STMoE: Multi-Scale Mixture-of-Experts for Single-Cell Gene Expression Prediction from Histology

Xi Li
University of California, Irvine
Department of Computer Science
Irvine, California, USA
xil43@uci.edu

Yaqi Hu
University of California, Irvine
Department of Computer Science
Irvine, California, USA
yaqih10@uci.edu

Ziheng Duan
University of California, Irvine
Department of Computer Science
Irvine, California, USA
zihend1@uci.edu

Chaoyang Wang
University of California, Irvine
Department of Computer Science
Irvine, California, USA
chaoyaw2@uci.edu

Jing Zhang
University of California, Irvine
Department of Computer Science
Irvine, California, USA
jingz31@uci.edu

Abstract

Predicting single-cell molecular state from routine histology could make spatial profiling more scalable, but existing histology-to-expression methods are mostly limited to spot-level targets, restricted gene panels, or organ-specific settings. This leaves three gaps for practical single-cell deployment: a lack of large aligned-cell resources, limited support for direct multi-organ prediction, and incomplete full-gene expression recovery from whole-slide images. We address these gaps by curating CellST-24M, a 23.94M-cell histology–transcriptomics resource spanning 66 studies, 25 organ categories, Visium HD, and Xenium, and by introducing STMoE, a WSI-only organ-routed prediction framework for full-gene single-cell expression inference. STMoE detects cells from H&E slides, encodes cell-centered and tissue-context crops with a frozen UNI2-h pathology foundation model, aligns image and expression embeddings with soft contrastive learning, and predicts full-gene profiles through organ-routed retrieval from training-only reference banks. On matched human lung benchmarks, STMoE achieves state-of-the-art cell-level performance, outperforming cell-adapted BLEEP, sCellST and GHIST under both in-domain and cross-patient settings. Across 22 organs, STMoE supports direct multi-organ prediction and achieves gene-wise Pearson up to 0.6451 and cell-wise Pearson up to 0.8379, while retaining approximately 60% predictive capability under cross-patient transfer. In downstream marker-derived analysis, predicted profiles recover coarse cell-state structure with 0.5935 accuracy and 0.5922 balanced accuracy. Together, CellST-24M and STMoE establish a benchmark and modeling framework for WSI-only, multi-organ, full-gene single-cell histology-to-expression prediction.

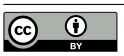
CCS Concepts

• **Computing methodologies** → **Supervised learning**; • **Applied computing** → **Bioinformatics**; • **Information systems** → *Data mining*.

Keywords

spatial transcriptomics, computational pathology, single-cell gene expression prediction, histology foundation models

ACM Reference Format:
Xi Li, Yaqi Hu, Ziheng Duan, Chaoyang Wang, and Jing Zhang. 2026. STMoE: Multi-Scale Mixture-of-Experts for Single-Cell Gene Expression Prediction from Histology. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 7–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799682.3841080



This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy.
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3841080

1 Introduction

Spatial transcriptomics (ST) preserves tissue coordinates, enabling tests of whether histology can recover molecular state. Histology-to-expression models could scale tissue screening, spatial hypothesis generation, and region prioritization without replacing molecular assays. At single-cell resolution, prediction must capture both target-cell morphology and local tissue context.

Current methods lack a practical single-cell deployment interface. Slide- and spot-level models predict aggregated expression [35, 50, 65, 67, 68], smoothing the cell-state heterogeneity that motivates spatial profiling. Recent single-cell methods including sCellST, GHIST, and DeepSpot2Cell [9, 24, 43] narrow this gap, but none simultaneously supports WSI-only inference, multi-organ deployment, and full-gene output (Table 1).

Here, we introduce STMoE, an organ-routed framework that aligns multi-scale image and expression embeddings and retrieves full-gene profiles from organ-specific training-only reference banks. Together with CellST-24M's fixed leakage-controlled splits, it provides WSI-only, multi-organ single-cell inference (Fig. 1). It combines WSI-derived cell detection with organ routing so slides from different organs share one interface while retaining specialized

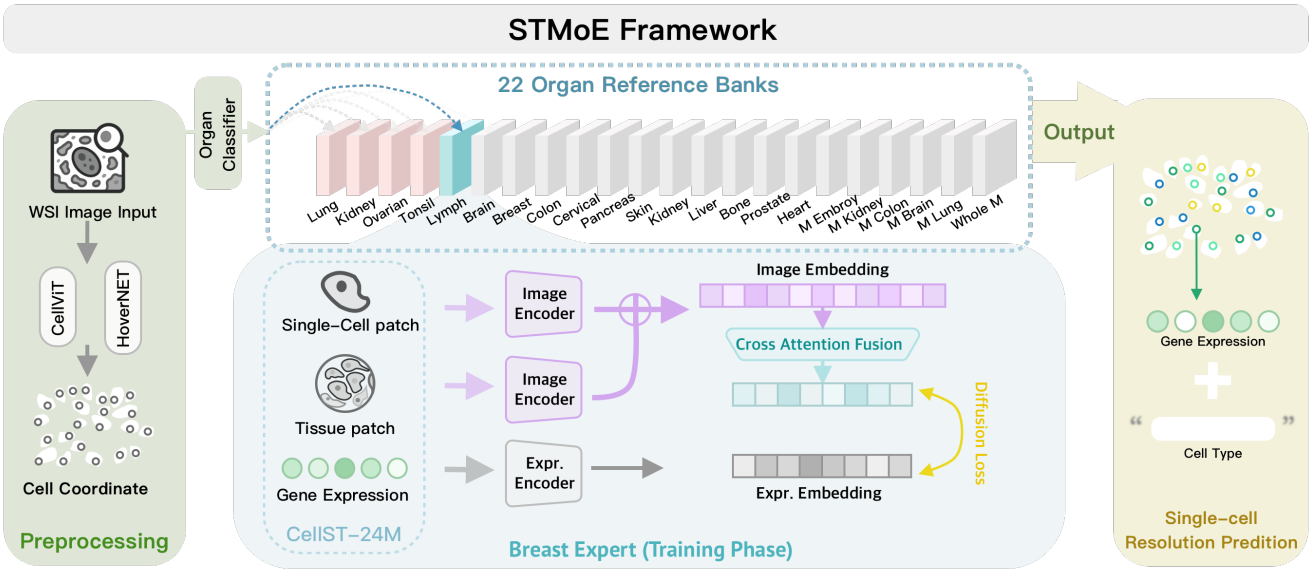


Figure 1: STMoE framework. During training, CellST-24M provides aligned cell records with local/context H&E crops and measured expression profiles. A frozen UNI2-h encoder and an expression branch are aligned with a soft contrastive objective. During WSI-only deployment, CellViT++ detects cells, an organ classifier estimates routing weights, and full-gene expression is predicted by organ-routed retrieval from 22 training-only reference banks. Lung, kidney, ovary, tonsil, and breast have the largest banks and most stable evaluated predictions.

reference banks. Soft contrastive alignment maps image and expression features into a shared retrieval space without a parametric full-gene decoder.

To summarize, we make four contributions to single-cell histology-to-expression prediction:

- **Large dataset curation.** We curate CellST-24M, a 23.94M-cell histology–transcriptomics resource spanning 66 studies, 25 organ categories, Visium HD, and Xenium, with aligned cell records designed for leakage-controlled single-cell benchmarking.
- **Multi-organ full-gene prediction framework.** We introduce STMoE, a WSI-only organ-routed prediction framework that detects cells from H&E slides, encodes cell-centered and tissue-context crops with UNI2-h, and predicts full-gene expression through organ-specific retrieval from training-only reference banks.
- **State-of-the-art strict cell-level performance.** On matched human lung benchmarks, STMoE outperforms representative baselines including cell-adapted BLEEP, sCellST and GHIST under in-domain and cross-patient evaluation.
- **Generalization and downstream validation.** Across 22 organs, STMoE supports direct multi-organ prediction, retains approximately 60% predictive capability under cross-organ transfer, and recovers marker-derived coarse cell-type structure with 0.5935 accuracy and 0.5922 balanced accuracy.

2 Related Work

Histology-to-expression and molecular prediction. Digital pathology has increasingly been used to infer molecular phenotypes from routine H&E images. Open-source slide analysis and weakly supervised WSI learning established scalable computational pathology workflows [4, 8, 42]. Early studies showed that histology can predict mutations, molecular subtypes, and transcriptomic signatures from weak slide- or patch-level supervision [17, 36, 50]. Spatial transcriptomics further enabled direct learning between local morphology and spatially indexed expression. ST-Net predicted breast-cancer gene expression from histology patches [27], while SpaCell integrated morphology and spatial expression for disease-cell prediction [57]. Later methods improved spot-level or super-resolved prediction using transformer, graph, contrastive, and multimodal architectures, including Hist2ST, HisToGene, BLEEP, TCGN, iSTAR, and THItoGene [35, 39, 45, 64, 65, 67, 68]. Recent

Table 1: Capability comparison with representative single-cell histology-to-expression methods. A red check mark denotes native support. A black cross denotes no native support. "Direct multi-organ" denotes native support for slides from multiple organs under one deployed interface. "Full-gene" denotes prediction over the full benchmark gene space rather than a restricted marker, panel, or spot-derived target subset.

Method	WSI-only inference	Multi-organ	Full-gene output
DeepSpot2Cell [43]	✗	✗	✗
sCellST [9]	✓	✗	✗
GHIST [24]	Partial	✗	✗
BLEEP (cell-adapted) [65]	✓	✗	✓
STMoE (ours)	✓	✓	✓

benchmarking further shows that histology-to-expression prediction remains sensitive to tissue context, evaluation protocol, and translational setting [60]. These works motivate our task, but most operate at the spot, tile, or super-resolution level rather than directly on aligned single cells.

Single-cell histology-to-expression prediction. Single-cell expression prediction is more challenging because the target vector is sparse, noisy, and only partly reflected in the visible morphology of the target cell. Recent methods have begun to address this setting, but they often rely on additional non-image information or indirect supervision. GHIST predicts spatially resolved single-cell expression by combining morphology with cell type, neighborhood composition, and expression-related objectives [24]. sCellST predicts single-cell expression from H&E images through weak supervision from spot-level spatial transcriptomics and multiple-instance learning [9]. In contrast, STMoE uses aligned cell-level records during training and requires only histology image inputs at inference: a cell-centered crop and a larger tissue-context crop. This keeps prediction cell-centered and image-only at deployment, while still allowing local tissue context to support genes that are not visually determined by the target cell alone.

Spatial transcriptomics platforms, atlases, and benchmarks. Our work is also related to spatial transcriptomics technologies and reference resources, whose rapid development is summarized in a broad survey [47]. Array-based, sequencing-based, and imaging-based platforms provide spatial expression at different resolutions and gene coverages [10, 12, 22, 33, 49, 53, 54, 58, 61]. Analysis frameworks support scalable exploration, spatial-domain discovery, subspot resolution, cell-type decomposition, multi-slice integration, and data completion [7, 19–21, 30, 44, 72]. Large paired histology–omics datasets such as HEST-1K and STimage-1K4M support systematic benchmarking [11, 34], while single-cell atlases and spatial integration methods provide biological references for annotation and interpretation [5, 18, 38, 48, 52]. These resources are complementary to our objective: rather than mapping reference cells into space, we predict expression directly from histology for aligned cells.

Pathology foundation models and multimodal pretraining. Self-supervised histopathology learning established transferable visual features from unlabeled images [16]. Large pretrained pathology encoders have since become important building blocks for downstream histology tasks. Representative models include CTransPath, HIPT, UNI, CONCH, GigaPath, Virchow, Virchow2, Phikon, H-optimus-0, CHIEF, and PRISM [6, 13, 14, 23, 40, 41, 59, 62, 63, 66, 73]. Vision–language pathology models and general medical-image pretraining further improve transferable and robust representation learning [3, 31, 32]. Broader LLM research examines robust rationalization, intent auditing, and cost-aware routing [69–71]. However, pathology encoders are not directly optimized for expression prediction, so we treat encoder choice as a separate scaling axis in our experiments.

3 Large Dataset Curation: CellST-24M Enables Leakage-Controlled Cell-Level Benchmarking

A central requirement for single-cell histology-to-expression modeling is a resource in which each training example is anchored to an individual cell rather than to a mixed spot or slide region. Existing resources partially support this goal: single-cell atlases provide cell-level molecular profiles but lack paired histology, pathology image–text corpora provide large visual pretraining data but lack expression targets, and spot-level ST resources provide paired histology and expression but aggregate multiple cells per target. CellST-24M closes this resource gap by curating a large cell-aligned histology–transcriptomics benchmark derived from public Visium HD and Xenium studies. The current release contains 23.94M aligned cell records across 66 studies and 25 organ categories, with each record linking a cell-level molecular target to a cell-centered H&E crop, a larger tissue-context crop, spatial coordinates, platform metadata, and sample-level biological context. These aligned records support leakage-controlled comparisons across platforms and preserve the heterogeneity that mixed spot-level targets obscure.

Table 2: Benchmark-oriented comparison with representative resources. "Split" indicates whether the resource natively supports predefined study-level or benchmark-level partitioning.

Dataset	Single Cell	Multi-Organ	Split
Single-cell-only atlases			
CELLxGENE Census [18]	Yes	Yes	Yes
Pathology image–text corpora			
OpenPath [31]	No	Partial	No
Quilt-1M [32]	No	Partial	No
PathGen-1.6M [56]	No	Partial	No
ST + histology resources			
HEST-1K [34]	No	Yes	Yes
STimage-1K4M [11]	No	Yes	Partial
CellST-24M (ours)	**Yes**	**Yes**	**Yes**

CellST-24M uses aligned cells as the benchmark unit (Table 2) and provides a manifest, platform-specific `.h5ad` files, fixed split indices, and benchmark configurations. The release contains 66 standardized studies and 23.94M aligned cells: 42 Xenium studies contribute 16.31M segmented human and mouse cells, while 24 Visium HD studies contribute 7.63M bin-to-cell aligned cells from human, mouse, and other species. These derived packages introduce no additional biological samples and support strict cell-level evaluation beyond spot-, tile-, or image–text-level benchmarks.

The split-indexed benchmark package supports in-domain evaluation across all organs, together with cross-patient and cross-platform protocols for eight multi-sample organs: human breast, ovary, lung, pancreas, and tonsil, and mouse brain, embryo, and kidney. It also includes context-shift evaluation configurations, with all remaining organs used for in-domain benchmarks.

CellST-24M combines two complementary spatial-omics assay families. Xenium provides native cell-level measurements through platform-defined cell boundaries and transcript coordinates. Visium HD provides dense spatial expression bins and whole-slide histology; we convert these bins into cell-level targets by registering molecular coordinates to the histology frame, segmenting candidate cell footprints, aggregating overlapping bins to cells, and filtering records without stable expression support or valid image

context. This curation step yields a common cell-aligned interface across assays: each retained cell has local morphology, neighborhood context, spatial coordinates, metadata, and an expression vector.

The construction pipeline follows the same record format across platforms. First, we audit each public source for tissue coverage, image availability, coordinate consistency, and usable molecular targets. Second, we define the cell unit: Xenium uses platform cell masks, whereas Visium HD uses a bin-to-cell conversion pipeline. Third, we extract two image views centered at each retained cell: a local crop for cell morphology and a larger crop for tissue context. Fourth, we harmonize expression targets within each evaluation regime. Within Visium HD, we use processed full-transcriptome targets; cross-platform experiments use only shared genes after symbol harmonization. Finally, we construct fixed benchmark splits for in-domain, cross-patient, cross-platform, and context-shift evaluation.

This curation also defines the prediction interface used in the rest of the paper. During training and evaluation, each retained cell provides a local crop I_i^c, a context crop I_i^t, an organ label o_i, and a processed expression vector g_i. The model learns an image–expression embedding space from the training split and predicts expression through retrieval from training-only reference banks:

$$f_\theta(I_i^c, I_i^t; \mathcal{B}_{\text{train}}^{(o_i)}) \to \hat{g}_i \in \mathbb{R}^G. \tag{1}$$

At WSI deployment, the query input is the slide image itself. Cell coordinates are obtained by CellViT++, building on established generalist and histology-specific cell-segmentation systems [25, 26, 29, 51, 55]; organ identity is estimated by an organ classifier, and the same local/context crop interface is generated automatically from histology. In-domain evaluation uses spatially separated train and test regions from the same sample, while cross-patient evaluation trains on one sample and tests on a different sample with the same target-gene definition. These regimes are defined from curated CellST-24M split files rather than sampled ad hoc, keeping model and baseline comparisons matched in cells, genes, preprocessing, and metrics.

4 Method

4.1 Overview

STMoE predicts single-cell gene expression from histology through a retrieval-based image-to-expression framework trained on CellST-24M. The training unit is an aligned cell record: each retained cell is paired with a cell-centered H&E crop, a larger tissue-context crop, and a measured expression profile. The two image crops are encoded by a shared frozen UNI2-h pathology foundation model and projected into a compact visual embedding. A paired expression branch maps reduced expression profiles into the same embedding space, and the two branches are aligned with a symmetric soft contrastive objective.

All training, validation, and test examples are defined by leakage-controlled CellST-24M split files. The split is fixed before model training, hyperparameter selection, and reference-bank construction. Test cells and their expression profiles are never used as retrieval references. In-domain, cross-patient, cross-study, and cross-platform regimes are therefore evaluated from predefined non-overlapping split files rather than from ad hoc sampling.

Unlike a parametric full-gene decoder, STMoE predicts expression through organ-routed retrieval. After contrastive alignment, a test cell is represented by its image embedding and retrieves similar expression-measured reference cells from the training split. The final full-gene expression vector is estimated as a similarity-weighted average of the retrieved reference profiles. Organ-specific reference banks allow the retrieval space to specialize across tissue contexts while preserving the same WSI-to-expression interface.

4.2 Cell-Aligned Training Records and Reference Banks

For each retained training cell i, CellST-24M provides a local crop I_i^c, a tissue-context crop I_i^t, a reduced expression representation x_i, a full-gene expression target g_i, an organ label o_i, and split metadata:

$$r_i = (I_i^c, I_i^t, x_i, g_i, o_i). \tag{2}$$

The local crop captures the target cell and immediate morphology, whereas the tissue-context crop captures nearby tissue organization and microenvironmental cues.

For each organ o, we define non-overlapping training, validation, and test sets:

$$\mathcal{T}_{\text{train}}^{(o)}, \mathcal{T}_{\text{val}}^{(o)}, \mathcal{T}_{\text{test}}^{(o)}. \tag{3}$$

The retrieval reference bank is constructed only from $\mathcal{T}_{\text{train}}^{(o)}$. Validation cells are used only for model selection, and test cells are excluded from both training and retrieval. This constraint is important for retrieval-based prediction because any leakage of test expression profiles into the reference bank would directly inflate performance.

4.3 Multi-Scale Image and Expression Embedding

For each cell i, we extract a local crop I_i^c and a tissue-context crop I_i^t centered at the same coordinate. In our implementation, the two crops are 64×64 and 256×256 pixels, respectively, and are resized to the UNI2-h input resolution. Both crops are encoded by the same frozen visual backbone:

$$z_i^c = E_{\text{UNI2-h}}(I_i^c), \qquad z_i^t = E_{\text{UNI2-h}}(I_i^t). \tag{4}$$

Using a shared frozen encoder keeps both image scales in the same representation space and limits trainable parameters to lightweight projection heads.

The two scale-specific embeddings are concatenated and passed through a residual projection head:

$$u_i = [z_i^c; z_i^t], \tag{5}$$

$$p_i = W_p u_i, \tag{6}$$

$$h_i^v = \text{LayerNorm}\left(p_i + \text{Dropout}\left(W_o\, \text{GELU}(p_i)\right)\right). \tag{7}$$

This produces a 256-dimensional visual embedding h_i^v.

During training, each cell also has a reduced expression representation x_i derived from its measured expression profile. We map

this vector into the same 256-dimensional embedding space using an expression projection head:

$$h_i^g = P_g(x_i). \tag{8}$$

The expression branch defines the molecular embedding space used for contrastive alignment and reference-cell retrieval. It is used only for cells with measured expression in the training/reference set; test-query cells are represented by image embeddings only.

4.4 Soft Contrastive Alignment Objective

We train the image and expression projection heads with a symmetric soft CLIP-style objective [15, 37, 46]. For a minibatch of n paired cells, image-expression logits are computed by dot-product similarity:

$$\ell_{ij} = \frac{(h_i^g)^\top h_j^v}{\tau}, \tag{9}$$

where τ is the temperature.

Instead of using only one-hot paired labels, we construct soft targets from the average of image-image and expression-expression similarities:

$$q_{ij} = \mathrm{softmax}_j\left(\frac{(h_i^v)^\top h_j^v + (h_i^g)^\top h_j^g}{2\tau}\right). \tag{10}$$

The final loss is the average of expression-to-image and image-to-expression cross entropy:

$$\mathcal{L}_{\mathrm{soft}} = \frac{1}{2}\left[\mathrm{CE}(\ell, q) + \mathrm{CE}(\ell^\top, q^\top)\right]. \tag{11}$$

This objective aligns matched image and expression embeddings while preserving neighborhood structure among biologically similar cells. Full-gene prediction is not learned through a parametric decoder; instead, the aligned embedding space is used for retrieval-based expression estimation.

4.5 Organ-Routed Full-Gene Prediction

After training, STMoE predicts full-gene expression by retrieving expression-measured reference cells from organ-specific training banks. For each organ o, we build an organ-specific reference bank:

$$\mathcal{B}^{(o)} = \{(h_j^g, g_j) : j \in \mathcal{T}_{\mathrm{train}}^{(o)}\}. \tag{12}$$

Here, h_j^g is the expression embedding of reference cell j, and g_j is its full-gene expression profile. The reference bank is constructed only from the training split, so validation and test cells are excluded from retrieval.

For a query cell i and organ o, we retrieve the K_r nearest reference cells by image-to-expression similarity:

$$\mathcal{N}_{K_r}^{(o)}(i) = \mathrm{TopK}_{j \in \mathcal{T}_{\mathrm{train}}^{(o)}}\left((h_i^v)^\top h_j^g\right). \tag{13}$$

The retrieval weights are computed by a softmax over the selected references:

$$w_{ij}^{(o)} = \frac{\exp((h_i^v)^\top h_j^g)}{\sum_{l \in \mathcal{N}_{K_r}^{(o)}(i)} \exp((h_i^v)^\top h_l^g)}. \tag{14}$$

The organ-specific retrieval prediction is:

$$\hat{g}_i^{(o)} = \sum_{j \in \mathcal{N}_{K_r}^{(o)}(i)} w_{ij}^{(o)} g_j. \tag{15}$$

When the organ is known from the benchmark split, we use the corresponding organ-specific reference bank. For WSI deployment, an organ classifier provides organ-routing weights $a_o(S)$, and the final prediction is:

$$\hat{g}_i = \sum_{o \in \mathcal{O}} a_o(S)\, \hat{g}_i^{(o)}. \tag{16}$$

This organ-routed retrieval formulation provides organ-specific specialization without introducing a parametric full-gene decoder.

4.6 WSI-Only Deployment Inference

At deployment time, the input is a whole-slide H&E image S. We first run two preprocessing modules. A cell-detection module, implemented with CellViT++, estimates cell locations:

$$\{(x_i, y_i)\}_{i=1}^{N_S} = D_{\mathrm{cell}}(S). \tag{17}$$

For each detected cell, we crop I_i^c and I_i^t centered at (x_i, y_i) and encode them with the frozen UNI2-h image branch.

In parallel, an organ classifier estimates the organ identity or organ probability vector:

$$a(S) = C_\eta(S), \qquad a(S) \in \Delta^{|\mathcal{O}|}. \tag{18}$$

The organ prediction selects or weights organ-specific reference banks for retrieval. The query WSI does not require molecular measurements, external cell-type annotations, or manually provided neighborhood labels. Thus, deployment inference is WSI-only: cell locations, image crops, organ routing, and single-cell expression predictions are all produced from the input histology slide.

5 Experiments

We organize the experiments around the claims needed to support WSI-only, multi-organ, full-gene single-cell prediction. First, we define matched cells, leakage-controlled splits, and training-only reference banks to ensure that retrieval-based evaluation does not leak test expression profiles. Second, we test whether STMoE works beyond a single tissue by comparing it with representative baselines across organs. Third, we evaluate in-domain and cross-patient transfer on matched human lung benchmarks. Finally, we analyze encoder choice, retrieval size, reference-bank coverage, target space, image scale, and downstream cell-state consistency to explain where the performance gains come from.

5.1 Data, Evaluation Protocol, and Training Details

The main human lung benchmark uses two public 10x Genomics Visium HD CytAssist lung cancer datasets:

- **Sample A** is the Visium HD Spatial Gene Expression Libraries, Post-Xenium, Human Lung Cancer (FFPE) dataset, analyzed with Space Ranger 3.0.0 [1].
- **Sample B** is the Visium HD Spatial Gene Expression Library, Human Lung Cancer (Fixed Frozen) dataset, analyzed with Space Ranger 3.1.1 [2].

We evaluate two matched regimes. In the in-domain A→A regime, the left two-thirds of sample A are used for training and retrieval reference construction, while the spatially held-out right third is used for testing. In the cross-patient A→B regime, the same sample-A training split is used, but evaluation is performed on sample B.

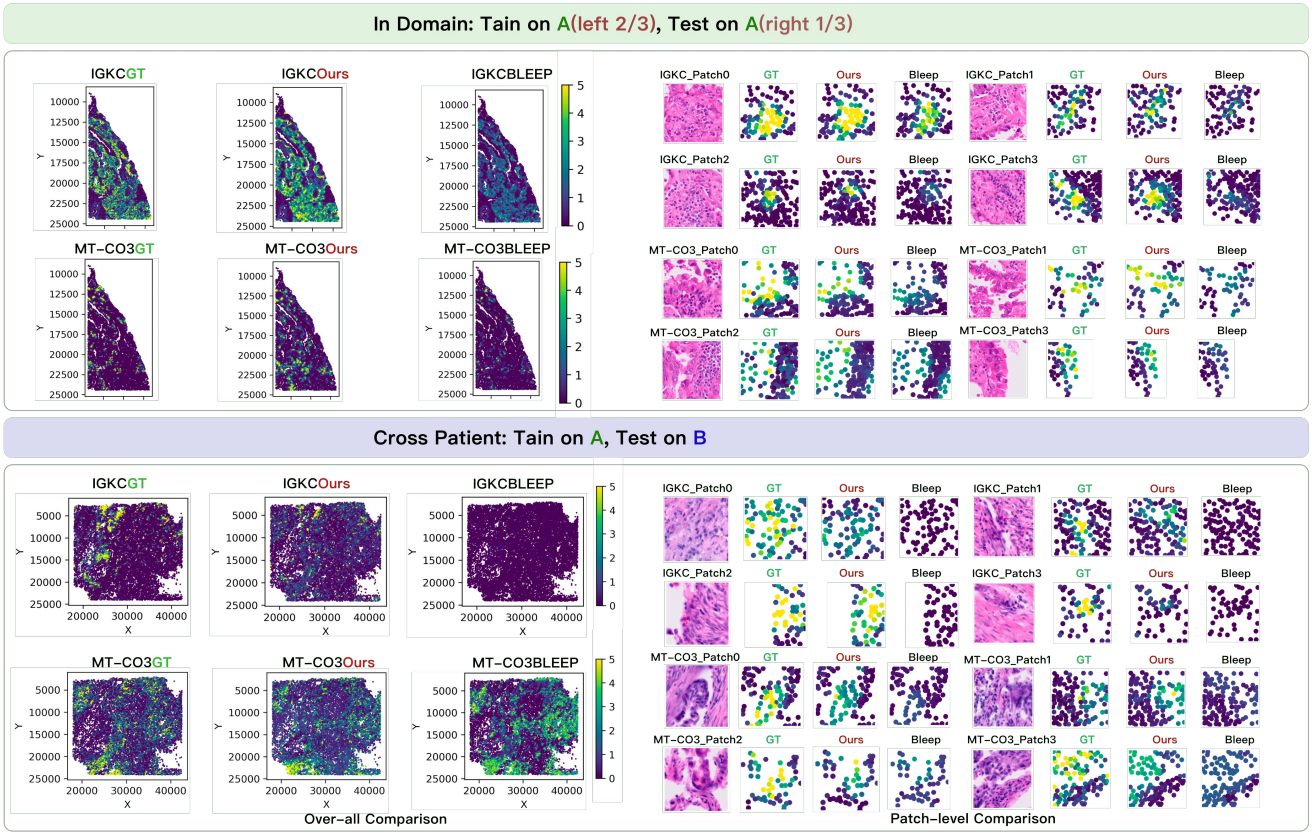


Figure 2: Qualitative single-cell image-to-gene prediction on the matched lung benchmark. Top: in-domain A→A evaluation. Bottom: cross-patient A→B evaluation. Left panels compare tissue-level ground-truth and STMoE expression maps; right panels show representative H&E neighborhoods with ground-truth and predicted expression patterns.

Table 3: Multi-organ strict cell-level comparison against BLEEP [65], GHIST [24], and sCellST [9]. Results are reported on the five organs for which all baselines are available. BL, GH, SC, and ST denote BLEEP, GHIST, sCellST, and STMoE, respectively. Higher is better for all metrics, and bold values indicate the best method for each organ and metric.

Organ	Gene Pearson				Gene Spearman				Cell Pearson			
	BLEEP [65]	GHIST [24]	sCellST [9]	STMoE (ours)	BLEEP [65]	GHIST [24]	sCellST [9]	STMoE (ours)	BLEEP [65]	GHIST [24]	sCellST [9]	STMoE (ours)
Ovary	0.4239	0.4523	0.0800	**0.6451**	0.4396	0.4813	0.0640	**0.6049**	0.8212	0.7382	0.6858	**0.8379**
Lung	0.4241	0.4754	0.1255	**0.5178**	0.4310	0.4814	0.1748	**0.5077**	0.5622	0.5787	0.4704	**0.6042**
Breast	0.0079	0.0487	0.0032	**0.5118**	0.0051	0.0495	0.0070	**0.5849**	0.1282	0.1832	0.1240	**0.5465**
Kidney	0.3356	0.3040	0.0236	**0.3643**	**0.3468**	0.2781	0.0264	0.3313	**0.5369**	0.5285	0.2601	0.5112
Tonsil	0.1809	0.1591	0.0088	**0.3817**	0.1794	0.1488	0.0084	**0.3354**	0.7808	0.6107	0.7199	**0.8193**

Unless otherwise stated, all methods are evaluated on matched cells, matched target genes, and the same preprocessing pipeline. **Cell-level target construction.** Visium HD measures expression on a dense 8 μm bin grid rather than directly on segmented cells. We construct cell-level targets by registering molecular coordinates to the histology frame, segmenting candidate cell footprints, assigning overlapping bins to segmented cells, resolving boundary cases, and retaining cells with valid expression support and aligned image crops. For platforms with native cell-boundary support, such as Xenium, the platform mask and transcript coordinates define the cell unit directly. Each retained cell contains a cell identifier, spatial coordinate, local crop, context crop, sample metadata, and expression vector.

Leakage-controlled splits. All training, validation, test, and retrieval reference sets are fixed before model fitting. In each benchmark regime, the retrieval reference bank is constructed only from

Table 4: Representative gene-level results by organ. Sample identifiers are omitted. Each row reports gene name, expression mean, expression variance, and Pearson correlation.

Ovary				Lung				Breast				Kidney				Tonsil			
Gene	Mean	Var.	Pearson	Gene	Mean	Var.	Pearson	Gene	Mean	Var.	Pearson	Gene	Mean	Var.	Pearson	Gene	Mean	Var.	Pearson
HVG (Highly Variable Gene)																			
MT-CO1	30.587	481.689	0.5211	IGKC	4.528	293.750	0.4483	EVL	0.134	0.254	0.2640	IGKC	0.480	30.080	0.4551	CD74	18.546	332.186	0.3205
MT-CO3	26.776	418.151	0.4869	IGHG1	1.858	65.348	0.4337	MYL9	0.121	0.226	0.3071	MT-CO3	2.024	7.732	0.4967	CLU	4.141	61.560	0.4522
MT-CYB	18.797	226.742	0.4963	MT-CO3	4.939	57.392	0.6521	TRPS1	0.103	0.220	0.3155	IGHG1	0.122	2.229	0.4973	TMSB4X	7.800	57.284	0.3285
HEG (Highly Expressed Gene)																			
CD9	15.464	150.632	0.5012	MT-ATP6	3.043	20.250	0.6376	TAGLN	0.125	0.202	0.2850	MT-CO1	1.041	2.037	0.3995	CCL19	2.975	31.486	0.3257
MT-ATP6	15.245	145.242	0.4772	TMSB4X	2.531	10.303	0.5497	SLC39A6	0.090	0.164	0.3313	MT-CYB	0.984	2.166	0.4517	CD79A	2.567	14.714	0.4282
ERBB2	14.218	117.295	0.5408	CD74	2.268	9.902	0.4967	ACTA2	0.068	0.121	0.3687	MT-ATP6	0.884	1.721	0.3820	CD37	2.513	10.586	0.3325

the corresponding training split. Validation cells are used only for checkpoint selection and hyperparameter selection. Test cells and their expression profiles are excluded from both training and retrieval. This is essential for retrieval-based prediction because leakage of test expression profiles into the reference bank would directly inflate performance.

Metrics. We report gene-wise Pearson and Spearman correlations, cell-wise Pearson and Spearman correlations, and sparse-expression F1 when applicable. Gene-wise metrics measure whether each gene's spatial variation is recovered across cells. Cell-wise metrics measure whether each cell's expression profile is recovered across genes. For sparse-expression recovery, we prefer rate-matched or row-wise top-k F1 in baseline comparisons because default thresholded F1 can be distorted by different predicted nonzero rates.

Baselines. We compare against a cell-adapted BLEEP baseline [65] under matched cell-level settings. In the expanded human lung benchmark, we additionally compare against DeepSpot2Cell [43], sCellST [9], and GHIST [24]. Because these methods differ in their native supervision and deployment assumptions, we report all methods under the same strict cell-level evaluation protocol, and separately note when a method uses spot-derived supervision, additional non-image information, or a more favorable aggregated evaluation unit.

Training details. For STMoE, we use MahmoodLab/UNI2-h as the frozen visual backbone with pretrained Hugging Face weights and an input resolution of 224×224. The local and tissue-context crops are 64×64 and 256×256 pixels, respectively, and are resized before encoding. Images are normalized with ImageNet mean and standard deviation. During training, we apply random horizontal flips, vertical flips, and rotations from $\{0°, 90°, 180°, -90°\}$.

The image and expression projection heads use a projection dimension of 256 and dropout of 0.1. We train the projection heads with AdamW using learning rate 1×10^{-3}, weight decay 1×10^{-4}, batch size 8, and temperature $\tau = 1.0$, while keeping the UNI2-h backbone fixed. Models are trained for 10 epochs, and the checkpoint with the lowest validation loss is selected. Unless otherwise stated, retrieval uses $K_r = 50$ nearest reference cells, selected from the retrieval-size sensitivity analysis in Table 8.

Table 5: Human lung benchmark under strict cell-level evaluation. All methods are evaluated on matched cells and target genes within each regime. All reported metrics are correlations; higher is better.

Method	Gene Pearson	Gene Spearman	Cell Pearson
In-domain A→A			
BLEEP [65]	0.4451	0.4066	0.5979
sCellST [9]	0.1281	0.1845	0.4832
GHIST [24]	0.4754	0.4814	0.5787
STMoE (ours)	**0.5117**	**0.5069**	**0.6017**
Cross-patient A→B			
BLEEP [65]	0.2248	0.2267	0.2592
sCellST [9]	0.0747	0.0681	0.1502
GHIST [24]	0.1959	0.1881	0.0844
STMoE (ours)	**0.2986**	**0.2405**	**0.3451**

5.2 STMoE Outperforms Baselines Across Most Multi-Organ Cell-Level Metrics

STMoE achieves the strongest overall multi-organ performance under strict cell-level evaluation, outperforming or matching representative baselines across most organs and metrics (Table 3). Across the five organs with complete baseline results, STMoE obtains the best score in 12 of 15 organ–metric comparisons. It is particularly strong for full-profile recovery, giving the best cell-wise Pearson in ovary, lung, breast, and tonsil. These results support the use of organ-routed retrieval as a shared multi-organ prediction interface rather than a single-organ model.

Representative gene-level results show that the recovered signal is not limited to a single marker family (Table 4). The retained examples include mitochondrial genes, immune-associated genes, epithelial markers, and tissue-structure-related genes, with several genes reaching strong Pearson correlations such as MT-CO3 in lung and ERBB2 in ovary. These examples support the main multi-organ result by showing that STMoE recovers heterogeneous gene-level signals rather than only one narrow marker type.

Table 6: Frozen pathology encoder comparison under a unified prediction protocol. H-optimus-0 is omitted. Bold values indicate the best result among the retained encoders for each metric. The highlighted row denotes the encoder selected for the final STMoE model.

Model	Mouse			Human Lung		
	Avg. Pearson	Sensitivity	Accuracy	Avg. Pearson	Sensitivity	Accuracy
ResNet50 [28]	0.4994 ± 0.0675	0.7927 ± 0.0140	0.6847 ± 0.0119	0.3629 ± 0.0860	0.6376 ± 0.0057	0.6222 ± 0.0004
CTransPath [62]	0.5058 ± 0.0611	0.7933 ± 0.0127	0.6865 ± 0.0097	0.3157 ± 0.1332	0.6287 ± 0.0146	0.6166 ± 0.0060
Phikon [23]	0.4961 ± 0.0708	0.7918 ± 0.0149	0.6828 ± 0.0138	0.3690 ± 0.0799	0.6351 ± 0.0082	0.6187 ± 0.0039
CONCH [40]	0.5145 ± 0.0524	0.7994 ± 0.0073	0.6905 ± 0.0061	0.3784 ± 0.0705	0.6361 ± 0.0072	0.6177 ± 0.0049
UNI [14]	**0.5549 ± 0.0120**	0.8051 ± 0.0016	0.6953 ± 0.0013	**0.4360 ± 0.0129**	0.6416 ± 0.0017	0.6218 ± 0.0008
UNI2-h [14]	0.5545 ± 0.0124	**0.8067 ± 0.0000**	**0.6966 ± 0.0000**	0.4352 ± 0.0137	**0.6425 ± 0.0008**	**0.6223 ± 0.0003**

Table 7: Reference-bank size, target-space, and image-scale ablations on the matched lung setting. Higher is better for all reported metrics. The highlighted row denotes the default configuration.

Variant	Gene P	Gene S	Cell P	Cell S	F1
Reference-bank size					
20k references	0.5318	0.5013	0.5901	0.5349	0.6291
50k references	0.5166	0.4911	0.5949	0.5332	**0.6330**
Full reference bank, $K_r = 50$	**0.5541**	0.4867	**0.6045**	0.5430	0.5757
Target space					
Train 50, eval 50	0.5178	**0.5077**	0.6042	**0.5452**	0.6179
Image scale					
No cell crop	0.4482	0.3951	0.4876	0.4372	0.4628
No context crop	0.2163	0.1895	0.2352	0.2089	0.2215

Table 8: Retrieval-size sensitivity on the matched lung setting. $K_r = 50$ is selected as the default because it improves continuous-expression correlation while avoiding the stronger over-smoothing observed at larger retrieval sizes.

K_r	Gene P	Gene S	Cell P	F1	Pred. rate
1	0.4033	0.4154	0.5038	0.6280	0.2790
5	0.4970	0.4820	0.5710	**0.6530**	0.4536
10	0.5221	**0.4915**	0.5848	0.6329	0.5199
20	0.5394	0.4909	0.5948	0.6080	0.5795
50	0.5541	0.4867	0.6045	0.5757	0.6477
75	0.5588	0.4846	0.6079	0.5622	0.6751
100	**0.5623**	0.4828	**0.6101**	0.5523	0.6950

5.3 STMoE Outperforms Cell-Level Baselines Under In-Domain and Cross-Patient Lung Transfer

We next evaluate STMoE on matched human lung benchmarks against representative single-cell histology-to-expression baselines. The in-domain A→A setting evaluates spatial generalization within the same sample, whereas the cross-patient A→B setting evaluates transfer to a different patient sample with the same target-gene definition. This comparison is stricter than spot-level or tile-level prediction because every method is evaluated against cell-level expression targets.

Table 5 summarizes the main human lung benchmark. In the current matched comparison with cell-adapted BLEEP, STMoE improves all three reported correlation metrics in the in-domain A→A regime: gene-wise Pearson increases from 0.4451 to 0.5117, gene-wise Spearman from 0.4066 to 0.5069, and cell-wise Pearson from 0.5979 to 0.6017. These results indicate that STMoE better recovers continuous expression structure when training and testing are performed within the same sample under spatially separated splits.

The cross-patient A→B setting is more challenging because it evaluates transfer to a different sample. In this regime, STMoE improves gene-wise Pearson from 0.2248 to 0.2986, gene-wise Spearman from 0.2267 to 0.2405, and cell-wise Pearson from 0.2592 to 0.3451. The gain is largest for cell-wise Pearson, suggesting that organ-routed retrieval preserves more profile-level expression correspondence under patient shift.

For the expanded benchmark, we evaluate DeepSpot2Cell, sCellST, GHIST, and cell-adapted BLEEP under the same in-domain and cross-patient protocol. We report the strict cell-level results as the primary comparison. When relevant, we also report spot-aggregated or native-granularity results as diagnostic analyses rather than as the main benchmark, because aggregation can smooth single-cell noise and produce a more favorable evaluation setting.

5.4 Pathology Encoder Selection: UNI2-h Provides the Strongest Encoder

Single-cell prediction depends strongly on the visual representation used by the downstream retrieval model. We therefore evaluate frozen histopathology encoders under a unified retrieval protocol on mouse brain and human lung benchmarks. Table 6 reports average Pearson correlation, sensitivity, and accuracy across the compared encoders.

UNI and UNI2-h form the strongest retained encoder tier. UNI gives the highest average Pearson correlation on both mouse and human lung, while UNI2-h achieves the best sensitivity and accuracy on both benchmarks and remains within a very small margin of UNI in average Pearson. This trade-off makes UNI2-h a stable

default encoder for the final STMoE model: it preserves near-best continuous-expression correlation while improving binary expression recovery and overall accuracy. We therefore use UNI2-h as the frozen pathology encoder in subsequent experiments.

5.5 Ablation: Retrieval Neighborhood, Reference Coverage, and Image Scale Shape Prediction Quality

Retrieval size controls a clear trade-off between continuous expression recovery and sparse nonzero-expression detection (Table 8). Small retrieval neighborhoods preserve sparsity better: $K_r = 5$ gives the highest F1 score, indicating the strongest nonzero-expression recovery, but its gene-wise and cell-wise Pearson correlations remain lower than larger retrieval neighborhoods. In contrast, increasing K_r improves continuous correlation by averaging over more reference profiles. For example, $K_r = 50$ improves gene-wise Pearson from 0.5221 to 0.5541 and cell-wise Pearson from 0.5848 to 0.6045 relative to $K_r = 10$, while avoiding the stronger over-prediction observed at very large retrieval sizes. An additional $K_r = 300$ run gives the highest Pearson values in the sweep, but its predicted nonzero rate rises to 0.7730 and F1 drops to 0.5147, indicating oversmoothing and excessive nonzero prediction. We therefore use $K_r = 50$ as the default setting because the paper prioritizes continuous full-gene expression recovery while still maintaining a reasonable sparsity trade-off.

Reference-bank coverage further affects the quality of organ-routed retrieval (Table 7). After fixing the same retrieval setting, using the full reference bank gives the strongest continuous-expression recovery, reaching gene-wise Pearson of 0.5541 and cell-wise Pearson of 0.6045. Smaller reference banks remain competitive for sparse nonzero detection, with the 50k-reference setting achieving the highest F1 score of 0.6330, but they do not match the full bank on Pearson correlation. This supports using the full available training reference bank as the default when the goal is correlation-oriented full-gene prediction.

The target-space ablation shows a complementary trade-off. Direct 50-gene training with the full reference bank gives stronger gene-wise Pearson, whereas training on the full gene space and evaluating on 50 genes improves rank correlation, cell-wise Spearman, and F1. These results suggest that target-restricted training better optimizes continuous recovery on the evaluated gene subset, while full-gene training preserves broader expression structure and sparsity behavior.

Image-scale ablations confirm that both cell-centered morphology and tissue context are needed for robust prediction (Table 7). Removing the cell crop reduces gene-wise Pearson from 0.5541 to 0.4482 and cell-wise Pearson from 0.6045 to 0.4876, indicating that target-cell morphology contributes useful expression signal. Removing the context crop causes a much larger drop, reducing gene-wise Pearson to 0.2163 and cell-wise Pearson to 0.2352. This suggests that local tissue organization and neighborhood morphology provide essential cues for single-cell expression recovery. The full model therefore retains both image scales.

Table 9: Global metrics for the representative sample-A downstream cell-state analysis. The analysis uses marker-derived coarse labels over a restricted 50-gene target subset.

Metric	Value
Cells	111,772
Clusters	10
Genes in downstream target	50
Cell-type accuracy	0.5935
Balanced accuracy	0.5922
Cluster–prediction ARI	0.1362
Cluster–prediction NMI	0.2154
Retained marker-derived classes	Club, AT2, Immune

5.6 STMoE shows strong performance on coarse cell-type prediction

Although the primary task is gene-expression prediction, the predicted profiles should preserve enough biological structure to support downstream summaries. We run a marker-derived downstream analysis on the representative sample A using 111,772 aligned cells, 10 unsupervised clusters, and a restricted 50-gene target subset. Because this setup preserves stable signal for three coarse classes—Club, AT2, and Immune—we interpret it as a coarse cell-state recovery use case rather than a core benchmark task.

As shown in Figure 3, the resulting coarse assignments reach 0.5935 accuracy and 0.5922 balanced accuracy. At the cluster level, all 10 clusters keep the same dominant coarse label between ground-truth-derived and prediction-derived assignments. The strongest agreement appears in a nearly pure Club cluster, cluster 3, with agreement 0.934. Several larger clusters also show robust agreement, including clusters 0, 2, 7, and 9, with agreement values of 0.639, 0.656, 0.550, and 0.670, respectively. The hardest cases are mixed Immune and AT2 clusters, especially clusters 1, 5, and 8, where agreement remains below 0.5 and the predicted composition becomes more diffuse. Table 9 summarizes the corresponding global statistics. The moderate cell-level accuracy and balanced accuracy indicate partial recovery of coarse marker-derived states, whereas the lower ARI and NMI show that prediction-derived labels do not fully reproduce the unsupervised cluster structure. These results suggest that STMoE can preserve coarse tissue organization even when fine-grained single-cell expression recovery remains imperfect.

6 Discussion

STMoE addresses three practical gaps in single-cell histology-to-expression prediction: the lack of large aligned-cell resources, limited support for direct multi-organ deployment, and the difficulty of full-gene prediction from WSI-only inputs. CellST-24M provides the aligned-cell benchmark substrate, while STMoE converts a WSI into cell-level predictions through CellViT++ cell detection, organ routing, multi-scale UNI2-h image encoding, soft image–expression alignment, and retrieval from training-only reference banks. This design keeps inference WSI-only while allowing organ-specific specialization without a parametric full-gene decoder.

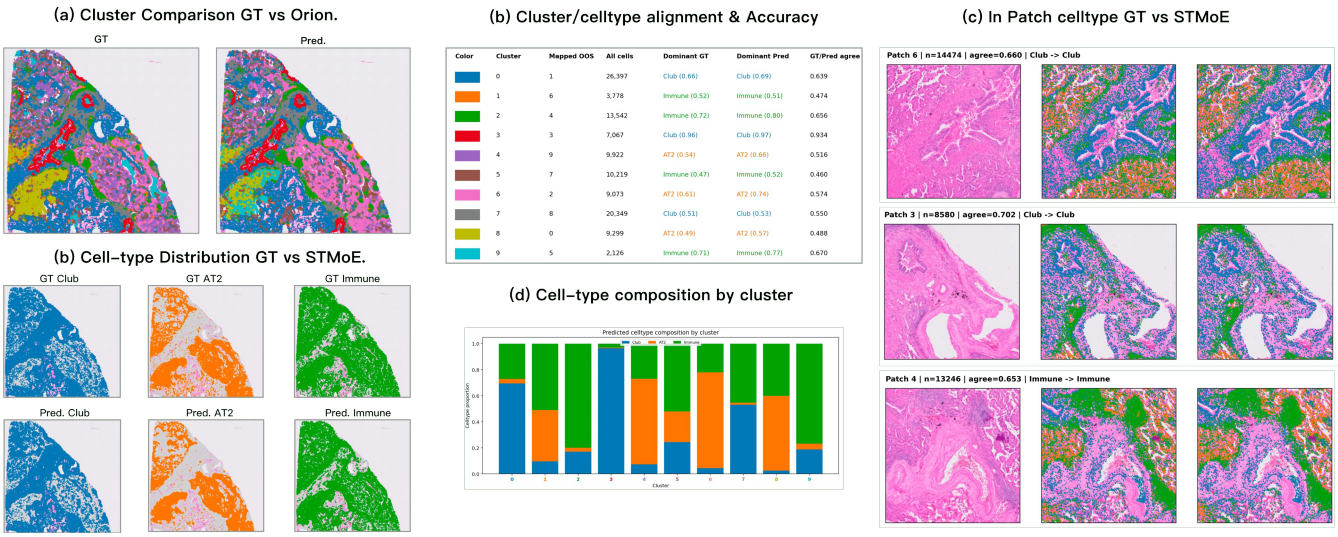


Figure 3: Marker-derived downstream cell-state consistency on sample A. Coarse cell-state assignments are compared between ground-truth-derived and prediction-derived profiles using 111,772 aligned cells, 10 unsupervised clusters, and a restricted 50-gene target subset. The analysis focuses on three stable coarse classes: Club, AT2, and Immune.

The results support this design across benchmark settings. In multi-organ evaluation, STMoE achieves the strongest overall performance across the five organs with complete baseline results, especially for cell-wise Pearson recovery. In matched human lung evaluation, STMoE outperforms cell-adapted BLEEP, sCellST, and GHIST under both in-domain and cross-patient regimes. The ablations show that retrieval size, reference-bank coverage, target space, and image scale all affect the balance between continuous-expression correlation and sparse-expression recovery. The downstream analysis further shows that predicted profiles preserve coarse cell-state structure, reaching 0.5935 accuracy and 0.5922 balanced accuracy.

Several limitations remain. Most importantly, cross-patient prediction is still less accurate than in-domain prediction. This gap likely reflects substantial patient-level context shift: differences in age, sex, comorbidities, treatment history, tissue processing, and disease state can change the global expression baseline even when the tissue morphology appears similar. Because current single-cell histology–transcriptomics resources remain limited in the number and diversity of matched patients, the reference banks do not yet fully cover these patient-specific expression shifts. Addressing this limitation will require larger multi-patient cohorts, better modeling of patient-level covariates, and calibration methods that separate morphology-driven signal from global patient-specific expression bias. This is a central direction for future work.

7 Conclusion

We presented STMoE, a WSI-only organ-routed framework for single-cell gene-expression inference from histology. We also curated CellST-24M, an aligned resource pairing histology and transcriptomics for 23.94M cells. STMoE combines cell-centered morphology, tissue context, soft image–expression alignment, and retrieval from training-only reference banks to predict full-gene profiles. Across multi-organ and matched human lung benchmarks, STMoE outperforms representative cell-level baselines and supports downstream coarse cell-state recovery. These results establish CellST-24M and STMoE as a benchmark and modeling framework for leakage-controlled, multi-organ, full-gene single-cell histology-to-expression prediction.

8 Data and Code Availability

The STMoE code, dataset manifest, and benchmark splits are publicly available at https://github.com/aicb-ZhangLabs/STMoE.

Acknowledgments

This work was supported by the National Institutes of Health under Award Numbers R01HG012572, R01NS128523, and R01DA063316.

Generative AI Usage Disclosure

Generative AI tools were used only for language polishing, structural editing, and LaTeX formatting assistance. All scientific claims, experimental design, data curation, model implementation, quantitative results, and conclusions were produced and verified by the authors. No patient-identifying information was provided to generative AI tools.

References

[1] 10x Genomics. 2024. Visium HD Spatial Gene Expression Libraries, Post-Xenium, Human Lung Cancer (FFPE). https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-lung-cancer-post-xenium-expt. HD Spatial Gene Expression dataset analyzed using Space Ranger 3.0.0.

[2] 10x Genomics. 2024. Visium HD Spatial Gene Expression Library, Human Lung Cancer (Fixed Frozen). https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-lung-cancer-fixed-frozen. HD Spatial Gene Expression dataset analyzed using Space Ranger 3.1.1.

[3] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering* 7 (2023), 756–779. doi:10.1038/s41551-023-01049-7

[4] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. 2017. QuPath: Open source software for digital pathology image analysis. *Scientific Reports* 7, 1 (2017), 16878. doi:10.1038/s41598-017-17204-5

[5] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. 2021. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods* 18, 11 (2021), 1352–1362. doi:10.1038/s41592-021-01264-7

[6] Bioptimus. 2024. H-optimus-0: Open-Source Foundation Model for Histology. Hugging Face model card. https://huggingface.co/bioptimus/H-optimus-0

[7] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. 2022. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology* 40, 4 (2022), 517–526. doi:10.1038/s41587-021-00830-w

[8] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* 25, 8 (2019), 1301–1309. doi:10.1038/s41591-019-0508-1

[9] Loïc Chadoutaud, Marvin Lerousseau, Daniel Herrero-Saboya, Julian Ostermaier, Jacqueline Fontugne, Emmanuel Barillot, Thomas Walter, et al. 2026. sCellST predicts single-cell gene expression from H&E images. *Nature Communications* 17 (2026), 1194. doi:10.1038/s41467-025-67965-1

[10] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Jing Yang, Wenlong Li, Jiarui Xu, Shijie Hao, et al. 2022. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 185, 10 (2022), 1777–1792.e21. doi:10.1016/j.cell.2022.04.003

[11] Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. 2024. STimage-1K4M: A Histopathology Image-Gene Expression Dataset for Spatial Transcriptomics. In *Advances in Neural Information Processing Systems*, Vol. 37. 35796–35823.

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

[13] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. 2022. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16144–16155.

[14] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, et al. 2024. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature Medicine* 30, 3 (2024), 850–862. doi:10.1038/s41591-024-02857-3

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607. https://proceedings.mlr.press/v119/chen20j.html

[16] Ozan Ciga, Tony Xu, and Anne Louise Martel. 2022. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications* 7 (2022), 100198. doi:10.1016/j.mlwa.2021.100198

[17] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. 2018. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* 24, 10 (2018), 1559–1567. doi:10.1038/s41591-018-0177-5

[18] CZI Cell Science Program. 2025. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research* 53, D1 (2025), D886–D900. doi:10.1093/nar/gkae1142

[19] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, Rani E. George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. 2021. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* 22, 1 (2021), 78. doi:10.1186/s13059-021-02286-2

[20] Ziheng Duan, Xi Li, Zhiqing Xiao, Rex Ying, and Jing Zhang. 2025. MUSE: A Multi-slice Joint Analysis Method for Spatial Transcriptomics Experiments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM, 625–634. doi:10.1145/3746252.3761240

[21] Ziheng Duan, Xi Li, Zhuoyang Zhang, James Song, and Jing Zhang. 2025. DISCO: A Diffusion Model for Spatial Transcriptomics Data Completion. In *2025 IEEE International Conference on Image Processing (ICIP)*. IEEE, 19–24. doi:10.1109/ICIP55913.2025.11084277

[22] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher J. Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 568, 7751 (2019), 235–239. doi:10.1038/s41586-019-1049-y

[23] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. 2023. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. *medRxiv* (2023), 2023.07.21.23292757. doi:10.1101/2023.07.21.23292757

[24] Xiaohang Fu, Yue Cao, Bin Bian, et al. 2025. Spatial gene expression at single-cell resolution from histology using deep learning with GHIST. *Nature Methods* 22 (2025), 1900–1910. doi:10.1038/s41592-025-02795-z

[25] Simon Graham, Quoc Dang Vu, Shan E. Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. 2019. HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* 58 (2019), 101563. doi:10.1016/j.media.2019.101563

[26] Noah F. Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J. McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, Cole Pavelchek, Sunny Cui, Isabella Camplisson, Omer Bar-Tal, Jaiveer Singh, Mara Fong, Gautam Chaudhry, Zion Abraham, Jackson Moseley, Shiri Warshawsky, Erin Soon, Shirley Greenbaum, Tyler Risom, Travis Hollmann, Sean C. Bendall, Leeat Keren, William Graf, Michael Angelo, and David Van Valen. 2022. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology* 40, 4 (2022), 555–565. doi:10.1038/s41587-021-01094-0

[27] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. 2020. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering* 4, 8 (2020), 827–834. doi:10.1038/s41551-020-0578-x

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[29] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. 2024. CellViT: Vision Transformers for precise cell segmentation and classification. *Medical Image Analysis* 94 (2024), 103143. doi:10.1016/j.media.2024.103143

[30] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. 2021. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods* 18, 11 (2021), 1342–1351. doi:10.1038/s41592-021-01255-8

[31] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. 2023. A visual-language foundation model for pathology image analysis using medical Twitter. *Nature Medicine* 29, 9 (2023), 2307–2316. doi:10.1038/s41591-023-02504-3

[32] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1M: One Million Image-Text Pairs for Histopathology. arXiv:2306.11207 [cs.CV] doi:10.48550/arXiv.2306.11207

[33] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, et al. 2023. High Resolution Mapping of the Tumor Microenvironment Using Integrated Single-Cell, Spatial and in Situ Analysis. *Nature Communications* 14 (2023), 8353. doi:10.1038/s41467-023-43458-x

[34] Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Perez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. 2024. HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis. In *Advances in Neural Information Processing Systems*, Vol. 37.

[35] Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. 2024. THItoGene: A Deep Learning Method for Predicting Spatial Transcriptomics from Histological Images. *Briefings in Bioinformatics* 25, 1 (2024), bbad464. doi:10.1093/bib/bbad464

[36] Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan Moritz Niehues, Kai A. J. Sommer, Peter Bankhead, et al. 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* 1, 8 (2020), 789–799. doi:10.1038/s43018-020-0087-6

[37] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html

[38] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. 2022. cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology* 40, 5 (2022), 661–671. doi:10.1038/s41587-021-01139-4

[39] Xi Li, Ziheng Duan, Yi Dai, Siwei Xu, and Jing Zhang. 2024. iHAST: Integrating Hybrid Attention for Super-Resolution in Spatial Transcriptomics. In *British Machine Vision Conference*. https://bmva-archive.org.uk/bmvc/2024/papers/

Paper_991/paper.pdf

[40] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. 2024. A Visual-Language Foundation Model for Computational Pathology. *Nature Medicine* 30, 3 (2024), 863–874. doi:10.1038/s41591-024-02856-4

[41] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. 2024. A multimodal generative AI copilot for human pathology. *Nature* 634, 8033 (2024), 466–473. doi:10.1038/s41586-024-07618-3

[42] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* 5, 6 (2021), 555–570. doi:10.1038/s41551-020-00682-w

[43] Kalin Nonchev, Glib Manaiev, Viktor H Koelzer, and Gunnar Rätsch. 2025. DeepSpot2Cell: Predicting Virtual Single-Cell Spatial Transcriptomics from H&E images using Spot-Level Supervision. *bioRxiv* (2025). doi:10.1101/2025.09.23.678121

[44] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. 2022. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods* 19, 2 (2022), 171–178. doi:10.1038/s41592-021-01358-2

[45] Minxing Pang, Kenong Su, and Mingyao Li. 2021. 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

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

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

[48] Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. 2017. The Human Cell Atlas. *eLife* 6 (2017), e27041. doi:10.7554/eLife.27041

[49] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. 2019. Slide-seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution. *Science* 363, 6434 (2019), 1463–1467. doi:10.1126/science.aaw1219

[50] Benoit Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maille, Julien Calderaro, Aurelie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. 2020. A Deep Learning Model to Predict RNA-Seq Expression of Tumours from Whole Slide Images. *Nature Communications* 11, 1 (2020), 3877. doi:10.1038/s41467-020-17678-4

[51] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. 2018. Cell Detection with Star-Convex Polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, Cham, 265–273. doi:10.1007/978-3-030-00934-2_30

[52] Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C. Strobl, Tessa E. Gillett, Luke Zappia, Elo Madissoon, Nikita S. Markov, Laure-Emmanuelle Zaragosi, Yan Ji, Meshal Ansari, et al. 2023. An integrated cell atlas of the lung in health and disease. *Nature Medicine* 29, 6 (2023), 1563–1577. doi:10.1038/s41591-023-02327-2

[53] Patrik L. Stahl, Fredrik Salmen, Sanja Vickovic, Anna Lundmark, Jose Fernandez Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Ake Borg, Fredrik Ponten, Paul Igor Costea, Pelin Sahlen, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisen. 2016. Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics. *Science* 353, 6294 (2016), 78–82. doi:10.1126/science.aaf2403

[54] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. 2021. Highly Sensitive Spatial Transcriptomics at Near-Cellular Resolution with Slide-seqV2. *Nature Biotechnology* 39, 3 (2021), 313–319. doi:10.1038/s41587-020-0739-1

[55] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. 2021. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* 18, 1 (2021), 100–106. doi:10.1038/s41592-020-01018-x

[56] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. 2024. PathGen-1.6M: 1.6 Million Pathology Image-text Pairs Generation through Multi-agent Collaboration. arXiv:2407.00203 [cs.CV] doi:10.48550/arXiv.2407.00203

[57] Xiao Tan, Andre Su, Mai Tran, and Quan Nguyen. 2020. SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics* 36, 7 (2020), 2293–2294. doi:10.1093/bioinformatics/btz914

[58] Sanja Vickovic, Gokcen Eraslan, Fredrik Salmen, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Aijo, Richard Bonneau, Ludvig Bergenstrahle, Jose Fernandez Navarro, Jeffrey Gould, Gareth K. Griffin, Ake Borg, Mostafa Ronaghi, Jonas Frisen, Joakim Lundeberg, Aviv Regev, and Patrik L. Stahl. 2019. High-Definition Spatial Transcriptomics for in Situ Tissue Profiling. *Nature Methods* 16, 10 (2019), 987–990. doi:10.1038/s41592-019-0548-y

[59] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. 2024. A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection. *Nature Medicine* 30 (2024), 2924–2935. doi:10.1038/s41591-024-03141-0

[60] Chuhan Wang, Adam S. Chan, Xiaohang Fu, Shila Ghazanfar, Jinman Kim, Ellis Patrick, Jean Y. H. Yang, et al. 2025. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications* 16 (2025), 1544. doi:10.1038/s41467-025-56618-y

[61] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361, 6400 (2018), eaat5691. doi:10.1126/science.aat5691

[62] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* 81 (2022), 102559. doi:10.1016/j.media.2022.102559

[63] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 634, 8035 (2024), 970–978. doi:10.1038/s41586-024-07894-z

[64] Xiao Xiao, Yan Kong, Ronghan Li, Zuoheng Wang, and Hui Lu. 2024. Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological image. *Medical Image Analysis* 91 (2024), 103040. doi:10.1016/j.media.2023.103040

[65] Ronald Xie, Kuan Pang, Sai W. Chung, Catia T. Perciani, Sonya A. MacParland, Bo Wang, and Gary D. Bader. 2023. Spatially Resolved Gene Expression Prediction from H&E Histology Images via Bi-modal Contrastive Learning. In *Advances in Neural Information Processing Systems*, Vol. 36.

[66] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier Gonzalez, Yu Gu, et al. 2024. A Whole-Slide Foundation Model for Digital Pathology from Real-World Data. *Nature* 630 (2024), 181–188. doi:10.1038/s41586-024-07441-w

[67] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. 2022. Spatial Transcriptomics Prediction from Histology Jointly through Transformer and Graph Neural Networks. *Briefings in Bioinformatics* 23, 5 (2022), bbac297. doi:10.1093/bib/bbac297

[68] Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle Y. Y. Lee, Kyung S. Cho, Katalin Susztak, George X. Xu, Michael D. Feldman, Edward B. Lee, Emma E. Furth, Linghua Wang, and Mingyao Li. 2024. Inferring Super-Resolution Tissue Architecture by Integrating Spatial Transcriptomics with Histology. *Nature Biotechnology* 42, 9 (2024), 1372–1377. doi:10.1038/s41587-023-02019-9

[69] Lijinghua Zhang, Michelle Yeasol Bang, and Hengrui Cai. 2026. Position: Prompting Intent Should Be Audited in LLM-Assisted Peer Review. In *Forty-third International Conference on Machine Learning Position Paper Track*. https://openreview.net/forum?id=HnB0EAQWYf

[70] Lijinghua Zhang and Hengrui Cai. 2025. Text Rationalization for Robust Causal Effect Estimation. *arXiv preprint arXiv:2512.05373* (2025).

[71] Wenbo Zhang, Lijinghua Zhang, Liner Xiang, and Hengrui Cai. 2026. Reasoning Is Not Free: Robust Adaptive Cost-Efficient Routing for LLM-as-a-Judge. In *Forty-third International Conference on Machine Learning*. https://openreview.net/forum?id=NoP2UQymHU

[72] Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. 2021. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology* 39, 11 (2021), 1375–1384. doi:10.1038/s41587-021-00935-2

[73] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. 2024. Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology. arXiv:2408.00738 [cs.CV]