CellHIST: A Cell-Centric Dataset and Benchmark for Weakly-Supervised Virtual Spatial Transcriptomics from Histopathology

Anonymous authors
Paper under double-blind review

Abstract

Spatial Transcriptomics (ST) at single-cell resolution has revolutionized our understanding of spatial biology. However, high experimental costs impede its widespread adoption, driving virtual ST from histopathology. Current approaches either require expensive single-cell labels for training or rely on paired spot labels at inference. An emerging paradigm, weakly-supervised learning (WSL), overcomes this by training a model on cost-effective spot-level ST data to infer single-cell spatial gene expression solely from images. Although offering a highly scalable alternative, a single-cell-level dataset and benchmark framework are still lacking. To bridge this gap, this paper presents CellHIST. It builds the first *cell-centric* dataset for virtual single-cell ST from histopathology. CellHIST-Data curates 169 samples across 14 organs, encompassing around 36 million cells and 1 million image patches, standardized to support model development across various paradigms. Furthermore, a comprehensive benchmark, CellHIST-Bench, is introduced to systematically evaluate weakly-supervised virtual single-cell ST in terms of predictive performance, data scaling behavior, single-cell prediction specificity, biological relevance, and downstream utility. Results uncover the practicability and limitations of WSL for this task. This cell-centric study can bridge an important gap and pave the way for histopathology-driven virtual single-cell ST.

1 Introduction

Spatial Transcriptomics (ST) technology enables *in situ* gene expression profiling, fundamentally transforming our understanding of spatial organization in biological systems (Ståhl et al., 2016; Chen et al., 2020; Marx, 2021). Early sequencing-based ST, *e.g.*, Visium, can only measure the overall gene expression within capture spots. Although they introduced spatial context into transcriptomics for the first time, they cannot directly test critical hypotheses in cell biology (Wu et al., 2021; Liu et al., 2026), as a spot typically contains multiple cells. For this reason, imaging-based ST, *e.g.*, Xenium, has emerged (Wang et al., 2025b). It enhances the resolution to single-cell, not only rendering previously intractable cell-level biological questions experimentally accessible (Gulati et al., 2025; Apostolov et al., 2026) but also laying the groundwork for building spatial virtual cell models (Bunne et al., 2024). Therefore, single-cell ST has gained popularity rapidly in recent years.

While advanced single-cell ST shows breakthrough capabilities, it often leads to high experimental costs. This motivates histopathology-driven virtual ST: cheap and clinically accessible histopathology images, which exhibit cell morphology and tissue phenotype at extremely high resolution, are leveraged to infer single-cell spatial gene expression (SGE) (Wang et al., 2025a; Chung et al., 2026). Two lines of research currently prevail for this task. (1) The first line is *fully-supervised learning* (FSL). It uses images paired with cell-level labels to train a model (Fu et al., 2025). Although straightforward, this approach relies heavily on expensive single-cell ST samples for training. (2) The second one is *super-resolution* (SR) (Zhang et al., 2024; 2025; Yang et al., 2025; Schroeder et al., 2025; Tang et al., 2026). This approach no longer relies on cell-level labels; instead, it takes the image and its spot-level label as input and infers single-cell SGE from the spot label using cellular histological features. Despite showing exciting progress, this approach still requires access to paired spot-level annotations at inference, limiting its applicability in broader scenarios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

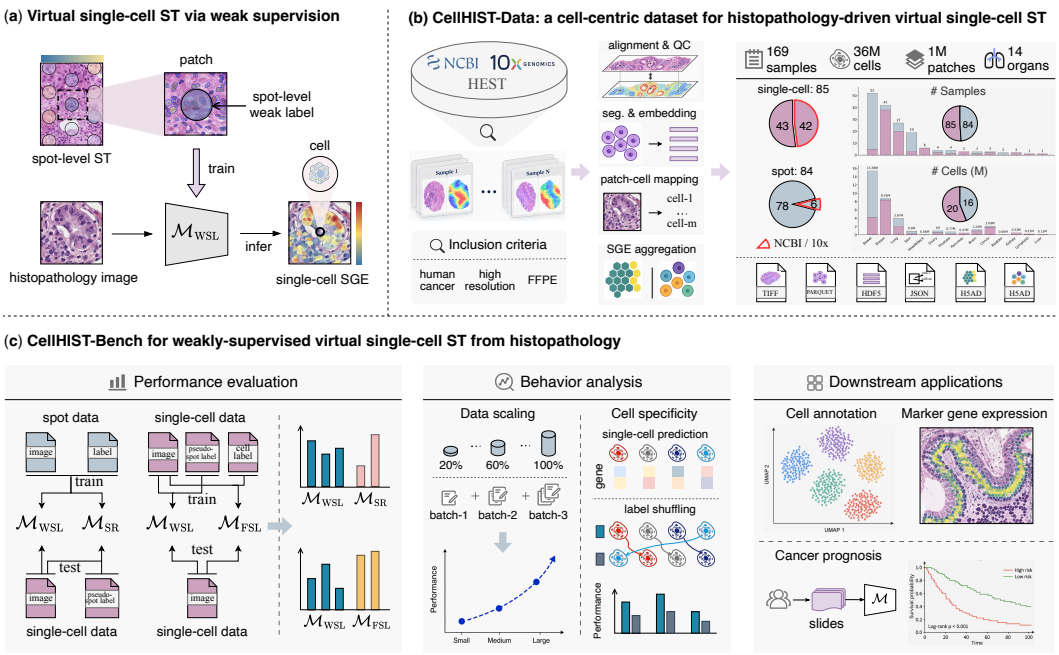

Figure 1: A diagram of weak supervision-based virtual ST and an overview of the proposed **CellHIST**. (**a**) An emerging WSL paradigm leverages cheaper, spot-level samples for training and infers single-cell SGE solely from images. (**b**) **CellHIST-Data**, the first cell-centric dataset for histopathology-driven virtual ST: it includes eligible samples from HEST (Jaume et al., 2024) and additionally collects 48 raw ST samples from public platforms, 42 of which are of single-cell resolution, with all samples processed mainly as described above to yield standardized files suitable for developing different cell-centric models. (**c**) **CellHIST-Bench** *i*) comprehensively evaluates WSL models against two mainstream paradigms, *ii*) conducts an in-depth analysis of WSL behaviors, and *iii*) examines the potential of virtual ST models in downstream single-cell applications.

To address these limitations, an emerging paradigm, *weakly-supervised learning* (WSL), has been explored recently and demonstrated promising performance in ST applications (Nonchev et al., 2025; Chadoutaud et al., 2026). As depicted in Figure 1(a), it leverages cheaper, weakly-annotated spot-level samples to train a WSL model that infers single-cell SGE solely from images, without relying on the spot label as a reference in prediction. Despite exhibiting appealing merits over previous approaches, many crucial questions remain to be systematically investigated for this new paradigm. For example, how does WSL perform compared with previous approaches? Can WSL models really infer cell-specific signals from images? Therefore, it is still uncertain whether WSL is a genuinely useful scheme. Moreover, existing public datasets are generally tailored for spot-resolution virtual ST (Stenbeck et al., 2021; Andersson et al., 2021; Chen et al., 2024a; Jaume et al., 2024). It is highly anticipated to develop a standardized dataset for virtual ST at single-cell resolution.

To bridge these gaps, this paper presents **CellHIST** (**Cell**-centric **HI**stopathology-driven virtual **ST**). It introduces *i*) the first *cell-centric* standardized dataset and *ii*) a comprehensive benchmark for the emerging WSL paradigm, as shown in Figure 1. (1) Concretely, **CellHIST-Data** comprises 169 samples covering 14 organs, with around 36 million cells and 1 million histopathology image patches. These samples are curated from public datasets and platforms and are processed mainly through image-ST alignment and quality control, nuclear segmentation and feature extraction, and single-cell SGE aggregation, ultimately yielding standardized data that supports the development of virtual single-cell ST models under various learning paradigms. (2) In **CellHIST-Bench**, we systematically evaluate WSL models in terms of predictive performance, data scaling behavior, cell specificity, and translational potential. The results uncover the performance of WSL models across five organs and their gap to previous approaches. Moreover, data scaling behavior, cell-specific expression pattern preservation, biological relevance, and downstream utility are observed in WSL models. We believe that this cell-centric work can bridge an important gap and lay a solid foundation for histopathology-

driven virtual single-cell ST. We will make our CellHIST-Data and CellHIST-Bench suites publicly available to facilitate future research.

2 Related work

Histopathology-driven virtual single-cell ST. H&E-stained histopathology images, also called whole-slide images (WSIs), are indispensable tools for cancer diagnosis. They are cost-effective to acquire clinically yet have extremely high resolution (*e.g.*, 0.25μm/pixel) and contain rich histology features, such as cellular organization and tissue phenotype (Chen et al., 2022; Song et al., 2023), that associate with specific SGE variations (Ståhl et al., 2016). Accordingly, histopathology images have been widely adopted to infer SGE for virtual ST. However, numerous earlier efforts generally focused on spot-level inference (Wang et al., 2025a; Chung et al., 2026). With the advent of high-resolution ST, single-cell-level studies have emerged. Among them, a WSL paradigm (Nonchev et al., 2025; Chadoutaud et al., 2026) shows appealing advantages over previous approaches. However, whether it is a practical solution remains to be systematically investigated. This paper bridges this gap by introducing a comprehensive benchmark.

Datasets and benchmarks for integrating ST with histopathology. Most existing datasets and benchmarks are designed for spot-level tasks, *e.g.*, STimage-1K4M (Chen et al., 2024a) and HEST-1K (Jaume et al., 2024). For any capture spot, a paired image patch and gene expression label are often provided for training spot-level predictive models (Ji et al., 2020; Stenbeck et al., 2021; Andersson et al., 2021; Wang et al., 2025a; Gindra et al., 2025; Chen et al., 2025; Liu et al., 2025). Although a few datasets contain single-cell-resolution ST samples, they often lack paired histopathology images (Tejada-Lapuerta et al., 2025; Zhao et al., 2025) or are processed as pseudo-spot-resolution data in benchmarking (Jaume et al., 2024). Consequently, there remains a substantial gap in the datasets and benchmarks for virtual single-cell ST.

Weakly-supervised learning approaches for computational pathology. WSL approaches, represented by multi-instance learning (MIL) (Dietterich et al., 1997; Carbonneau et al., 2018), have achieved considerable success in computational pathology in past years (Lu et al., 2023; Liu et al., 2024b). They are adopted to model gigapixel WSIs for slide-level or patch-level tasks: a WSI is divided into multiple patches; given a WSI and its overall label, *e.g.*, cancer or normal tissue, a model is trained by MIL and is required to yield slide-level or patch-level predictions, termed bag-level and instance-level MIL, respectively (Ilse et al., 2018). Most approaches are proposed for *bag-level* tasks, such as WSI classification (Lu et al., 2021; Li et al., 2021; Shao et al., 2021; Zhang et al., 2022; Shao et al., 2025) and survival prediction (Yao et al., 2020; Liu et al., 2024a), aiming at learning better bag representations. A few studies address *instance-level* inference. They train an instance-level estimator under weak supervision to predict whether the patch is cancerous or normal (Shi et al., 2020; Qu et al., 2022; Cui et al., 2023; Liu & Ji, 2024; Chen et al., 2024b). WSL-based virtual single-cell ST shares the same problem formulation. However, it faces distinct issues and challenges, as it is required to infer molecular states from histology.

3 Problem formulation

WSL-based single-cell SGE inference is illustrated in Figure 1(a). In this section, we provide its mathematical formulation, followed by a task analysis in the context of MIL.

Data. Let $\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^{N}$ denote weakly-annotated data from spot-resolution ST samples, where $\boldsymbol{X}_i$ is the histopathology image associated with a spot s_i, $\boldsymbol{Y}_i \in \mathbb{R}^k$ is the spot-level label of s_i, and k is the number of measured genes. $\boldsymbol{X}_i$ is usually generated by cropping a patch from a high-resolution WSI, where s_i is located at the center of $\boldsymbol{X}_i$ and is completely encompassed by $\boldsymbol{X}_i$. Furthermore, we assume there are m_i cells in $\boldsymbol{X}_i$. We write them as $\{c_{i,j} \mid j \in [1, m_i]\}$ and use $I_{i,j} \in \{0, 1\}$ to indicate whether the cell $c_{i,j}$ falls within s_i. By definition, there is $\boldsymbol{Y}_i = \sum_{j \in [1,m_i]} I_{i,j} \boldsymbol{y}_{i,j}$, where $\boldsymbol{y}_{i,j}$ is the gene expression of $c_{i,j}$ but it is *not* accessible in WSL-based model training.

Model. Let $\mathcal{M}(c_{i,j} \mid \boldsymbol{X}_i)$ be a model that infers the SGE of $c_{i,j}$ from $\boldsymbol{X}_i$, expressed by

$$\hat{\boldsymbol{y}}_{i,j} = \mathcal{M}(c_{i,j} \mid \boldsymbol{X}_i). \tag{1}$$

Since only the spot-level weak label $\boldsymbol{Y}_i$ is available in training, single-cell-level predictions need to be aggregated to obtain a mean-based spot-level estimation for optimization, *i.e.*,

$$\hat{\boldsymbol{Y}}_i = \frac{1}{\sum_j I_{i,j}} \sum_j I_{i,j} \hat{\boldsymbol{y}}_{i,j}. \tag{2}$$

$\hat{\boldsymbol{Y}}_i$ is optimized through a learning objective $\mathcal{L}_{\text{wsl}} = \mathcal{L}(\hat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i)$.

Framed as MIL. Taking $c_{i,j}$ as an instance and $\{c_{i,j} \mid j \in [1, m_i], I_{i,j} = 1\}$ as a bag, the above task can be framed as an *instance-level* MIL problem (Ilse et al., 2018). It requires constructing and training an instance-level estimator under the weak supervision of bag-level labels. Conventional MIL often assumes that instances are *i.i.d.* Under this assumption, each cell is encoded independently from its neighbors; cell-level predictions are aggregated into a spot-level prediction for optimization, as defined in Equation 2. However, the SGE of single cells can be affected by cell-to-cell interactions and tissue microenvironment (Fu et al., 2025). In this case, context-aware or correlated MIL architectures, which explicitly model instance dependencies via self-attention or graph convolution (Tu et al., 2019; Shao et al., 2021), could be introduced to improve WSL models.

4 CELLHIST-DATA

We introduce the first *cell-centric* dataset, CellHIST-Data, for histopathology-driven virtual single-cell ST. It covers both spot-resolution and single-cell-resolution standardized data, supporting the development of virtual single-cell ST models under various learning paradigms. Concretely, as shown in Figure 1(b), we construct CellHIST-Data by first collecting publicly available ST samples from large-scale data repositories (10×Genomics and NCBI) and an existing resource HEST-1K (Jaume et al., 2024), followed by removing duplicates. To enable more reliable cell-level prediction from histopathology, we retain human cancer tissue samples with paired H&E-stained images at a resolution below 0.75μm/pixel and exclude frozen fixed slides following Chadoutaud et al. (2026). Then, we align images with ST for samples without alignment files, perform nuclear segmentation and feature extraction on all images, and derive single-cell and spot labels from imaging-based ST samples. This pipeline yields 169 samples, where half of the single-cell ST samples are newly collected. Additional dataset details are provided in Appendix A.1.

Dataset overview. CellHIST-Data provides the following essential data. For **histopathology images**, we detect nuclei and extract nuclear morphology embeddings (§4.1), which serve as a foundation for cell-level tasks. Moreover, we process all ultra-large images into spot-centered patches (§4.2) to enable feasible and efficient patch-based training. For **spatial gene expression labels**, we prepare native spot labels and construct pseudo-spot labels from the ST samples without explicit spots to provide available weak labels for all samples, while aggregating captured mRNA transcripts into single-cell-level labels for model evaluation (§4.3). We elaborate the details below.

4.1 NUCLEAR SEGMENTATION AND CELL EMBEDDING

Nuclear segmentation. For each H&E-stained image, we adopt CellViT++ (Hörst et al., 2025), pretrained on PanNuke (Gamper et al., 2019; 2020), for nuclear segmentation, owing to its strong performance and computational efficiency on ultra-large WSIs. The outputs are stored as a PARQUET file for each sample, where each segmented cell is provided with nucleus identifier (*cell_id*), geometry (*geometry*), and predicted cell class (*class*). There are approximately 36 million nuclei detected, averaging about 210K per sample. Since segmentation quality is crucial for this task, we provide a comprehensive quality analysis of cell segmentation in Appendix A.3.

Cell embedding. Besides nuclear segmentation, we employ CellViT++ as a cell feature encoder to extract nuclear representations, as nuclear morphology provides valuable cues for single-cell SGE inference (Fu et al., 2025; Chadoutaud et al., 2026). Specifically, we follow the default settings in CellViT++: for each detected nucleus, its feature map from the final ViT layer is pooled to obtain a 1280-dimensional vector as a cell embedding, where the pretrained SAM-H (Kirillov et al., 2023) is the network backbone. For each sample, the embeddings and coordinates of all detected nuclei are stored in an HDF5 file under the key *cell_embedding* and *cell_coords*, respectively.

4.2 IMAGE PATCHING AND CELL RE-IDENTIFICATION

Image patching. Following HEST-1K, for each given spot, we crop a 224×224 patch from images at a resolution of 0.5μm/pixel. For the ST samples without explicit spots, we manually construct pseudo-spots to provide available patch data for all samples and enable patch-based training. Specifically, we strictly follow Visium settings: circular spots with a diameter of 55μm, arranged on a hexagonal grid with 100μm spacing between spots. Each pseudo-spot is assigned a unique barcode. The barcodes and 2D coordinates of all spots are stored in an HDF5 file. These steps provide a unified protocol for spot and patch setting across different ST platforms.

Re-identification of cells in patches. Since original images are organized as spot-centered patches, we re-identify the detected cells within each patch by *i*) recording their index in the above cell embedding data and *ii*) measuring whether a cell falls within this spot. This cell re-identification step yields a JSON file for each sample. In this dictionary-formatted data, each key is a spot barcode, and each value contains the indices and in-spot indicators of all detected cells within the patch. Using this file, the model can efficiently retrieve cell embeddings and coordinates for any given spot.

4.3 SPATIAL GENE EXPRESSION

Spot-resolution SGE. For ST samples with native spots, we directly use the provided spot-level SGE with spot barcodes as indices. For samples without native spots, *e.g.*, Xenium ST, we construct pseudo-spots as described in Section 4.2 and then derive spot labels by aggregating the captured mRNA transcripts within each pseudo-spot. This step provides each sample with available spot labels, so single-cell-resolution samples can also be treated as spot-level data for WSL.

Single-cell-resolution SGE. For imaging-based ST samples, we obtain single-cell-level labels using the original transcript assignments and cell boundaries offered by ST platforms. We do not generate single-cell labels for bin-resolution ST samples without reliable cell boundaries, *e.g.*, Visium HD, since boundary estimates may deviate from ground truth (Polański et al., 2024; Blampey et al., 2024). Accordingly, Visium HD samples are viewed as spot-resolution data. Moreover, for imaging-based ST samples, we match each segmented nucleus to its corresponding platform-defined cell and store the matching result as *st_cell_id* in a PARQUET file that shares the row index with the PARQUET file generated in cell embedding. We report cell-matching quality and coverage in Appendix A.4 and examine the sensitivity of models to cell-matching confidence in Appendix A.5.

5 CELLHIST-BENCH

We present CellHIST-Bench to systematically evaluate the emerging WSL paradigm for single-cell SGE inference from histopathology. It mainly addresses the following understudied questions. First, how do WSL models compare with super-resolution (§5.2) and fully-supervised learning (§5.3) under comparable settings? Second, will WSL models perform better as the size of weakly-annotated data increases (§5.4)? Third, do WSL models really infer cell-specific signals (§5.5)? Last, we examine the biological relevance of these models in downstream ST applications (§5.6).

5.1 EXPERIMENTAL SETUP

Datasets. We define five organ-specific tasks: Bowel, Breast, Lung, Ovary, and Skin. For each task, its dataset comprises both spot-resolution and single-cell-resolution samples from CellHIST-Data. As described in Section 4, for spot-resolution samples, each spot-centered patch has a spot label; for single-cell-resolution ones, each patch carries both spot-level and single-cell-level labels. This enables *comparable* data settings in benchmarking models with different learning paradigms. (1) For the comparison between WSL and super-resolution, models are trained and evaluated on the same spot-resolution and single-cell-resolution data, respectively. We call this data setting **standard-spot**, as WSL uses spot labels from spot-resolution data. (2) For the comparison between WSL and fully-supervised models, the identical single-cell-resolution data are used by them. The only difference lies in the label granularity for training. We call this setting **pseudo-spot**, as WSL uses pseudo-spot labels from single-cell-resolution data. Refer to Appendix B.2 for more details.

Baselines. We include **sCellST** (Chadoutaud et al., 2026) and **DeepSpot2Cell** (Nonchev et al., 2025), two WSL approaches recently proposed for this task. In addition, we adapt prevalent and compatible MIL approaches for comparison. They cover classical MIL assumptions and various schemes for modeling instance correlations. (1) Attention-based MIL (Ilse et al., 2018), *i.e.*, **ABMIL**, is a frequently-adopted method for bag-level prediction. We adapt it for single-cell prediction by multiplying its spot-level prediction by cell-level gene-wise attention scores. A gated version, **ABMIL-gated**, is also included. (2) Graph-based MIL (Tu et al., 2019), termed **GraphMIL**, which treats instances as nodes and leverages a graph to model instance correlations. We adapt it by constructing a graph within each spot and using node-level features after graph convolution for single-cell prediction. **MS-GraphMIL**, a multi-scale variant that uses node features from different graph convolution layers, is also compared. (3) **Transformer** (Vaswani et al., 2017; Wagner et al., 2023), originally designed for sequences, is adapted by utilizing token features after self-attention for single-cell prediction. Note that these baselines only use spot-level labels in training.

Training and evaluation details. We reproduce all baselines following their official implementations, retaining method-specific training settings where applicable. For sCellST, graph-based models, and Transformer, cell-level predictions are mean-pooled within each spot to obtain spot-level targets for optimization. In particular, ABMIL and ABMIL-gated use gene-specific attention weights for cell prediction, as stated above. If not specified, the model we evaluate is the one trained under the standard-spot data setting. For evaluation metrics, we adopt gene-wise Pearson correlation coefficients (PCC). We repeat each experiment three times and report the mean and standard deviation across three runs. For more relevant details, please refer to Appendix B.3.

5.2 Evaluating weakly-supervised models for single-cell SGE inference

We first evaluate WSL methods under the standard-spot data setting, where the spot-resolution data from CellHIST-Data are used for training. This setting is also applicable to super-resolution methods. Therefore, we also evaluate a recent state-of-the-art super-resolution model, Thor (Zhang et al., 2025), for comparison, and show its performance on the same test data for reference.

Table 1: **Evaluating weakly-supervised models for single-cell SGE inference from histopathology.** † Reference only, as it requires spot labels at inference. Best performance in **bold**. Mean and standard deviation are reported over three runs.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
- Super-resolution reference						
Thor †	0.490	0.353	0.258	0.295	0.458	0.371
- Weak supervision baselines						
ABMIL	0.288 ± 0.010	0.144 ± 0.004	$\mathbf{0.237} \pm 0.002$	0.074 ± 0.001	0.014 ± 0.006	0.151
ABMIL-gated	0.289 ± 0.008	0.152 ± 0.008	0.236 ± 0.009	0.075 ± 0.010	0.016 ± 0.001	0.154
GraphMIL	0.348 ± 0.004	0.122 ± 0.012	0.217 ± 0.004	0.059 ± 0.005	0.018 ± 0.014	0.153
MS-GraphMIL	0.267 ± 0.005	0.167 ± 0.016	0.103 ± 0.009	$\mathbf{0.136} \pm 0.005$	$\mathbf{0.051} \pm 0.023$	0.145
Transformer	0.325 ± 0.003	0.128 ± 0.010	0.212 ± 0.002	0.068 ± 0.005	0.012 ± 0.014	0.149
DeepSpot2Cell	0.335 ± 0.002	0.122 ± 0.007	0.190 ± 0.001	0.061 ± 0.010	0.049 ± 0.005	0.151
sCellST	$\mathbf{0.363} \pm 0.003$	$\mathbf{0.171} \pm 0.007$	0.229 ± 0.004	0.097 ± 0.008	0.019 ± 0.008	**0.176**

As shown in Table 1, sCellST obtains the highest overall PCC in WSL baselines. However, it leaves a notable gap to the super-resolution method, with a drop of 52.6% in overall PCC. Besides, MS-GraphMIL performs significantly better than others on Ovary and Skin. All WSL methods perform best on Bowel and worst on Skin, reflecting intrinsic heterogeneity across organs.

Our further analysis is as follows. (1) **Cross-cancer generalization challenges**: all WSL methods perform poorly on Skin, since training and test samples represent different cancer types, squamous cell carcinoma and melanoma, respectively. This reveals the challenges of cross-cancer generalization for this task. (2) **Tissue-dependent benefits of correlation modeling**: cell-to-cell correlation modeling performs best in Ovary and Skin, whereas instance-independent MIL leads in others, indicating inconsistent advantages from explicit correlation modeling across different tissue origins.

5.3 COMPARISON BETWEEN WEAKLY-SUPERVISED AND FULLY-SUPERVISED MODELS

We show the performance gap between WSL and fully-supervised learning to assess the potential of WSL in single-cell SGE inference. For this purpose, we adopt the pseudo-spot data setting. Specifically, we use the same single-cell training cohort and identical patient-level train-test splits (k-fold) for two learning paradigms. This enables us to evaluate the competitiveness of weak supervision with full supervision when the data discrepancy is controlled. We compare three fully-supervised learning baselines, HisToGene (Pang et al., 2021) and ST-Net (He et al., 2020) adapted for this task, and the recently proposed GHIST (Fu et al., 2025).

Table 2: **Comparison with fully-supervised models.** The pseudo-spot data setting is applied for head-to-head comparisons, *i.e.*, two different models use the same single-cell-resolution data in training and test. Best performance in **bold**. Standard deviation is reported over three runs.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
- Weak supervision baselines w/ pseudo-spot labels						
ABMIL	**0.388** $\pm$ 0.003	0.244 $\pm$ 0.003	0.175 $\pm$ 0.000	0.203 $\pm$ 0.001	0.074 $\pm$ 0.008	0.217
ABMIL-gated	0.387 $\pm$ 0.001	0.234 $\pm$ 0.003	0.175 $\pm$ 0.002	**0.206** $\pm$ 0.001	0.077 $\pm$ 0.018	0.216
GraphMIL	0.372 $\pm$ 0.003	**0.288** $\pm$ 0.001	0.178 $\pm$ 0.001	0.187 $\pm$ 0.009	0.083 $\pm$ 0.020	0.222
MS-GraphMIL	0.294 $\pm$ 0.001	0.279 $\pm$ 0.004	0.124 $\pm$ 0.000	**0.206** $\pm$ 0.002	**0.243** $\pm$ 0.008	0.229
Transformer	0.357 $\pm$ 0.006	0.285 $\pm$ 0.003	**0.187** $\pm$ 0.001	0.201 $\pm$ 0.006	0.146 $\pm$ 0.034	**0.235**
DeepSpot2Cell	0.369 $\pm$ 0.001	0.237 $\pm$ 0.002	0.178 $\pm$ 0.002	0.153 $\pm$ 0.006	0.163 $\pm$ 0.001	0.220
sCellST	0.382 $\pm$ 0.002	0.271 $\pm$ 0.000	0.176 $\pm$ 0.002	0.205 $\pm$ 0.005	0.086 $\pm$ 0.008	0.224
- Full supervision baselines w/ single-cell labels						
GHIST	0.409 $\pm$ 0.011	0.342 $\pm$ 0.002	0.105 $\pm$ 0.005	0.224 $\pm$ 0.009	**0.163** $\pm$ 0.030	0.249
ST-Net	0.468 $\pm$ 0.000	0.347 $\pm$ 0.000	0.227 $\pm$ 0.000	0.266 $\pm$ 0.001	0.061 $\pm$ 0.000	0.274
HisToGene	**0.495** $\pm$ 0.001	**0.389** $\pm$ 0.001	**0.255** $\pm$ 0.001	**0.271** $\pm$ 0.002	0.128 $\pm$ 0.004	**0.308**

We present this benchmarking result in Table 2. Our main observations are as follows. (1) **Objective performance gap**: the best fully-supervised method outperforms the best WSL baseline by 31% in terms of average PCC across five organs. Moreover, on four organs other than Skin, HisToGene performs better than the best WSL method, with an average improvement of 33% in PCC. (2) **Competitiveness of weak supervision**: on Skin, two WSL methods, MS-GraphMIL and DeepSpot2Cell, are comparable to or even better than the best fully-supervised model, GHIST. We note that, although weak supervision is often worse than full supervision, it could still be a competent candidate because it leverages much cheaper ST data yet obtains a modest overall performance, a PCC of 0.235. We will further assess the translational potential of these WSL models in Section 5.6.

5.4 EMPIRICAL STUDIES ON DATA SCALING

Data scaling tests are critical for developing better deep learning models (Kaplan et al., 2020; Jaume et al., 2024). Consequently, in this section, we adopt sCellST and investigate the models in two data expansion settings: *i*) the slide fraction increased from 20% to 100% (8 to 40 slides), and *ii*) the number of data sources from 1 to 7, with other configurations kept unchanged.

As shown in Figure 2, PCCs generally increase with the expansion of spot-level training samples. This trend persists across both experimental settings, underscoring the potential of leveraging larger and more diverse spot ST data to improve WSL models for this task.

5.5 ANALYZING THE SPECIFICITY OF SINGLE-CELL PREDICTIONS

To assess whether the prediction of WSL models captures cell-specific expression differences within individual spots, we perform a cell label shuffling test on Bowel. For each test spot, we randomly permute the labels of cells within this spot while keeping the predictions fixed. This preserves the expression profiles within each spot while disrupting their correspondence with individual cells.

Figure 3 shows substantial and consistent performance drops across different WSL methods and data settings. Specifically, shuffling decreases PCC by approximately 12.0% and 13.4% on average

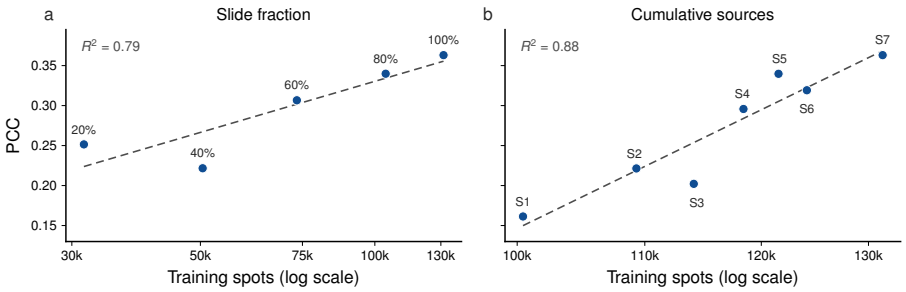

Figure 2: **Data scaling studies on weakly-supervised models.** We expand the spot-resolution training data cumulatively by two strategies: (a) from 8 to 40 slides and (b) from 1 to 7 data sources in Bowel. Dashed lines show the result of log-transformed linear regression.

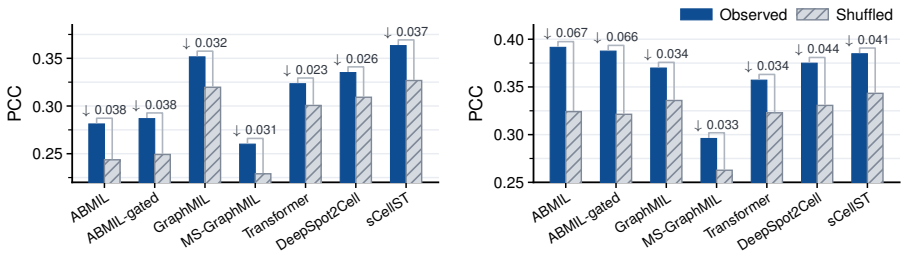

Figure 3: **Cell label shuffling within spots**: standard-spot (left) and pseudo-spot (right) data settings.

in standard-spot and pseudo-spot data settings, respectively. These drops suggest the presence of cell-specific signals in WSL models. In other words, WSL could often capture expression differences among cells within the same spot. Additional results are provided in Appendix C.4.

5.6 Downstream applications in virtual single-cell spatial transcriptomics

Beyond conventional quantitative metrics, we further test whether WSL-based single-cell SGE predictions retain the biologically relevant signals supporting downstream ST analyses. We compare WSL models with fully-supervised ones and consider three representative use cases: cell annotation, marker-gene expression visualization, and cancer survival analysis.

Cell annotation. Following Chadoutaud et al. (2026), we obtain cell types from inferred single-cell expression profiles on a held-out breast cancer sample. Figure 4 presents the result of a malignant-enriched region, ROI (region of interest) a. It is found that DeepSpot2Cell recovers major cell populations verified by the measured SGE. Nevertheless, discrepancies remain in local cell-type assignments and closely-related cell populations. We provide the results for two additional ROIs, a malignant-fibroblast interface and an immune-enriched region, in Appendix C.5.

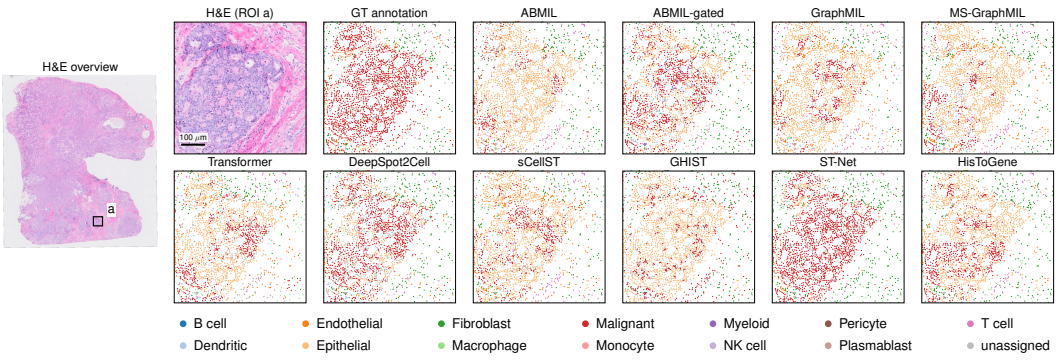

Figure 4: **Cell-type annotation at a malignant-enriched region.** Regional cell-type maps are derived from ground truth (GT) and predicted SGE in a held-out breast cancer sample.

Marker-gene expression analysis. We further visualize ERBB2, a well-established cancer marker gene, to inspect whether WSL models can infer biologically relevant spatial expression from histopathology for important genes. As visualized in Figure 5, the predicted maps retain part of the spatial heterogeneity observed in ground truth, including locally enriched and depleted regions, although the discrepancies in local details remain. These results suggest that WSL-based single-cell SGE from histopathology could provide an exploratory *in silico* view of marker-gene expression.

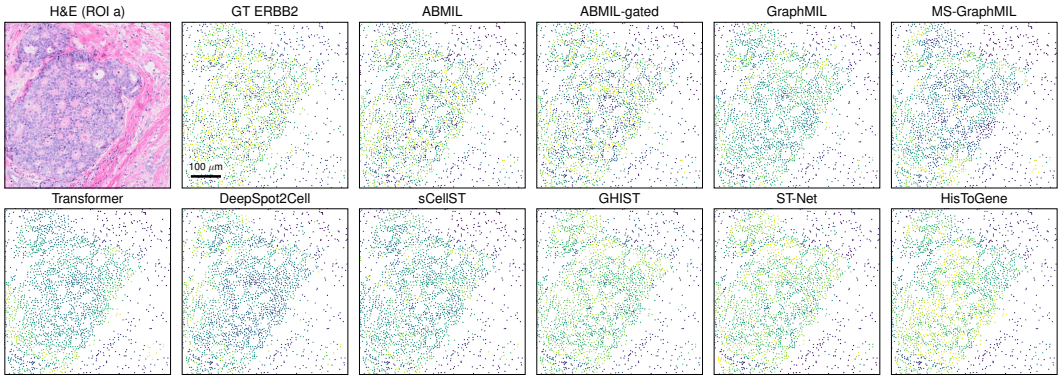

Figure 5: **Spatial gene expression (ERBB2) at single-cell resolution.** Ground truth and predictions are transformed into percentiles for visualization, from purple (low) to yellow (high).

Cancer survival analysis. We explore whether predicted SGE enables patient-level prognostic stratification. Following Fu et al. (2025), we use a cohort of 92 TCGA-BRCA patients with nine observed deaths and infer single-cell SGE from the H&E-stained tissue slides of these patients. Then, we aggregate single-cell predictions into patient-level expression profiles and fit penalized Cox regression models using selected genes. Figure 6 exhibits the Kaplan-Meier curves for risk-stratified patients. It shows the translational potential of inferred SGE ($P < 0.01$) in cancer prognosis.

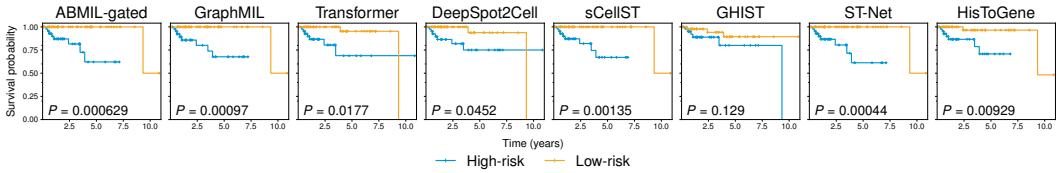

Figure 6: **Cancer survival analysis with gene expression inferred from histopathology slides.** Kaplan–Meier curves for TCGA-BRCA patients stratified into high-risk and low-risk groups according to the median risk score estimated by the multivariate Cox regression model for each method.

6 Conclusion

This paper presents CellHIST, a cell-centric study of histopathology-driven virtual single-cell ST. We construct CellHIST-Data, the first standardized cell-centric dataset for this task, comprising 169 samples across 14 organs with around 36 million cells and 1 million image patches. Using this dataset, we establish CellHIST-Bench to systematically evaluate the emerging WSL paradigm under a unified setting. Through extensive experiments, we show that WSL often recovers meaningful single-cell SGE from histopathology without requiring paired single-cell supervision, and that its performance can be further improved by scaling weakly-annotated spot-level data. Beyond predictive accuracy, the cell label shuffling tests and downstream analyses provide evidence that WSL models can capture cell-specific and biologically relevant signals. Moreover, our benchmark reveals an objective gap between weak and full supervision, as well as remaining challenges in cross-cancer generalization and robust single-cell inference. Together, these findings suggest that large-scale WSL could provide a promising route toward scalable virtual single-cell ST, while substantial room remains for improving biological fidelity and generalizability. CellHIST can pave the way for future research on more accurate, scalable, and biologically grounded virtual single-cell ST models.

AI USE STATEMENT

In this work, we used large language models (LLMs) solely to assist with language editing, including correcting grammar, refining phrasing, and improving the clarity and readability of the manuscript. We did not use generative AI tools for other research tasks requiring disclosure under the ICLR 2027 AI Policy for Authors. All AI-assisted edits were reviewed and revised by the authors to ensure that they preserved the intended technical meaning and did not introduce unsupported claims or inaccuracies. We take full responsibility for the final content of this work, including all AI-assisted text.

ETHICS STATEMENT

CellHIST is a dataset and benchmark for histopathology-driven virtual single-cell spatial transcriptomics, constructed from existing public resources, including **HEST-1K, 10x Genomics, and NCBI**. The downstream survival analysis uses existing **TCGA-BRCA** data. Use and redistribution of these human tissue-derived data should respect the original data providers' usage conditions and applicable privacy protections. Our experiments reveal variation in predictive performance across organs and limitations in cross-cancer generalization, which warrant caution when extending the findings to other tissues or patient populations. The cell annotation, marker-gene visualization, and survival analyses are exploratory evaluations and do not establish clinical validity. CellHIST is intended to support research and methodological development; its predictions should not be used as a substitute for experimental measurements or as a basis for clinical decisions without further independent validation.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 formalizes the prediction task and the weak-supervision framework. Section 4 details the data sources, inclusion criteria, and dataset construction pipeline, including nuclear segmentation, cell embedding extraction, image patching, cell re-identification, and the preparation of spot-level and single-cell gene expression labels. Section 5.1 describes the benchmark settings, baseline adaptations, evaluation metric, and reporting protocol across repeated runs. Section 5.3 specifies the shared patient-level train–test splits used to compare weakly and fully supervised models. Sections 5.4–5.6 describe the data scaling experiments, within-spot label shuffling tests, and downstream analyses. The anonymous project page is available at `https://anonymous.4open.science/r/CellHist_Bench-7BDF/`. For anonymous review, the raw and processed data used in CellHIST-Data are also available at `https://anonymous-hf.com/a/qzg0wl2p83tq/`.

REFERENCES

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1):6012, 2021.

Eva Apostolov, Daniel L Roden, Holly Holliday, Aurélie Cazet, Kate Harvey, Hanyun Zhang, Sunny Z Wu, Sophie van der Leij, Hani Jieun Kim, Luke A Selth, et al. Single-cell and spatial transcriptomic profiling reveals epithelial functional states and fibroblast phenotypes in hormone therapy–naïve localized prostate cancer. *Cancer Research*, 86(8):1836–1853, 2026.

Quentin Blampey, Kevin Mulder, Margaux Gardet, Stergios Christodoulidis, Charles-Antoine Dutertre, Fabrice André, Florent Ginhoux, and Paul-Henry Cournède. Sopa: a technology-invariant pipeline for analyses of image-based spatial omics. *Nature Communications*, 15(1):4981, 2024.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353, 2018.

Loïc Chadoutaud, Marvin Lerousseau, Daniel Herrero-Saboya, Julian Ostermaier, Jacqueline Fontugne, Emmanuel Barillot, and Thomas Walter. scellst predicts single-cell gene expression from h& e images. *Nature Communications*, 17(1):1194, 2026.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 35796–35823. Curran Associates, Inc., 2024a.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16144–16155, 2022.

Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, et al. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991, 2020.

Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, et al. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, 2025.

Yihang Chen, Tsai Hor Chan, Guosheng Yin, Yuming Jiang, and Lequan Yu. cdp-mil: Robust multiple instance learning via cascaded dirichlet process. In *European conference on computer vision*, pp. 232–250. Springer, 2024b.

Youngmin Chung, Ji Hun Ha, Andrew H Song, Cristina Almagro-Pérez, Chaeyoung Seo, Won Jun Suh, Jeong Won Beom, Kyoung Bin Oh, Eytan Ruppin, Faisal Mahmood, et al. Stp-bench: A unified systematic benchmark for virtual spatial transcriptomics from histopathology images. *arXiv preprint arXiv:2609.05956*, 2026.

Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *The Eleventh International Conference on Learning Representations*, 2023.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Xiaohang Fu, Yue Cao, Beilei Bian, Chuhan Wang, Dinny Graham, Nirmala Pathmanathan, Ellis Patrick, Jinman Kim, and Jean Yee Hwa Yang. Spatial gene expression at single-cell resolution from histology using deep learning with ghist. *Nature methods*, 22(9):1900–1910, 2025.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European congress on digital pathology*, pp. 11–19. Springer, 2019.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.

Rushin H Gindra, Giovanni Palla, Mathias Nguyen, Sophia J Wagner, Manuel Tran, Fabian J Theis, Dieter Saur, Lorin Crawford, and Tingying Peng. A large-scale benchmark of cross-modal learning for histology and gene expression in spatial transcriptomics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1182–1192, 2025.

Gunsagar S Gulati, Jeremy Philip D'Silva, Yunhe Liu, Linghua Wang, and Aaron M Newman. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nature reviews Molecular cell biology*, 26(1):11–31, 2025.

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.

Fabian Hörst, Moritz Rempe, Helmut Becker, Lukas Heine, Julius Keyl, and Jens Kleesiek. Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models, 2025. URL https://arxiv.org/abs/2501.05269.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

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. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, 2024.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021.

Junzhuo Liu, Markus Eckstein, Zhixiang Wang, Friedrich Feuerhake, and Dorit Merhof. Spatial transcriptomics expression prediction from histopathology based on cross-modal mask reconstruction and contrastive learning. *Medical Image Analysis*, pp. 103889, 2025.

Pei Liu and Luping Ji. Weakly-supervised residual evidential learning for multi-instance uncertainty estimation. In *International Conference on Machine Learning*, pp. 31262–31292. PMLR, 2024.

Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91:103020, 2024a.

Pei Liu, Luping Ji, Xinyu Zhang, and Feng Ye. Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. *IEEE Transactions on Medical Imaging*, 43(5): 1841–1852, 2024b.

Yunhe Liu, Yibo Dai, and Linghua Wang. Spatial omics at the forefront: emerging technologies, analytical innovations, and clinical applications. *Cancer cell*, 44(1):24–49, 2026.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19764–19775, 2023.

Xin Luo, Yicheng Tao, Haoxuan Zeng, Suyuan Wang, Chenzi Ouyang, Meiqi Zhu, Kai Liu, Shuibing Chen, and Jie Liu. Voice: A vision-omics foundation model integrating direct and retrieval-based prediction of in-situ single-cell gene expression. *arXiv preprint arXiv:2608.08366*, 2026.

Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

Kalin Nonchev, Glib Manaiev, Viktor H Koelzer, and Gunnar Rätsch. Deepspot2cell: predicting virtual single-cell spatial transcriptomics from h&e images using spot-level supervision. *bioRxiv*, pp. 2025–09, 2025.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pp. 2021–11, 2021.

Krzysztof Polański, Raquel Bartolomé-Casado, Ioannis Sarropoulos, Chuan Xu, Nick England, Frode L Jahnsen, Sarah A Teichmann, and Nadav Yayon. Bin2cell reconstructs cells from high resolution visium hd data. *Bioinformatics*, 40(9):btae546, 2024.

Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022.

Amelia Schroeder, Melanie L Loth, Chunyu Luo, Sicong Yao, Hanying Yan, Daiwei Zhang, Sarbottam Piya, Edward Plowey, Wenxing Hu, Jean R Clemenceau, et al. Scaling up spatial transcriptomics for large-sized tissues: uncovering cellular-level tissue architecture beyond conventional platforms with iscale. *Nature methods*, 22(9):1911–1922, 2025.

Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? In *International Conference on Machine Learning*, pp. 54219–54238. PMLR, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5742–5749, 2020.

Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016.

Linnea Stenbeck, Ludvig Bergenstråhle, Joakim Lundeberg, and Åke Borg. Human breast cancer in situ capturing transcriptomics. *Mendeley Data*, 2, 2021.

Junjie Tang, Zihao Chen, Kun Qian, Siyuan Huang, Yang He, Shenyi Yin, Xinyu He, Buqing Ye, Yan Zhuang, Hongxue Meng, et al. The interpretable multimodal dimension reduction framework spahdmap enhances resolution in spatial transcriptomics. *Nature Cell Biology*, pp. 1–15, 2026.

Alejandro Tejada-Lapuerta, Anna C Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, et al. Nicheformer: a foundation model for single-cell and spatial omics. *Nature methods*, pp. 1–14, 2025.

Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, Piet A van den Brandt, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer cell*, 41(9):1650–1661, 2023.

Chuhan Wang, Adam S Chan, Xiaohang Fu, Shila Ghazanfar, Jinman Kim, Ellis Patrick, and Jean YH Yang. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16(1):1544, 2025a.

Huan Wang, Ruixu Huang, Jack Nelson, Ce Gao, Miles Tran, Anna Yeaton, Sachi Krishna, Kristen Felt, Kathleen L Pfaff, Teri Bowman, et al. Systematic benchmarking of imaging spatial transcriptomics platforms in ffpe tissues. *Nature Communications*, 16(1):10215, 2025b.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021.

Jiyuan Yang, Ziqian Zheng, Yun Jiao, Kaiwen Yu, Sheetal Bhatara, Xu Yang, Sivaraman Natarajan, Jiahui Zhang, Qingfei Pan, John Easton, et al. Spotiphy enables single-cell spatial whole transcriptomics across an entire section. *Nature Methods*, 22(4):724–736, 2025.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65:101789, 2020.

Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle YY Lee, Kyung S Cho, Katalin Susztak, George X Xu, Michael D Feldman, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature biotechnology*, 42(9): 1372–1377, 2024.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18802–18812, June 2022.

Pengzhi Zhang, Weiqing Chen, Tu N Tran, Minghao Zhou, Kaylee N Carter, Ibrahem Kandel, Shengyu Li, Xen Ping Hoi, Yuxing Sun, Li Lai, et al. Thor: a platform for cell-level investigation of spatial transcriptomics and histology. *Nature Communications*, 16(1):7178, 2025.

Suyuan Zhao, Yizhen Luo, Ganbo Yang, Yan Zhong, Hao Zhou, and Zaiqing Nie. Stofm: a multi-scale foundation model for spatial transcriptomics. In *International Conference on Machine Learning*, pp. 77589–77607. PMLR, 2025.

APPENDIX

A DATASET

A.1 DATA SOURCES AND INCLUSION CRITERIA

Data sources. We curate CellHIST-Data from publicly available spatial transcriptomics (ST) resources, including HEST-1K and additional datasets from 10× Genomics and NCBI. To support weakly supervised single-cell SGE inference, we apply consistent, task-specific inclusion criteria across data sources. The resulting cohort comprises 169 samples, including 84 with native spot- or bin-resolution measurements and 85 with native single-cell-resolution measurements. Table 3 summarizes the cohort composition by source.

Table 3: **Data composition of CellHIST-Data under our task-specific inclusion criteria.** Among the 169 eligible samples, 121 are obtained from HEST-1K and 48 additional samples are curated from public repositories by CellHIST-Data. The 48 samples curated from other public resources include 42 with single-cell-resolution measurements, expanding the single-cell-resolution subset from 43 to 85 samples.

Data source	Spot-/bin-resolution	Single-cell-resolution	Total
HEST-1K eligible samples	78	43	121
Additional samples curated by CellHIST	6	42	48
NCBI	0	38	38
10× Genomics	6	4	10
CellHIST-Data	**84**	**85**	**169**

Inclusion criteria. We retain samples that meet the following requirements: (1) human cancer tissue prepared as formalin-fixed paraffin-embedded (FFPE) specimens; (2) paired H&E-stained whole-slide images (WSIs) and ST measurements; (3) an image resolution finer than $0.75\,\mu$m/pixel to support nuclear segmentation and cellular feature extraction; and (4) sufficient spatial metadata for histology–transcriptomics alignment, with the alignment passing quality control. Samples that do not meet these requirements are excluded. Together, these criteria yield a cohort with high-resolution histology and spatially aligned gene-expression measurements for constructing cell-centric inputs and spot-level weak supervision.

A.2 SPOT-LEVEL LABEL HARMONIZATION AND PSEUDO-SPOT CONSTRUCTION

Samples in CellHIST-Data span ST platforms with different spatial measurement units. To enable consistent weak supervision across platforms, we provide spot-level SGE labels for every sample. We retain native spot geometry and expression measurements where compatible and construct pseudo-spots otherwise, as described in Section 4.2. Pseudo-spots follow the 10× Visium layout: circular regions approximately $55\,\mu$m in diameter, arranged on a hexagonal grid with $100\,\mu$m center-to-center spacing. Expression measurements within each region are aggregated to obtain its spot-level label.

Sources of spot-level supervision. The 169 samples fall into two groups according to how their spot-level labels are obtained. (1) Native spot labels are retained for 69 samples, including 57 Visium and 12 Spatial Transcriptomics samples. (2) Pseudo-spot labels are constructed for the remaining 100 samples, including 85 single-cell-resolution samples from Xenium, Xenium 5K, and Atera, and 15 Visium HD samples. For Visium HD, we aggregate expression measurements from the underlying high-resolution spatial grids. For the single-cell-resolution platforms, we accumulate captured transcripts within each pseudo-spot boundary. Table 4 summarizes this composition.

Support for different learning paradigms. For the 85 single-cell-resolution samples, we preserve the original cell-level expression labels alongside the constructed pseudo-spot labels. The same samples can therefore support weakly supervised training with pseudo-spot labels, as well as fully supervised training or evaluation with cell-level labels. This enables matched-data comparisons across learning paradigms, with the training labels determined by the experimental setting. Further processing details are provided in Sections 4.2 and 4.3.

Table 4: **Composition of spot-level supervision in CellHIST-Data.** Of the 169 samples, 69 retain native spot labels and 100 use pseudo-spot labels constructed from higher-resolution measurements. Xenium, Xenium 5K, and Atera together contribute 85 single-cell-resolution samples, for which cell-level labels are also preserved.

Spot-label source	Platform	Samples	Subtotal
Native spots	Visium	57	69
	Spatial Transcriptomics	12	
Pseudo-spots	Xenium, Xenium 5K, Atera	85	100
	Visium HD	15	
Total			**169**

A.3 Cell segmentation quality control

To assess the quality of CellViT++ segmentation, we visually inspect whether predicted nuclear polygons align with the corresponding nuclear structures in H&E images. This provides a qualitative assessment of spatial alignment and segmentation plausibility for cell-level representation extraction, complementing the quantitative analysis of geometric agreement with ST-provided reference geometries in Section A.4. We apply a consistent ROI selection and visualization procedure across samples, as detailed below.

ROI selection. For each sample, we load the level-0 H&E whole-slide image (WSI) and the corresponding segmentation polygons. We sample up to 64 candidate ROIs of 1024×1024 pixels in the level-0 coordinate system and select one for visualization, prioritizing regions containing at least 50 segmented nuclei. If no candidate meets this criterion, we select the region with the most segmented nuclei.

Visualization analysis. For each selected ROI, we display the original H&E image alongside the same region overlaid with predicted nuclear polygons. These paired views allow us to assess the correspondence between predicted boundaries and visible nuclear structures and identify gross spatial misalignment or apparent segmentation errors. Figure 7 presents representative examples from multiple ST platforms.

A.4 Cell matching quality and evaluation coverage

To assess cell-level correspondence and evaluation coverage, we compare CellViT++-predicted nuclei with ST reference cells across 62 test slides from five organs: bowel, breast, lung, ovary, and skin. We describe the matching criteria and coverage metrics below.

Matching criteria. We match a predicted nucleus to an ST reference nucleus only when their centroids are mutual nearest neighbors and lie within $5\ \mu\mathrm{m}$ of each other. We exclude cases with exact distance ties or geometries that fail the matching pipeline's validity checks. Each matched predicted nucleus is assigned the ST cell identifier of its reference nucleus, establishing a strict one-to-one correspondence.

Coverage metrics. Let N_{ST} denote the number of ST cells in the label file, N_{pred} the number of predicted nuclei across the WSI, N_{match} the number of one-to-one matched pairs, and N_{eval} the number of cells retained for final evaluation. We report the ST-cell matching rate ($R_{1:1}$), the unmatched fractions of ST cells (U_{ST}) and predicted nuclei (U_{pred}), and the evaluation coverage relative to ST cells (R_{eval}):

$$R_{1:1} = \frac{N_{\mathrm{match}}}{N_{\mathrm{ST}}}, \qquad U_{\mathrm{ST}} = 1 - \frac{N_{\mathrm{match}}}{N_{\mathrm{ST}}}, \tag{3}$$

$$U_{\mathrm{pred}} = 1 - \frac{N_{\mathrm{match}}}{N_{\mathrm{pred}}}, \qquad R_{\mathrm{eval}} = \frac{N_{\mathrm{eval}}}{N_{\mathrm{ST}}}. \tag{4}$$

For each organ, we compute these rates from pooled cell counts across its slides rather than averaging slide-level percentages. Tables 5 and 6 summarize cohort sizes and coverage rates, respectively.

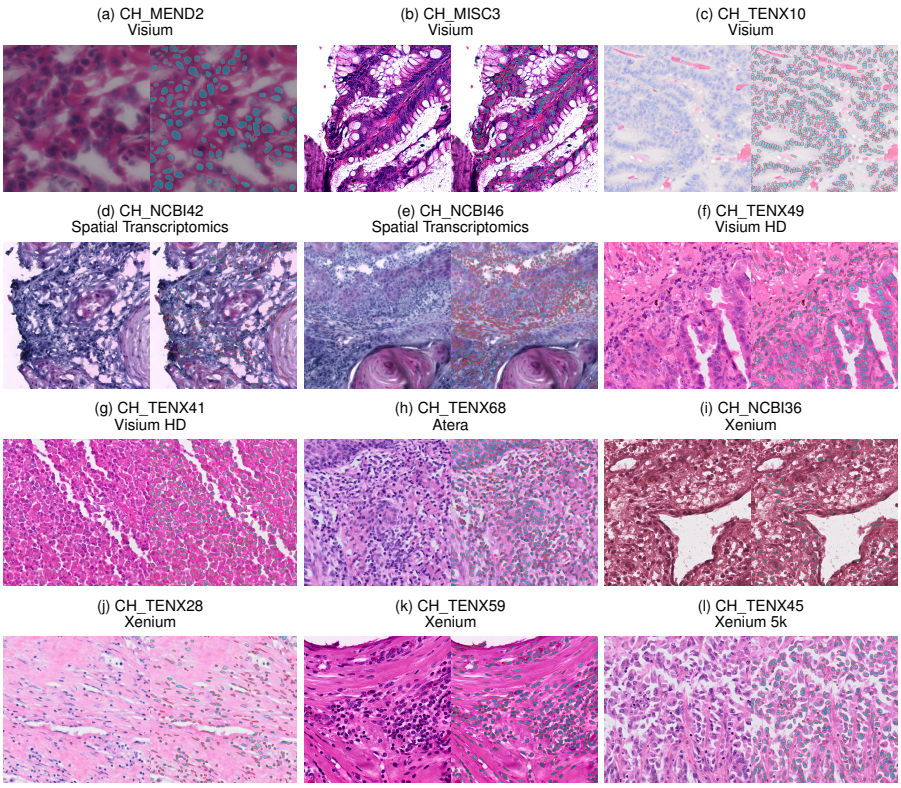

Figure 7: **Visual quality assessment of CellViT++ segmentation across spatial transcriptomics platforms.** Each pair shows an original H&E ROI (left) and the same region overlaid with predicted nuclear boundaries (right), allowing visual inspection of segmentation quality and spatial alignment.

Table 5: **Cell counts for matching and evaluation by organ.** Cell and nucleus counts are reported in thousands (k), rounded to one decimal place. ST cells are counted from reference label files, while predicted nuclei are counted across the entire WSI, including regions outside ST assay coverage. Matched pairs satisfy the mutual nearest-neighbor criterion with a centroid distance of at most $5\,\mu\mathrm{m}$. Evaluated cells are those retained for final evaluation.

Organ	number of Slides	ST cells (k)	Pred. nuclei (k)	Matched pairs (k)	Eval. cells (k)
Bowel	5	1,858.4	4,200.2	1,354.2	1,354.0
Breast	32	7,622.2	6,980.0	5,400.5	5,393.6
Lung	20	3,687.2	2,062.2	1,731.9	1,731.4
Ovary	2	649.0	550.3	497.0	496.8
Skin	3	304.3	308.8	231.0	231.0

Interpretation of unmatched cells. Predicted nuclei are counted across the entire WSI, including regions outside the ST assay coverage. Accordingly, U_{pred} represents the fraction of predicted nuclei without an ST match and should not be interpreted as a segmentation false-positive rate. Unmatched ST cells may reflect missed or inaccurate nuclear segmentation, spatial registration errors, or differences in tissue coverage between the two modalities.

Candidate ambiguity and geometric agreement. To assess matching ambiguity, we first assign each predicted nucleus to its nearest ST reference nucleus within $5\,\mu\mathrm{m}$, without requiring mutual nearest-neighbor selection. We define the candidate duplicate rate as the proportion of ST cells assigned multiple predicted nuclei among those assigned at least one. This metric captures ambiguity before reciprocal matching; the final strict correspondence contains no duplicate assignments by construction.

Table 6: **Cell-matching rates and evaluation coverage by organ.** All values are percentages calculated from unrounded counts. Matching rate, unmatched ST cells, and evaluation coverage use all ST cells as the denominator. Unmatched predicted nuclei use all predicted nuclei in the WSI, including those outside ST assay coverage.

Organ	Matching rate	Unmatched		Eval. coverage
		ST cells	Pred. nuclei	
Bowel	72.87	27.13	67.76	72.86
Breast	70.85	29.15	22.63	70.76
Lung	46.97	53.03	16.02	46.96
Ovary	76.57	23.43	9.69	76.55
Skin	75.92	24.08	25.18	75.91

For strictly matched pairs, we measure the distance from the predicted nuclear centroid to both the reference nuclear centroid and the matched ST cell centroid. Where valid polygons are available, we further assess geometric agreement using nucleus–nucleus intersection over union (IoU), nucleus–cell IoU, and the fraction of the predicted nuclear area contained within the matched ST cell:

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \qquad C_{\mathrm{nuc}\to\mathrm{cell}} = \frac{|P_{\mathrm{nuc}} \cap P_{\mathrm{cell}}|}{|P_{\mathrm{nuc}}|}. \tag{5}$$

Here, nucleus–nucleus IoU compares the predicted nuclear polygon with the ST reference nuclear polygon, whereas nucleus–cell IoU compares it with the matched ST cell polygon. Because a nucleus occupies only part of a cell, nucleus–cell IoU is not expected to approach one even for a geometrically consistent match.

Polygon statistics are conditional on valid matched polygon pairs. Ambiguous duplicated cell-boundary identifiers were excluded from cell-overlap and containment calculations: seven identifiers in `CH_TENX58` and two in `CH_TENX63`. These exclusions do not alter nuclear matching or PCC evaluation. Table 7 summarizes the geometric results.

Table 7: **Candidate ambiguity and geometric agreement of strict matches.** Duplicates denotes the percentage of ST cells assigned multiple predicted nuclei among those receiving at least one candidate assignment within 5 μm. All other values are medians over valid matched pairs. d_{nuc} and d_{cell} denote distances from the predicted nuclear centroid to the reference nuclear and cell centroids, respectively. Containment is the percentage of predicted nuclear area inside the matched ST cell. Geometric measures quantify agreement with ST-provided references, not accuracy against independent manual annotations.

Organ	Duplicates (%)	Distance (μm)		IoU		Containment (%)
		d_{nuc}	d_{cell}	Nuc.–nuc.	Nuc.–cell	
Bowel	5.79	0.700	1.252	0.654	0.382	98.50
Breast	3.32	0.856	1.700	0.628	0.328	99.52
Lung	4.96	1.328	1.999	0.482	0.251	97.94
Ovary	3.07	0.617	1.106	0.701	0.401	99.73
Skin	2.78	0.424	1.299	0.760	0.339	100.00

Interpretation. Strict one-to-one matching retains 70.85–76.57% of ST cells in Bowel, Breast, Ovary, and Skin, compared with 46.97% in Lung. Among matched pairs, the median distance between predicted nuclear centroids and ST cell centroids ranges from 1.106 to 1.999 μm across organs, while the median fraction of predicted nuclear area contained within the matched cell ranges from 97.94% to 100.00%. These small centroid distances and high containment fractions indicate close spatial agreement for the retained pairs, although matching coverage varies across organs.

Table 8: **Sensitivity of gene-wise PCC to matching-confidence filtering.** The 75% and 50% subsets retain the corresponding fractions of strictly matched evaluation cells with the smallest nuclear centroid distances within each slide. Ranges show the minimum and maximum $\Delta\mathrm{PCC}$ across six methods relative to their full strict evaluation sets. Positive counts indicate the number of methods with $\Delta\mathrm{PCC} > 0$, without implying statistical significance. All results use training seed 0.

Organ	High-confidence 75%		High-confidence 50%	
	ΔPCC range	Positive	ΔPCC range	Positive
Bowel	$[-0.0072, +0.0015]$	2/6	$[-0.0154, -0.0008]$	0/6
Breast	$[-0.0003, +0.0066]$	5/6	$[+0.0011, +0.0099]$	6/6
Lung	$[-0.0044, +0.0011]$	2/6	$[-0.0056, +0.0019]$	2/6
Ovary	$[-0.0080, -0.0021]$	0/6	$[-0.0129, -0.0028]$	0/6
Skin	$[-0.0027, +0.0014]$	2/6	$[-0.0063, +0.0007]$	2/6

A.5 Sensitivity to cell-matching confidence

Experimental setup. We examine whether restricting evaluation to geometrically higher-confidence matches changes prediction performance. We use fixed predictions from six models trained under pseudo-spot weak supervision: ABMIL, ABMIL-gated, GraphMIL, MS-GraphMIL, Transformer, and sCellST. All models use training seed 0, and no retraining is performed.

The analysis includes 62 test slides, yielding 372 slide–method combinations. Re-evaluation on the original strict matching set reproduces the reference PCC results for all combinations. Each slide uses the same evaluation gene set across all subset comparisons.

Confidence subsets. Starting from the original strict one-to-one evaluation set, we construct two nested subsets by retaining the 75% and 50% of cells with the smallest distances between predicted and reference nuclear centroids within each slide. Subset sizes are rounded down, and distance ties are resolved deterministically by ST cell identifier. Selection depends solely on matching distance, without using expression values or model predictions.

For a retained fraction $q \in \{0.75, 0.50\}$, we define

$$\Delta\mathrm{PCC}_q = \mathrm{PCC}(\mathcal{S}_q) - \mathrm{PCC}(\mathcal{S}_{\mathrm{strict}}), \tag{6}$$

where $\mathcal{S}_{\mathrm{strict}}$ denotes the full strict evaluation set and $\mathcal{S}_q$ is the subset retained by distance-based filtering. Thus, the full set comprises all strictly matched evaluation cells, rather than all ST cells in the tissue.

Metrics and aggregation. We evaluate both gene-wise and cell-wise PCC using the same validity criteria as in the original evaluation. Gene-wise PCC is computed across cells for each gene and macro-averaged over genes with defined correlations. Cell-wise PCC is computed across genes for each cell and macro-averaged over cells with defined correlations.

For each organ and method, we average slide-level scores equally within each fold, followed by an equal-weight average across folds. Tables 8 and 9 summarize the resulting organ-level changes relative to each method's full strict evaluation set. We report the minimum and maximum $\Delta\mathrm{PCC}$ across the six methods, together with the number of methods showing a positive change. These ranges reflect variation across methods rather than confidence intervals, and positive changes do not imply statistical significance.

Results and limitations. Restricting evaluation to closer matches yields organ-dependent changes in PCC. In Breast, retaining the closest 50% of matches improves both metrics for all six methods, with increases of 0.0011–0.0099 in gene-wise PCC and 0.0014–0.0037 in cell-wise PCC. In Bowel, the same restriction reduces both metrics for all six methods, by 0.0008–0.0154 and 0.0011–0.0094, respectively. Lung, Ovary, and Skin exhibit predominantly negative or mixed changes. These observations indicate that closer geometric correspondence does not consistently translate into higher prediction performance.

However, the absence of consistent improvement does not establish that residual matching errors are negligible. Distance-based filtering may alter the distributions of cell types, morphology, local

Table 9: **Sensitivity of cell-wise PCC to matching-confidence filtering.** Subset construction and score aggregation follow Table 8. Ranges show the minimum and maximum $\Delta\mathrm{PCC}$ across six methods relative to their full strict evaluation sets. Positive counts describe the direction of change without implying statistical significance. All results use training seed `0`.

Organ	High-confidence 75%		High-confidence 50%	
	ΔPCC range	Positive	ΔPCC range	Positive
Bowel	$[-0.0062, -0.0011]$	0/6	$[-0.0094, -0.0011]$	0/6
Breast	$[+0.0001, +0.0016]$	6/6	$[+0.0014, +0.0037]$	6/6
Lung	$[-0.0024, -0.0007]$	0/6	$[-0.0033, -0.0011]$	0/6
Ovary	$[-0.0023, -0.0014]$	0/6	$[-0.0011, +0.0003]$	1/6
Skin	$[-0.0045, +0.0038]$	1/6	$[-0.0084, +0.0088]$	1/6

density, and expression variability, so the observed PCC changes cannot isolate the effect of matching error.

The analysis is restricted to the existing strict matching set and does not assess expression recovery for unmatched cells. It also uses a single training seed and includes only two Ovary slides and three Skin slides. We therefore interpret these findings as a descriptive assessment of sensitivity to distance-based filtering.

B Experimental details

B.1 Computational resources

All experiments were conducted on a local computing cluster equipped with four NVIDIA RTX 4090D GPUs (24 GB memory each), two NVIDIA H20 GPUs (96 GB memory each), and 24 CPU cores. The complete data preprocessing pipeline, including histology image processing, cell segmentation, feature extraction, and processed-data organization, required approximately one week to complete. Reproducing the benchmark methods and processing the TCGA data for the downstream tasks required approximately two additional weeks. The above computational resources were used for all data preprocessing, model training, benchmarking, and downstream experiments reported in this study.

B.2 Training and test set splits

Benchmark eligibility. The benchmark includes 126 samples from five organs: 64 spot-/bin-resolution samples for weakly supervised training under the standard-spot setting and 62 FFPE single-cell-resolution samples for evaluation (Table 10). For standard weak supervision, results on these 62 samples are reported separately for the 34-sample Xenium group in the main text and the 28-sample post-Xenium group in Appendix C.1. Replicates and samples that fail benchmark quality control are excluded from these experiments but retained in the released dataset. The training and evaluation pools have disjoint `patient_split_group` values to prevent patient-level data leakage.

Patient-level outer splits. For each organ, we partition single-cell samples into outer folds using `patient_split_group`, keeping all samples from the same patient group in a single fold. To limit training costs, we use at most six folds per organ, with patient groups assigned deterministically in a round-robin manner. Each patient group serves as test data in exactly one fold. We do not reserve a separate slide-level validation set; instead, models are trained for a fixed number of epochs and evaluated using the final checkpoint.

Training settings. We evaluate all methods at single-cell resolution on the same organ-specific, patient-level test folds. We consider three training settings that differ in the training data and supervision used.

Table 10: **Platform composition of samples from CellHIST-Data used in the benchmark.** The benchmark comprises 64 spot-/bin-resolution samples and 62 single-cell-resolution samples across five organs. Sample counts are reported for each pool, with the contributing platforms listed.

Sample pool	ST platform	# Samples
Spot-/bin-resolution	Spatial Transcriptomics Visium Visium HD Visium HD 3′	64
Single-cell-resolution	Xenium Xenium 5K Atera	62
Total		**126**

(1) Standard weak supervision. For each organ, we train on a fixed pool of spot-/bin-resolution ST samples that is separate from the test cohort, using the harmonized spot-level labels described in Appendix A.2 as weak supervision.

(2) Pseudo-spot weak supervision. We train on single-cell-resolution samples outside the current test fold, using only aggregated pseudo-spot expression labels for supervision.

(3) Full supervision. We use the same training samples as in setting (2), but supervise models directly with single-cell expression labels. Settings (2) and (3) therefore enable head-to-head comparisons under identical training samples and patient-level splits, with only the supervision granularity differing.

Evaluation aggregation. For each organ and training seed, we first average test-sample metrics within each outer fold, then average the fold-level scores with equal weights. This gives each fold equal weight regardless of its sample count. We report the mean and standard deviation across three training seeds, using `ddof=0` to compute the standard deviation.

Table 11: **Organ-specific training pools and patient-level evaluation splits.** Standard weak supervision uses a fixed spot-/bin-resolution training pool for each organ. Pseudo-spot weak supervision and full supervision use identical outer folds of the single-cell-resolution pool, with patient groups kept intact. In each outer fold, the remaining folds provide the training samples for these two settings. All three settings are evaluated on the same held-out single-cell samples.

Organ	Spot-level train	Single-cell pool	Patient groups	Outer folds
Bowel	40	5	5	5
Breast	3	32	22	6
Lung	3	20	11	6
Ovary	2	2	2	2
Skin	16	3	3	3
Total	**64**	**62**	**43**	**22**

B.3 Additional training and evaluation details

Weakly supervised training targets. For weakly supervised models in CellHIST, we construct spot-level targets from raw expression counts. We first merge columns with duplicate gene symbols by summing their counts and remove mitochondrial, technical, and other non-gene-expression features. We then normalize each spot to a total count of 10^4 and apply a $\log(1+x)$ transformation. Normalization uses all retained biological gene-expression features measured by the assay, before alignment to the model vocabulary. Specifically, the target for gene g in spot s is

$$y_{s,g} = \log\left(1 + \frac{10^4 x_{s,g}}{L_s}\right), \qquad L_s = \sum_{h \in \mathcal{G}_{\text{assay}}} x_{s,h}, \tag{7}$$

where $x_{s,g}$ denotes the raw count and $\mathcal{G}_{\text{assay}}$ is the retained assay-wide gene set. Finally, we align the normalized targets to the fixed training vocabulary. Genes not measured by a given assay are filled with zeros for alignment and excluded from the training loss using a gene mask.

Weakly supervised outputs and pooling. Cell-level outputs are learned indirectly from spot-level supervision without access to single-cell expression labels. We therefore interpret them as model-derived expression scores rather than raw single-cell transcript counts. sCellST, Transformer, GraphMIL, and MS-GraphMIL produce non-negative gene scores for individual cells and aggregate them by mean pooling within each spot. ABMIL and ABMIL-gated instead use gene-specific attention weights to construct spot-level representations for expression prediction. Their cell-level outputs are obtained by applying the inverse log transformation to spot-level predictions and distributing the resulting values across cells according to the learned attention weights. All models are optimized using a gene-masked mean squared error (MSE) loss between spot-level predictions and normalized targets. MS-GraphMIL additionally incorporates a morphological-class-based representation regularizer using nuclear classes predicted by CellViT++.

Cell-level ground truth and assay-normalized evaluation. We evaluate all methods against a common reference constructed from raw single-cell assay counts, using the *assay-normalized evaluation* protocol. We first align assay cells with detected cells through mutual nearest-neighbor matching, retaining only one-to-one pairs within $5\,\mu\text{m}$.

For each matched reference cell, we normalize counts to a total of 10^4 over the complete retained assay-wide gene set and apply a $\log(1+x)$ transformation. The reference expression for gene g in cell c is

$$z_{c,g} = \log\left(1 + \frac{10^4 x_{c,g}}{L_c}\right), \qquad L_c = \sum_{h \in \mathcal{G}_{\text{assay}}} x_{c,h}, \tag{8}$$

where $x_{c,g}$ denotes the raw single-cell count and $\mathcal{G}_{\text{assay}}$ is the retained assay-wide gene set. We compute the normalization denominator before selecting the TargetH50 evaluation panel and perform no further normalization within this panel.

For model predictions, we retain log-like outputs in their original representation and apply $\log(1+x)$ to linear or count-like outputs. We then align predictions with the assay-normalized reference by gene symbol. This protocol provides a method-independent reference for all main benchmark results, without assuming that model outputs share the same absolute expression scale.

Evaluation metrics. We primarily assess how well models capture expression variation using correlation-based metrics. For gene-wise evaluation, we compute Pearson correlation coefficients (PCC) and Spearman correlation coefficients (SCC) across matched cells for each gene, then macro-average over genes with defined correlations. For cell-wise evaluation, we compute PCC and SCC across genes for each cell, then macro-average over cells with defined correlations. We additionally report root mean squared error (RMSE) in the appendix under the same assay-normalized evaluation protocol. Since model outputs may differ in absolute expression scale, RMSE reflects both prediction errors and scale differences relative to the common reference. Scores are aggregated across samples, folds, and random seeds as described above.

Training details. We train ABMIL, ABMIL-gated, sCellST, Transformer, GraphMIL, and MS-GraphMIL for 100 epochs using AdamW with a learning rate of 1×10^{-4} and a weight decay of 10^{-2}, without learning-rate scheduling or early stopping. Batch sizes are set by model: 1024 for sCellST, 512 for ABMIL and ABMIL-gated, 384 for Transformer under standard weak supervision and 640 under pseudo-spot weak supervision, and 256 for the graph-based models (GraphMIL and MS-GraphMIL). DeepSpot2Cell retains its method-specific training configuration following its original implementation. We repeat the main benchmark experiments with three random seeds (`0`, `1`, and `2`). Other external baselines retain their method-specific benchmark configurations for optimization, training duration, and loss functions.

Panel-aware output vocabulary and TargetH50 evaluation. To accommodate differences in gene panels across samples, we follow the panel-aware supervision strategy of VOICE (Luo et al., 2026). To reduce computational cost, we construct a separate output vocabulary for each organ and outer fold using the genes measured in Xenium samples within the corresponding single-cell training partition.

Specifically, for organ o and fold f, we define

$$\mathcal{V}_{o,f} = \bigcup_{d \in \mathcal{D}_{o,f}^{\text{train,Xenium}}} \mathcal{G}_d, \qquad m_{d,g} = \mathbb{I}[g \in \mathcal{G}_d], \quad g \in \mathcal{V}_{o,f}, \tag{9}$$

where $\mathcal{D}_{o,f}^{\text{train,Xenium}}$ denotes the Xenium subset of the single-cell training partition for organ o and fold f, and $\mathcal{G}_d$ is the set of biological genes measured in sample d. No samples from the corresponding held-out test fold are used to construct the vocabulary.

The vocabulary remains fixed throughout each organ–fold training run. Standard and pseudo-spot weak supervision share this organ- and fold-specific vocabulary, with observation masks determined by the genes measured in their respective training samples. All spots from the same sample share its mask. Genes in the vocabulary that are not measured in a given sample are zero-filled for tensor alignment and excluded from the regression loss.

For each evaluation sample e in organ o and fold f, we select a TargetH50 panel using its highly variable gene (HVG) ranking. Let $\boldsymbol{\pi}_e^{\text{HVG}}$ denote the measured genes ranked by decreasing HVG score. We retain genes present in $\mathcal{V}_{o,f}$ and select the top 50 without changing their order:

$$\mathcal{T}_{o,f,e}^{\text{H50}} = \text{First}_{50}\left(\boldsymbol{\pi}_e^{\text{HVG}} \cap_{\text{ord}} \mathcal{V}_{o,f}\right), \tag{10}$$

where $\cap_{\text{ord}}$ denotes intersection while preserving the HVG ranking. We align cell-level predictions and assay-normalized ground truth by gene symbol and evaluate them on this sample-specific panel. The normalization denominator is computed from the complete retained assay-wide gene set before restriction to either TargetH50 or the model vocabulary. The evaluation-sample HVG ranking is used solely to select genes for scoring and does not affect model training.

Single-cell features and model inputs. We precompute a 1,280-dimensional embedding for each detected cell using the SAM-H encoder in the CellViT++ pipeline. These shared embeddings provide the base visual representation for the seven weakly supervised methods and our ST-Net and single-cell HisToGene adaptations. Additional spatial, morphological, and patch-level information remains method-specific, as summarized in Table 12. GHIST instead extracts features from RGB H&E patches using its own CNN, whereas Thor uses PCA-reduced versions of the shared cell embeddings. All configurations described here refer to our benchmark implementations.

Expression supervision and inference. Weakly supervised methods use spot expression under the standard-spot setting and aggregated pseudo-spot expression under the pseudo-spot setting as training labels, rather than expression inputs at inference. Fully supervised baselines are trained with single-cell expression labels. GHIST additionally uses class-wise mean-expression references computed from training cells during prediction. In contrast, Thor requires the observed spot expression of the slide being reconstructed and operates on each slide separately. It is therefore included as a spatial super-resolution reference, rather than an image-only predictor on held-out slides.

Pseudo-spot weak supervision. For each outer fold, we construct pseudo-spots from the same single-cell training samples used for full supervision. We normalize the aggregated raw counts to a total of 10^4 over the complete retained assay-wide gene set and apply a $\log(1+x)$ transformation, following the procedure used for standard weak supervision. The two weak supervision settings thus share the same target normalization procedure but differ in training samples and the genes measured in those samples. Both are evaluated against the same assay-normalized single-cell reference described above.

C Additional experimental results

C.1 Additional evaluation on post-Xenium data

The 62 single-cell evaluation samples are divided into Xenium (34 samples) and post-Xenium (28 samples) groups according to H&E–Xenium registration provenance, as summarized in Table 13. The post-Xenium group contains samples whose registration was reconstructed within our project, either manually or automatically.

To complement the main-text results on the Xenium group, we further evaluate standard weakly supervised models on the 28 post-Xenium samples, comprising 12 Breast and 16 Lung samples.

Table 12: **Model inputs in the benchmark implementations.** Cell embeddings denote the shared, precomputed 1,280-dimensional SAM-H representations. The final column lists expression information used during prediction or reconstruction, excluding training labels.

Method	Visual representation	Additional information and processing	Expression at inference
Weakly supervised methods			
ABMIL	Cell embeddings	Spot–cell membership for bag construction; no explicit coordinate or morphological-class inputs.	None
ABMIL-gated	Cell embeddings	Same inputs as ABMIL, with gated attention.	None
sCellST (Mean-MIL)	Cell embeddings	Spot–cell membership for aggregating cell predictions.	None
Transformer	Cell embeddings	Within-bag cell sequences and bag/token masks; no explicit two-dimensional positional encoding.	None
GraphMIL	Cell embeddings	Spatial kNN graphs constructed from cell centroids and associated edge information.	None
MS-GraphMIL	Cell embeddings	Spatial graphs and morphological-class information; fusion of hidden features across graph layers.	None
DeepSpot2Cell	Cell embeddings and SAM-H patch features	Image context from the current and neighboring spots; membership and neighborhood masks.	None
Fully supervised methods			
GHIST	RGB H&E patches encoded by its own CNN	Nuclear instance masks; morphological-class supervision during training.	Class-wise mean-expression references computed from training cells
ST-Net	Cell embeddings	Linear expression prediction head; no coordinate inputs.	None
HisToGene	Cell embeddings	Cell-group Transformer with learnable x/y positional embeddings derived from normalized and discretized cell coordinates.	None
Spatial super-resolution reference			
Thor	PCA-reduced cell embeddings (at most 16 components by default)	Cell and spot coordinates and spot-coverage geometry; slide-specific reconstruction.	Observed spot expression of the slide being reconstructed

Table 13: Registration-based grouping of the single-cell-resolution evaluation cohort.

Group	Bowel	Breast	Lung	Ovary	Skin	Total
Xenium	5	20	4	2	3	34
post-Xenium	0	12	16	0	0	28
Total	**5**	**32**	**20**	**2**	**3**	**62**

Table 14 reports gene-wise PCC using the same assay-normalized evaluation protocol. The Average column presents the unweighted mean across the two organs.

Table 14: **Standard weak supervision on post-Xenium data.** Results are assay-normalized gene-wise PCC. The first line shows the mean, and the second line shows the standard deviation across three training seeds in parentheses. Unavailable results (–). Average is the unweighted mean across Breast and Lung. Higher is better.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
ABMIL	–	0.0367 (0.0056)	0.0747 (0.0038)	–	–	0.0557
ABMIL-gated	–	0.0333 (0.0089)	0.0755 (0.0047)	–	–	0.0544
GraphMIL	–	0.0066 (0.0114)	0.0633 (0.0198)	–	–	0.0350
MS-GraphMIL	–	0.1337 (0.0104)	0.0234 (0.0026)	–	–	0.0786
Transformer	–	0.0403 (0.0096)	0.0846 (0.0128)	–	–	0.0625
DeepSpot2Cell	–	0.0583 (0.0087)	0.0771 (0.0078)	–	–	0.0657
sCellST	–	0.0149 (0.0051)	0.0576 (0.0100)	–	–	0.0363

Table 15: **Prediction representations and evaluation-time processing.** Assay-CPM10k denotes expression normalized to a total count of 10^4 over the retained assay-wide gene set; assay-CPM10k-log1p additionally applies $\log(1+x)$. Linear outputs are clipped at zero before log transformation. Predictions are not re-normalized over the model vocabulary or evaluation panel.

Method group	Methods	Exported representation	Evaluation processing
Weak supervision	ABMIL	Linear assay-CPM10k values allocated from spot-level predictions	Apply $\log(1+z)$
Weak supervision	ABMIL-gated	Linear assay-CPM10k values allocated from spot-level predictions	Apply $\log(1+z)$
Weak supervision	GraphMIL	Assay-CPM10k-log1p	Retain output
Weak supervision	MS-GraphMIL	Assay-CPM10k-log1p	Retain output
Weak supervision	Transformer	Assay-CPM10k-log1p	Retain output
Weak supervision	DeepSpot2Cell	Log-like generic model score	Retain score
Weak supervision	sCellST	Assay-CPM10k-log1p	Retain output
Full supervision	GHIST	Log-like generic model score	Retain score
Full supervision	ST-Net; HisToGene	Assay-CPM10k-log1p	Retain output
Spatial super-resolution	Thor	Assay-CPM10k-log1p	Retain output

C.2 Prediction representations and supplementary RMSE evaluation

We follow the assay-normalized evaluation protocol described in Section B.3. For each reference cell, raw counts are normalized to a total of 10^4 over the complete retained assay-wide gene set and transformed using $\log(1+x)$. Normalization is performed before selecting evaluation genes. Neither reference values nor predictions are re-normalized over the model vocabulary or the evaluation panel.

Prediction representations. Methods produce outputs with different numerical representations. Table 15 summarizes these representations and their evaluation-time processing. For linear outputs, we clip negative values to zero and apply $\log(1+z)$. Log-transformed expression values and log-like model scores are retained without further transformation. These operations preserve each method's output interpretation without assuming that generic model scores are calibrated to assay-level expression units.

Table 16: **Gene-wise RMSE under standard weak supervision.** Predictions and single-cell reference values are compared in assay-CPM10k-log1p space. Each entry reports the mean across three training seeds, with the population standard deviation in parentheses below. Average is the unweighted mean across five organs. Lower is better.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
ABMIL	2.8721 (0.0123)	3.2977 (0.0030)	2.2619 (0.0041)	2.1987 (0.0038)	2.8280 (0.0149)	2.6917
ABMIL-gated	2.8763 (0.0104)	3.2866 (0.0068)	2.2540 (0.0046)	2.1928 (0.0041)	2.8404 (0.0272)	2.6900
GraphMIL	2.7575 (0.0056)	3.2543 (0.0020)	2.2511 (0.0042)	2.2165 (0.0025)	2.8209 (0.0137)	2.6601
MS-GraphMIL	2.8452 (0.0029)	3.0800 (0.0075)	2.2698 (0.0046)	2.1664 (0.0028)	2.7888 (0.0019)	2.6300
Transformer	2.6988 (0.0147)	3.2235 (0.0185)	2.2442 (0.0131)	2.2210 (0.0035)	2.8120 (0.0273)	2.6399
DeepSpot2Cell	—	—	—	—	—	—
sCellST	2.7771 (0.0078)	3.3178 (0.0120)	2.2475 (0.0042)	2.1994 (0.0037)	2.8362 (0.0144)	2.6756

DeepSpot2Cell produces generic log-like scores without verified assay-level expression units; its RMSE is therefore not reported (—).

Supplementary RMSE evaluation. To complement the correlation-based metrics in the main benchmark, we report gene-wise RMSE for methods whose processed outputs have verified assay-CPM10k-log1p semantics. For each sample, we compute RMSE across matched cells for each evaluation gene and then macro-average over genes:

$$\mathrm{RMSE} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \sqrt{\frac{1}{N} \sum_{c=1}^{N} (\hat{y}_{cg} - y_{cg}^{\mathrm{assay}})^2}. \tag{11}$$

Here, $\mathcal{G}$ denotes the evaluation gene set, N is the number of matched cells, and $\hat{y}_{cg}$ and y_{cg}^{assay} are the processed prediction and assay-normalized reference for gene g in cell c, respectively. RMSE is computed directly in assay-CPM10k-log1p space, without inverse transformation. Lower values indicate closer agreement with the reference.

We do not report RMSE for GHIST or DeepSpot2Cell because their generic model scores do not have verified absolute assay-level units. Both methods remain included in the correlation-based evaluations. Tables 16 and 17 report results under standard and pseudo-spot weak supervision, respectively. Table 18 provides the fully supervised and spatial super-resolution reference results.

Score aggregation. For each organ and training seed, we first average sample scores equally within each fold and then average equally across folds. We report the mean and population standard deviation across random seeds `0`, `1`, and `2`. Thor is evaluated with seed `0` only. The Average column reports the unweighted mean of the five organ-level scores.

C.3 Additional evaluation on the H50 gene panel

We additionally evaluate all methods on H50, a target-defined gene panel restricted to genes supported by both supervision settings in each comparison. This complements TargetH50 evaluation by assessing performance on a shared set of supported genes. H50 is used solely to define the evaluation subset: test expression informs gene selection for scoring but is not used for model training or hyperparameter tuning.

H50 construction. For each test slide in organ o and fold f, H50 contains up to 50 highly variable genes from the output vocabulary $\mathcal{V}_{o,f}$ that are supported by both supervision settings in the comparison. We first identify their common training support and then select genes using the test slide's expression variability.

Table 17: **Gene-wise RMSE under pseudo-spot weak supervision.** Predictions and single-cell reference values are compared in assay-CPM10k-log1p space. Each entry reports the mean across three training seeds, with the population standard deviation in parentheses below. Average is the unweighted mean across five organs. Lower is better.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
ABMIL	2.5079 (0.0132)	2.5521 (0.0050)	2.2651 (0.0018)	2.8991 (0.0072)	2.9325 (0.0253)	2.6313
ABMIL-gated	2.5178 (0.0044)	2.5658 (0.0085)	2.2643 (0.0007)	2.8789 (0.0187)	2.9339 (0.0325)	2.6321
GraphMIL	2.6588 (0.0076)	2.5544 (0.0047)	2.3518 (0.0056)	2.9808 (0.0108)	2.8325 (0.0207)	2.6757
MS-GraphMIL	2.7675 (0.0150)	2.5845 (0.0089)	2.4383 (0.0189)	2.9654 (0.0157)	2.5444 (0.0051)	2.6600
Transformer	2.8016 (0.0056)	2.7744 (0.0106)	2.4527 (0.0124)	2.9670 (0.0023)	2.8876 (0.0470)	2.7767
DeepSpot2Cell	—	—	—	—	—	—
sCellST	2.5495 (0.0041)	2.5249 (0.0036)	2.2772 (0.0041)	2.8519 (0.0215)	3.0177 (0.0055)	2.6442

DeepSpot2Cell produces generic log-like scores without verified assay-level expression units; its RMSE is therefore not reported (—).

Table 18: **Gene-wise RMSE for fully supervised and spatial super-resolution references.** Predictions and single-cell reference values are compared in assay-CPM10k-log1p space. ST-Net and HisToGene results report the mean across three training seeds, with the population standard deviation in parentheses below. Thor is evaluated with seed `0` only. Average is the unweighted mean across five organs. Lower is better.

Category	Method	Bowel	Breast	Lung	Ovary	Skin	Average
Full supervision	GHIST	—	—	—	—	—	—
Full supervision	ST-Net	2.3177 (0.0009)	2.2994 (0.0013)	2.0266 (0.0005)	2.5260 (0.0024)	2.4169 (0.0006)	2.3173
Full supervision	HisToGene	2.2831 (0.0020)	2.2688 (0.0067)	2.0112 (0.0014)	2.5319 (0.0109)	2.4936 (0.0117)	2.3177
Spatial super-resolution	Thor	2.8710	2.7916	2.8151	3.0806	2.3269	2.7770

GHIST produces generic log-like scores without verified assay-level expression units; its RMSE is therefore not reported (—). No across-seed standard deviation is reported for Thor.

Common training support. Let $\mathcal{V}_{o,f}$ denote the organ- and fold-specific output vocabulary defined in Section B.3, constructed as the union of biological genes measured in Xenium samples within the corresponding single-cell training partition. This vocabulary remains fixed throughout each organ–fold training run, and no samples from the corresponding held-out test fold are used in its construction. For supervision setting t, organ o, and fold f, we define the training-supported gene set as

$$S_{t,o,f} = \mathcal{V}_{o,f} \cap \left(\bigcup_{s \in \mathcal{D}^{\text{train}}_{t,o,f}} G_s \right), \tag{12}$$

where $\mathcal{D}^{\text{train}}_{t,o,f}$ is the corresponding training set and G_s contains the biological genes measured in training slide s. A gene in $\mathcal{V}_{o,f}$ is considered supported under setting t if it is measured in at least one slide in this training set. For each comparison, the common-support set is

$$C_{o,f} = \begin{cases} S_{\text{standard},o,f} \cap S_{\text{full},o,f}, & \text{standard weak versus full,} \\ S_{\text{pseudo},o,f} \cap S_{\text{full},o,f}, & \text{pseudo-spot weak versus full.} \end{cases} \tag{13}$$

Pseudo-spot training support is determined from the pseudo-spot training matrices.

Gene variability. For each test slide s, we retrieve expression values from `layers["counts"]` when available. Otherwise, we use `raw.X` if it contains the requested genes and fall back to `X` if it does not. We exclude technical features, canonicalize gene symbols, and sum entries with duplicate symbols. Let x_{ig} denote the resulting expression value for gene g in cell i. We rank genes by the variance of their log-transformed expression:

$$z_{ig} = \log(1 + x_{ig}), \qquad v_g = \frac{1}{N_s} \sum_{i=1}^{N_s} (z_{ig} - \bar{z}_g)^2, \tag{14}$$

where N_s is the number of cells in the test slide's H5AD file and $\bar{z}_g$ is the mean log-transformed expression of gene g. We apply no library-size normalization when computing this selection statistic. Gene selection is therefore distinct from the assay-normalized expression processing used for scoring.

Panel selection. Eligible genes must have nonzero expression in at least 20 cells and variance greater than 10^{-12}:

$$E_s = \left\{ g : n_g^+ \geq 20, \quad v_g > 10^{-12} \right\}, \tag{15}$$

where n_g^+ denotes the number of cells with nonzero expression for gene g. We intersect eligible genes with the common-support set before selecting the panel:

$$\mathrm{H50}_{o,f,s} = \mathrm{Top}_{50,v} \left(E_s \cap C_{o,f} \right). \tag{16}$$

where $\mathrm{Top}_{50,v}$ selects up to 50 genes in descending order of v_g, with ties resolved by gene-name ordering. If fewer than 50 eligible genes remain, we retain all of them.

H50 is specific to the test slide, fold, and comparison setting. Within each comparison, all methods and training seeds share the same panel. Test expression is used solely to select evaluation genes, not for model training or hyperparameter tuning.

Worked example: Bowel, fold 0. In the standard weak-versus-full comparison for Bowel fold 0, the organ- and fold-specific output vocabulary $\mathcal{V}_{o,f}$ contains 840 genes, of which 837 are supported by both training settings. For the held-out sample `CH_TENX39`, 419 genes remain after biological-feature filtering, gene-symbol processing, and expression quality control. Restricting these genes to the common-support set $C_{o,f}$ removes *RPS4Y1* and *TRGV4*, leaving 417 candidates. We then select the 50 genes with the highest variance in $\log(1 + x)$-transformed expression to form the H50 panel (Table 19). The complete panel and its gene ranking are recorded in the sample-specific panel file.

Table 19: **H50 construction for Bowel fold 0, test sample `CH_TENX39`.** Training support is restricted to the organ- and fold-specific output vocabulary. Gene selection uses log1p-expression variance without library-size normalization; subsequent evaluation uses assay-normalized expression.

Gene set	Count
Bowel fold-0 output vocabulary	840
Standard-weak training support	837
Fully supervised training support	840
Common training support	837
Test genes passing preprocessing and QC	419
Eligible test genes within common support	417
Final H50 panel	50

Evaluation and aggregation. We evaluate H50 using the same assay-normalized ground truth as the main benchmark. Reference counts are normalized to a total of 10^4 over the complete retained assay-wide gene set and transformed using $\log(1 + x)$ before selecting H50 genes. Predictions undergo the method-specific processing summarized in Table 15, with no additional normalization over H50.

For each test sample s in organ o and fold f, we compute PCC across matched cells for each H50 gene and macro-average over genes with defined correlations:

$$\mathrm{PCC}_{o,f,s} = \frac{1}{|\mathcal{H}^{\mathrm{valid}}_{o,f,s}|} \sum_{g \in \mathcal{H}^{\mathrm{valid}}_{o,f,s}} \mathrm{PCC} \left(\hat{\mathbf{y}}_{s,:,g}, \mathbf{y}^{\mathrm{assay}}_{s,:,g} \right), \tag{17}$$

Table 20: **H50 gene-wise PCC under standard weak supervision.** All results use assay-normalized ground truth. Values are mean±population standard deviation across three training seeds, with folds equally weighted within each seed. Average is the unweighted mean across organs. The highest mean in each column is shown in bold. Higher is better. All values are rounded to three decimal places.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
ABMIL	0.274 ± 0.008	0.148 ± 0.004	$\mathbf{0.240} \pm 0.003$	0.072 ± 0.001	0.011 ± 0.005	0.149
ABMIL-gated	0.274 ± 0.006	0.156 ± 0.007	0.239 ± 0.007	0.072 ± 0.007	0.014 ± 0.003	0.151
GraphMIL	0.328 ± 0.003	0.130 ± 0.007	0.225 ± 0.002	0.054 ± 0.006	0.016 ± 0.012	0.150
MS-GraphMIL	0.260 ± 0.004	0.146 ± 0.012	0.104 ± 0.007	$\mathbf{0.139} \pm 0.005$	0.043 ± 0.017	0.138
Transformer	0.306 ± 0.002	0.132 ± 0.005	0.218 ± 0.001	0.063 ± 0.002	0.010 ± 0.010	0.146
DeepSpot2Cell	$\mathbf{0.351} \pm 0.001$	0.140 ± 0.007	0.228 ± 0.000	0.125 ± 0.002	$\mathbf{0.052} \pm 0.003$	**0.179**
sCellST	0.343 ± 0.001	$\mathbf{0.170} \pm 0.006$	0.236 ± 0.004	0.096 ± 0.008	0.015 ± 0.006	0.172

Table 21: **H50 gene-wise PCC under pseudo-spot weak supervision.** Models are trained using aggregated pseudo-spot labels and evaluated against assay-normalized single-cell ground truth. Values are mean±population standard deviation across three training seeds, with folds equally weighted within each seed. Average is the unweighted mean across organs. † The Transformer result on Breast is a historical-reproduction result. Higher is better.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
ABMIL	0.3702 ± 0.0025	0.2443 ± 0.0034	0.1726 ± 0.0002	0.1889 ± 0.0008	0.0779 ± 0.0069	0.2108
ABMIL-gated	0.3701 ± 0.0016	0.2335 ± 0.0029	0.1732 ± 0.0023	0.1925 ± 0.0023	0.0839 ± 0.0177	0.2106
GraphMIL	0.3534 ± 0.0033	0.2876 ± 0.0014	0.1766 ± 0.0014	0.1734 ± 0.0061	0.0869 ± 0.0189	0.2156
MS-GraphMIL	0.2759 ± 0.0022	0.2785 ± 0.0044	0.1230 ± 0.0003	0.1857 ± 0.0019	0.2329 ± 0.0075	0.2192
Transformer	0.3434 ± 0.0031	$0.2854 \pm 0.0027^{\dagger}$	0.1857 ± 0.0012	0.1865 ± 0.0057	0.1476 ± 0.0307	0.2297
DeepSpot2Cell	0.3543 ± 0.0011	0.2374 ± 0.0018	0.1766 ± 0.0015	0.1411 ± 0.0042	0.1579 ± 0.0015	0.2135
sCellST	0.3645 ± 0.0019	0.2710 ± 0.0004	0.1748 ± 0.0020	0.1888 ± 0.0050	0.0882 ± 0.0069	0.2175

where $\mathcal{H}^{\text{valid}}_{o,f,s}$ denotes the subset of genes in the H50 panel for sample s with defined PCC values, and the two vectors contain the model predictions and assay-normalized reference values, respectively, across matched cells.

For each organ and training seed, we first average sample scores equally within each fold and then average the fold-level scores with equal weights. We report the mean and population standard deviation across random seeds `0`, `1`, and `2`; Thor is evaluated with seed `0` only. The Average column reports the unweighted mean across the five organ-level scores. Values in Table 20 are rounded to three decimal places, whereas Tables 21 and 22 are reported to four decimal places. Reported standard deviations of zero at the displayed precision may therefore reflect rounding rather than exactly zero variation.

Results. Tables 20–22 summarize performance on H50. DeepSpot2Cell achieves the highest average PCC under standard weak supervision (0.179), while Transformer leads under pseudo-spot weak supervision (0.2297). Among fully supervised baselines, HisToGene performs best on both comparison-specific panels, with average PCCs of 0.2995 for the standard weak-versus-full comparison and 0.2985 for the pseudo-spot weak-versus-full comparison.

C.4 Additional within-spot label-shuffling results

We extend the within-spot label-shuffling analysis to Breast and Ovary using the protocol described in Section 5.5. In each permutation, we randomly reassign complete assay-normalized ground-truth expression vectors among cells within each test spot while keeping model predictions fixed. This preserves the expression profiles present in each spot but disrupts their original correspondence with individual cells.

Table 22: **H50 gene-wise PCC for fully supervised and spatial super-resolution references.** All results use assay-normalized ground truth. Standard-comparison panels use common training support between standard weak and full supervision; pseudo-spot-comparison panels use common support between pseudo-spot weak and full supervision. Values are mean±population standard deviation across three seeds, except for Thor, which uses seed 0 only. Average is the unweighted mean across organs. Higher is better.

Method	Bowel	Breast	Lung	Ovary	Skin	Average
Fully supervised references: standard-comparison H50						
GHIST	0.3807 ± 0.0089	0.3433 ± 0.0018	0.1057 ± 0.0049	0.2085 ± 0.0091	0.1578 ± 0.0318	0.2392
ST-Net	0.4432 ± 0.0002	0.3479 ± 0.0002	0.2275 ± 0.0003	0.2497 ± 0.0004	0.0649 ± 0.0000	0.2667
HisToGene	0.4698 ± 0.0006	0.3901 ± 0.0011	0.2564 ± 0.0011	0.2546 ± 0.0017	0.1263 ± 0.0037	0.2995
Fully supervised references: pseudo-spot-comparison H50						
GHIST	0.3807 ± 0.0089	0.3420 ± 0.0017	0.1031 ± 0.0049	0.2085 ± 0.0091	0.1578 ± 0.0318	0.2384
ST-Net	0.4432 ± 0.0002	0.3469 ± 0.0003	0.2247 ± 0.0003	0.2497 ± 0.0004	0.0649 ± 0.0000	0.2659
HisToGene	0.4698 ± 0.0006	0.3886 ± 0.0011	0.2530 ± 0.0012	0.2546 ± 0.0017	0.1263 ± 0.0037	0.2985
Spatial super-resolution reference: standard-comparison H50						
Thor	0.4640	0.3487	0.2541	0.2827	0.4480	0.3595

We perform 9,999 permutations and compare the observed gene-wise PCC with the resulting permutation null distribution. Figures 8 and 9 show the Breast and Ovary results, respectively. The observed PCC is generally higher than the mean shuffled PCC, with the difference varying across organs, models, and training settings. Nonzero shuffled PCC is consistent with a contribution from shared tissue- and spot-level expression structure to the overall cell-level correlation. Together with the Bowel results in the main text, the reductions after shuffling provide additional evidence that predictions retain information associated with the original cell–expression correspondence beyond shared spot-level signals.

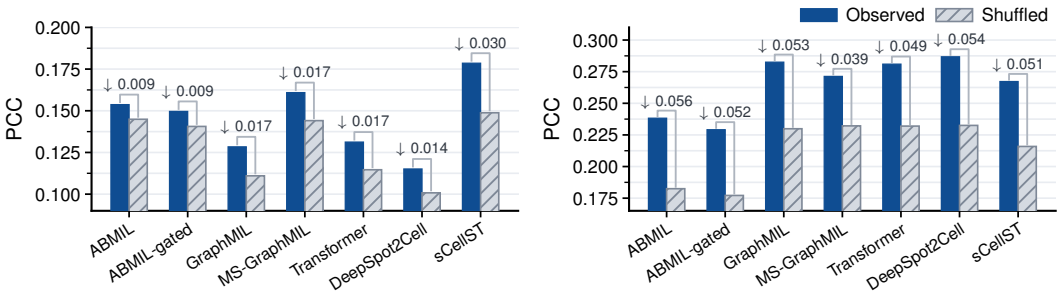

Figure 8: **Cell label shuffling within spots on Breast:** standard-spot (left) and pseudo-spot (right) data settings. Bars compare observed gene-wise PCC with the mean PCC over 9,999 within-spot permutations of assay-normalized cell expression labels, with predictions held fixed. Annotations show the PCC decrease relative to the observed result.

C.5 Additional cell-type annotation visualizations

To complement the main-text results, we visualize cell-type annotations in two additional regions from the same held-out breast cancer sample: a malignant–fibroblast interface (ROI b) and an immune-enriched region (ROI c). Figures 10 and 11 compare annotations derived from measured single-cell expression with those obtained from expression predicted by seven weakly supervised and three fully supervised methods.

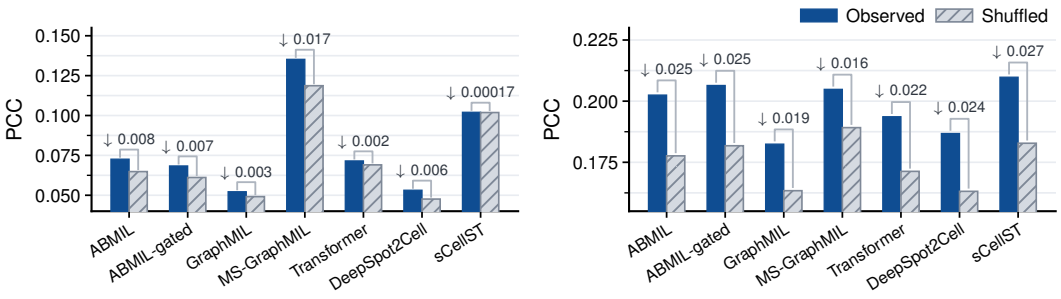

Figure 9: **Cell label shuffling within spots on Ovary:** standard-spot (left) and pseudo-spot (right) data settings. Bars compare observed gene-wise PCC with the mean PCC over 9,999 within-spot permutations of assay-normalized cell expression labels, with predictions held fixed. Annotations show the PCC decrease relative to the observed result.

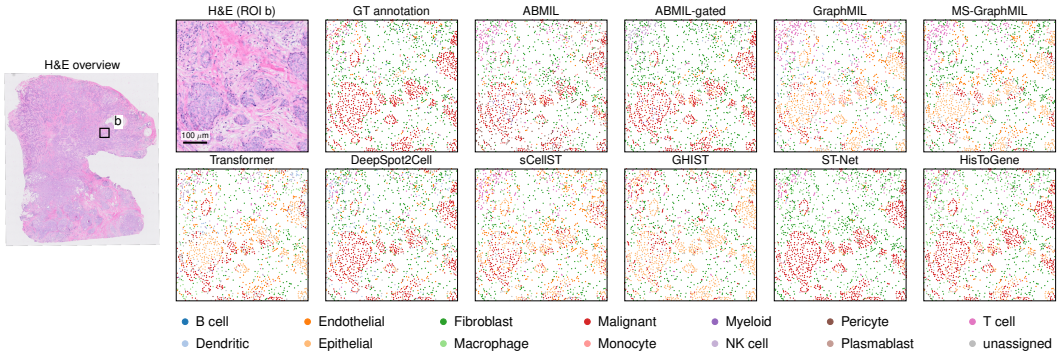

Figure 10: **Cell-type annotations at the malignant–fibroblast interface (ROI b).** Annotations inferred from predicted single-cell expression are shown for seven weakly supervised and three fully supervised methods, alongside the measured-expression reference. This $600 \times 600\,\mu\mathrm{m}^2$ region from a held-out breast cancer sample captures the interface between malignant cells and fibroblasts. Reference annotations were derived from measured expression, and region selection was independent of model performance.

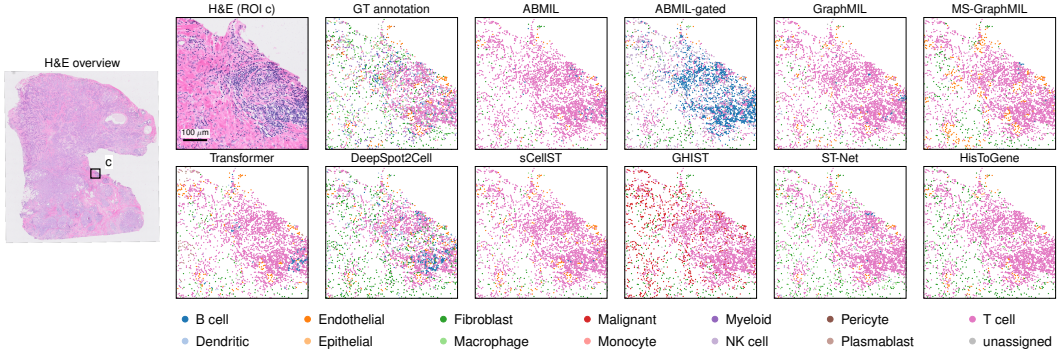

Figure 11: **Cell-type annotations in an immune-enriched region (ROI c).** Annotations inferred from predicted single-cell expression are shown for seven weakly supervised and three fully supervised methods, alongside the measured-expression reference. This $600 \times 600\,\mu\mathrm{m}^2$ region from the same held-out breast cancer sample contains 2,698 cells, approximately 79.7% of which are annotated as immune cells in the reference. The ROI was selected based on reference cell-type composition within a $300\,\mu\mathrm{m}$ window and expanded to $600\,\mu\mathrm{m}$ with its center fixed. Reference annotations were derived from measured expression, and region selection was independent of model performance.