNA-seq) could further improve the biological interpretability of the inferred representations.

A.6 IMPLEMENTATION DETAILS

The core software libraries used to implement *Square* and conduct the experiments are listed in Table 11.

Table 11: *Square*'s software dependencies.

Library	Version	GitHub / Source
Scanpy – Single-Cell Analysis in Python[3]	1.10.3	`https://github.com/scverse/scanpy`
PyTorch[4]	2.8.0	`https://github.com/pytorch/pytorch`
CUDA Toolkit	12.8	`https://developer.nvidia.com/cuda-toolkit-archive`
cuDNN	9.10.2	`https://docs.nvidia.com/deeplearning/cudnn/latest/`
BayesSpace[5]	1.5.1	`https://github.com/edward130603/BayesSpace`
CARD[6]	–	`https://github.com/YMa-lab/CARD`
STAGE[7]	1.0.1	`https://github.com/zhanglabtools/STAGE`

The set of parameters used to run the competing methods is gathered in Table 12.

Method	Parameters
BayesSpace	PCs $= 7$, HVGs $= 100$, $\gamma = 2$, Jitter prior $= 0.3$, Jitter scale $= 3.5$, Error distr.: Student's t
CARD	Number of neighbors $= 5$
STAGE	$d = 0.5 \cdot N_g, \alpha_1 = 0.1, \alpha_2 = 0.1, \alpha_3 = 0.1$

Table 12: Set of parameters used to run *Square*'s competing methods.

The computational requirements to run *Square* on all datasets in the experiments are specified in Table 13. The reported computational resources correspond to a single-node Slurm job executed on an NVIDIA A100-PCIe GPU 80GB, using 8 tasks per node and 4 CPUs per task.

Dataset	Epochs	Spots or Bins	Genes	Time (seconds)				Memory (GBs)
				(1)	(2)	(3)	(4)	
Simulated	200	8,100 (2,500)	100	0.10	17.61	42.51	7.27	19.59
Mouse olfactory bulb	3,000	278	182	0.08	22.52	88.06	18.38	17.01
Glioblastoma	200	3,468	562	0.17	47.12	56.73	28.42	34.57
Mouse small intestine	300	765	662	0.32	18.13	72.9	22.45	15.16

Table 13: Pipeline runtime and memory usage across datasets. Time subcolumns correspond to: (1) data loading and preprocessing, (2) training set construction, (3) model training, (4) resolution enhancement. Note that gene counts refer only to the subset of highly variable and/or marker genes retained after preprocessing, rather than the full raw gene set.

[3]Sultana, S. (2017). SCANPY: large-scale single-cell gene expression data analysis. Environ. Microbiol., 19, 2926.

[4]Paszke, A. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

[5]Zhao, E. et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. Nature biotechnology, 39(11), 1375-1384.

[6]Ma, Y., & Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. Nature biotechnology, 40(9), 1349-1359.

[7]Li, S., Gai, K., Dong, K., Zhang, Y., & Zhang, S. (2024). High-density generation of spatial transcriptomics with STAGE. Nucleic Acids Research, 52(9), 4843-4856.