Adapting a Pre-trained Single-Cell Foundation Model to Spatial Gene Expression Generation from Histology Images

Supplementary Material

A. Experimental Setup Details

This section provides an expanded description of the experimental setup, complementing the settings outlined in the main paper.

A.1. Datasets Description

We employed three publicly available spatial transcriptomics (ST) datasets covering distinct tissue types, disease contexts, and experimental platforms (as summarized in Table S1).

(1) **cSCC (ST, Cutaneous Squamous Cell Carcinoma).** The cSCC dataset [8] consists of formalin-fixed paraffin-embedded (FFPE) cutaneous squamous cell carcinoma samples from four patients, profiled with the Spatial Transcriptomics platform using a spot grid with 110 μm center-to-center spacing and 150 μm spot diameter. The slices exhibit highly heterogeneous tumor microenvironments, including keratinized tumor nests, stromal regions, and immune cell infiltrates. Although FFPE processing typically leads to reduced RNA integrity, this dataset offers a realistic and diverse benchmark for assessing model robustness under degraded molecular quality.

(2) **Her2ST (ST, HER2$^+$ Breast Cancer).** The Her2ST dataset [1] comprises spatial transcriptomics measurements of HER2-positive invasive ductal carcinoma (IDC) from eight patients. Each slice was captured using the original ST protocol on fresh-frozen breast tissue, with a spot diameter of 100 μm and an inter-spot distance of 200 μm. In total, 36 slices were included, covering tumor core and peritumoral regions. The dataset provides high-quality histology images aligned with corresponding spot-level gene expression matrices ($\approx$15k detected genes per slice).

(3) **Human Kidney (Visium ST).** The Kidney dataset [9] represents a large-scale Visium Spatial Gene Expression collection containing 23 patient samples spanning both healthy and diseased conditions (Diabetic Kidney Disease and Acute Kidney Injury). All slices were obtained from fresh-frozen human Kidney tissue with a 55 μm spot diameter and 100 μm inter-spot distance. Each slice contains between 0.3k–4k spots with over 33k expressed genes, capturing both cortical and medullary regions at high molecular depth. This dataset provides extensive biological and technical variability, serving as the primary benchmark for assessing generalization across tissues and disease states.

A.2. Dataset Partitioning

Each histology slice is held out in turn as the test set, while the remaining slices are used for model training and validation. From the training pool, 10% of samples are randomly reserved as a validation subset to monitor model performance and trigger early stopping. This design ensures that model assessment is always performed on unseen tissue slices, thereby preventing data leakage and providing a reliable measure of generalization across tissue slices.

Specially for the Her2ST dataset, which provides 3–6 serial slices per patient across eight patients, we designate the first slice from each patient (A1, B1, C1, D1, E1, F1, G1, H1) as a fixed pool of evaluation candidates. In each evaluation fold, one of these eight slices is held out as the test slice, while all remaining slices, including other slices from the same patient, are used for training, with 10% of the training samples reserved for validation. This fixed pool with one representative slice per patient keeps the difficulty of different folds comparable and guarantees that every patient is evaluated on its own test slice.

For the cSCC and Kidney datasets, we adopt the same slice-wise training–testing scheme. In each fold, one histology slice is held out for testing, and the remaining slices form the training pool, from which 10% of samples are reserved for validation. This unified evaluation protocol provides a consistent and fair basis for assessing model generalization across heterogeneous tissues, disease states, and ST platforms.

A.3. Genes Selection

Since HINGE adapts CellFM, which was trained on a fixed 24,078-gene vocabulary, we first intersect each ST dataset's gene list with this vocabulary to ensure compatibility. We then follow the same criteria as in Stem [12], selecting the intersection of genes ranked in the top 300 for both mean expression and variance across training slices, which forms the Highly Mean–High Variance Gene (HMHVG) set used for training, validation, and evaluation. The selected genes are summarized in Fig. S1.

A.4. Histology Feature Extraction

Our pipeline operates on spot-level spatial transcriptomics data, where each spot corresponds to a spatial location on an H&E-stained tissue slide with paired gene expression and histology signals.

Concretely, each spot is associated with an RGB image patch centered at its spatial coordinates (x_c, y_c) on the reg-

Table S1. Summary of spatial transcriptomics datasets aggregated by patient. Each patient entry reports spot- and gene-level ranges across multiple slices, together with the corresponding platform.

Dataset	Platform	Patient	Tissue	Condition	Samples	Inter-spot Dist (μm)	Spot Diameter (μm)	Spots under Tissue	Genes per slice	Preservation	Reference
Her2ST	ST	Patient A	Breast	Cancer	6	200	100	325–360	15,045–15,645	Fresh Frozen	PMID: 34650042
	ST	Patient B	Breast	Cancer	6	200	100	270–295	15,109–15,387	Fresh Frozen	same as above
	ST	Patient C	Breast	Cancer	6	200	100	176–187	15,557–15,842	Fresh Frozen	same as above
	ST	Patient D	Breast	Cancer	6	200	100	301–315	15,396–15,666	Fresh Frozen	same as above
	ST	Patient E	Breast	Cancer	3	200	100	570–587	15,097–15,701	Fresh Frozen	same as above
	ST	Patient F	Breast	Cancer	3	200	100	691–712	14,861–15,067	Fresh Frozen	same as above
	ST	Patient G	Breast	Cancer	3	200	100	441–467	14,992–15,258	Fresh Frozen	same as above
	ST	Patient H	Breast	Cancer	3	200	100	510–613	14,873–15,029	Fresh Frozen	same as above
cSCC	ST	Patient 2	Skin	Cancer	3	110	150	638–666	17,138–17,883	FFPE	PMID: 7391009
	ST	Patient 5	Skin	Cancer	3	110	150	521–590	16,959–17,689	FFPE	same as above
	ST	Patient 9	Skin	Cancer	3	110	150	1071–1182	17,823–19,314	FFPE	same as above
	ST	Patient 10	Skin	Cancer	3	110	150	462–621	15,383–17,047	FFPE	same as above
Kidney	Visium ST	Patient 1	Kidney	Healthy	1	100	55	3007	33538	Fresh Frozen	PMID: 10356613
	Visium ST	Patient 2	Kidney	Healthy	1	100	55	3627	36601	Fresh Frozen	same as above
	Visium ST	Patient 3	Kidney	Healthy	1	100	55	4166	36601	Fresh Frozen	same as above
	Visium ST	Patient 4	Kidney	Healthy	1	100	55	2627	36601	Fresh Frozen	same as above
	Visium ST	Patient 5	Kidney	Healthy	1	100	55	956	36601	Fresh Frozen	same as above
	Visium ST	Patient 6	Kidney	Healthy	1	100	55	1034	36601	Fresh Frozen	same as above
	Visium ST	Patient 7	Kidney	Diseased	1	100	55	1322	36601	Fresh Frozen	same as above
	Visium ST	Patient 8	Kidney	Diseased	1	100	55	673	36601	Fresh Frozen	same as above
	Visium ST	Patient 9	Kidney	Diseased	1	100	55	673	36601	Fresh Frozen	same as above
	Visium ST	Patient 10	Kidney	Diseased	1	100	55	560	36601	Fresh Frozen	same as above
	Visium ST	Patient 11	Kidney	Diseased	1	100	55	534	36601	Fresh Frozen	same as above
	Visium ST	Patient 12	Kidney	Diseased	1	100	55	453	36601	Fresh Frozen	same as above
	Visium ST	Patient 13	Kidney	Diseased	2	100	55	461-904	36601	Fresh Frozen	same as above
	Visium ST	Patient 14	Kidney	Diseased	1	100	55	601	36601	Fresh Frozen	same as above
	Visium ST	Patient 15	Kidney	Diseased	1	100	55	787	36601	Fresh Frozen	same as above
	Visium ST	Patient 16	Kidney	Diseased	1	100	55	407	36601	Fresh Frozen	same as above
	Visium ST	Patient 17	Kidney	Diseased	1	100	55	317	36601	Fresh Frozen	same as above
	Visium ST	Patient 18	Kidney	Diseased	1	100	55	645	36601	Fresh Frozen	same as above
	Visium ST	Patient 19	Kidney	Diseased	1	100	55	673	36601	Fresh Frozen	same as above
	Visium ST	Patient 20	Kidney	Diseased	1	100	55	640	36601	Fresh Frozen	same as above
	Visium ST	Patient 21	Kidney	Diseased	1	100	55	507	36601	Fresh Frozen	same as above
	Visium ST	Patient 22	Kidney	Diseased	1	100	55	370	36601	Fresh Frozen	same as above

istered whole-slide image (WSI). Patch extraction proceeds as follows:

1. **Coordinate-based cropping**: Given (x_c, y_c), we extract a square region centered at the spot on the H&E WSI.
2. **Pixel normalization**: Raw pixel intensities in $[0, 255]$ are linearly rescaled to $[0, 1]$ before feeding patches into the encoders.

We adopt a dual-encoder framework to jointly capture complementary visual information from histopathology images. Each normalized image patch is independently encoded by two pre-trained vision backbones:

- **UNI** [2]: A pathology-domain vision transformer pre-trained on large-scale histopathology datasets. It produces embeddings $\mathbf{h}_{\text{uni}} \in \mathbb{R}^{1024}$ that emphasize detailed morphology, including cellular topology, nuclear texture, and tissue microarchitecture.
- **CONCH** [10]: A contrastive vision–language foundation model trained to align histopathology images with expert pathology reports. It outputs embeddings $\mathbf{h}_{\text{conch}} \in \mathbb{R}^{512}$ that capture high-level semantic tissue context and pathological attributes.

The two encoders provide complementary representations—UNI focuses on structural and morphological fidelity, whereas CONCH encodes semantic and contextual information from cross-modal supervision. Their outputs are concatenated to form a unified visual representation for each spot:

$$\mathbf{v} = \phi(\mathbf{c}) = [\mathbf{h}_{\text{uni}}; \mathbf{h}_{\text{conch}}] \in \mathbb{R}^{1536}, \tag{1}$$

which serves as the histology feature conditioning the spatial expression generation network.

A.5. Baselines

ST-Net [5] integrates spatial transcriptomics and histology through a deep convolutional network to predict gene expression directly from H&E-stained images. It employs a DenseNet-121 backbone [6] pre-trained on ImageNet, with a fully connected regression head for gene-level prediction. Each 224×224 patch centered on a spatial spot serves as input. Our implementation retains the original network configuration and normalization strategy to ensure faithful reproduction of the published framework.

BLEEP [11] proposes a bi-modal embedding framework for spatial gene expression prediction from H&E-stained histology images. It aligns image and expression modalities through contrastive learning to construct a shared embedding space, using a ResNet-50 encoder for images and an MLP for expression features. Gene expression is inferred via query–reference imputation based on the proximity of embeddings in the joint space. We adopt the same dual-encoder architecture and contrastive alignment scheme as described in the original work.

TRIPLEX [3] introduces a multi-resolution deep learning framework for predicting spatial gene expression from whole-slide histology images. The model captures complementary information at three hierarchical levels—the target spot, its local neighborhood, and the global tissue context—using independent ResNet-based encoders followed by a transformer-based fusion layer. These representations are integrated through an efficient fusion mechanism to jointly model fine-grained morphology and global organization. In our reimplementation, we preserve the multi-resolution design and fusion strategy to maintain methodological fidelity to the original model.

MERGE [4] introduces a graph-based framework for spatial gene expression prediction from whole-slide histology images. It constructs a multi-faceted hierarchical graph where nodes represent tissue patches, and edges capture both spatial and morphological relationships. Using a ResNet18 encoder to extract patch features, MERGE employs a Graph Attention Network (GAT) to jointly model short- and long-range dependencies across the tissue. The hierarchical graph integrates intra-cluster and inter-cluster connections, enabling efficient information propagation between morphologically similar but spatially distant regions. We follow the original multi-faceted graph design and SPCS-based gene smoothing to reproduce its morphology-aware prediction behavior

Stem [12] introduces a diffusion-based generative framework for predicting spatially resolved gene expression from H&E-stained histology images. Instead of treating prediction as deterministic regression, Stem models the conditional distribution of gene expression given image features, enabling one-to-many mappings that capture biological heterogeneity. The model leverages pretrained pathology foundation encoders (UNI [2], CONCH [10]) to derive image embeddings and conditions a DiT-based diffusion network for expression generation. This design allows Stem to generate biologically diverse yet accurate predictions across spatial locations. In our implementation, we maintain the same conditional diffusion formulation and foundation-model conditioning strategy as described in the original paper.

STFlow [7] formulates spatial gene expression prediction as a generative modeling problem via whole-slide flow matching. Instead of independent spot-level regression, it models the joint distribution of gene expressions across all spatial locations, capturing cell–cell interactions and global dependencies. The model employs an $E(2)$-invariant Transformer denoiser with local spatial attention and leverages pretrained pathology foundation encoders for feature extraction. We adopt the same flow matching formulation and spatial attention architecture as described in the original work to ensure methodological consistency across baselines.

Variant	PCC-50 ↑	PCC-200 ↑	MSE ↓	MAE ↓
Scratch	0.2511	0.1804	1.9176	1.1215
Decoder-Tune	0.4726	**0.3374**	0.9814	0.7716
Backbone-LoRA	0.4599	0.3163	0.9975	0.7774
HINGE	**0.4801**	0.3355	**0.9481**	**0.7638**
Gauss-Diff	0.3437	0.2084	1.8403	1.0447
Mask-Diff (NoCurr)	0.4702	0.3203	0.9641	0.7691
Mask-Diff (RandMask)	**0.4889**	**0.3427**	0.9500	0.7729
Mask-Diff (HINGE)	0.4801	0.3355	**0.9481**	**0.7638**
Hist-Affine-LN	0.2892	0.2187	1.9070	1.1176
SoftAdaLN (NoSoftNorm)	0.3707	0.2432	1.3417	0.9050
SoftAdaLN (NoIdInit)	0.4008	0.2873	1.2219	0.8592
SoftAdaLN (Full)	**0.4801**	**0.3355**	**0.9481**	**0.7638**
ResNet-50	0.4348	0.2956	0.9945	0.8001
UNI	0.4668	0.3208	0.9530	0.7667
CONCH	0.4091	0.2630	1.0895	0.8133
UNI + CONCH	**0.4801**	**0.3355**	**0.9481**	**0.7638**

Table S2. Ablations on Her2ST. Component-wise analysis of HINGE variants on the Her2ST (A1) dataset.

Variant	PCC-50 ↑	PCC-200 ↑	MSE ↓	MAE ↓
Scratch	0.2517	0.2072	1.7576	1.1192
Decoder-Tune	0.4702	0.3627	0.9066	0.7552
Backbone-LoRA	0.4558	0.3271	1.0054	0.7984
HINGE	**0.4871**	**0.3815**	**0.8956**	**0.7467**
Gauss-Diff	0.2592	0.1673	1.7559	1.0054
Mask-Diff (NoCurr)	0.4725	0.3735	0.9086	0.7591
Mask-Diff (RandMask)	0.4707	0.3703	0.8985	0.7514
Mask-Diff (HINGE)	**0.4871**	**0.3815**	**0.8956**	**0.7467**
Hist-Affine-LN	0.2285	0.1869	1.7018	1.0957
SoftAdaLN (NoSoftNorm)	0.3962	0.2899	1.3680	0.9486
SoftAdaLN (NoIdInit)	0.4160	0.3030	0.9644	0.8015
SoftAdaLN (Full)	**0.4871**	**0.3815**	**0.8956**	**0.7467**
ResNet-50	0.4455	0.3247	0.9811	0.8008
UNI	0.4699	0.3642	0.9675	0.7884
CONCH	0.4535	0.3557	1.0675	0.8151
UNI + CONCH	**0.4871**	**0.3815**	**0.8956**	**0.7467**

Table S3. Ablations on Kidney. Component-wise analysis of HINGE variants on the Kidney (IU-F52) dataset.

A.6. Implementation Details

Our approach is implemented using PyTorch (version 2.1.0) with Python 3.9, and models are trained on NVIDIA A800 GPUs with CUDA 12.1. We employ mixed precision training, utilizing PyTorch's native Automatic Mixed Precision (AMP) for computational efficiency. To ensure reproducibility, the random seed is consistently set at 42 across all experiments. The model is optimized using AdamW with a learning rate of 1×10^{-4}, weight decay of 0.0, and a global batch size of 32. We adopt a MultiStepLR scheduler with decay milestones at epochs [20, 30] and a decay factor of 0.2. The training process is capped at a maximum of 50 epochs, with an early stopping mechanism triggered if there is no improvement in validation MSE for 5 consecutive epochs after an initial validation warmup period of 15

Methods	cSCC				Her2ST				Kidney			
	PCC-50 ↑	PCC-200 ↑	MSE ↓	MAE ↓	PCC-50 ↑	PCC-200 ↑	MSE ↓	MAE ↓	PCC-50 ↑	PCC-200 ↑	MSE ↓	MAE ↓
Linear (ζ=1)	0.8728	0.7957	1.2067	0.8627	0.4922	0.3524	1.0802	0.8220	0.4817	0.3751	0.8986	0.7495
Cosine	0.8641	0.7868	1.1760	0.8392	0.4652	0.3294	1.1293	0.8383	0.4579	0.3469	1.0026	0.7999
$\zeta = 0.5$	0.8743	0.7964	1.2196	0.8576	**0.5090**	**0.3578**	1.1948	0.8619	0.4567	0.3485	0.9886	0.7948
$\zeta = 2$	0.8657	0.7828	1.3044	0.8990	0.4913	0.3450	1.0377	0.7888	0.4773	0.3729	0.9504	0.7829
$\zeta = \log_T G$	**0.8755**	**0.8021**	**1.0096**	**0.7793**	0.4817	0.3751	**0.8976**	**0.7445**	**0.4871**	**0.3815**	**0.8956**	**0.7467**

Table S4. Masking schedules. Results under different masking schedules on representative slices from the cSCC (P2_ST_rep3), Her2ST (A1), and Kidney (IU-F52) datasets.

epochs. To stabilize early-stage training, we implement a curriculum learning scheme where the first 5 epochs are restricted to mask ratios $\leq 20\%$ before exposing the model to the full diffusion schedule.

B. Additional Visualization Results

In this section, we present additional visualizations of marker-gene spatial expression predictions across all three datasets used in our experiments (Figs. S2–S4). These qualitative results complement the quantitative metrics in the main paper by providing a more fine-grained view of spatial localization patterns across datasets and tissue sections.

For the cSCC dataset, Fig. S2 shows spatial expression maps for *KRT6A*, *KRT10*, and *GJB2* on multiple tissue sections. Beyond the P2_ST_rep3 slice shown in the main text, we include additional cSCC slices to illustrate how the predicted localization patterns of these markers behave across different sections rather than on a single example.

We follow the same protocol on the Her2ST and Kidney datasets. For Her2ST, Fig. S3 visualizes predictions for *GNAS*, *ERBB2*, and *FASN* on multiple tissue sections. For the Kidney dataset, Fig. S4 shows *FXYD2*, *ATP1B1*, and *PODXL* on representative slices. These examples are meant to complement the quantitative results in the main paper by visually inspecting whether the predicted marker-gene maps preserve the expected spatial structures and localization patterns across datasets and sections.

To further quantify spatial coherence beyond point-wise errors, we additionally report gene-wise structural similarity (SSIM) between the predicted and ground-truth 2D expression maps. Specifically, for each slice and each gene, we compute SSIM on the corresponding 2D spatial expression maps, and then aggregate the results per slice and finally average across slices. We highlight representative marker genes (*KRT6A*, *GNAS*, *FXYD2*) in Fig. S5 as references from cSCC, Her2ST, and Kidney, respectively. This SSIM-based evaluation provides a complementary perspective to MSE/MAE/PCC by emphasizing structural agreement of spatial patterns (e.g., contiguous regions and tissue-level organization) rather than purely per-spot deviations.

In addition, for the Kidney dataset we visualize gene–gene correlation matrix heatmaps computed across multiple slices (Fig. S6). These co-expression maps provide a complementary perspective to the marker-gene expression plots by highlighting gene–gene dependencies at the tissue level.

C. Additional Ablation Studies

We extend the ablation analysis from the main paper to the Her2ST (A1) and Kidney (IU-F52) datasets to verify that our design choices are not specific to cSCC. In all cases, we reuse the same protocol, metrics, and variant definitions as in the main-text ablations.

On the Her2ST (Table S2) and the Kidney dataset (Table S3), we report the same suite of ablations as in the main text. Each table includes the Scratch baseline and all variants that reuse the pre-trained CellFM backbone, compares Gaussian and masked diffusion objectives (including the HINGE schedule with curriculum), and lists conditioning designs such as Hist-Affine-LN and the full SoftAdaLN module, along with different histology encoders (UNI, CONCH, and UNI+CONCH). The experiments follow the same protocol and metrics as the cSCC ablations in the main paper.

We further study different masking schedules on representative slices from the cSCC, Her2ST, and Kidney datasets (Table S4). In this experiment, we vary the forward masking schedule $\bar{\alpha}_t$ while keeping the rest of the setup fixed. We consider a linear schedule (**Linear**), a cosine schedule (**Cosine**), and two power-law schedules of the form $\bar{\alpha}_t = \left(1 - \frac{t}{T}\right)^{\zeta}$ with $\zeta \in \{0.5, 2\}$. In addition, we include a variant where the exponent is set to $\zeta = \log_T G$ for a prescribed global masking level G, which serves as the default schedule in our other experiments. Results are reported for all three datasets using the same evaluation metrics as in the main text.

With a full T-step run, we also evaluate the intermediate estimate $\hat{x}_0(t)$ at several t values. MSE/MAE decrease and PCC increases until convergence, without late-step degradation (Fig. S7(a)), suggesting progressive refinement rather than error accumulation under our progressive unmasking scheme. In addition, fixing the trained model (with T), we vary the inference budget K and report final metrics vs. K.

Method	Type	Steps	Time/slide (s) ↓	spots/s ↑	PCC@50 ↑	Time@HR (s) ↓
ST-Net	Reg.	1	0.7162	890.82	0.739	13.97
BLEEP	Reg.	1	48.1776	13.24	0.785	991.29
TRIPLEX	Reg.	1	10.8633	58.45	0.805	OOM
Stem	Gen.	1000	230.1942	2.77	0.823	4138.51
STFlow	Gen.	10	0.2431	2624.83	0.692	OOM
HINGE (scGPT)	Gen.	5	1.1643	547.97	0.806	37.3242
HINGE (CellFM)	Gen.	5	23.3428	26.46	0.874	453.06
HINGE (CellFM)	Gen.	50	257.7769	2.48	0.877	4411.02

Table S5. Inference efficiency (OOM: Out of Memory).

A small K already approaches the full-T result, giving a clear quality–speed trade-off (Fig. S7(b)).

Finally, we examine the effect of the masking horizon T on the same three datasets (Fig. S8). For each dataset, we fix the masking schedule and vary T over several values, and then record the corresponding performance metrics. We visualize these results as curves of each metric versus T, providing a summary of how the choice of masking horizon interacts with our masked diffusion formulation on cSCC, Her2ST, and Kidney slices. Unless otherwise specified, we set $T = 50$ as the default masking horizon in our experiments.

D. Inference efficiency

(i) Inference latency. We add a runtime comparison of regression and generative baselines (Table S5). At our default setting (HINGE (with-CellFM), T=50), throughput is 2.48 spots/s, comparable to Stem (2.77 spots/s) while achieving higher PCC. **(ii) Quality–speed trade-off (T).** HINGE exposes the denoising steps as a practical test-time scaling: reducing T from 50 to 5 increases throughput by $\sim$11$\times$ (2.48$\rightarrow$26.46 spots/s, exceeding BLEEP in this setting) while keeping PCC nearly unchanged (0.877$\rightarrow$0.874; Table S5, Figure S7). This shows HINGE reaches near-full accuracy with few steps, enabling practical inference throughput. **(iii) High-resolution case.** Inference is batched over spots and scales approximately linearly. Time@HR in Table S5 summarizes this regime, where some baselines are OOM while HINGE completes inference.

References

[1] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of HER2-positive breast tumors reveals novel intercellular relationships. *bioRxiv*, pages 2020–07, 2020. 1

[2] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 2, 3

[3] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024. 3

[4] Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. MERGE: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15611–15620, 2025. 3

[5] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8): 827–834, 2020. 2

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[7] Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *Forty-second International Conference on Machine Learning*, 2025. 3

[8] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2): 497–514, 2020. 1

[9] Blue B Lake, Rajasree Menon, Seth Winfree, Qiwen Hu, Ricardo Melo Ferreira, Kian Kalhor, Daria Barwinska, Edgar A Otto, Michael Ferkowicz, Dinh Diep, et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594, 2023. 1

[10] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 2, 3

[11] Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bimodal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023. 2

[12] Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3

Dataset	Selected genes per dataset
cSCC	A2ML1, ACTB, ACTN1, ACTN4, ACTR2, AEBP1, ALDOA, ANXA1, ANXA8, ANXA8L1, AP2S1, APCDD1, APRT, AQP3, ARPC2, ARPC5, BGN, BTF3, BTG1, C1R, CALM1, CALML5, CAP1, CAPN2, CAPNS1, CAPZB, CASP14, CAST, CCT6A, CD24, CD74, CDH3, CDSN, CFL1, CHCHD2, CLCA2, CLIC1, CLTC, CNFN, COL17A1, COL18A1, COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL6A1, COL6A2, COL6A3, COL7A1, COX6C, COX7C, CRABP2, CST3, CSTA, CSTB, CTNNA1, CTSB, CXCL14, CYCS, DBI, DCN, DDX17, DDX5, DEFB103A, DEFB103B, DMKN, DSC2, DSG1, DSG3, DSP, DST, ECM1, EEF1A1, EEF1B2, EIF1, EIF2S2, EIF3F, EIF3K, EIF5, EIF6, ELL2, ENAH, FABP5, FGFBP1, FGFR3, FLG, FN1, FTH1, FTL, GAPDH, GDI2, GJA1, GJB2, GLO1, GLTP, GNAI2, GNB1, GPNMB, HDGF, HIF1A, HINT1, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DRA, HMGB1, HNRNPA1, HNRNPD, HNRNPM, HOPX, HSP90AA1, HSP90B1, IFI27, IFI6, IGFBP3, IGFBP4, IGFBP7, IGFL1, IL1RN, ITGA6, ITGB4, ITM2B, IVL, JUNB, JUP, KLF5, KLK10, KLK11, KLK5, KLK7, KRT1, KRT10, KRT14, KRT15, KRT16, KRT17, KRT2, KRT5, KRT6A, KRT6B, KRT6C, KRT75, KRTDAP, KTN1, LAMB3, LAMC2, LAMP1, LCE3D, LGALS1, LGALS3, LGALS3BP, LGALS7, LGALS7B, LMNA, LUM, LYPD3, MAF, MAFB, MARCKS, MCL1, MKNK2, MMP1, MORF4L1, MUCL1, MYL12B, MZT2B, NCCRP1, NCL, NDRG1, NDUFA4, NDUFB9, NUCKS1, ODC1, PFN1, PGAM1, PI3, PKP1, PLS3, POLR2L, POMP, PPDPF, PPP1CA, PPP1CB, PRDX1, PRDX6, PRELID1, PRNP, PSMA7, PSMB7, PSME2, PTGES3, PTMA, PTPRF, RAB10, RAC1, RACK1, RAD23B, RAN, RBM3, RHOA, RNASE1, RPN2, RTN4, S100A14, S100A2, S100A6, S100A7, S100A7A, S100A8, S100A9, SAT1, SBSN, SEC61G, SERPINB1, SERPINB13, SERPINB3, SERPINB4, SERPINB5, SET, SFN, SFPQ, SKP1, SLC25A6, SLPI, SNRPD2, SOD2, SPARC, SPINK5, SPINK6, SPRR1A, SPRR1B, SPRR2A, SPRR2B, SPRR2D, SPRR2E, SPRR2F, SPRR2G, SRSF2, SSR4, SUB1, SUMO2, SYNGR2, TACSTD2, TGFBI, TGM1, TIMM13, TIMP1, TMED2, TMEM45A, TNC, TXNDC17, TYMP, UBA52, UBC, UBE2D3, UQCR11, UQCRB, UQCRC1, VAMP8, VCP, VDAC1, VIM, YBX1, YWHAB, YWHAE, YWHAG, ZFP36L1, ZFP36L2
Her2ST	A2M, ACTB, ACTG1, ACTN4, ADAM15, AEBP1, AES, ALDOA, AP000769.1, AP2S1, APOC1, APOE, ARHGDIA, ATG10, ATP5B, ATP5E, ATP5G2, ATP6AP1, ATP6V0B, AZGP1, B2M, BEST1, BGN, BSG, BST2, C12orf57, C1QA, C1QB, C1orf122, C3, C4orf48, CALM2, CALML5, CALR, CCND1, CCT3, CD24, CD63, CD74, CFL1, CHCHD2, CHPF, CIB1, CLDN3, CLDN4, CLDN7, CNN3, COL18A1, COL1A1, COL1A2, COL3A1, COL6A2, COMP, COPE, COPS9, COX4I1, COX5B, COX6B1, COX6C, COX7C, CRIP2, CST3, CTSB, CTSD, CTTN, CYBA, DBI, DDIT4, DDX5, DHCR24, EDF1, EEF1D, EEF2, EIF3B, EIF4G1, ELOVL1, ENO1, ERBB2, ERGIC1, FADS2, FASN, FAU, FKBP2, FLNA, FN1, FNBP1L, FTH1, FTL, GAPDH, GNAI2, GNAS, GPX4, GRB7, GRINA, GRN, GUK1, H1F0, H2AFJ, HINT1, HLA-A, HLA-B, HLA-C, HLA-DRA, HLA-E, HM13, HN1, HNRNPA2B1, HSP90AA1, HSP90AB1, HSP90B1, HSPA8, HSPB1, IDH2, IFI27, IFI6, IGFBP2, IGFBP7, IGHA1, IGHG1, IGHG3, IGHG4, IGHM, IGKC, IGLC2, IGLC3, INTS1, ISG15, JTB, KDELR1, KRT18, KRT19, KRT7, KRT8, KRT81, LAPTM4A, LAPTM5, LASP1, LGALS1, LGALS3, LGALS3BP, LLGL2, LMAN2, LMNA, LSM7, LUM, LY6E, MAPKAPK2, MDK, MGP, MIDN, MIEN1, MLLT6, MMACHC, MMP14, MRPL12, MUC1, MUCL1, MYL6, MYL9, MZT2B, NACA, NBL1, NDUFA3, NDUFB7, NDUFB9, NUCKS1, NUPR1, ORMDL3, P4HB, PCGF2, PCSK7, PEBP1, PERP, PFDN5, PFKL, PGAP3, PHB, PIP4K2B, PKM, PLD3, PNMT, POSTN, PPDPF, PPP1CA, PPP1R14B, PPP1R1B, PRDX1, PRRC2A, PRSS8, PSMB1, PSMB3, PSMB4, PSMD3, PSMD8, PTBP1, PTGES3, PTMA, PTMS, PTPRF, RABAC1, RACK1, ROMO1, RRBP1, S100A10, S100A14, S100A6, S100A8, S100A9, SCAND1, SCD, SDC1, SEC61A1, SEPW1, SERF2, SERINC2, SF3B5, SH3BGRL3, SLC2A4RG, SLC44A2, SLC9A3R1, SMARCD2, SNRPB, SPARC, SPDEF, SPINT2, SREBF1, SRRM2, SSR2, SSR4, STARD10, STARD3, SUPT6H, SYNGR2, TAGLN, TAPBP, TCEB2, TFF3, TIMP1, TMBIM6, TMED9, TMSB10, TMSB4X, TPT1, TRIM28, TSPO, TUBB, TXNIP, TYMP, UBA52, UBB, UBC, UBE2M, UBL5, UQCR11, UQCRQ, VCP, VIM, ZBTB7B, ZFP36L1, ZYX
Kidney	A2M, ACADVL, ACAT1, ACTA2, ACTB, ACTG1, ADGRG1, ADIRF, AEBP1, ALDOB, ANPEP, ANXA2, ANXA5, APOE, APP, AQP1, AQP2, ASAH1, ASS1, ATP1A1, ATP1B1, ATP5F1D, ATP5MC3, ATP5ME, ATP5MF, ATP6V0C, ATP6V1F, B2M, BBOX1, BCAM, BGN, BSG, C1QA, C1R, C7, CA2, CALB1, CALD1, CALM1, CALM2, CANX, CAPN2, CD151, CD24, CD74, CD81, CD9, CDKN1C, CFL1, CHCHD10, CIRBP, CKB, CLCNKB, CLU, COL1A2, COL3A1, COL4A1, COL4A2, COX5A, COX5B, COX6B1, COX6C, COX7A2, COX7B, COX7C, CRIM1, CRIP2, CRYAB, CST3, CTSB, CTSH, CXCL12, CXCL14, CYC1, CYCS, CYSTM1, DCN, DDT, DDX17, DDX5, DEFB1, DSTN, DUSP1, DYNLL1, EEF1D, EEF1G, EEF2, EFHD1, EIF3K, EIF4A1, EIF4A2, ENG, EPAS1, EZR, FABP1, FLNA, FTH1, FTL, FXYD2, FXYD4, GABARAP, GATM, GHITM, GPX3, GSN, GSTP1, GTF2I, HINT1, HNRNPA1, HNRNPA2B1, HSD11B2, HSPA8, HSPB1, HTRA1, IDH2, IFITM2, IFITM3, IGFBP2, IGFBP4, IGFBP5, IGFBP7, IGHA1, IGHG1, IGHG3, IGHG4, IGKC, IGLC1, IGLC2, IGLC3, ITGA3, ITGB1, ITM2B, IVNS1ABP, KCNJ1, KCNJ15, KNG1, KRT8, LAMP1, LAMTOR5, LAPTM4A, LDHA, LGALS1, LRP2, LUM, MAL, MALAT1, MGP, MGST1, MGST3, MIOX, MMP7, MUC1, MYL12A, MYL6, MYL9, MZT2B, NAT8, NDRG1, NDUFA1, NDUFA13, NDUFA2, NDUFA4, NDUFA6, NDUFB2, NDUFB7, NDUFB8, NDUFB9, NDUFC1, NDUFS6, NDUFV1, NEAT1, NME2, NPC2, NPHS2, OAZ1, OGDHL, OST4, P4HB, PCBP1, PCK1, PDZK1IP1, PEBP1, PEPD, PFN1, PGK1, PIGR, PODXL, PPP1R1A, PTGDS, PTH1R, REN, RHCG, RHOA, RNASE1, ROMO1, RTN4, S100A10, S100A2, S100A6, SAT1, SCNN1A, SDC1, SELENOM, SELENOP, SERPINA1, SERPINA5, SFRP1, SLC12A1, SLC12A3, SLC13A3, SLC25A3, SLC25A5, SLC25A6, SLC3A1, SLC5A12, SMIM24, SNHG25, SOD1, SOD2, SPARC, SPINK1, SPP1, SRP14, SSR4, SUCLG1, TAGLN, TAGLN2, TGFBR2, THY1, TIMP1, TIMP2, TIMP3, TINAGL1, TMA7, TMEM176A, TMSB10, TMSB4X, TPI1, TPM1, TPT1, TSC22D1, TSPAN1, TUBA1A, TUBB, TXN, UBA52, UGT2B7, UMOD, UQCRB, UQCRC1, UQCRFS1, VIM, WFDC2

Figure S1. Selected genes used across different datasets.

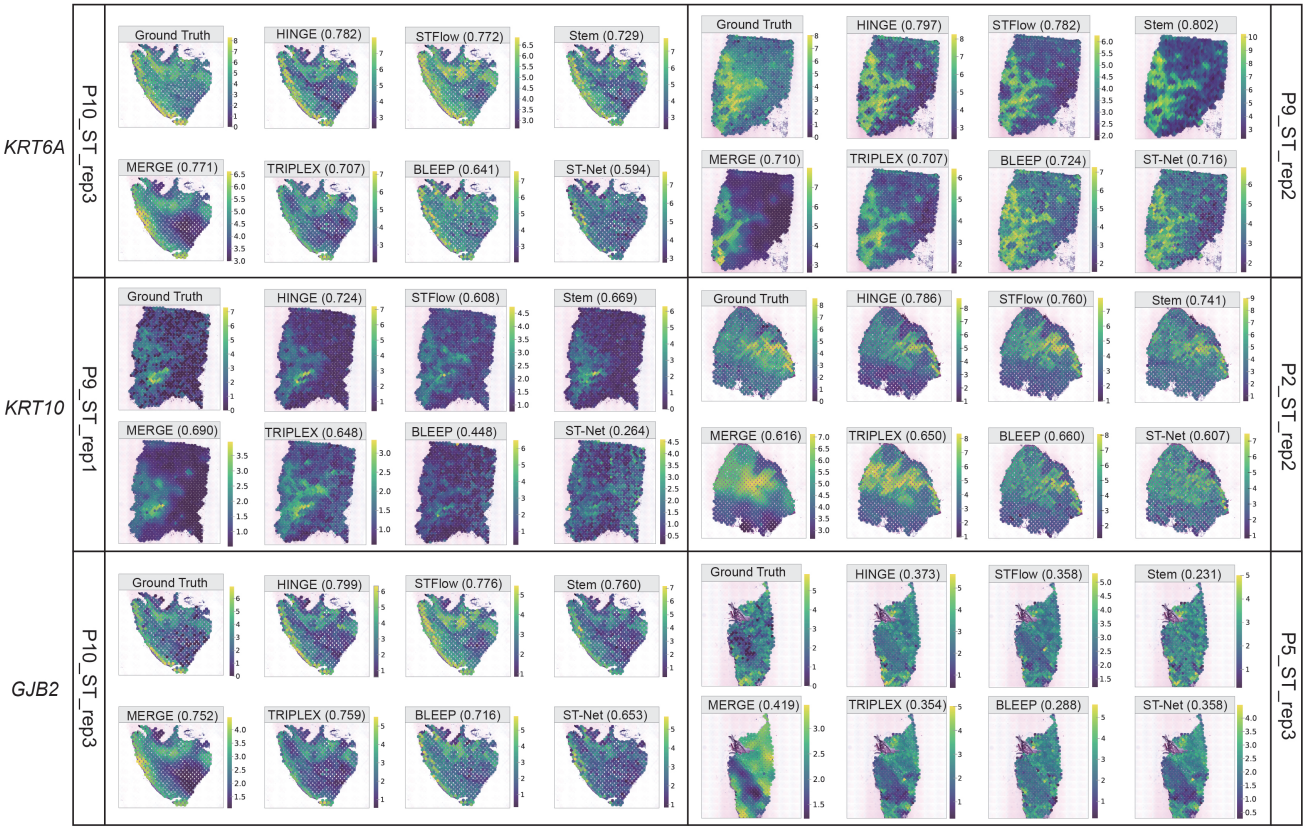

Figure S2. Spatial gene expression predictions for *KRT6A*, *KRT10*, and *GJB2* on the **cSCC** dataset.

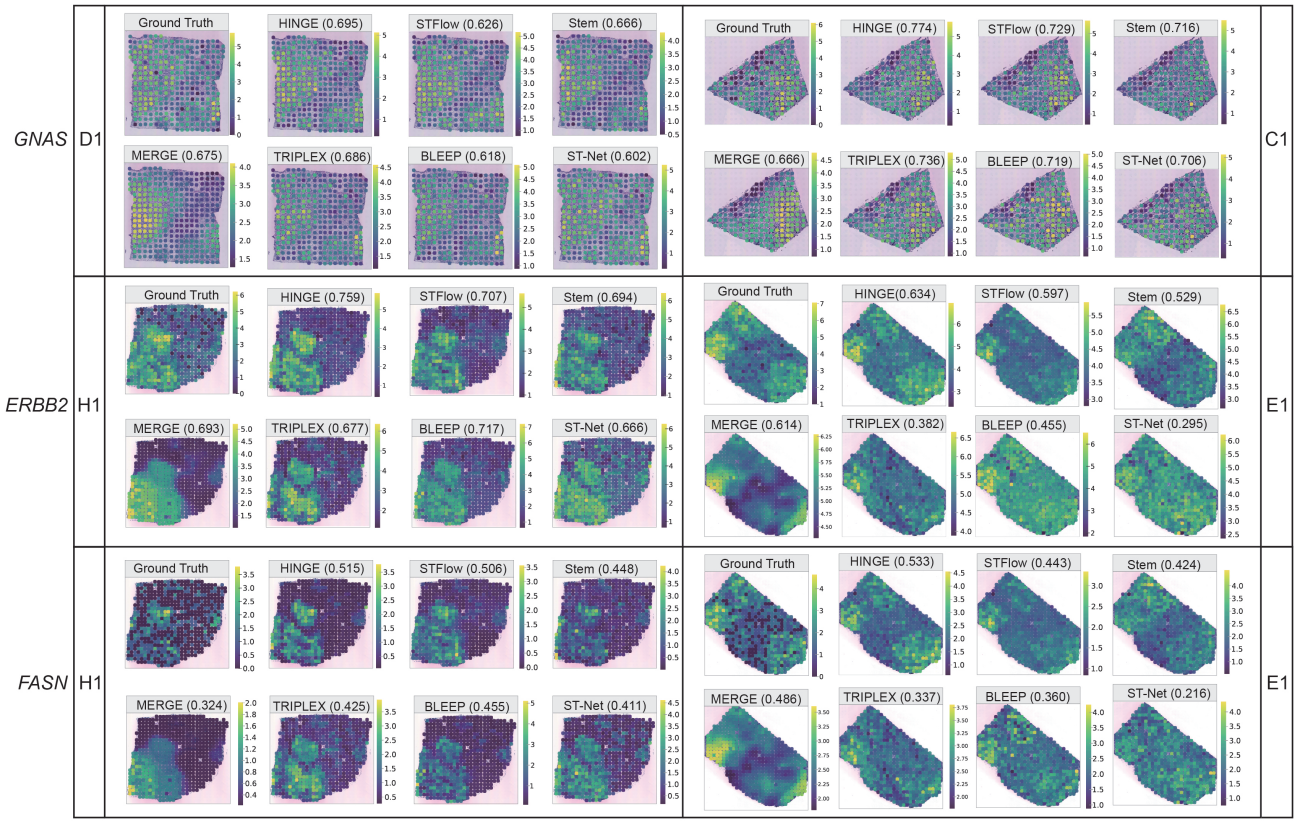

Figure S3. Spatial gene expression predictions for *GNAS*, *ERBB2*, and *FASN* on the **Her2ST** dataset.

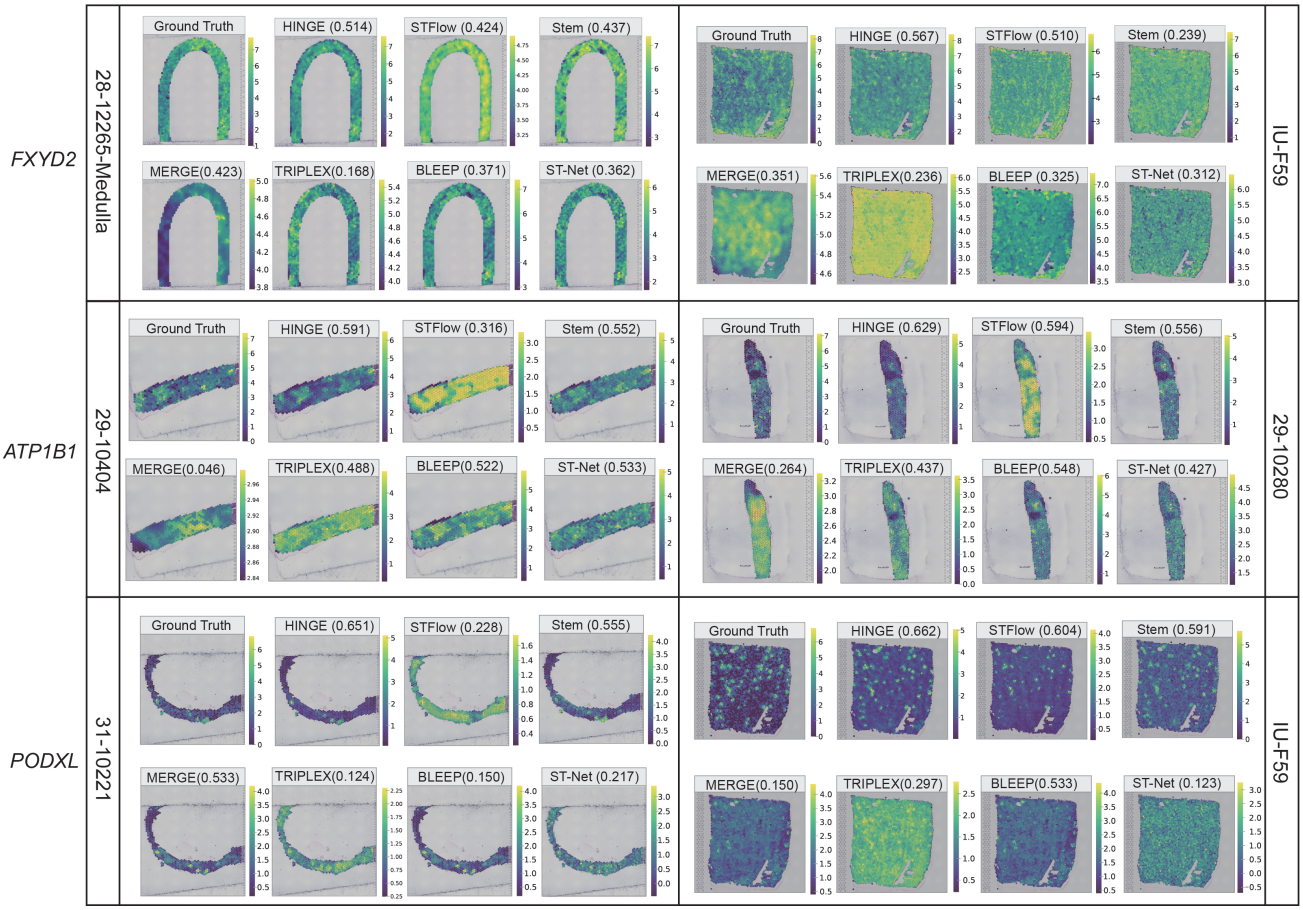

Figure S4. Spatial gene expression predictions for *FXYD2*, *ATP1B1*, and *PODXL* on the **Kidney** dataset.

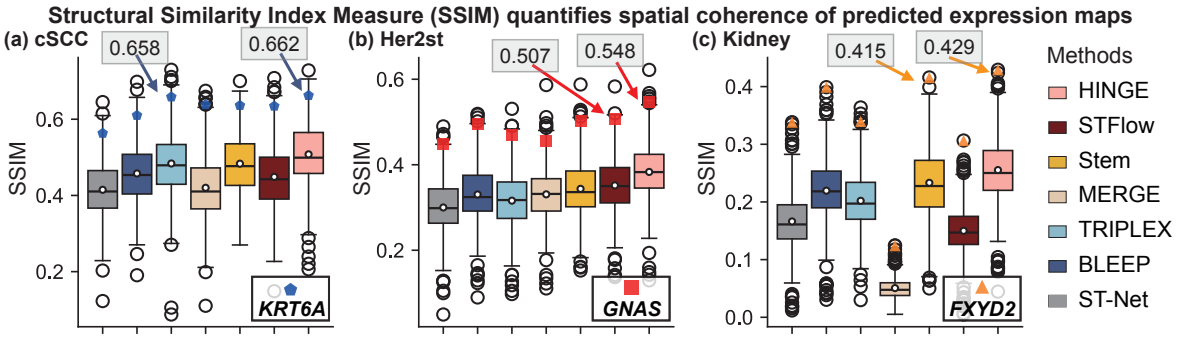

Figure S5. Gene-wise SSIM with marker genes highlighted.

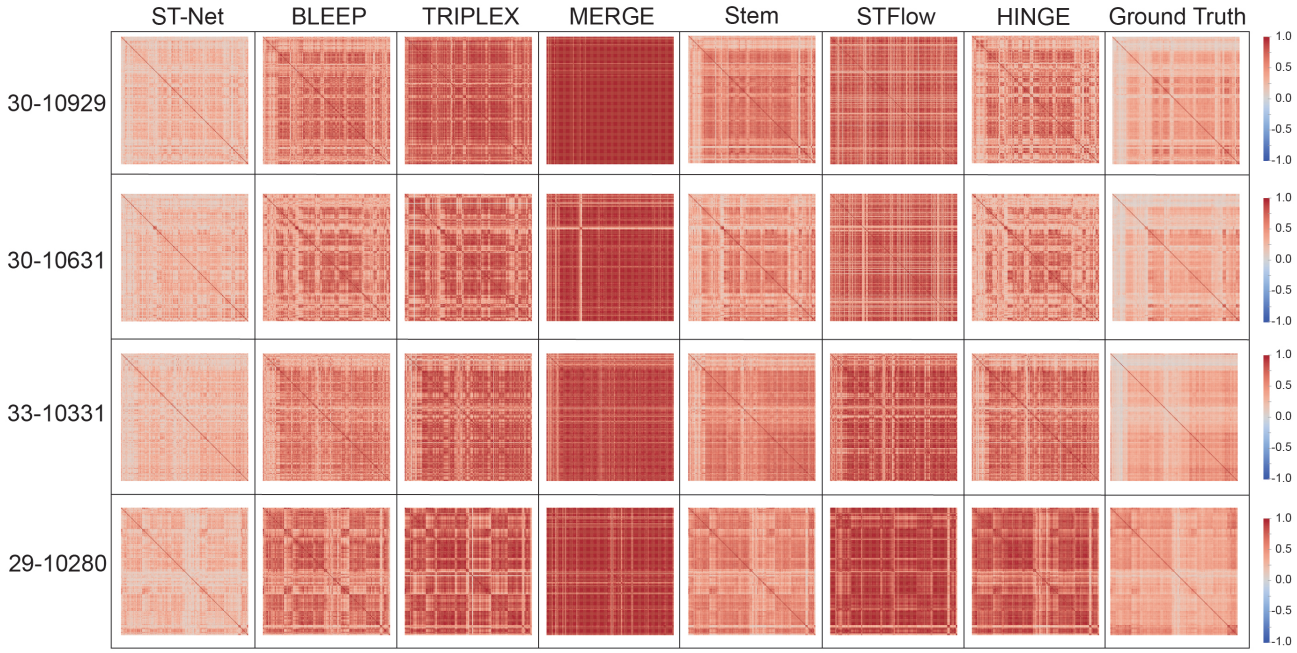

Figure S6. Gene–gene correlation heatmaps on the **Kidney** dataset.

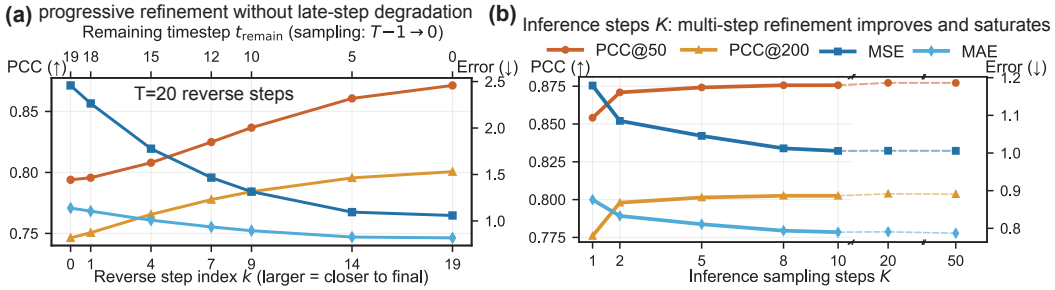

Figure S7. Step-wise analysis of masked diffusion sampling.

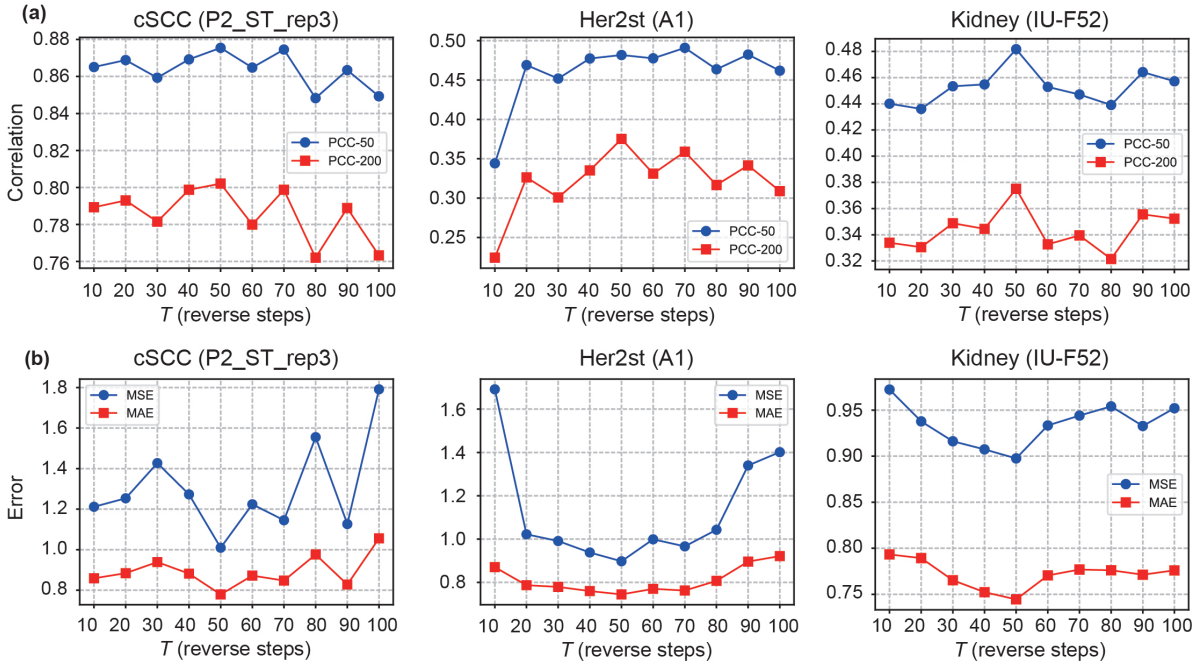

Figure S8. Masking horizon T. Evaluation metrics as functions of the masking horizon T on representative slices from the cSCC, Her2ST, and Kidney datasets.