Pathway-Aligned Regulator Tokens for Interpretable Spatial Gene Expression Prediction from Histology

Hyun Namgung
Graduate School of Artificial Intelligence
POSTECH
Pohang, Republic of Korea
hyun.namgung@postech.ac.kr

Sang Hyun Park*
Graduate School of Artificial Intelligence
Department of Computer Science and Engineering
POSTECH
Pohang, Republic of Korea
sanghyunpark@postech.ac.kr

Abstract

Predicting spatial gene expression from histopathology images enables transcriptomic analysis on archival H&E samples where matched spatial transcriptomics data is unavailable. Gene expression is organized into co-regulated modules: transcription factor complexes coordinate the co-expression of many downstream targets (e.g., the ISGF3 complex coordinating hundreds of interferon-stimulated genes through shared cis-regulatory elements). Existing approaches to histology-to-expression prediction, both deterministic and generative, employ generic architectures that ignore this structure. We introduce regulator attention, a modular co-expression prior that pools gene representations into a small set of content-dependent tokens via a learned selector and routes them back through cross-attention to modulate gene-level features, in contrast to learnable token arrays whose identity is independent of the input. Within a flow matching framework conditioned on pathology foundation model embeddings, our method outperforms deterministic and generative baselines on correlation-based metrics across four tissue types spanning human cancer, normal tissue, and cross-species samples, with the largest gains on highly variable genes. Isolated under the final flow matching backbone on Mouse Brain, regulator attention improves over full gene–gene self-attention, whereas input-agnostic bottlenecks fall below it, identifying content-dependent token generation as the source of the gain. The learned tokens attend to distinct gene sets, and their pathway alignment exceeds expression-, variance-, and detection-matched null gene sets. The bottleneck also removes the quadratic memory cost of gene–gene attention, reducing peak memory by $5.3\times$ at the Xenium Prime 5K panel size.

1 Introduction

Spatial transcriptomics enables comprehensive gene expression profiling while preserving tissue architecture, providing unprecedented insights into cellular organization in development [3], neurodegeneration [29], and cancer [21]. The technology remains costly and technically demanding, motivating computational approaches that predict spatial gene expression directly from histology images [13, 45].

Gene expression is not produced by independent programs at each gene. It is organized into co-regulated modules in which transcription factor complexes, themselves products of regulated genes, coordinate the co-expression of many downstream targets [1, 38] (Fig. 1c). The ISGF3 complex illustrates this organization: a heterotrimer of STAT1, STAT2, and IRF9 that coordinately regulates

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

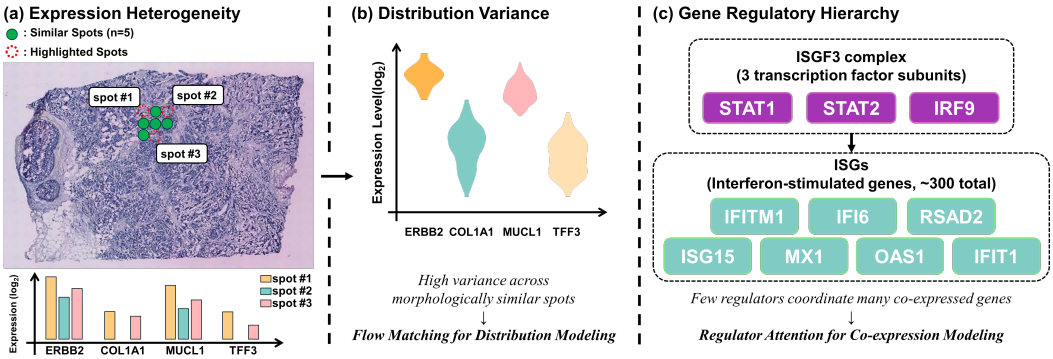


Figure 1: **Motivation for distributional and structural modeling.** (a) Morphologically similar regions exhibit heterogeneous gene expression profiles. (b) Aggregated distributions show substantial variance that point-prediction methods cannot capture. (c) Gene expression follows a regulatory hierarchy: a small number of transcription factor complexes (e.g., ISGF3, comprising STAT1/STAT2/IRF9 subunits) coordinate the co-expression of hundreds of target genes (interferon-stimulated genes, ISGs) [23, 35, 36], motivating an architecture that respects this structure.

hundreds of interferon-stimulated genes through a single class of cis-regulatory elements [35, 36]. Existing approaches to histology-to-expression prediction, spanning both deterministic regressors [13, 31, 45] and recent generative methods [15, 47], rely on generic backbones that treat genes as independent outputs of patch-level features, with no architectural prior reflecting this modular organization.

Encoding this modular organization architecturally requires latent regulators whose identity derives from the gene content they aggregate, rather than from input-independent parameters. We introduce *regulator attention*, an architectural mechanism that realizes this design through a small set of content-dependent tokens generated from gene representations themselves via learned weighted aggregation. Each token aggregates signals from gene-level features and then modulates co-expressed gene programs through a two-stage attention path (genes $\rightarrow$ regulators $\rightarrow$ genes), so that its identity is anchored to the gene content at that layer rather than fixed in advance. The path is feed-forward within each layer and encodes no regulatory feedback or dynamics. What it imposes is a modular co-expression prior. We embed this mechanism in a flow matching framework conditioned on pathology foundation model embeddings [6, 27] through two complementary conditioning pathways, providing the distributional substrate needed to model the substantial expression variance that arises among morphologically similar tissue regions [10] (Fig. 1a,b).

Experiments across four tissue types show consistent improvements over deterministic and generative baselines, including three recent methods that we reimplement under our protocol on two of the datasets [8, 12, 39], with the largest gains on highly variable genes. Under the final flow matching backbone, regulator attention raises PCC@200 over full gene–gene self-attention by 0.0676 on Mouse Brain, while input-agnostic bottlenecks fall below it. The learned tokens attend to distinct gene sets whose pathway alignment exceeds matched null gene sets, the method outperforms the cross-patient configurations of super-resolution methods under leave-one-patient-out evaluation, and the bottleneck removes the quadratic memory term of gene–gene attention.

2 Related Work

Spatial Gene Expression Prediction from Histology. Spatial transcriptomics produces paired observations of histology images and gene expression vectors, creating a structured prediction problem with two distinctive characteristics: outputs exhibit strong inter-gene correlations reflecting biological pathways, and identical morphology can yield variable expression due to transcriptional stochasticity [10, 17]. ST-Net [13] first demonstrated this prediction task using convolutional networks, with subsequent work exploring graph-based spatial modeling [45], multi-scale features [7, 31], contrastive learning [42], exemplar-based retrieval [43, 44], and, more recently, hierarchical spot graphs [12], multi-level contrastive discrimination [39], and gene-embedding priors [8]. These regression-based

methods predict point estimates that cannot represent expression stochasticity, motivating recent shifts toward distributional modeling. Stem [47] applies diffusion to spatial transcriptomics and demonstrates improved capture of biological heterogeneity, while STFlow [15] introduces flow matching with graph-based message passing for whole-slide context. These advances confirm that distributional formulations are viable. Concurrent efforts develop spatial transcriptomics-aware foundation models that integrate ST data into pretraining itself [14, 16], which is complementary to our focus on architectural design for histology-only prediction with frozen pathology encoders. A related but distinct line of work performs spatial super-resolution within-tissue by integrating measured low-resolution expression data with matched histology [37, 46]; we discuss the differences in protocol and provide direct comparisons in Appendix C.

Conditioning on Pathology Foundation Models. Pathology foundation models such as UNI [6], CONCH [27], and Phikon [11] produce embeddings that encode information at multiple biological scales, from sub-cellular morphology to tissue-level architecture, all of which can correlate with gene expression. Existing spatial transcriptomics models, including the recent generative approaches above, inject such embeddings through a single conditioning pathway, either as global modulation [33] or as keys for cross-attention [34]. How to jointly exploit the multi-scale richness of these embeddings for gene expression prediction remains an open design question.

Gene Regulatory Structure as a Design Prior. Gene expression is governed by hierarchical regulatory networks: master transcription factors coordinate downstream targets into modular structures of co-expressed genes, and this co-regulatory organization is a defining property of transcriptomic data [4, 24, 40]. Most computational pathology models treat the gene panel as an unstructured output vector [13, 15, 47]. Three recent predictors introduce relational structure. MERGE [12] encodes spatial and feature-space structure with a hierarchical graph attention network over spots, so its prior is over spots rather than genes. Gene-DML [39] aligns multi-scale image representations to expression through contrastive discrimination and has no explicit gene-relational structure. HAGE [8] is closest in spirit: it uses co-expression embeddings from Gene2Vec [9] as keys and values in cross-attention, but these embeddings are fixed, so the same gene relations are applied to every spot. Regulator attention instead generates its latent tokens from the current gene representation, so the induced grouping is conditioned on the input. A natural way to introduce structure is through a fixed set of latent tokens that summarize a variable-sized input, as used in Perceiver [18] and Set Transformer [25]; in these architectures, however, token identity is initialized at random and emerges purely from optimization, with no mechanism to anchor each token to a specific gene subset. How to translate the co-regulatory structure of gene expression into a token-level architectural prior remains open.

3 Method

Problem Formulation Given a histology image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and spatial coordinates $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ where $\mathbf{s}_i \in \mathbb{R}^2$, we predict the gene expression matrix $\mathbf{X} \in \mathbb{R}^{N\times G}$ where N denotes spatial spots and G denotes target genes. Unlike deterministic approaches that predict a point estimate $\hat{\mathbf{X}}$, we learn the conditional distribution $p(\mathbf{X}|\mathbf{I}, \mathbf{S})$, enabling sampling of multiple plausible expression profiles for the same histological pattern.

Overview The architecture has one primary contribution and two supporting design choices. *Regulator attention* (§3.1) is the core mechanism: latent tokens generated from gene representations themselves via a learned selector, routed back through cross-attention to modulate gene-level features, imposing a modular co-expression prior. *Dual conditioning* on pathology foundation model embeddings (§3.2) routes the same morphological signal through both AdaLN and cross-attention, so that global modulation and selective retrieval act as complementary pathways. The architecture is embedded in a *flow matching framework* (§3.3) that models gene expression as a conditional distribution, motivated by the substantial expression variance across morphologically similar tissue regions that point predictions cannot represent.

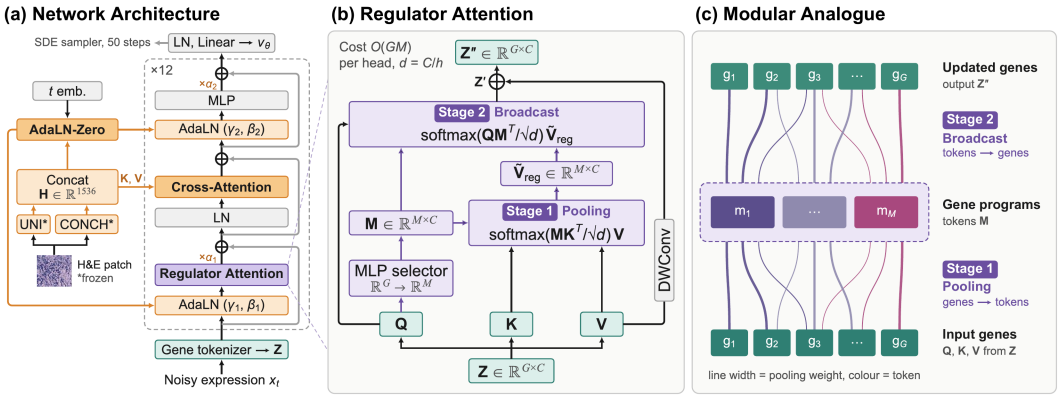


Figure 2: **Network architecture and modular analogue.** Amber marks morphology conditioning, purple regulator attention, and teal gene representations. **(a)** Flow matching network. The frozen UNI+CONCH feature $\mathbf{H}$ conditions each of the 12 blocks through AdaLN-Zero and as cross-attention keys and values (dual conditioning), and regulator attention replaces gene–gene self-attention. **(b)** Regulator attention. Tokens $\mathbf{M}$ generated from the gene queries by an MLP selector (Eq. 1) pool the genes in stage 1 and are read back by them in stage 2, at cost $\mathcal{O}(GM)$. **(c)** Modular analogue. Each token pools a soft, input-dependent subset of genes and broadcasts the summary back to them, so genes in the same program receive coordinated updates through a feed-forward path.

3.1 Regulator Attention

To impose this modular co-regulatory structure (Fig. 2c), we introduce M regulator tokens whose identity is generated from the gene representations themselves at each layer rather than initialized as a separate learnable array. This content-dependent design contrasts with input-agnostic learnable queries, where token identity is fixed prior to optimization. Given gene embeddings $\mathbf{Z} \in \mathbb{R}^{B\times G\times C}$ at a network layer, we compute projections $\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{ZW}_Q, \mathbf{ZW}_K, \mathbf{ZW}_V$. Regulator tokens are derived by aggregating gene representations across the gene dimension:

$$\mathbf{M} = \mathrm{MLP}(\mathbf{Q}^T)^T \in \mathbb{R}^{B\times M\times C}, \tag{1}$$

where $\mathrm{MLP} : \mathbb{R}^G \to \mathbb{R}^M$ learns weighted combinations across genes. Each row of $\mathbf{M}$ thus corresponds to a learned weighted pooling of the current gene representations, anchoring the token's identity to the gene content at that layer. For multi-head attention with h heads and dimension $d = C/h$, the tensors are reshaped to $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times h\times G\times d}$ and $\mathbf{M} \in \mathbb{R}^{B\times h\times M\times d}$.

The mechanism then operates in two stages (Figure 2b). Each regulator token first aggregates information from all genes:

$$\tilde{\mathbf{V}}_{\text{reg}} = \mathrm{softmax}\left(\frac{\mathbf{MK}^T}{\sqrt{d}}\right)\mathbf{V} \in \mathbb{R}^{B\times h\times M\times d}, \tag{2}$$

and each gene then queries the regulator representations:

$$\mathbf{Z}' = \mathrm{softmax}\left(\frac{\mathbf{QM}^T}{\sqrt{d}}\right)\tilde{\mathbf{V}}_{\text{reg}} \in \mathbb{R}^{B\times h\times G\times d}. \tag{3}$$

This two-stage path imposes a modular co-expression prior (Fig. 2c). Co-expressed genes form programs whose shared variation is carried by a small number of latent factors, the logic behind module eigengenes in co-expression analysis [24]. Each regulator token pools a soft subset of genes into one such factor in the first stage, and each gene reads back the factors of the programs it belongs to in the second, so genes in the same program receive coordinated updates while genes in different programs interact only through the M tokens. Because the pooling weights are generated from the current gene representation (Eq. 1), the active programs can differ across spots and layers, unlike a fixed set of learned queries. Because each gene interacts with M tokens rather than with all G genes, the attention cost is $\mathcal{O}(GM)$ instead of $\mathcal{O}(G^2)$ (§4.5). We additionally apply a depthwise convolution on $\mathbf{V}$ to capture local co-expression patterns between adjacent gene channels: $\mathbf{Z}'' = \mathbf{Z}' + \mathrm{DWConv}(\mathbf{V})$.

3.2 Dual Conditioning on Pathology Embeddings

We argue that the two standard conditioning strategies for foundation embeddings, AdaLN and cross-attention, are complementary rather than substitutes, and route the same embedding through both within each transformer block. AdaLN [33] compresses the embedding into per-spot scalar shift and scale factors that uniformly modulate all gene tokens at each layer, anchoring the generative trajectory in morphological context but unable to express gene-specific routing. Cross-attention [34] retrieves embedding components selectively on a gene-specific basis but lacks the layer-wide modulation that AdaLN provides. Pathology foundation models [6, 27] encode information at multiple biological scales (cellular composition to tissue architecture), and different aspects of this signal are relevant to different genes; using both pathways extracts both layer-wide and gene-specific signal from the same embedding.

Following Stem [47], we extract morphological features by concatenating frozen UNI [6] and CONCH [27] embeddings of histology patches, yielding $\mathbf{H} \in \mathbb{R}^{N \times D_H}$ aligned with the N spatial spots. AdaLN is applied as in DiT, where $\mathbf{H}$ is projected to per-spot shift and scale parameters that modulate the layer normalization output. Cross-attention is computed at each layer ℓ using gene-side hidden representations $\mathbf{Z}^{(\ell)} \in \mathbb{R}^{N \times G \times C}$ as queries:

$$\text{CrossAttn}(\mathbf{Z}^{(\ell)}, \mathbf{H}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \tag{4}$$

where $\mathbf{Q} = \mathbf{Z}^{(\ell)}\mathbf{W}_Q^{(\ell)}$ and $\mathbf{K}, \mathbf{V} = \mathbf{H}\mathbf{W}_K^{(\ell)}, \mathbf{H}\mathbf{W}_V^{(\ell)}$. Because the projection matrices are layer-specific, attention patterns can vary across depth, allowing different layers to emphasize different morphological aspects of the same embedding.

3.3 Flow Matching Framework

We use flow matching as the generative framework to model gene expression as a conditional distribution rather than as a point estimate (Fig. 1a,b). Compared to diffusion, flow matching with linear interpolation paths admits straighter transport trajectories that converge in fewer sampling steps; we verify this empirically in our ablation (Table 2a). Flow matching learns a time-dependent vector field $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{C})$ that defines a continuous trajectory from a Gaussian prior $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the target distribution $p(\mathbf{X}|\mathbf{I}, \mathbf{S})$, where $\mathbf{C} \in \mathbb{R}^{N \times D}$ denotes the morphological conditioning. We construct training paths via linear interpolation $\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ with $t \sim \mathcal{U}[0, 1]$, yielding the constant target vector field $\mathbf{u}_t = \mathbf{x}_1 - \mathbf{x}_0$ (optimal transport along straight lines). The model is trained to regress this target:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1, \mathbf{C}}\left[\left\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{C}) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right]. \tag{5}$$

At inference, we sample using the corresponding SDE

$$d\mathbf{x}_t = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{C})\, dt + g(t)\, d\mathbf{W}_t, \tag{6}$$

where $\mathbf{W}_t$ is a Wiener process and $g(t) = \eta(1-t)$ linearly decreases noise injection as the trajectory approaches the target distribution. The hyperparameter η controls stochasticity: $\eta = 0$ recovers deterministic ODE sampling, while $\eta > 0$ generates diverse samples reflecting expression heterogeneity. We integrate using Heun's second-order method [22] with 50 sampling steps.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on five spatial transcriptomics datasets from the HEST-1K benchmark [19] spanning human cancer (Breast Cancer I [13], Breast Cancer II [2], Squamous Cell Carcinoma [20]), human normal tissue (Human Bowel), and cross-species (Mouse Brain), all acquired on the 10x Visium platform with paired H&E. Following standard preprocessing [30, 41], we adopt a leave-one-slide-out evaluation protocol where one slide per dataset is held out as the test set. On Breast Cancer I and Squamous Cell Carcinoma the test patient contributes other sections to training, so we additionally report leave-one-patient-out evaluation (§4.5, Appendix C). Each dataset uses a panel

of highly variable genes (195–218 genes after filtering) shared by all methods. This panel lies at the 95.0–98.5th percentile of genome-wide spatial autocorrelation (Moran's I), and 77–91% of it falls within the genome-wide top-2000 spatially variable genes (Appendix J), so dispersion-based selection closely tracks spatial variability. Detailed dataset descriptions are provided in Appendix A.

Baseline Methods. We compare against representative deterministic predictors (ST-Net [13], EGN [43], BLEEP [42], HistoGene [31], TRIPLEX [7]) and recent generative approaches (the diffusion-based Stem [47] and the flow-based STFlow [15]), all of which operate in the histology-only setting shared by our method. A separate line of work performs within-tissue spatial super-resolution by integrating measured low-resolution expression with histology [37, 46]; we evaluate this family separately under cross-patient generalization in Appendix C, as their training assumptions differ from the histology-only setting considered here. All baselines are re-trained on our datasets using their official implementations and the hyperparameters from their original publications. We additionally reimplement three recent methods, MERGE [12], Gene-DML [39], and HAGE [8], under our exact protocol (identical slide splits, gene panels, $\log_2(\text{count}+1)$ targets, and frozen UNI+CONCH features) with their published hyperparameters and three seeds each (Appendix I).

Evaluation Metrics. We adopt Pearson correlation as the primary metric, reporting PCC@k averaged over the top-k highly variable genes, ranked by ground-truth variance on the held-out slide and therefore identical for all methods, for $k \in \{10, 50, 200\}$ (on Human Bowel, whose panel has 195 genes, PCC@200 covers the full panel). Mean Squared Error (MSE) and Mean Absolute Error (MAE) are reported as supporting metrics for pointwise error. We additionally evaluate biological structure preservation via unsupervised Leiden clustering on predicted expression, scored against ground truth by Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The full evaluation protocol, including the Leiden resolution selection procedure, is provided in Appendix B.

Implementation Details. We implement our model in PyTorch and train on a single NVIDIA RTX A5000 GPU. The transformer backbone has 12 layers, 6 attention heads, and hidden dimension 768. We use $M = 64$ regulator tokens. Foundation model features are obtained by concatenating frozen UNI [6] and CONCH [27] embeddings, yielding $\mathbf{H} \in \mathbb{R}^{N\times 1536}$, with both encoders kept frozen. Training targets are $\log_2(\text{count}+1)$. We train for 200 epochs with AdamW at learning rate 1e^{-4}. Inference uses Heun's second-order solver with 50 sampling steps, the SDE noise form follows the SiT defaults [28], and predictions are the mean of 20 samples per spot. $M{=}64$ is selected by the token-count sweep in Table 2b. The Layer-6, $t{=}0.4$ analysis point and the Leiden resolution enter only the interpretability and clustering analyses and affect no predictive metric. Computational cost comparisons against baselines are reported in Appendix H. Code will be released upon acceptance.

4.2 Main Results

Prediction performance across tissues. Our method achieves the highest correlation scores on all four benchmark datasets (Table 1), with the largest margins on highly variable genes where distributional modeling matters most. The three reimplemented methods are strong baselines: each exceeds Stem, the strongest previously reported baseline, on PCC@200 on both datasets where they are evaluated, by 0.077–0.126 on Breast Cancer I and 0.035–0.048 on Squamous Cell Carcinoma. Against the strongest of them, our PCC@200 is 0.3418 versus 0.2947 for MERGE on Breast Cancer I (+16.0%) and 0.7932 versus 0.7384 for Gene-DML on Squamous Cell Carcinoma (+7.4%). On Human Bowel and Mouse Brain, PCC@200 improves over Stem by 12.0% and 28.9%. The gap over earlier deterministic baselines is larger still: TRIPLEX reaches PCC@200 of only 0.0511 on Breast Cancer I.

MSE/MAE versus correlation. STFlow attains the lowest MSE and MAE on most datasets, and MERGE (both datasets) and Gene-DML (Breast Cancer I) also attain lower MSE than our method, while all of them reach lower correlation. We report this as observed. Squared-error training drives a point predictor toward the conditional mean $\mathbb{E}[Y \mid X]$, whose aggregate error is dominated by the genes and spots it fits well, whereas correlation on highly variable genes measures the focal patterns that the aggregate obscures. In Figure 3, TRIPLEX and MERGE smooth out the focal *EDF1* spots and the *CDSN* boundary that our method preserves. On the ten highest-correlation Squamous Cell Carcinoma markers, STFlow's per-gene MSE is $5.9\times$ ours and its per-gene correlations average -0.07, and when the error is restricted to each gene's top-k spots by ground-truth expression, STFlow is the worst method on all three datasets tested (Appendix F). The MSE gap of our method is not a Monte-Carlo artefact of the 20-sample mean, which explains 0.9% of the gap to the strongest

Table 1: **Quantitative comparison across four diverse tissue types.** We evaluate on human cancer (Breast Cancer I, Squamous Cell Carcinoma), human normal tissue (Human Bowel), and cross-species (Mouse Brain). Best results are shown in **bold** and second-best are underlined. † denotes deterministic ODE sampling. MERGE, Gene-DML, and HAGE are reimplemented under our protocol and reported as means over three seeds (standard deviations in Appendix I). ‡ HAGE trains its predictor with a correlation objective only, so its outputs carry no calibrated absolute scale and its MSE/MAE are excluded from ranking. Multi-seed results for our method are reported in Appendix E.

Method	Breast Cancer I					Squamous Cell Carcinoma				
	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓
ST-Net [13]	.3343	.2046	.0246	4.6487	1.7887	.2865	.2135	.0404	5.0162	1.8608
EGN [43]	.3805	.2236	.0332	4.7383	1.8075	.3066	.2330	.0449	5.5302	1.9620
BLEEP [42]	.3179	.2173	.0275	3.1445	1.4234	.2558	.2051	.0548	4.6867	1.6513
HistoGene [31]	.3415	.2241	.0391	4.7252	1.8058	.3314	.2580	.1107	5.7498	2.0164
TRIPLEX [7]	.3784	.2398	.0511	4.7592	1.8142	.3131	.2342	.0494	5.1512	1.8850
Stem [47]	.4689	.3252	.1689	1.7856	1.0295	.8159	.7689	.6901	1.3121	.8385
STFlow† [15]	.3809	.2490	.1150	**.7309**	**.6669**	.5924	.4650	.2953	**.6555**	**.6217**
MERGE [12]	.4962	.3528	.2947	1.2255	.8529	.8435	.8162	.7372	.9677	.7691
Gene-DML [39]	.4665	.3038	.2456	1.3784	.9082	.8494	.8222	.7384	1.0948	.8216
HAGE [8]	.4588	.3288	.2626	5.3169‡	1.8754‡	.8051	.7845	.7249	7.2620‡	1.9994‡
Ours	**.5453**	**.4191**	**.3418**	1.6719	.9982	**.8703**	**.8278**	**.7932**	1.0145	.7461

Method	Human Bowel					Mouse Brain				
	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓
ST-Net [13]	.0813	.0537	.0089	6.2004	2.1493	.1625	.0935	.0283	2.4738	1.2533
EGN [43]	.0712	.0491	.0050	6.1257	2.1268	.1916	.1047	.0291	2.6500	1.2828
BLEEP [42]	.0604	.0325	.0082	4.1441	1.7017	.2055	.0974	.0517	1.8150	1.0593
HistoGene [31]	.0803	.0524	.0068	6.7936	2.2536	.1778	.1040	.0274	2.3856	1.2276
TRIPLEX [7]	.0610	.0383	.0013	6.1023	2.1136	.2055	.0974	.0517	1.8150	1.0593
Stem [47]	.5267	.4277	.2384	2.0573	1.0789	.6269	.5354	.3737	.9049	.7208
STFlow† [15]	.5366	.4339	.2190	**1.0113**	**.7944**	.6097	.4667	.2892	**.4152**	**.4907**
Ours	**.6009**	**.4974**	**.2671**	2.1912	1.1139	**.6398**	**.5497**	**.4818**	.8794	.7095

baseline on Breast Cancer I and 10.1% on Squamous Cell Carcinoma. The same appendix identifies under-dispersed predictive intervals as a shared limitation of current generative baselines.

Generalization across tissues and species. On Squamous Cell Carcinoma, a morphologically heterogeneous tissue, the gap between generative methods and deterministic baselines is dramatic (Ours PCC@10 0.8703 vs. ST-Net 0.2865, TRIPLEX 0.3131), underscoring the importance of variance modeling in complex morphology. On Human Bowel, where morphological cues are subtler, our method still improves over Stem by 14.1% on PCC@10. The Mouse Brain result further indicates that learned representations transfer across species rather than capturing tissue-specific artifacts.

4.3 Ablation Study

We progressively ablate each architectural component starting from the diffusion baseline (Stem): adding dual conditioning (B+C: AdaLN with cross-attention from the same foundation embedding), content-dependent regulator tokens (B+C+R), and flow matching (Ours), evaluated on three datasets (Table 2a). Adding cross-attention alongside AdaLN (B $\rightarrow$ B+C) raises PCC@200 from 0.1689 to 0.2212 on Breast Cancer I (+31.0%), indicating that AdaLN's global modulation and cross-attention's selective retrieval extract complementary information from the same foundation embedding. Under the diffusion backbone, regulator tokens (B+C $\rightarrow$ B+C+R) raise PCC@200 on Breast Cancer I (+12.5%) and Mouse Brain (+4.3%) but not on Squamous Cell Carcinoma (-0.0059). Replacing diffusion with flow matching (B+C+R $\rightarrow$ Ours) yields the largest PCC@200 improvement on Breast Cancer I (+37.4%) and Mouse Brain (+26.9%), consistent with the benefit of straighter transport trajectories. Because this order adds regulator attention before the final backbone is in place, §4.5 isolates it under flow matching.

Table 2b shows performance generally improves with token count, with $M{=}64$ achieving the best correlation across datasets (PCC@200 from 0.2376 to 0.3418 on Breast Cancer I). MSE/MAE marginally favor smaller counts, and we select $M{=}64$ on the basis of correlation.

Table 2: **Ablation studies on the proposed components.** (a) Component decomposition across three tissues. (b) Performance scales with regulator token count, with $M = 64$ achieving the best correlation.

(a) Component ablation. **B**: diffusion baseline (Stem [47]); **C**: dual conditioning (AdaLN + cross-attention from the same foundation embedding); **R**: content-dependent regulator tokens. *Ours* additionally replaces diffusion with flow matching.

Method	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓
Breast Cancer I					
B	.4689	.3252	.1689	1.786	1.030
B+C	.5119	.3875	.2212	**1.431**	**.917**
B+C+R	.5425	.4184	.2488	1.590	.973
Ours	**.5453**	**.4191**	**.3418**	1.609	.980
Squamous Cell Carcinoma					
B	.8159	.7689	.6901	1.312	.839
B+C	.8371	.7941	.7128	1.135	.785
B+C+R	.8293	.7837	.7069	1.220	.803
Ours	**.8703**	**.8278**	**.7932**	**1.015**	**.746**
Mouse Brain					
B	.6269	.5354	.3737	.905	.721
B+C	.6325	.5389	.3638	.890	.720
B+C+R	.6353	.5436	.3796	.882	.709
Ours	**.6398**	**.5497**	**.4818**	**.879**	**.709**

(b) Impact of regulator token count $M \in \{4, 8, 16, 32, 64\}$. The number of tokens controls the bottleneck capacity of the regulator attention mechanism. BC: Breast Cancer; HB: Human Bowel.

Data	Metric	4	8	16	32	64
BC-I	PCC@50	.4140	.4016	.4174	.4156	**.4191**
	PCC@200	.2376	.2382	.2480	.2505	**.3418**
	MSE	1.556	1.583	**1.528**	1.622	1.672
	MAE	.9575	.9663	**.9487**	.9784	.9982
BC-II	PCC@50	.5966	.5932	.5934	.5951	**.6022**
	PCC@200	.4306	.4218	.4284	.4232	**.4369**
	MSE	1.284	1.223	1.230	1.236	**1.212**
	MAE	.8283	.8092	.8194	.8141	**.8053**
HB	PCC@50	.4897	.4890	.4909	.4913	**.4974**
	PCC@200	.2650	.2666	.2670	.2621	**.2671**
	MSE	2.114	2.164	2.115	**2.098**	2.191
	MAE	1.114	1.115	1.111	**1.089**	1.114

4.4 Beyond Predictive Accuracy

We score Leiden clustering of predicted expression against ground truth via ARI and NMI, averaged over 10 Leiden seeds (Table 3a, Figure 4a). Our method exceeds both generative baselines on every dataset and metric, and this ordering holds at every target cluster count from 4 to 10 (Appendix B). On Squamous Cell Carcinoma it attains the highest NMI, with ARI within one standard deviation of Gene-DML. It is the only method ranked in the top two by NMI on both datasets, and on Squamous Cell Carcinoma it leads in NMI at every target count from 7 to 10 (Figure 5), indicating that its predictions retain the finer subdivisions of the tissue. On Breast Cancer I, MERGE attains the highest ARI and NMI, a consequence of its design rather than of more accurate expression. Its graph attention aggregates features over spatial neighbors and cluster centroids, which smooths predictions into

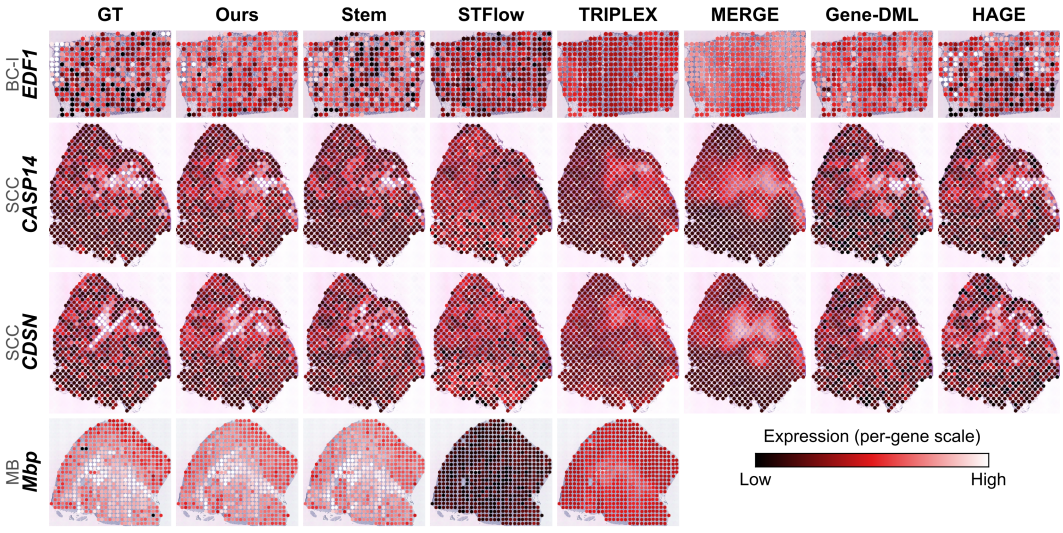


Figure 3: **Qualitative comparison of spatial gene expression predictions.** Each row shows one gene on the held-out slide of Breast Cancer I (BC-I), Squamous Cell Carcinoma (SCC), or Mouse Brain (MB), with colors scaled per gene. MERGE, Gene-DML, and HAGE are trained on BC-I and SCC only, and HAGE is shown after per-gene calibration (Appendix I).

Table 3: **Beyond predictive accuracy.** (a) Predicted expression preserves tissue organization, with our method exceeding both generative baselines on cluster correspondence in every setting. (b) Regulator tokens attend to distinct gene sets, and their pathway alignment exceeds matched null gene sets.

(a) Spatial clustering. ARI/NMI between Leiden clusters of predicted and ground-truth expression, mean$_{\pm \text{s.d.}}$ over 10 Leiden seeds. BC-I: Breast Cancer I. SCC: Squamous Cell Carcinoma.

	BC-I		SCC	
Method	ARI↑	NMI↑	ARI↑	NMI↑
Stem	$.061_{\pm.006}$	$.143_{\pm.009}$	$.210_{\pm.009}$	$.336_{\pm.012}$
STFlow	$.069_{\pm.015}$	$.119_{\pm.013}$	$.218_{\pm.015}$	$.343_{\pm.009}$
MERGE	$\mathbf{.168}_{\pm.017}$	$\mathbf{.235}_{\pm.018}$	$.199_{\pm.020}$	$.360_{\pm.015}$
Gene-DML	$\underline{.116}_{\pm.010}$	$.173_{\pm.010}$	$\mathbf{.278}_{\pm.027}$	$\underline{.414}_{\pm.018}$
HAGE	$.065_{\pm.018}$	$.134_{\pm.013}$	$.257_{\pm.032}$	$.395_{\pm.023}$
Ours	$.105_{\pm.010}$	$\underline{.176}_{\pm.011}$	$\underline{.269}_{\pm.037}$	$\mathbf{.429}_{\pm.014}$

(b) Token specificity and null-controlled pathway alignment (Breast Cancer I). Uniq.: one minus mean pairwise cosine similarity of attention. Jaccard: mean pairwise overlap of top-30 gene sets. Rate: fraction of tokens enriched for a Hallmark set [26] (panel background, BH $p < 0.05$). Nulls: 100 draws of size-matched gene sets.

Gene sets	**Uniq.↑**	**Jaccard↓**	**Rate**
Regulator tokens (Ours)	**.575**	**.114**	.234
Learned queries	.052	.244	.219
Uniform random	—	—	.016
Expression-matched	—	—	$.068_{\pm.029}$
Variance-matched	—	—	$.023_{\pm.019}$
Detection-matched	—	—	$.072_{\pm.033}$

homogeneous regions (Figure 3). In the seed-0 partitions, the mean spot-to-centroid distance of MERGE is 39% (Breast Cancer I) and 28% (Squamous Cell Carcinoma) below that of the ground truth, whereas ours lies within 3% of it on both. MERGE thus produces domains more contiguous than those of the tissue itself, which raises agreement on Breast Cancer I but yields the lowest ARI on Squamous Cell Carcinoma, and its PCC@200 remains below ours on both datasets (Table 1). Our model predicts each spot independently from its own morphology (Appendix L), so its cluster agreement reflects the predicted expression rather than a spatial prior, which is what downstream analysis of predicted tissue structure relies on.

We extract regulator-to-gene attention weights from Layer 6 at timestep $t = 0.4$, where token specialization is maximal (Appendix G). Specificity is the primary evidence (Table 3b): learned queries converge on largely the same genes in every token, whereas regulator tokens attend to distinct gene programs, such as immunoglobulins for M14 and *FNBP1L*/*PERP* for M2 (Figure 4b). Regulator tokens reach a uniqueness of 0.575 against 0.052 for learned queries, an 11-fold difference, and their top-30 gene sets overlap less than half as much (Jaccard 0.114 versus 0.244). Testing each token's top-30 genes against MSigDB Hallmark [26] with the gene panel as the Fisher background, 15 of 64 tokens (23.4%) are enriched at adjusted $p < 0.05$, above every matched null (empirical $p < 0.01$), so the alignment is not an artefact of panel composition. The rate does not separate regulator tokens from learned queries (21.9%), which is why specificity rather than hit rate is the discriminating quantity. Attention is also not drawn to abundant genes: per-gene total attention correlates negatively with mean expression (Pearson -0.48), variance (-0.29), and detection frequency (-0.58).

4.5 Further Analyses

Table 4: **Further analyses.** (a) Gene-mixing variants under flow matching with dual conditioning on Mouse Brain, three seeds at a reduced budget. †Original run, single seed. (b) SPA154 (Patient A). LOSO keeps five other Patient A sections in training, LOPO holds out all six (three seeds).

(a) Isolation of regulator attention.

Gene mixing	**PCC@10↑**	**PCC@50↑**	**PCC@200↑**
Full self-attention	$.5333_{\pm.0057}$	$.5112_{\pm.0086}$	$.4129_{\pm.0080}$
ISAB [25]	—	—	$.3906_{\pm.0023}$
Learned queries†	—	—	.388
Regulator tokens (Ours)	$\mathbf{.6397}_{\pm.0016}$	$\mathbf{.5497}_{\pm.0025}$	$\mathbf{.4805}_{\pm.0053}$

(b) Patient-level holdout (Appendix C).

Method (protocol)	**PCC@10↑**	**PCC@200↑**
Ours (LOSO)	.545	.342
Ours (LOPO)	$\mathbf{.382}_{\pm.021}$	$\mathbf{.141}_{\pm.013}$
iStar, best cross-patient	.223	.102
iSCALE, best cross-patient	.167	.100
iStar, best same-patient	.368	.271

Replacing regulator attention in our model with full gene–gene self-attention on Mouse Brain (Table 4a) lowers PCC@200 by 0.0676 (9.9σ of the pooled seed standard deviation), $4.3\times$ the contribution measured under the diffusion backbone (Table 2a). Content-dependence, not compression, is

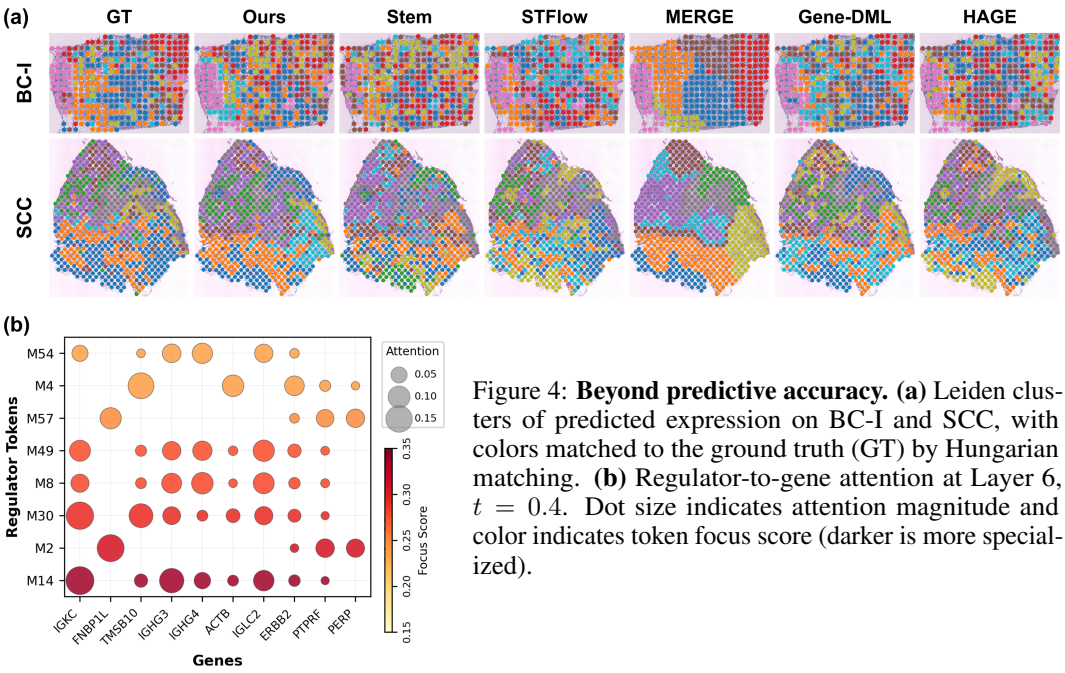


Figure 4: **Beyond predictive accuracy. (a)** Leiden clusters of predicted expression on BC-I and SCC, with colors matched to the ground truth (GT) by Hungarian matching. **(b)** Regulator-to-gene attention at Layer 6, $t = 0.4$. Dot size indicates attention magnitude and color indicates token focus score (darker is more specialized).

the source of this gain: learned queries that differ from our module only in the latent array [18] and a faithful ISAB [25] both fall below the model without any bottleneck (Appendix D).

On Breast Cancer I, the test slide SPA154 belongs to Patient A, who contributes six sections. With all six held out (LOPO), our method exceeds the best cross-patient configuration of iStar [46] and iSCALE [37] by $1.4\times$ on PCC@200 (Table 4b), and their cross-patient means by $3.5\times$ and $5.6\times$. It falls below iStar's best same-patient configuration on PCC@200 while remaining above it on PCC@10. On Squamous Cell Carcinoma the same holdout yields PCC@200 of 0.411 (single seed). Slide-level numbers therefore include same-patient continuity, which likely applies to every method evaluated under those splits.

Full gene–gene self-attention materializes a $B \times h \times G \times G$ attention tensor, whereas regulator attention scales with GM. On a 24 GB GPU (Appendix K) the two cost the same at the current panel size ($G \approx 210$). At our inference batch (B=200) and G=2000, full self-attention fails on a 17.9 GiB attention tensor while regulator attention runs in 13.2 GB. At Xenium Prime 5K scale (G=5000) that tensor reaches 111.8 GiB, and at a batch where both variants run (B=8) regulator attention uses $5.3\times$ less peak memory.

5 Conclusion

We presented regulator attention, an architectural mechanism whose latent tokens are generated from gene representations themselves, imposing a modular co-expression prior within a flow matching framework with dual conditioning on pathology foundation embeddings. Across four tissue types, the method outperforms deterministic and generative baselines, including three recent methods reimplemented under our protocol on two of them, with the largest gains on highly variable genes. Isolated under the final backbone on Mouse Brain, content-dependent tokens improve over full self-attention while input-agnostic bottlenecks fall below it, the learned tokens attend to distinct gene sets whose pathway alignment exceeds matched nulls, and the bottleneck removes the quadratic memory term of gene–gene attention. Domain-informed design can thus yield models whose internal representations are interpretable in the language of the underlying biology. The evaluation covers the marker-gene regime of Visium panels, and accuracy on wider spatially variable gene sets, on high-plex platforms, and under patient-level splits for every cohort remains to be established (Appendix L).

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) in 2026 (No. RS-2026-25611334, AI Mechanician: A Physics- and Math-Based AI that Identifies and Unifies Multi-Transport Phenomena in Irregular Porous Media from Visualization), the NRF grant funded by the Korea government (MSIT) (No. RS-2025-00516124), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. IITP-2026-25616313, AI Star Fellowship Support Program), the Basic Science Research Program through the NRF funded by the Ministry of Education (No. RS-2025-25431811), and the Technology Development Program funded by the Ministry of SMEs and Startups (MSS, Korea) (No. 2420016015).

References

[1] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6): 450–461, 2007. doi: 10.1038/nrg2102.

[2] A. Andersson, L. Larsson, L. Stenbeck, F. Salmén, A. Ehinger, S. Z. Wu, G. Al-Eryani, D. Roden, A. Swarbrick, Å. Borg, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):6012, 2021. doi: 10.1038/s41467-021-26271-2.

[3] M. Asp, S. Giacomello, L. Larsson, C. Wu, D. Fürth, X. Qian, E. Wärdell, J. Custodio, J. Reimegård, F. Salmén, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660.e19, 2019. doi: 10.1016/j.cell.2019.11.025.

[4] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. doi: 10.1038/nrg1272.

[5] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14:128, 2013. doi: 10.1186/1471-2105-14-128.

[6] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

[7] Y. Chung, J. H. Ha, K. C. Im, and J. S. Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11591–11600, 2024. doi: 10.1109/CVPR52733.2024.01101.

[8] T. M. Dang, H. Li, Y. Guo, H. Ma, F. Jiang, Y. Miao, Q. Zhou, J. Gao, and J. Huang. HAGE: Hierarchical alignment gene-enhanced pathology representation learning with spatial transcriptomics. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15960 of *Lecture Notes in Computer Science*, pages 228–238. Springer, 2025. doi: 10.1007/978-3-032-04927-8_22.

[9] J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, and D. Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(Suppl 1):82, 2019. doi: 10.1186/s12864-018-5370-x.

[10] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002. doi: 10.1126/science.1070919.

[11] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. Camara, A. Mac Kain, C. Saillard, and J.-B. Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757.

[12] A. Ganguly, D. Chatterjee, W. Huang, J. Zhang, A. Yurovsky, T. S. Johnson, and C. Chen. MERGE: Multi-faceted hierarchical graph-based GNN for gene expression prediction from whole slide histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15611–15620, 2025. doi: 10.1109/CVPR52734.2025.01455.

[13] B. He, L. Bergenstråhle, L. Stenbeck, A. Abid, A. Andersson, Å. Borg, J. Maaskola, J. Lundeberg, and J. Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834, 2020. doi: 10.1038/s41551-020-0578-x.

[14] K. Hemker, A. H. Song, C. Almagro-Pérez, G. Jaume, S. J. Wagner, A. Vaidya, N. Simidjievski, M. Jamnik, and F. Mahmood. Towards spatial transcriptomics-driven pathology foundation models. *arXiv preprint arXiv:2602.14177*, 2026. doi: 10.48550/arXiv.2602.14177.

[15] T. Huang, T. Liu, M. Babadi, W. Jin, and R. Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 25550–25565, 2025.

[16] T. Huang, T. Liu, M. Babadi, R. Ying, and W. Jin. STPath: a generative foundation model for integrating spatial transcriptomics and whole-slide images. *npj Digital Medicine*, 8(1):659, 2025. doi: 10.1038/s41746-025-02020-3.

[17] M. V. Hunter, R. Moncada, J. M. Weiss, I. Yanai, and R. M. White. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nature Communications*, 12(1):6278, 2021. doi: 10.1038/s41467-021-26614-z.

[18] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664, 2021.

[19] G. Jaume, P. Doucet, A. H. Song, M. Y. Lu, C. Almagro-Pérez, S. J. Wagner, A. J. Vaidya, R. J. Chen, D. F. K. Williamson, A. Kim, and F. Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, volume 37, pages 53798–53833, 2024.

[20] A. L. Ji, A. J. Rubin, K. Thrane, S. Jiang, D. L. Reynolds, R. M. Meyers, M. G. Guo, B. M. George, A. Mollbrink, J. Bergenstråhle, L. Larsson, Y. Bai, B. Zhu, A. Bhaduri, J. M. Meyers, X. Rovira-Clavé, S. T. Hollmig, S. Z. Aasi, G. P. Nolan, J. Lundeberg, and P. A. Khavari. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514.e22, 2020. doi: 10.1016/j.cell.2020.05.039.

[21] Y. Jin, Y. Zuo, G. Li, W. Liu, Y. Pan, T. Fan, X. Fu, X. Yao, and Y. Peng. Advances in spatial transcriptomics and its applications in cancer research. *Molecular Cancer*, 23(1):129, 2024. doi: 10.1186/s12943-024-02040-9.

[22] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577, 2022.

[23] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018. doi: 10.1016/j.cell.2018.01.029.

[24] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008. doi: 10.1186/1471-2105-9-559.

[25] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753, 2019.

[26] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

[27] M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

[28] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, volume 15135 of *Lecture Notes in Computer Science*, pages 23–40. Springer, 2024. doi: 10.1007/978-3-031-72980-5_2.

[29] S. Maniatis, T. Äijö, S. Vickovic, C. Braine, K. Kang, A. Mollbrink, D. Fagegaltier, Ž. Andrusivová, S. Saarenpää, G. Saiz-Castro, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435):89–93, 2019. doi: 10.1126/science.aav9776.

[30] G. Palla, H. Spitzer, M. Klein, D. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg, I. Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022. doi: 10.1038/s41592-021-01358-2.

[31] M. Pang, K. Su, and M. Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021. doi: 10.1101/2021.11.28.470212.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[33] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

[35] J. W. Schoggins. Interferon-stimulated genes: What do they all do? *Annual Review of Virology*, 6:567–584, 2019. doi: 10.1146/annurev-virology-092818-015756.

[36] J. W. Schoggins, S. J. Wilson, M. Panis, M. Y. Murphy, C. T. Jones, P. Bieniasz, and C. M. Rice. A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, 472(7344):481–485, 2011. doi: 10.1038/nature09907.

[37] A. Schroeder, M. L. Loth, C. Luo, S. Yao, H. Yan, D. Zhang, S. Piya, E. Plowey, W. Hu, J. R. Clemenceau, et al. Scaling up spatial transcriptomics for large-sized tissues: uncovering cellular-level tissue architecture beyond conventional platforms with iSCALE. *Nature Methods*, 22(9):1911–1922, 2025. doi: 10.1038/s41592-025-02770-8.

[38] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, 2002. doi: 10.1038/ng881.

[39] Y. Song, J. Fan, H. Chang, and W. Cai. Gene-DML: Dual-pathway multi-level discrimination for gene expression prediction from histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5090–5099, 2026. doi: 10.1109/WACV61042.2026.00494.

[40] F. Spitz and E. E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012. doi: 10.1038/nrg3207.

[41] F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

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

[43] Y. Yang, M. Z. Hossain, E. A. Stone, and S. Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5039–5048, 2023. doi: 10.1109/WACV56688.2023.00501.

[44] Y. Yang, M. Z. Hossain, E. Stone, and S. Rahman. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognition*, 145:109966, 2024. doi: 10.1016/j.patcog.2023.109966.

[45] Z. Zeng, Y. Li, Y. Li, and Y. Luo. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biology*, 23(1):83, 2022. doi: 10.1186/s13059-022-02653-7.

[46] 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. doi: 10.1038/s41587-023-02019-9.

[47] S. Zhu, Y. Zhu, M. Tao, and P. Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *International Conference on Learning Representations (ICLR)*, 2025.

A Dataset Details

We evaluate on five spatial transcriptomics datasets spanning different tissue types, pathological conditions, and species, all sourced from the HEST-1K benchmark [19]. Breast Cancer I [13] comprises invasive ductal carcinoma samples with heterogeneous tumor architectures including invasive carcinoma, ductal carcinoma in situ, and tumor-infiltrating lymphocytes. Breast Cancer II [2] focuses on HER2-positive breast tumors, providing a more homogeneous molecular context for assessing subtle morphological-molecular variations within a specific subtype. Squamous Cell Carcinoma [20] contains cutaneous skin cancer samples with distinct keratinization patterns and spatial organization of tumor, fibroblast, and immune populations. Human Bowel represents normal tissue architecture with organized crypt structures. The Mouse Brain dataset enables cross-species validation on neural tissue with highly organized cortical structure. All datasets are acquired on the 10x Visium platform with paired H&E images and gene expression measurements at predefined spots. We follow standard preprocessing including quality filtering, normalization, and log-transformation [30, 41], and adopt a leave-one-slide-out evaluation protocol in which one slide per dataset is held out as the test set while the remaining slides are used for training. Under this protocol other sections of the test patient can remain in training. On Breast Cancer I, the test slide SPA154 belongs to Patient A, who contributes six of the 36 sections from eight patients. On Squamous Cell Carcinoma, the 12 sections come from four patients with three sections each, and the other two sections of the test slide's patient are in the training set. Leave-one-patient-out results for both are reported in Appendix C. Gene panel sizes after filtering are 210 (Breast Cancer I), 218 (Squamous Cell Carcinoma), 195 (Human Bowel), and 209 (Mouse Brain), with mitochondrial genes excluded. Training targets are $\log_2(\text{count} + 1)$.

B Evaluation Protocol Details

Pointwise metrics. PCC@k, MSE, and MAE are computed per-gene and averaged across the test set, following established protocols [13, 47]. The k genes for PCC@k are the top-k by ground-truth variance on the held-out slide, so the gene set is identical for every method. For sampling-based methods all metrics are computed on the mean of 20 samples per spot.

Clustering pipeline. For both ground truth and predicted expression, we construct a 50-component PCA representation followed by a k-nearest neighbor graph ($k_{\text{NN}} = 15$), then apply Leiden clustering. The Leiden resolution is selected by binary search over $[0.01, 3.0]$ for up to 25 iterations: at each step the midpoint resolution is evaluated, and the interval is narrowed toward smaller resolutions if the resulting cluster count exceeds the target k_{tgt} and toward larger resolutions otherwise. The resolution achieving the cluster count closest to k_{tgt} is retained. The target k_{tgt} is derived from the known anatomical structure of each tissue and is applied consistently to ground truth and all methods so that all partitions being compared have the same target granularity.

Preprocessing. Ground truth expression is preprocessed with total-count normalization (target sum 10^4) followed by log1p transformation prior to clustering. Predicted expression is clustered without further normalization, since all predictions are already on the log-normalized scale matching the training target. NaN and Inf values, if any, are replaced with zero before PCA.

Cluster correspondence and metrics. ARI and NMI are computed between the resulting cluster assignments on predicted and ground truth profiles using their standard scikit-learn implementations [32]. Table 3a reports the mean and standard deviation over Leiden random seeds 0–9, with ground truth and predictions clustered under the same seed, at k_{tgt}=7 on Breast Cancer I and k_{tgt}=8 on Squamous Cell Carcinoma. Figure 5 repeats the analysis for k_{tgt} from 4 to 10. For the spatial overlay visualization (Figure 4a, seed 0), predicted cluster labels are remapped to ground truth labels via Hungarian matching on the contingency matrix to enable consistent coloring across methods. This remapping is purely cosmetic and does not affect the ARI/NMI values.

Spot alignment. For each test slide, the number of evaluation spots is set to the minimum across ground truth and all method predictions to ensure all comparisons are computed on identical spot indices. Predictions are gene-aligned to the held-out gene panel by selecting the columns corresponding to genes present in both the prediction and the ground truth.

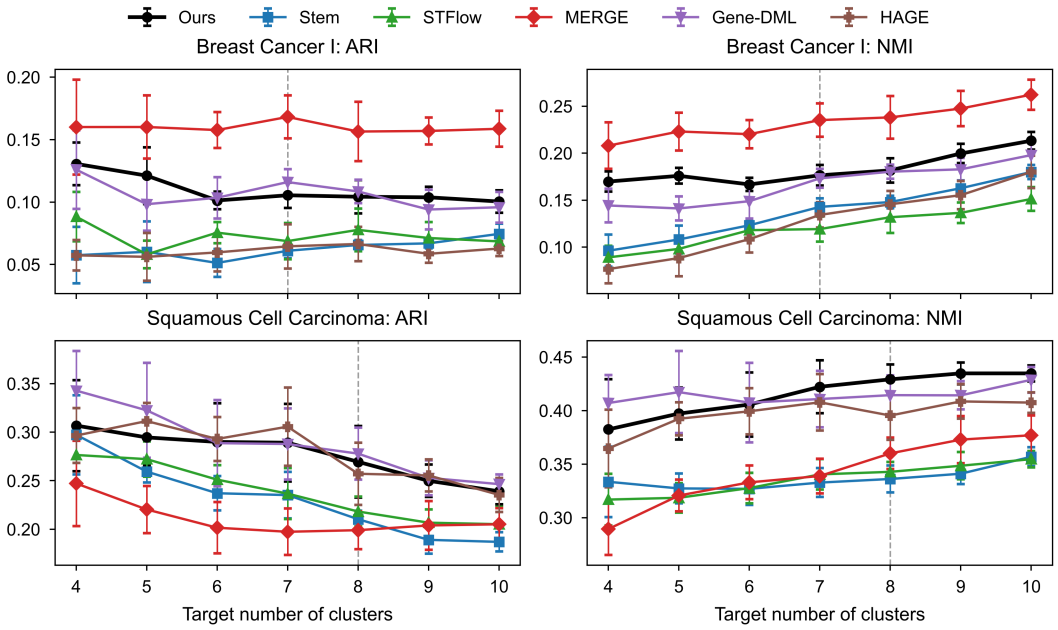


Figure 5: **Clustering agreement across target cluster counts.** ARI and NMI between Leiden clusters of predicted and ground-truth expression for k_{tgt} from 4 to 10, mean $\pm$ s.d. over 10 Leiden seeds. Dashed lines mark the target used in Table 3a.

C Cross-Patient Generalization Analysis

C.1 Motivation and Protocol

Recent super-resolution methods iStar [46] and iSCALE [37] address a related problem: given a low-resolution Visium slide and its matched histology, they upscale expression to higher resolution. Although their primary use case requires ST data as input, both methods explicitly support *out-of-sample prediction*, in which a model trained on one slide is evaluated on another slide using only its histology image and no ST data at inference [37, 46]. This setting overlaps with the de novo prediction task addressed in our paper, and we therefore include them as additional baselines in this appendix.

The Andersson Breast Cancer dataset [2] contains 36 slides from 8 patients (A–H). Our test slide SPA154 belongs to Patient A, who contributes 6 slides (SPA149–154). iStar and iSCALE are designed for single-slide training [46]: each model is trained on exactly one slide and evaluated on another. We therefore enumerate all 35 possible single-slide training configurations and stratify them by patient identity:

- **Same-patient** (5 configurations): training slides SPA149, 150, 151, 152, 153, all from Patient A. These represent adjacent or near-adjacent tissue sections from the same tissue block, which is the original use case of iStar and iSCALE.
- **Cross-patient** (30 configurations): training slides from Patients B–H. These reflect the clinically relevant scenario in which the test patient is unseen during training.

For each baseline we report (i) the best single-slide configuration in each stratum and (ii) the aggregate mean across all configurations in the stratum, a protocol that is maximally favorable to the baselines, since within each stratum we let them select their own optimal training slide *per metric*. We report our method under two protocols. Leave-one-slide-out, shared with Stem [47], STFlow [15], and TRIPLEX [7], trains on the remaining 35 slides, which include the five other sections of Patient A. Leave-one-patient-out trains on the 30 slides of Patients B–H with three seeds, so that no slide from the test patient is seen in training. It matches the condition of the cross-patient stratum and is the protocol that supports a cross-patient comparison.

C.2 Best-Case Comparison

Table 5 reports the best single-slide configuration in each stratum.

Table 5: **Best-case single-slide configurations of iStar and iSCALE, stratified by patient identity, versus our leave-one-slide-out and leave-one-patient-out results on SPA154.** *Same-patient* configurations train on a slide from Patient A (the test patient), *cross-patient* configurations train on a slide from Patients B–H. Leave-one-patient-out: mean $\pm$ std over three seeds. Best in **bold**.

Method	Training slide	PCC@10↑	PCC@200↑
Ours	Leave-one-slide-out (35 slides)	**0.545**	**0.342**
Ours	Leave-one-patient-out (30 slides)	0.382 ± 0.021	0.141 ± 0.013
iStar (best, same-patient)	SPA152 (Patient A)	0.368	0.271
iStar (best, cross-patient)	SPA138 (Patient C)	0.223	0.102
iSCALE (best, same-patient)	SPA152 (Patient A)	0.259	0.148
iSCALE (best, cross-patient)	SPA138 (Patient C)	0.167	0.100

Under leave-one-slide-out, our method outperforms even the most favorable same-patient setting, training on an adjacent section from the same tissue block, by $+26\%$ over iStar and $+131\%$ over iSCALE on PCC@200. Because that protocol includes other Patient A sections, the cross-patient comparison rests on leave-one-patient-out. There, our method exceeds the best cross-patient configuration of iStar by $1.71\times$ and of iSCALE by $2.29\times$ on PCC@10, and by $1.38\times$ and $1.41\times$ on PCC@200. Under strict holdout it no longer exceeds iStar's best same-patient configuration on PCC@200 (0.141 against 0.271), while remaining above it on PCC@10 (0.382 against 0.368).

C.3 Aggregate Stratum Statistics

Best-case numbers may overstate baseline reliability because they cherry-pick the optimal training slide per metric. Table 6 reports the mean across all configurations within each stratum, which is the deployment-relevant statistic: a practitioner cannot select the optimal training slide a priori without ground-truth ST data on the test slide.

Table 6: **Aggregate stratum statistics across all 35 single-slide training configurations.** Mean PCC@200 within each stratum. The same-patient to cross-patient ratio is approximately $5\times$ for both methods.

Method	Stratum	N configs	Mean PCC@200
iStar	Same-patient	5	0.202
iStar	Cross-patient	30	0.040
iSCALE	Same-patient	5	0.105
iSCALE	Cross-patient	30	0.025
Ours	Leave-one-slide-out	1	**0.342**
Ours	Leave-one-patient-out	1 (3 seeds)	0.141

C.4 Discussion

The patient stratification reveals a structural distinction. iStar and iSCALE achieve their strongest performance when the training and test slides belong to the same tissue block, which is consistent with their original publications targeting within-tissue super-resolution [37, 46]. When required to generalize to unseen patients, both methods degrade by approximately $5\times$, indicating that they exploit local tissue continuity rather than learn patient-invariant histology-to-expression mappings. Under leave-one-patient-out, our method exceeds the cross-patient stratum means of iStar and iSCALE by $3.5\times$ and $5.6\times$ on PCC@200 and the best cross-patient configuration of each, which is selected per metric with access to the test slide, supporting its applicability to retrospective analysis of archival samples without matched ST data. Its PCC@200 drops from 0.342 to 0.141 when the other Patient A sections are removed from training, so the leave-one-slide-out figure includes same-patient continuity,

which likely applies to every method evaluated under that split. The same holdout on Squamous Cell Carcinoma (Patient 2, sections NCBI768–770 held out, evaluated on NCBI770) yields PCC@200 of 0.411 with a single seed. No cross-patient baseline is available for that cohort, and the remaining datasets use slide-level splits only (Appendix L).

D Bottleneck Variants

We compare regulator tokens against two input-agnostic bottlenecks and against no bottleneck, all under flow matching with dual conditioning. The learned-query variant, in the spirit of Perceiver [18], shares the QKV projections, the two-stage genes $\rightarrow$ tokens $\rightarrow$ genes path, and the depthwise convolution with our module. The only difference is the source of the token identities: generated from gene representations via Eq. 1 (ours) versus a separate learnable array. A faithful ISAB [25] uses separate projections for the latent and input streams and therefore changes several factors at once. We report it as an additional point rather than as the controlled comparison. The no-bottleneck variant replaces the two-stage path with full gene–gene self-attention while keeping the QKV projections and the depthwise convolution. All variants share the token count (M=64), hidden dimensionality, and conditioning pathways.

Table 7: **Bottleneck variants under flow matching with dual conditioning.** Left: token specificity on Breast Cancer I. Uniqueness follows Appendix G, Jaccard is the mean pairwise overlap of top-30 gene sets, and the enrichment rate uses MSigDB Hallmark with the gene panel as Fisher background. Right: Mouse Brain PCC@200, mean $\pm$ std over three seeds at a reduced training budget, at which regulator tokens reproduce Table 1 within 0.002. †Original training run, single seed.

Breast Cancer I	**Regulator**	**Learned q.**
Uniqueness↑	**.575**	.052
Jaccard↓	**.114**	.244
Enrichment rate	.234	.219

Mouse Brain	**PCC@200↑**
Regulator tokens (Ours)	**.4805**$_{\pm.0053}$
No bottleneck	.4129$_{\pm.0080}$
ISAB	.3906$_{\pm.0023}$
Learned queries†	.388

On Mouse Brain, both input-agnostic bottlenecks fall below the model without any bottleneck, while regulator tokens exceed it by 0.0676 PCC@200. Compression alone therefore does not help, and the gain comes from content-dependent generation. The specificity metrics separate the variants more sharply than prediction does: learned queries converge on largely the same genes in every token (uniqueness 0.052 and Jaccard 0.244 on Breast Cancer I, against 0.575 and 0.114 for regulator tokens).

Enrichment rate. With the gene panel as the Fisher background and Hallmark alone, the enrichment rate is 23.4% for regulator tokens and 21.9% for learned queries. Enrichr's default genome-wide background [5] yields rates above 84% for both variants because it inflates enrichment for any subset of a panel pre-selected for biological coherence, so all rates in this paper use the panel background and Hallmark alone. Regulator tokens exceed every matched null under this background (Appendix G.3), but the rate does not separate them from learned queries. The discriminating quantity is whether each token captures a distinct gene program, which uniqueness and Jaccard overlap measure directly.

Mouse Brain enrichment. We omit pathway enrichment results for Mouse Brain because MSigDB Hallmark [26], KEGG, and Reactome are predominantly curated for human genes. Cross-species enrichment introduces orthology-mapping artifacts that confound the comparison; we therefore restrict the enrichment analysis to BC-I.

E Multi-Seed Stability

To assess the variance of our reported metrics, we re-ran our method with three independent random seeds on Breast Cancer I and Mouse Brain (representative of human cancer and cross-species normal tissue, respectively). All other settings (data splits, hyperparameters, training schedule) are identical across seeds.

Table 8: **Multi-seed stability of our method.** Mean $\pm$ standard deviation across three independent random seeds on Breast Cancer I and Mouse Brain. Dashes indicate metrics for which a multi-seed run is not available.

Dataset	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓
Breast Cancer I	0.5453 ± 0.0041	0.4191 ± 0.0014	0.3418 ± 0.0091	—	—
Mouse Brain	0.6398 ± 0.0015	0.5497 ± 0.0024	0.4818 ± 0.0066	0.8794 ± 0.0060	0.7095 ± 0.0021

The standard deviation is below 0.01 PCC@200 on both datasets. On Breast Cancer I, the margins over the strongest baseline (MERGE) are 12.0σ, 47.4σ, and 5.2σ of our seed standard deviation on PCC@10, PCC@50, and PCC@200, and 1.9σ, 2.8σ, and 1.9σ of MERGE's own seed spread (Appendix I). On Mouse Brain, the margins over Stem lie well outside our noise band on all three correlation metrics. Squamous Cell Carcinoma and Human Bowel have no multi-seed run of our method, and on Squamous Cell Carcinoma the PCC@50 margin over Gene-DML (+0.0056) is 1.3σ of Gene-DML's seed spread.

F Per-Gene Marker Analysis and Distributional Behavior

F.1 Motivation

Section 4.4 adopts Pearson correlation as the primary metric for spatial pattern fidelity, with MSE/MAE reported as supporting pointwise error measures. This appendix provides three complementary pieces of evidence: (i) a per-gene breakdown on biologically reliable marker genes that quantitatively explains the apparent discrepancy between STFlow's competitive aggregate MSE and its qualitative under-performance on visualized high-variance genes (Figure 3); (ii) a per-spot complement that restricts MSE to per-gene high-expression spot subsets, isolating the hotspot regime where conditional-mean predictors over-smooth; and (iii) per-sample distributional measurements of sample variance and predictive interval coverage for the generative baselines that admit multi-sample inference (our method and Stem). All pointwise metrics are computed on mean predictions (averaged over 20 samples for sampling-based methods), matching the aggregation convention used in the main results.

F.2 Per-Gene Marker Gene Analysis on Squamous Cell Carcinoma

We select the ten marker genes with the highest per-gene Pearson correlation under our method on the Squamous Cell Carcinoma test slide (NCBI770). These genes are predominantly keratin (*KRT5*, *KRT16*, *KRT6C*) and S100 family members (*S100A2*, *S100A7*, *S100A8*, *S100A9*, *SFN*), characteristic of squamous epithelial differentiation in this dataset. Together they constitute the high-variance subset whose spatial expression patterns are visualized in qualitative comparisons.

Two patterns emerge. First, STFlow attains low aggregate MSE on this dataset under standard evaluation but exhibits dramatically inflated per-gene MSE on these marker genes (geometric mean $5.9\times$ ours, with individual ratios up to $11.2\times$ on *KRT16*, $10.4\times$ on *KRT5*, and $10.2\times$ on *S100A8*, Table 9). Second, STFlow's per-gene Pearson correlations on the same genes are consistently near zero or negative (mean -0.07, range -0.30 to 0.20), in contrast to the strong positive correlations achieved by both Stem (mean 0.31) and our method (mean 0.36). Together these results indicate that STFlow's aggregate MSE advantage comes from the genes and spots that a conditional-mean predictor fits well, while the marker genes that drive biological interpretation are not well recovered. The panel itself contains no large low-variance pool: in the training target space the maximum-to-median variance ratio within the panel is 1.8–2.9 across the four datasets, with only 0–11 effectively constant genes, so the effect concerns difficulty rather than variance. This reconciles STFlow's qualitative under-performance (Figure 3) with its quantitative aggregate score and supports our adoption of correlation as the primary metric.

F.3 Per-Spot High-Expression Subset Analysis

Section F.2 filters by gene (restricting to high-variance markers, MSE averaged across all spots). Here we report a per-spot complement: for each gene we restrict the pointwise error computation to the

Table 9: **Per-gene Pearson correlation and MSE on Squamous Cell Carcinoma marker genes.** STFlow's correlations are near zero or negative on most of the ten genes and its per-gene MSE is consistently inflated (geometric mean $5.9\times$ ours), in stark contrast to its competitive aggregate MSE.

	Pearson correlation↑			MSE↓		
Gene	**Ours**	**Stem**	**STFlow**	**Ours**	**Stem**	**STFlow**
GJB2	**0.40**	0.35	−0.28	**2.83**	4.86	10.74
KRT16	**0.27**	0.22	0.06	**2.52**	4.09	28.15
KRT5	**0.36**	0.35	−0.30	**2.33**	3.44	24.13
KRT6C	**0.25**	0.21	0.20	**2.60**	4.10	4.66
S100A2	**0.48**	0.43	0.06	**2.86**	5.16	17.46
S100A7	**0.40**	0.33	0.03	**2.98**	4.26	12.79
S100A8	**0.30**	0.23	−0.13	**2.38**	3.85	24.38
S100A9	**0.38**	0.33	−0.10	**2.58**	4.18	23.14
SFN	**0.37**	0.34	0.07	**2.74**	4.48	11.16
SPRR1B	**0.42**	0.34	−0.30	**2.86**	4.59	17.54
Mean	**0.36**	0.31	−0.07	**2.67**	4.30	17.41

top-k spots by GT expression for that gene, then average across the entire gene panel. This isolates the per-spot regime where biological signal concentrates and tests whether each method matches GT magnitude on focal hotspots specifically. The filter uses GT expression (not predicted), avoiding selection bias.

Table 10: **Per-spot high-expression subset MSE (Hi-Exp Top-k).** For each gene, the k spots with highest GT expression are selected; MSE is computed on those spots for that gene and averaged across the gene panel. BC-I: Breast Cancer I (SPA154); SCC: Squamous Cell Carcinoma (NCBI770); MB: Mouse Brain (NCBI379). Lower is better. Best in **bold**.

	Breast Cancer I			SCC			Mouse Brain		
Method	k=10	k=20	k=50	k=10	k=20	k=50	k=10	k=20	k=50
Ours	**4.63**	**3.96**	**3.15**	**8.36**	**7.51**	**6.35**	2.36	2.16	1.90
Stem	6.35	5.58	4.61	10.52	9.64	8.31	**1.99**	**1.81**	**1.60**
STFlow	10.15	9.31	8.04	15.69	14.43	12.57	7.89	7.35	6.59

Two patterns emerge. STFlow's Hi-Exp MSE is the highest across all three datasets and all three subset sizes, by a substantial margin. On Breast Cancer I at k=10, STFlow's Hi-Exp MSE (10.15) is over $2\times$ ours (4.63) and $1.6\times$ Stem's (6.35), despite STFlow's aggregate MSE being the lowest of all methods in Table 1. This is the direct quantitative confirmation of the conditional-mean argument in Section 4.1: under deterministic ODE sampling STFlow approximates $\mathbb{E}[Y \mid X]$ and over-smooths the focal hotspots where biological signal concentrates, inflating Hi-Exp MSE even though aggregate MSE remains low.

Our method achieves the lowest Hi-Exp MSE on Breast Cancer I and SCC across all three subset sizes (a 27% reduction over Stem on BC-I k=10, 21% reduction on SCC k=10). On Mouse Brain, Stem achieves marginally lower Hi-Exp MSE than our method, consistent with the overall pattern in Table 1 where Ours and Stem are closest on this dataset; both methods nonetheless decisively outperform STFlow, with the catastrophic failure mode shared across all three tissue types. Specific failure magnitudes on individual high-variance genes are striking: on Breast Cancer I *ERBB2*, STFlow's per-gene MSE is 24.91 versus our 0.38 ($66\times$ ratio); on Mouse Brain *Mbp*, STFlow's per-gene MSE is 17.45 versus our 0.46 ($38\times$ ratio).

Monte-Carlo error of the sample mean. Our point predictions average S=20 samples, so a natural alternative explanation of our higher aggregate MSE is Monte-Carlo error. Subsampling the stored draws and fitting $\mathrm{MSE}(S) = a + b/S$ gives an excellent fit ($R^2 > 0.999$) but a small sampling term. At S=20 it accounts for 0.9% of our MSE gap to the strongest baseline on Breast Cancer I and 10.1% on Squamous Cell Carcinoma, and extrapolating to $S \to \infty$ leaves the gap essentially unchanged. The MSE gap is therefore not a sampling artefact, and we report it as observed.

F.4 Sample Variance and Predictive Interval Coverage

For our method and the diffusion-based Stem baseline, we draw 20 independent samples per spot under each method's standard inference protocol and report two per-sample distributional quantities: (i) sample variance, computed per spot and averaged over spots and genes, characterizing the magnitude of stochasticity in each method's predictive distribution; and (ii) empirical coverage of nominal 80% predictive intervals, defined as the fraction of ground-truth values that fall within the empirical 10th–90th percentile of the 20 samples at each spot–gene pair, quantifying calibration of the predictive intervals.

Table 11: **Sample variance and 80% predictive interval coverage across three datasets.** Sample variance is averaged over spots and genes; coverage is the fraction of ground-truth values within the empirical 10th–90th percentile of 20 samples (target 0.80). BC-I: Breast Cancer I (SPA154); SCC: Squamous Cell Carcinoma (NCBI770); MB: Mouse Brain (NCBI379).

	Sample Variance			Coverage @ 80%		
Method	**BC-I**	**SCC**	**MB**	**BC-I**	**SCC**	**MB**
Ours	0.13	0.20	0.15	0.26	0.29	0.25
Stem	0.28	0.15	0.09	0.27	0.25	0.23

Both methods produce moderate, data-adaptive sample variance (Ours: 0.13–0.20 across datasets; Stem: 0.09–0.28), reflecting per-spot uncertainty that varies with input morphology. However, both are uniformly under-dispersed: empirical 80% interval coverage falls in the 0.23–0.29 range, far below the nominal 0.80 target. This indicates that while both methods produce non-trivial sample-to-sample variation, the magnitude of that variation systematically underestimates the true conditional spread of gene expression at each spot.

F.5 Calibration as a Limitation

The under-dispersion documented above is shared by both our method and the diffusion-based Stem [47] baseline, indicating that it is a property of the current generation of generative spatial-transcriptomics predictors rather than specific to our framework. Although our method matches or exceeds prior generative baselines on point-prediction metrics and qualitative spatial pattern fidelity, the per-spot sample distributions should not be interpreted as well-calibrated predictive intervals in their current form. Improving calibration through post-hoc methods (conformal prediction, isotonic re-calibration of per-spot quantiles) or training-time CRPS-aware objectives is an orthogonal direction that we leave to future work.

G Detailed Regulator Token Analysis

G.1 Analysis Methodology

We performed regulator token analysis on the Breast Cancer I dataset [13], specifically on the held-out test slide SPA154. This dataset was selected for interpretability analysis due to its heterogeneous tumor microenvironment containing diverse cell populations including invasive carcinoma, stromal tissue, and tumor-infiltrating lymphocytes, providing rich biological context for evaluating whether regulator tokens capture meaningful gene regulatory programs.

We systematically analyzed regulator token attention patterns across all transformer layers and flow matching timesteps to identify where biological specialization emerges most clearly. For each layer $\ell \in \{1, \ldots, 12\}$ and timestep $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, we computed two metrics:

Focus Score. Measures how concentrated each token's attention is on specific genes:

$$\text{Focus}(m) = \max_g A_{m,g} - \frac{1}{G} \sum_g A_{m,g}, \tag{7}$$

where $A_{m,g}$ is the attention weight from regulator token m to gene g. Higher focus scores indicate more specialized tokens.

Token Uniqueness. Measures how distinct each token's attention pattern is from others:

$$\text{Uniqueness} = 1 - \frac{1}{M(M-1)} \sum_{i \neq j} \cos(\mathbf{a}_i, \mathbf{a}_j), \tag{8}$$

where $\mathbf{a}_i$ is the attention vector of token i. Higher uniqueness indicates more diverse specialization across tokens.

Layer 6 at $t = 0.4$ was selected for detailed analysis based on maximum focus score (mean: 0.115, max: 0.336) and token uniqueness (0.57). This intermediate layer and timestep appears to capture the most biologically meaningful representations before the final prediction layers. The selection enters only this analysis and affects no predictive metric.

G.2 Pathway Enrichment Analysis

For each regulator token, we identified the top-30 genes with highest attention weights and performed pathway enrichment analysis using MSigDB Hallmark gene sets [26] with a one-sided Fisher exact test whose background is the gene panel of the dataset, not the genome. P-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

G.3 Null-Controlled Enrichment and Abundance

A panel pre-selected for variability is enriched for coherent programs, and attention could in principle follow abundant or variable genes. We therefore compare the enrichment rate against null gene sets of the same size drawn from the panel, either uniformly or matched to the token gene sets in mean log-expression, variance, or detection frequency (100 draws each, Table 12).

Table 12: **Enrichment rate against null gene sets** (Breast Cancer I, Hallmark, panel background). σ: standard deviation of the null over 100 draws. Empirical p: fraction of null draws whose rate reaches the observed rate.

Gene sets	**Rate**	**Observed − null**	**σ units**	**Empirical p**
Regulator tokens (observed)	.234	—	—	—
Uniform random	.016	+.218	—	—
Expression-matched	.068 ± .029	+.166	5.7	< 0.01
Variance-matched	.023 ± .019	+.211	11.1	< 0.01
Detection-matched	.072 ± .033	+.162	4.9	< 0.01

No null draw reaches the observed rate in any matched condition. Per-gene total attention, summed over the 64 tokens, correlates negatively with every abundance statistic: mean log-expression (Pearson -0.481, Spearman -0.545), variance (-0.288, -0.303), and detection frequency (-0.582, -0.557). Attention is thus allocated away from abundant, widely detected genes in aggregate. This does not preclude individual tokens from attending to abundant families. The immunoglobulin tokens M14, M8, M49, and M54 are such a case, and they remain distinct from one another in their per-gene weight profiles (Section G.5).

G.4 Top Tokens by Focus

Table 13 lists the regulator tokens with the highest focus scores and their most attended genes. Of the 64 tokens, 15 (23.4%) are enriched for at least one Hallmark pathway at adjusted $p < 0.05$ with the panel background, a rate above every matched null (Section G.3).

G.5 Full Attention Pattern Visualization

Figure 6 shows the attention pattern for top-ranked tokens as a dot plot, where dot size indicates attention magnitude and color indicates token focus score. This visualization highlights the sparse and specialized nature of attention, with Token M14 exhibiting the strongest specialization on immunoglobulin gene *IGKC*.

Table 13: **Top regulator tokens by focus score.** Top genes are the four highest-attended genes of each token.

Token	Focus	Top Genes
M14	0.336	*IGKC, IGHG3, IGLC2, IGHG4*
M2	0.309	*FNBP1L, PTPRF, PERP, MIEN1*
M30	0.293	*IGKC, TMSB10, IGLC2, IGHG3*
M8	0.279	*IGHG4, IGLC2, IGHG3, IGLC3*
M49	0.276	*IGLC2, IGKC, IGHG4, IGHG3*
M57	0.251	*FNBP1L, PERP, SPINT2, PTPRF*
M4	0.240	*TMSB10, ACTB, ERBB2, MIEN1*
M54	0.237	*IGHG4, IGHG3, IGLC2, IGHG1*
M37	0.231	*TMSB10, ERBB2, ACTB, MIEN1*
M21	0.222	*IGHG4, IGHG1, S100A9, S100A8*
M29	0.218	*PERP, SF3B5, FNBP1L, TXNIP*
M58	0.207	*TMSB10, ACTB, ERBB2, BEST1*

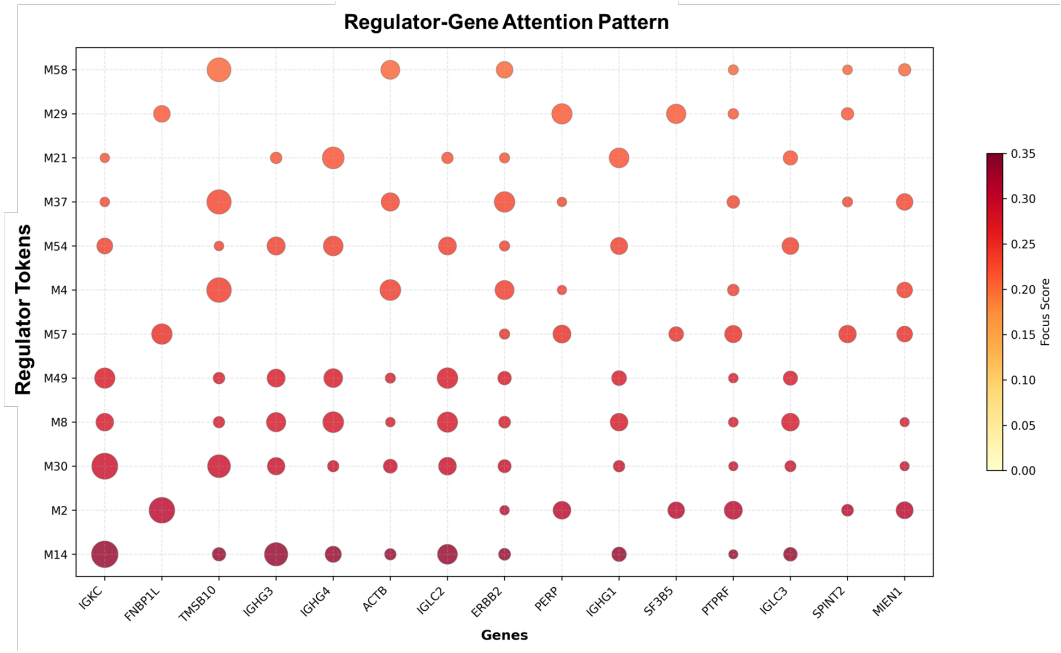


Figure 6: **Per-token attention specialization on Breast Cancer I (SPA154).** Dot plot of attention weights from selected regulator tokens (rows) to selected genes (columns), evaluated on the SPA154 test slide. The 15 genes shown are the top-15 by total attention summed across all 64 regulator tokens; the 12 tokens shown are a top-ranked subset by focus score and do not exhaust all 64 regulator tokens. Each dot's area is proportional to the attention weight from one token to one gene; color encodes the per-token focus score, with darker red indicating attention concentrated on fewer genes. Token M14 is the most specialized, placing nearly all of its attention on the immunoglobulin gene *IGKC*; M8, M49, and M54 share immunoglobulin attention but with distinct per-gene weight profiles.

Attention Pattern Observations. Several key patterns emerge from the visualizations:

Immune Gene Clustering. Tokens M14, M8, M49, and M54 show concentrated attention on immunoglobulin genes (*IGKC*, *IGHG3*, *IGHG4*, *IGLC2*, *IGLC3*). Hierarchical clustering of the attention vectors groups these tokens together, reflecting the coordinated nature of humoral immune response in the tumor microenvironment.

Cytoskeletal Module. Tokens M4, M37, and M58 preferentially attend to cytoskeletal and structural genes including *TMSB10* (thymosin beta-10), *ACTB* (beta-actin), and *MIEN1*. These genes are associated with cell motility and morphological changes relevant to tumor progression.

Signaling Genes. Token M29 shows distinct attention to *PERP*, *SF3B5*, and *TXNIP*. Token M2 attends to *FNBP1L* and *PTPRF*, which are involved in apoptotic signaling.

Oncogene Attention. ERBB2 (HER2) receives moderate attention from multiple tokens (M4, M37, M58), reflecting its importance in breast cancer biology despite this being a heterogeneous cohort rather than HER2-specific.

Sparse and Specialized Attention. The attention patterns are notably sparse rather than uniform, with each token concentrating on a subset of genes. This sparsity emerges without explicit regularization, suggesting that the model naturally discovers modular gene organization through the prediction task. The mean attention entropy across tokens is 3.2 (compared to the maximum entropy of 5.3 for the 210-gene panel), quantifying this specialization.

Token Redundancy Analysis. While some tokens attend to overlapping gene sets (e.g., M8 and M54 both focus on immunoglobulins), their attention weight distributions differ. M54 shows higher attention to *IGHG4* and *IGHG1*, while M8 emphasizes *IGLC2* and *IGLC3*. This partial overlap with distinct emphasis allows the model to capture both shared module membership and gene-specific emphasis within the same gene family.

G.6 Functional Specialization Analysis

To quantify whether tokens capture distinct biological functions rather than redundant patterns, we computed the Jaccard similarity between top-30 gene sets for each token pair. The mean pairwise Jaccard similarity was 0.114, against 0.244 for learned queries (Table 7), indicating substantial specialization. Tokens focused on the same gene family (e.g., M8 and M54, both attending to immunoglobulins) still showed distinct gene attention patterns (Jaccard $= 0.31$), suggesting they capture complementary aspects of the same program.

H Computational Cost Analysis

We compare wall-clock and memory cost against Stem [47], our diffusion-based predecessor that shares the same UNI + CONCH foundation model conditioning. All measurements are conducted on a single NVIDIA RTX A5000 GPU with batch size 1 (forward pass, inference) and batch size 64 (training, memory).

Table 14: **Computational cost comparison against Stem.** All measurements on a single NVIDIA RTX A5000 GPU.

	Stem (baseline)	**Ours**
Parameters	133.8M	165.9M (+24%)
Forward pass (bs=1)	5.6 ms	12.3 ms ($\times$2.2)
Training throughput	105 samples/sec	85 samples/sec ($-$19%)
GPU memory (bs=64)	942 MB	1,065 MB (+13%)
Sampling steps	250 (diffusion)	50 (flow matching)
Inference per spot (SDE, used in Table 1)	$\sim$1,400 ms	1,274 ms ($\times$1.1 faster)
Inference per spot (ODE)	$\sim$1,400 ms	618 ms ($\times$2.3 faster)

The comparison in Table 14 is not step-matched by construction. Stem is a diffusion model and uses 250 sampling steps, whereas flow matching along linear interpolation paths uses 50. Our architecture is the more expensive one: 165.9M parameters against 133.8M and 12.3 ms against 5.6 ms per forward pass. Any inference advantage over Stem therefore comes from requiring $5\times$ fewer sampling steps, a property of the generative formulation, and not from a cheaper architecture. In SDE mode, which the main results use, the net effect is a $1.1\times$ speedup. We do not claim efficiency as a contribution.

To make the step dependence explicit, we measure both methods at a matched sampling budget on the same held-out slide (SPA154, 346 spots, 20 samples per spot, batch 200, Euler 50 steps) and report the number of model calls alongside wall clock (Table 15). At an identical model-call count

of 1,715 in ODE mode, our architecture costs $1.41\times$ Stem's wall clock, above the $1.24\times$ parameter ratio. Our SDE sampler evaluates the network twice per step, so SDE mode raises the call count to 3,500 and the wall clock by $2.02\times$ relative to our ODE run, giving the measured $2.84\times$ overall. Halving our sampling steps to 25 brings the model-call count to 1,750, 2% above Stem's, and costs 1.3% of PCC@200 (both settings evaluated with a single script on the same slide). At that budget our PCC@200 remains roughly twice the value Stem reports in Table 1. STFlow is far cheaper than both because it is a 4.5M-parameter graph model that processes the whole slide in each forward pass.

Table 15: **Inference cost at a matched sampling budget** on SPA154 (346 spots, 20 samples per spot, batch 200, Euler solver). Model calls count forward passes over all spot chunks. STFlow forwards the whole slide in each call, so its call count is not comparable to the per-spot methods. *Extrapolated from 5 of 35 chunks.

Method	Mode	Steps	Model calls	Wall clock (s/slide)	Relative to Stem
Stem	ODE	50	1,715	491.6	1.00×
Ours	ODE	50	1,715	691.9*	1.41×
Ours	SDE	50	3,500	1,397.6	2.84×
STFlow [15]	ODE	50	1,000	4.1	0.008×

Training throughput drops by 19% (105 → 85 samples/sec), reflecting the additional cross-attention layers, and total training cost remains within the standard ST training budget on a single consumer-grade GPU.

I Reimplemented Baselines

We reimplement MERGE [12], Gene-DML [39], and HAGE [8] under our exact protocol: identical slide splits, gene panels, $\log_2(\text{count}+1)$ targets, and the same frozen UNI+CONCH features in place of each method's original image encoder. We follow each paper's published hyperparameters. MERGE uses a four-layer GAT over a hierarchical graph built from spatial and feature-space clusters of size 100. Gene-DML uses its dataset-specific group count (k=18 on Breast Cancer I, k=25 on Squamous Cell Carcinoma) with λ=0.8 and τ=0.07. HAGE uses Gene2Vec-initialized gene embeddings [9] with its correlation objective (λ_1=λ_4=10, λ_2=λ_3=1). All three are deterministic predictors evaluated by a single forward pass over the held-out slide with the PCC@k gene selection of Appendix B. Table 16 reports mean ± std over three seeds. In the qualitative figures (Figures 3 and 4a), each method is shown with the training seed whose PCC@200 is closest to its three-seed mean. Because its correlation objective leaves the output scale uncalibrated, HAGE's expression maps are shown after matching each gene's mean and standard deviation to the ground truth, which leaves its correlation unchanged.

Table 16: **Reimplemented baselines** (mean ± std over three seeds). ‡ HAGE trains its predictor with a correlation objective only (its Eq. 2), so its outputs carry no calibrated absolute scale. Its seed spread on MSE (±0.004 on Breast Cancer I, ±0.007 on Squamous Cell Carcinoma) indicates a systematic offset rather than noise.

Dataset	Method	PCC@10↑	PCC@50↑	PCC@200↑	MSE↓	MAE↓
BC-I	MERGE	.4962 ± .0253	.3528 ± .0239	.2947 ± .0253	1.2255 ± .1098	.8529 ± .0412
	Gene-DML	.4665 ± .0175	.3038 ± .0183	.2456 ± .0191	1.3784 ± .0665	.9082 ± .0201
	HAGE	.4588 ± .0079	.3288 ± .0066	.2626 ± .0106	5.3169 ± .0037‡	1.8754 ± .0005‡
	Ours	**.5453** ± .0041	**.4191** ± .0014	**.3418** ± .0091	1.6719	.9982
SCC	MERGE	.8435 ± .0100	.8162 ± .0086	.7372 ± .0079	.9677 ± .0181	.7691 ± .0063
	Gene-DML	.8494 ± .0059	.8222 ± .0042	.7384 ± .0051	1.0948 ± .0149	.8216 ± .0070
	HAGE	.8051 ± .0033	.7845 ± .0026	.7249 ± .0030	7.2620 ± .0066‡	1.9994 ± .0009‡
	Ours	**.8703**	**.8278**	**.7932**	1.0145	.7461

All three reimplementations outperform Stem, the strongest baseline in the earlier comparison, on PCC@200 on both datasets, so the comparison does not rest on weak reimplementations. Against the strongest baseline in each dataset, our margins are +0.0491/+0.0663/+0.0471 on PCC@10/50/200

over MERGE on Breast Cancer I and +0.0209/+0.0056/+0.0548 over Gene-DML on Squamous Cell Carcinoma. MERGE attains lower MSE than our method on both datasets.

J Spatial Variability of the Gene Panel

On each held-out slide we compute Moran's I for every gene on a k-nearest-neighbor spot graph (k=6) after total-count normalization to 10^4 and log1p, with mitochondrial genes excluded from both the panel and the background, and locate our panel in the genome-wide ranking (Table 17).

Table 17: **Spatial autocorrelation of the gene panel.** Median Moran's I of the panel and of all genes, the percentile of the panel in the genome-wide ranking, and the number of panel genes inside the genome-wide top-2000 by Moran's I.

Dataset	Median I, panel	Median I, all genes	Panel percentile	Panel in top-2000
Breast Cancer I	.0859	.0074	95.0%	161/210 (77%)
SCC	.2166	.0297	97.5%	195/218 (89%)
Human Bowel	.2752	.0164	98.5%	169/195 (87%)
Mouse Brain	.2176	.0124	98.1%	191/209 (91%)

In the training target space, $\log_2(\text{count}+1)$, the panel sits between the 97.6th and 98.2nd percentiles and 90–100% of it falls inside the genome-wide top-2000. Within the panel, per-gene variance and Moran's I correlate at 0.41–0.85 in that space. Dispersion-based selection is therefore, on these data, a close proxy for spatial variability at the 2–3k scale, and selecting by Moran's I would return largely the same genes. The panel describes the marker-gene regime that every compared method shares, and performance on wider spatially variable gene sets is not covered by Table 1.

K Memory Scaling to High-Plex Panels

We measure the forward-pass peak memory of the full model with regulator attention and of the same model with every regulator attention module replaced by full gene–gene self-attention, while sweeping the panel size $G \in \{210, 500, 1000, 2000, 5000\}$ on a 24 GB GPU. The two models differ only in the attention module, and weights are randomly initialized because the cost depends only on tensor shapes. Out-of-memory failures are recorded together with the size of the failing allocation, and peak memory is reported at the production inference batch (B=200) and at a small batch (B=8) where both variants run.

The requested allocations of full self-attention equal $B{\cdot}h{\cdot}G^2{\cdot}4$ bytes (17.9 GiB at G=2000, 111.8 GiB at G=5000), so the measurement follows the analytical quadratic cost exactly. At G=5000 and B=200 neither variant fits in 24 GB. The regulator model fails on a 2.9 GiB hidden-state allocation ($B \cdot G \cdot C \cdot 4$ bytes), which shrinks linearly with batch size, whereas full self-attention fails on a single 111.8 GiB attention tensor. At B=8, where both run up to G=5000, regulator attention uses $5.26\times$ less peak memory at G=5000 and the same memory at G=210. We report memory rather than speed: the measured speed ratio at B=8 grows from $0.98\times$ to $2.61\times$, well below the attention-only ratio $G/2M$, because the per-token feed-forward layers, whose cost is linear in G, dominate the forward pass at these sizes.

L Limitations and Future Work

Gene panel scope. Our experiments use a panel of 195–218 highly variable genes per dataset, shared by all compared methods. The panel lies within the genome-wide top spatially variable genes (Appendix J), so the reported numbers describe the marker-gene regime rather than the full spatially variable set. How performance changes as the panel widens toward 2–3k spatially variable genes, with all methods retrained on a common footing, is not covered by Table 1 and is left to future work. Regulator attention keeps the memory of gene-token models tractable at such sizes (Appendix K), but the optimal token count M may need to scale with G, and pathway-level interpretability becomes more sensitive to noise.

Platform resolution. All experiments use Visium-resolution data with 55-μm spots, where each spot averages over 5–20 cells. Our predictions therefore reflect spot-level rather than cellular-level expression. Newer platforms such as Visium HD (2-μm bins), Stereo-seq, Xenium, and the Xenium Prime 5K panel provide near-cellular or sub-cellular resolution and higher plex. Regulator attention removes the quadratic memory term at Xenium-scale panel sizes (Appendix K), but accuracy at that resolution requires a matched high-plex cohort for these tissues, which is not available to us, and cross-platform generalization is not evaluated. Whether our regulator tokens, currently trained at spot resolution, generalize to these finer scales without modification is unclear: the conditioning pathway through patch-level foundation model features may need reformulation to resolve sub-spot heterogeneity.

Patient-level splits. Leave-one-patient-out results are reported for Breast Cancer I (three seeds) and Squamous Cell Carcinoma (single seed). Human Bowel and Mouse Brain use slide-level splits only, and their reported numbers may include same-patient or adjacent-section continuity for every compared method.

Ablation scope. The isolation of regulator attention under the flow matching backbone and its comparison against input-agnostic bottlenecks are measured on Mouse Brain. The decomposition on the other tissues uses the diffusion backbone (Table 2a), and extending the flow matching isolation to them is left to future work.

Generative form for zero-inflated expression. We model $\log_2(\text{count} + 1)$ expression with a continuous Gaussian-prior flow. The log transform reduces but does not remove the point mass at zero in sparse genes, and whether a continuous flow with noise injection is the most suitable generative form for zero-inflated expression remains an open question.

SDE noise schedule sensitivity. The choice of SDE noise schedule controls the trade-off between predictive accuracy and sample diversity. We use a fixed schedule across all datasets, which yields consistent results in our experiments but may not be optimal for tissues with different expression variance profiles. Adaptive or learned noise schedules (conditioned on tissue type, gene-specific variance, or reverse-time uncertainty) are a natural extension.

Computational overhead at WSI scale. Compared to deterministic regression baselines (TRIPLEX requires 9.5 ms/spot in a single forward pass), our SDE sampling at 50 steps takes approximately 1.27 s/spot. For a typical WSI containing 10,000–50,000 Visium spots, this translates to 3–18 hours of inference per slide. ODE sampling halves this cost (618 ms/spot, 1.7–9 hours/slide) at the cost of distributional information. Practical deployment on full WSIs would benefit from amortized inference techniques such as consistency distillation or single-step generators, both of which are compatible with the flow matching framework.

Spatial coordinates not modeled. By design, our model predicts each spot independently from local morphology, without modeling spatial dependencies between neighboring spots. This makes the method applicable to scenarios where spatial continuity is absent (tissue microarrays, fragmented samples, sparse sampling) but ignores potentially useful spatial smoothness priors. Incorporating spot-to-spot graph structure via GNN layers or spatial transformer attention is an orthogonal direction we leave to future work.

Broader Impacts. On the positive side, accurate histology-to-expression prediction would lower the cost barrier of spatial transcriptomic analysis on archival H&E samples, broadening access to molecular-level interpretation in research and retrospective clinical studies. On the negative side, predicted expression should not be used as a substitute for measured molecular data in clinical decision-making without further validation; over-reliance on model outputs in diagnostic settings, particularly across patient cohorts that differ from the training distribution, could lead to systematic errors. We see calibration improvements (Appendix F) and broader-cohort validation as prerequisites for any clinical translation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions stated in the abstract and §1 (regulator attention, dual conditioning, flow matching framework, predictive gains over deterministic and generative baselines, and 87.5% unsupervised pathway alignment) are supported by Tables 1 and 2, and §4.4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix L discusses gene panel size, platform resolution, SDE noise schedule sensitivity, computational overhead at WSI scale, and the absence of explicit spatial dependency modeling. Calibration of predictive intervals is additionally identified as a shared limitation in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce new theoretical results. The flow matching framework in §3.3 relies on established formulations cited from prior work.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §4.1 (Experimental Setup) and Appendices A and B provide the architecture, optimizer, training schedule, dataset preprocessing, leave-one-slide-out protocol, and Leiden clustering pipeline. All datasets are publicly available through HEST-1K [19], and code will be released upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All datasets are publicly accessible through HEST-1K [19]. Source code is not released at submission time to preserve anonymity but will be released upon acceptance, as stated in §4.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: §4.1 reports the architecture (12 layers, 6 heads, hidden dimension 768, $M{=}64$ regulator tokens), optimizer (AdamW, learning rate 10^{-4}, 200 epochs), and inference protocol (Heun's second-order solver, 50 sampling steps). Appendix B details the evaluation protocol including the Leiden resolution selection procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Appendix E reports mean $\pm$ standard deviation over three independent random seeds on Breast Cancer I and Mouse Brain, confirming that all reported gains over the strongest baseline lie outside the model's seed-level variance band. Pathway enrichment significance is reported with adjusted p-values (Benjamini–Hochberg) in Tables 3b and 13.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: §4.1 specifies a single NVIDIA RTX A5000 GPU. Appendix H provides a detailed cost comparison against the diffusion baseline, including parameter count, forward-pass latency, training throughput, GPU memory at batch size 64, and per-spot inference time under both ODE and SDE sampling.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly available, de-identified spatial transcriptomics datasets aggregated in HEST-1K [19], with no new human subjects research and no use of sensitive personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix L discusses both positive (lowered cost barrier for transcriptomic analysis on archival samples) and negative (clinical over-reliance, cross-cohort calibration) societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifact is a spatial gene expression prediction model trained on publicly available academic datasets and does not pose misuse risks comparable to large language models, image generators, or scraped web datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (HEST-1K [19] and its constituent studies [13, 20**?**]), pretrained pathology foundation models (UNI [6], CONCH [27]), pathway resources (MSigDB Hallmark [26], Enrichr [5]), and baseline implementations are used in accordance with their respective academic-use licenses and properly credited via citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or pretrained models at submission time. Source code with documentation will be released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or new human subjects research. All tissue samples were collected under the original studies aggregated by HEST-1K [19].

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work re-analyzes publicly released spatial transcriptomics datasets that were collected under IRB approvals (or international equivalents) of their original studies aggregated in HEST-1K [19]. No new human subjects research is conducted in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as components of the core methodology. The pathology foundation models employed (UNI [6], CONCH [27]) are vision and vision-language encoders, not generative LLMs, and they are used as frozen feature extractors only.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.