SINGLE-CELL SPATIAL TRANSCRIPTOMICS PREDICTION FROM HISTOLOGY WITH MULTI-SCALE EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) offers complementary insights to underlying tissue morphology, yet the acquisition cost limits adoption in large cohorts. Virtual ST has demonstrated its cost-effective potential by predicting ST from H&E morphology, with recent methods extending prediction to single-cell resolution. At the cell level, sparse and heterogeneous gene expression creates a trade-off between preserving localized signals and capturing spatially coherent expression patterns. In this work, we introduce TIGER, a *single-cell* virtual ST framework that leverages pathology foundation model token embeddings at the spatial scale of individual cells. TIGER integrates local morphologic evidence with tissue-level context through two complementary branches: a *focal branch* that independently captures fine-grained morphomolecular signals and a *context branch* that integrates local cellular and global tissue context. A gene- and cell-specific routing mechanism adaptively combines the two. We evaluate TIGER on 43 H&E-paired Xenium tissue sections across diverse cancer types, comprising over 10 million cells. TIGER improves correlation by over 30% and significantly outperforms five state-of-the-art virtual ST baselines across a comprehensive set of conventional and sparsity-aware metrics. Furthermore, TIGER shows stronger generalization performance when evaluated on two external breast and lung cohorts (36 Xenium samples). The application of TIGER-predicted ST on downstream tasks, such as cell phenotyping and treatment response prediction, shows improved performance, underlining the biological and clinical utility of single-cell virtual ST. Code and preprocessed data are available: https://anonymous-hf.com/a/xgn0n1j4ylek/.

1 INTRODUCTION

Hematoxylin and eosin (H&E)–stained whole-slide images (WSIs) are the cornerstone of diagnostic surgical pathology and are available for almost every cancer patient, providing rich morphological information at scale and at a fraction of the cost of molecular assays. Rapid advances in pathology deep learning models have shown capacities to extract clinically actionable signals directly from histology (Song et al. (2023); Campanella et al. (2019); Lu et al. (2021); Campanella et al. (2025); Yao et al. (2020)). In parallel, spatial transcriptomics (ST), which measures spatially-resolved gene expressions, has emerged as a complementary modality, enabling researchers to interrogate tumor microenvironments and cell–cell interactions in their native spatial context (Ståhl et al. (2016); Marx (2021); Moses & Pachter (2022)). However, the substantial sequencing costs have hindered its broader application in large patient cohorts. The natural pairing between H&E images and ST, which provide co-registered morphological and molecular views of the same tissue, has motivated numerous methods that predict ST from histology for scalable ST, termed *virtual* ST.

Existing studies predominantly focus on *spot-level* expression prediction of Visium-style data, with each capture spot ($\sim$55 μm in diameter) aggregating transcripts over several cells. With pathology AI models primarily operating on patch-level, most works predict spot-level expressions from morphological patches centered around each spot, showing evidence that morphology carries substantial information about local transcriptional state. However, the spatial granularity of predictions is fundamentally bound by the assay, as each spot blends signals from tumor cells, stroma, and immune populations into a single bulk-like measurement. Predictions are thus averaged across heterogeneous cellular communities and cannot resolve the fine-grained spatial organization such as rare cell states and immune infiltration patterns (Kiviaho et al. (2024); Liu et al. (2025); Oliveira et al. (2025)).

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.

With the increasing availability of subcellular-resolution ST platforms, such as Xenium, recent works have tackled single-cell virtual ST (Zhu et al. (2025b); Nonchev et al. (2025b); Fu et al. (2025)). However, they do not explicitly account for sparse gene expression at cellular resolution, where many genes are expressed only in subsets of cells, resulting in highly localized expression patterns (**Fig. A1**). Such sparsity necessitates rethinking how spatial context should be used for prediction. While additional morphological context improves spot-level ST prediction by capturing spatially coherent gene expression (Chung et al. (2024); Nonchev et al. (2025a)), the same strategy cannot be directly applied at single-cell resolution. Spatial aggregation can oversmooth sparse, localized signals, whereas prediction based on local morphology can better preserve these signals but fails to exploit spatial coherence. Single-cell virtual ST therefore requires balancing two competing objectives: preserving spatially sparse expression and capturing spatially coherent patterns.

We introduce **TIGER** (**T**ranscriptomics **I**nference at single-cell resolution with fully adaptive **G**ene-level **E**xpression **R**outing), which generates cell representations from pathology foundation model token embeddings and adaptively combines two complementary experts. $\text{TIGER}_{\text{focal}}$ utilizes fine-grained morphological features to preserve localized expressions, whereas $\text{TIGER}_{\text{context}}$ integrates local and global spatial context to capture coherent expression patterns. An adaptive gene- and cell-specific gating mechanism combines them, which allows TIGER to faithfully capture heterogeneous spatial expression patterns. Evaluated on a large cohort of 79 Xenium samples comprising nearly 15 million cells across lung, breast, and colorectal cancers, we show that our hybrid approach significantly outperforms state-of-the-art single-cell virtual ST models in ST prediction accuracy and other biological and clinical downstream tasks. In addition, to properly evaluate such single-cell virtual ST, for which existing spot-level metrics are not always suitable, we introduce a new set of evaluation metrics tailored to the complexity and sparsity of single-cell ST data.

To summarize, our contributions are (1) an adaptive *single-cell* virtual ST framework that balances focal and spatial-context modeling, (2) evaluation metrics tailored for single-cell ST prediction that account for gene expression sparsity, and (3) demonstration of out-of-the-box generalization and downstream utility in cell phenotyping and patient-level treatment response prediction.

2 Related Work

2.1 Pathology Foundation Models

Pathology Foundation Models (PFMs), predominantly based on a Vision Transformer (ViT) backbone, are pretrained on millions of H&E histology image patches and generate general-purpose image embeddings for a wide range of downstream tasks. To enable large-scale pretraining, PFMs are trained with self-supervised learning objectives, either using vision-only objectives such as DINOv2 (Chen et al. (2024); Zimmermann et al. (2024); Aben et al. (2024)) or multi-modal contrastive learning with image and complementary modalities, such as texts or gene expression data (Ding et al. (2025); Lu et al. (2024); Vaidya et al. (2025); Hemker et al. (2026); Chen et al. (2025)). Since gene expressions are tightly correlated with underlying cell and tissue morphology, PFM embeddings have recently been used for spot-level ST prediction (Jaume et al. (2024)). However, these works are limited to utilizing the CLS embeddings for each patch, which mixes information from multiple cells and does not provide fine-grained resolution required for cellular-level ST prediction.

2.2 Morphology-based spot-level ST prediction

Prediction of ST from tissue images is based on the biological principle that underlying gene expressions are reflected as histologically discernible morphological patterns. With initial ST measurements predominantly based on spot-level aggregates of transcripts (Ståhl et al. (2016)), earlier works used tissue image patches that contained multiple cells. This was formulated as a regression problem with gene expressions predicted directly from H&E image patches encoded with vision encoders (He et al. (2020a); Pang et al. (2021); Yang et al. (2023); Zhu et al. (2025a)). Further works extended this framework to include tissue spatial contexts with graph neural networks or multi-scale approaches (Zeng et al. (2022); Chung et al. (2024); Nonchev et al. (2025a)). TIGER inherits the multi-scale regression framework for *cellular-level* ST prediction while adapting the contribution of spatial context on a per-gene and per-cell basis.

2.3 MORPHOLOGY-BASED CELL-LEVEL ST PREDICTION

Recent technologies now provide transcripts at (sub-)cellular resolution, leading to models capable of higher resolution prediction (Janesick et al. (2023); Tran et al. (2026)). To address the lack of available high-resolution ST data, early works adopted weakly-supervised approaches, where models were trained on spot-based expression (Visium or Xenium aggregated per spot) and subsequently distributed transcripts to individual cells (Zhang et al. (2024); Nonchev et al. (2025b); Chadoutaud et al. (2026); Zhang et al. (2026)). With more available data, recent methods have predicted single-cell ST directly. These include GHIST (Fu et al. (2025)), which simultaneously predicts and utilizes cell types and neighborhood composition for ST prediction, or Img2ST-Net (Zhu et al. (2025b)) that reformulates the task as a high-resolution image translation task. However, these often fall short in predictive performance and do not explicitly account for the wide variability in gene sparsity, with genes ranging from extremely sparse to broadly expressed. We leverage token-level representations of PFMs to handle cellular-level ST directly and develop a biologically informed dual-branch architecture that can simultaneously capture spatially sparse and smooth gene expressions adaptively.

3 METHODS

We present TIGER for single-cell ST prediction from histomorphology (**Figure 1**), which adaptively combines fine-grained and context-level predictions using two complementary expert modules $\text{TIGER}_{\text{focal}}$ and $\text{TIGER}_{\text{context}}$. We first define the problem (**Section 3.1**), describe the individual prediction modules (**Sections 3.2,3.3**), and the integration and training procedures (**Section 3.4, 3.5**).

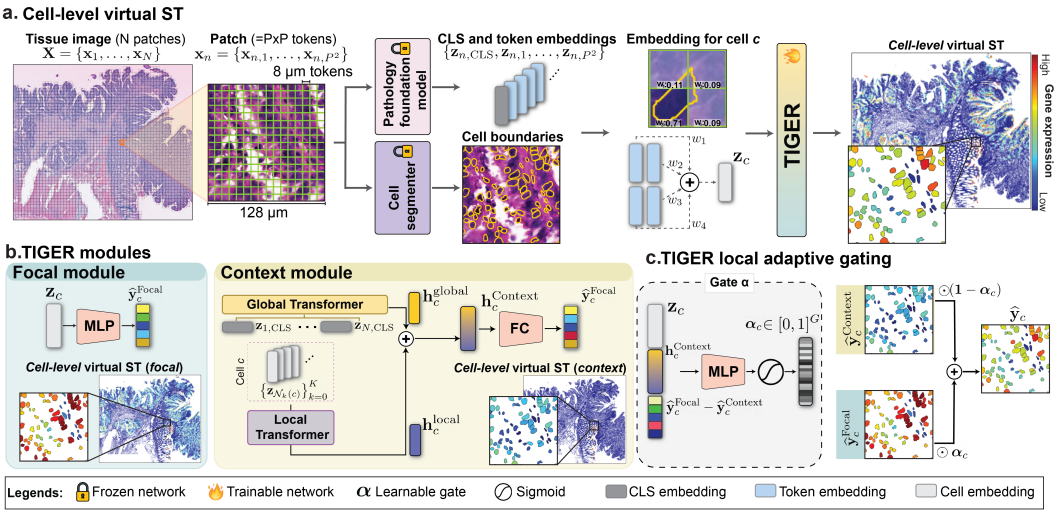


Figure 1: **TIGER architecture. a.** TIGER partitions the H&E WSI into N patches, each divided into P^2 tokens (8 µm bins). A ViT pathology foundation model yields CLS (patch) and token (bin) embeddings, and a cell segmenter delineates cell boundaries. Fine-grained cellular representations are obtained by pooling the token embeddings that fall within each segmented cell boundary, weighted by their area of overlap. **b.** TIGER consists of two complementary modules. $\text{TIGER}_{\text{focal}}$ predicts localized ST independently for each cell. $\text{TIGER}_{\text{context}}$ enriches cell representations with multi-scale spatial context based on *inter-patch* (global) and *inter-cell* (local) integration over each cell and its K nearest cells. **c.** TIGER integrates them with a per-cell and gene adaptive gating.

3.1 PROBLEM SETUP AND CELL-LEVEL MORPHOLOGICAL EMBEDDINGS

High-resolution spatial transcriptomics (e.g., Xenium) detects individual mRNA transcripts at cellular resolution. To obtain matching cell-level morphological representations from vision transformers (ViT Dosovitskiy et al. (2020))-based PFMs, we extract token features on a fixed grid over the slide and pool them within each cell boundary, weighting every token by the area it shares with the cell.

In more detail, given a H&E tissue image, we divide it into N smaller non-overlapping patches $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with $\mathbf{x}_n \in \mathbb{R}^{W \times H \times 3}$, where W and H denote width and height. Each image

patch is further divided into P^2 bins $\mathbf{x}_n = \{\mathbf{x}_{n,1}, \ldots, \mathbf{x}_{n,P^2}\}$ with $\mathbf{x}_{n,p} \in \mathbb{R}^{(W/P)\times(H/P)\times 3}$, each corresponding to a single token in ViT. We use $W = H = 256$ and $P = 16$ bins per side at $0.5\mu m$/pixel resolution, with the bin size of 8×8 μm roughly matching a single cell. The bins for each patch $\mathbf{x}_n$ are then encoded with pretrained PFM based on ViT, $f_{\text{enc.}}$, yielding token embeddings $\{\mathbf{z}_{n,p}\}_{p=1}^{P^2}$ and CLS embedding $\mathbf{z}_{n,\text{CLS}}$, both in $\mathbb{R}^D$ and contextualized through self-attention.

Let $w_{n,p}^c$ denote the fraction of the area overlapped by bin p in patch n, such that $\sum_{n,p=1}^{N,P^2} w_{n,p}^c = 1$. In practice, typically less than five $w_{n,p}^c$ are non-zero for a given cell c. The cell-level morphological embedding is computed as the weighted average $\mathbf{z}_c = \sum_{n,p=1}^{N,P^2} w_{n,p}^c \mathbf{z}_{n,p}$, where cell centroids and boundaries are obtained from the H&E image with a cell segmenter (i.e., Hörst et al. (2024)). This yields paired cell embeddings and cellular ST signatures $\mathbf{y}_c \in \mathbb{R}_+^G$ over the G genes. Further details on *cell-embedding* and ST preprocessing can be found in **Appendix A.2** and **Table A1**.

3.2 Context module for spatially coherent expressions

The first module, $\text{TIGER}_{\text{context}}$, enriches each cell with multi-scale spatial context $\mathbf{h}_c^{\text{Context}}$ for cell-level ST prediction $\widehat{\mathbf{y}}_c^{\text{Context}}$. Specifically, this is achieved by combining *local* context, $\mathbf{h}_c^{\text{local}}$, based on immediate cellular neighborhoods and *global* context, $\mathbf{h}_c^{\text{global}}$, based on the entire tissue,

$$\widehat{\mathbf{y}}_c^{\text{Context}} = f_{\text{lin.}}^{\text{gene}}(\mathbf{h}_c^{\text{Context}}) \quad \text{,where } \mathbf{h}_c^{\text{Context}} = \text{LayerNorm}(\mathbf{h}_c^{\text{local}} + f_{\text{lin.}}(\mathbf{h}_c^{\text{global}})). \tag{1}$$

The contextualization produces smoothed and spatially coherent ST predictions. We explain how the local and global contexts are derived below.

Local context branch. The local context branch encodes interactions between a cell and its cellular neighborhood. For cell c, we identify K closest neighboring cells $\{\mathbf{z}_{\mathcal{N}_k(c)}\}_{k=0}^K$ where $\mathcal{N}_k(c)$ denotes k-th closest neighbor to cell c and $k = 0$ denotes the cell itself. In TIGER, we use $K = 9$. The embeddings are linearly projected and combined with learnable embedding $\mathbf{e}_k$ based on neighbor rank, such that $\widetilde{\mathbf{z}}_{\mathcal{N}_k(c)} = f_{\text{lin.}}^{\text{local}}(\mathbf{z}_{\mathcal{N}_k(c)}) + \mathbf{e}_k$. They are then passed through a two-layer local transformer module $f_{\text{Trans.}}^{\text{local}}(\cdot)$ to integrate the local tumor microenvironment of the cell by attending to its neighboring cells. The local embedding $\mathbf{h}_c^{\text{local}}$ is defined as the post-attended $\widetilde{\mathbf{z}}_0$:

$$\mathbf{h}_c^{\text{local}} = [\text{LayerNorm}\left(f_{\text{Trans.}}^{\text{local}}\left(\{\widetilde{\mathbf{z}}_{\mathcal{N}_k(c)}\}_{k=0}^K\right)\right)]_{k=0}, \text{ where } \mathbf{h}_c^{\text{local}} \in \mathbb{R}^{512}. \tag{2}$$

Global context branch. The global context branch captures inter-patch context across the tissue by enabling attention-based interactions between patch-level CLS embeddings $\mathbf{z}_{n,\text{CLS}}$. First, $\mathbf{z}_{n,\text{CLS}}$ are linearly projected to a low-dimensional space and combined with coordinate-based positional encodings $\boldsymbol{\phi}(\mathbf{p}_n)$ based on patch coordinates in the tissue. This leads to a position-aware embedding $\widetilde{\mathbf{z}}_n = f_{\text{lin.}}^{\text{global}}(\mathbf{z}_{n,\text{CLS}}) + \boldsymbol{\phi}(\mathbf{p}_n) \in \mathbb{R}^{128}$. The embeddings $\{\widetilde{\mathbf{z}}_n\}_{n=1}^N$ are passed through a single-layer transformer module $f_{\text{Trans.}}^{\text{global}}(\cdot)$ for the inter-patch self-attention. We obtain the embeddings contextualized with global spatial context $\{\mathbf{h}_n^{\text{global}}\}_{n=1}^N$ as:

$$\{\mathbf{h}_n^{\text{global}}\}_{n=1}^N = \text{LayerNorm}(f_{\text{Trans.}}^{\text{global}}(\{\widetilde{\mathbf{z}}_n\}_{n=1}^N)), \text{ where } \mathbf{h}_n^{\text{global}} \in \mathbb{R}^{128}. \tag{3}$$

To avoid cells in patch n receiving the same context $\mathbf{h}_n^{\text{global}}$, we construct the global context for cell c within patch n as follows. We identify the four patches, denoted as $\mathcal{N}_4^{\text{patch}}(c)$, whose centers form the square enclosing the cell centroid. We then define $\mathbf{h}_c^{\text{global}}$ as the weighted average of $\mathbf{h}_n^{\text{global}}$:

$$\mathbf{h}_c^{\text{global}} = \sum\nolimits_{n\in\mathcal{N}_4^{\text{patch}}(c)} w_{c,n}\, \mathbf{h}_n^{\text{global}}, \text{ where } \sum\nolimits_{n\in\mathcal{N}_4^{\text{patch}}(c)} w_{c,n} = 1, \tag{4}$$

with the weights $w_{c,n}$ determined by the offset of the cell centroid within the quad of patch centers (**Appendix A.3.1**). The context vector $\mathbf{h}_c^{\text{global}}$ summarizes the broader tissue environment surrounding each cell and provides additional slide-level context.

3.3 Focal module for spatially sparse expressions

The second module, $\text{TIGER}_{\text{focal}}$, predicts cell-level ST directly from token-level representations encoded by a PFM. We use a two-layer multi-layer perceptron (MLP), $f_{\text{MLP}}(\cdot)$, such that $\widehat{\mathbf{y}}_c^{\text{Focal}} =$

$f_{\text{MLP}}(\mathbf{z}_c)$ (**Figure 1**). $\text{TIGER}_{\text{focal}}$ serves as a simple yet effective model for predicting localized ST independently for each cell, which is in contrast to $\text{TIGER}_{\text{context}}$ that produces smoothed ST prediction due to correlated cell embeddings.

3.4 Merging context and focal modules

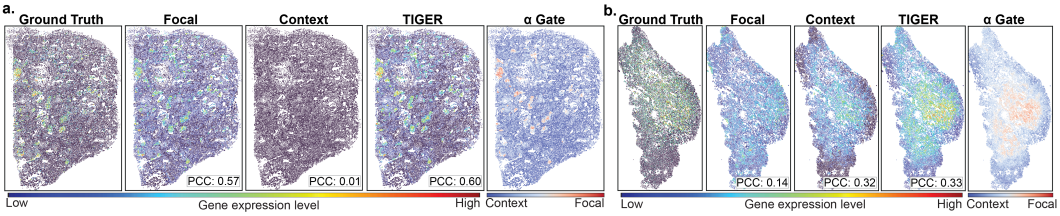


Figure 2: **TIGER components for *single-cell* ST prediction.** Each panel shows a representative gene in two lung cancer samples (**a**: *PLA2G7* gene; **b**: *MUC5AC* gene). Ground-truth expression, along with ST predictions from the different TIGER components and the spatial α gate is shown.

While $\text{TIGER}_{\text{focal}}$ and $\text{TIGER}_{\text{context}}$ both predict *cell-level* ST expression, they exhibit complementary strengths. Since $\text{TIGER}_{\text{focal}}$ processes each cell independently, it recovers spatially sparse expression patterns, such as the immune-niche gene *PLA2G7* (**Figure 2a**). However, this independence also makes it more prone to false positives, with non-zero expression predictions in cells without a true signal. In contrast, $\text{TIGER}_{\text{context}}$ captures spatially coherent expression patterns, such as *MUC5AC* in the predominant tumor region (**Figure 2b**). Its reliance on neighboring cells, however, can smooth out spatially sparse signals, resulting in false negative predictions.

To combine their complementary strengths, TIGER uses a *fully adaptive* gating mechanism. A gating score $\boldsymbol{\alpha}_c \in [0,1]^G$ determines the relative contribution of each module for every cell–gene pair. The prediction of TIGER is given by:

$$\begin{aligned} \boldsymbol{\alpha}_c &= \sigma\left(f_{\text{gate}}\left(\text{sg}\left[\mathbf{z}_c; \mathbf{h}_c^{\text{Context}}; \widehat{\mathbf{y}}_c^{\text{Focal}} - \widehat{\mathbf{y}}_c^{\text{Context}}\right]\right)\right), \\ \widehat{\mathbf{y}}_c &= \widehat{\mathbf{y}}_c^{\text{Context}} + \boldsymbol{\alpha}_c \odot \left(\widehat{\mathbf{y}}_c^{\text{Focal}} - \widehat{\mathbf{y}}_c^{\text{Context}}\right), \end{aligned} \tag{5}$$

where $\mathbf{h}_c^{\text{Context}} \in \mathbb{R}^{512}$ is the $\text{TIGER}_{\text{context}}$ embedding of the cell, $f_{\text{gate}} : \mathbb{R}^{512+D+G} \to \mathbb{R}^G$ is a two-layer MLP, and σ, $\odot$ and $\text{sg}[\cdot]$ denote the sigmoid function, element-wise multiplication and the stop-gradient operator, respectively. A high $\boldsymbol{\alpha}_c$ score indicates that the cell's own morphology is informative, whereas a low score defers to the surrounding microenvironment. The gating is performed *per-cell* and *per-gene*, which allows it to capture both sparse and spatially-coherent genes.

3.5 Training

We jointly train the TIGER prediction modules while keeping the image encoder frozen. The stop-gradient blocks gating gradients to the experts, while both experts are optimized through the fused prediction. We supervise the fused prediction with a gene-wise scaled mean squared error, $\mathcal{L} = \frac{1}{G}\sum_{g=1}^{G} \text{MSE}_g/\sigma_g$, where σ_g is the standard deviation of gene g over the training cells, partially balancing genes spanning several orders of magnitude. Details can be found in **Appendix A.3**.

4 Experiments

Datasets. We assembled five Xenium ST datasets paired with H&E-stained WSIs spanning 79 tissue images with 14.7×10^6 cells from lung, breast, and colorectal tissues, predominantly comprising cancer samples. We designated three cohorts for internal cross-validation. The colorectal cohort consists of 16 sections from distinct patients (11 internal and 5 public samples) with $G = 322$ genes. The lung cohort consists of 16 tissue sections (16 patients) and $G = 302$ genes, and the breast cohort consists of 11 sections (11 patients) and $G = 280$ genes, all from public datasets (Takano et al. (2024); Janesick et al. (2025)). We used 5-fold cross-validation partitioned at the patient level. In addition, we designated two independent lung and breast cohorts (Bilous et al. (2026)) with 17 and 19 Xenium samples for external validation, where the models trained on the internal breast and lung cohorts are reused. Further details are provided in **Appendix A.4**.

Baselines. We compare TIGER against five baselines for high-resolution virtual ST, which differ across supervision and resolution (**Table 1**): 1) *GHIST* (Fu et al. (2025)), a fully supervised multitask framework that jointly predicts cell types, neighborhood composition, and single-cell expression from paired cellular ST data; 2) *sCellST* (Chadoutaud et al. (2026)), a weakly supervised framework that assigns per-cell expression scores aggregated to spot-level supervision; 3) *DeepSpot2Cell* (Nonchev et al. (2025b)), which aggregates multi-scale PFM embeddings per-cell under weakly supervised spot-level training; 4) *Img2ST-Net* (Zhu et al. (2025b)), a fully supervised UNet-based architecture that generates dense expression maps through image-to-image translation; and 5) *iStar* (Zhang et al. (2024)), that uses weakly supervised learning with hierarchical histology features. For *Img2ST-Net* and *iStar*, which produce bin-level predictions, each cell-level prediction is assigned by aggregating bin expressions overlapping with each cell (Zhang et al. (2024)). TIGER is trained with UNIv2 (Chen et al. (2024)), with baseline models also using the same encoder whenever possible for fair comparison. All baselines have been retrained on the same datasets as TIGER. In the external cohorts, we additionally evaluate *Phoenix* (Tran et al. (2026)), a latent flow-matching generative model. Phoenix is excluded from the internal cross-validation as it is pretrained on 22 million cell image-expression pairs and designed for zero-shot use without further fine-tuning. Additional details can be found in **Appendix A.5**.

Table 1: **Model taxonomy** by supervision (Full: cell-level GT, Weak: spot-level GT) and prediction granularity (Bin: fixed grid, Cell: per-cell).

	Full	**Weak**
Bin	Img2ST-Net	iStar
Cell	GHIST, **TIGER**	sCellST, DeepSpot2Cell

Evaluation metrics. We use seven metrics for comprehensive assessment. First, spot-based metrics (Chung et al. (2024); He et al. (2020a)) were adapted for *cell-level* evaluation. These include the Pearson Correlation Coefficient (PCC), calculated for each gene across all cells in each sample. We report both the mean PCC for all genes and for the top 50 highly predictive genes (PCC_{50}), as independently identified for each fold. Next, we introduce five sparsity-aware metrics that distinguish expressing cells (those with non-zero ground-truth expression) from non-expressing cells: (1) PCC+, PCC restricted to locations with non-zero ground-truth gene expressions; (2) AUC-PR, the area under the precision-recall curve, formulating gene detection as a binary classification, where positives are defined as cells with non-zero expressions; (3) Recall, which measures whether the model assigns high predicted expression to the true expressing cells; (4) bMAE (*balanced* Mean Absolute Error), that equally weights the absolute error on expressing and non-expressing cells, penalizing both near-zero and globally inflated predictions; and (5) bRMSE (*balanced* Root Mean Squared Error) with a similar formulation but penalising large deviations more strongly. Finally, we assess how the models recapitulate the *variance* of gene expression, which reflects cellular heterogeneity. Additional details can be found in **Appendix A.6** and **A.7.2**.

5 Results

5.1 Single-cell level spatial gene expression prediction

***Internal cross-validation*. Table 2** summarizes performance averaged across the three cohorts (lung, colorectal, and breast; $N_{\text{samples}} = 43$). TIGER achieves the best performance on every metric, and its advantage over every baseline is statistically significant (one-sided paired t-test over 43 patients, all 35 comparisons at $q < 0.001$ after Benjamini–Hochberg correction; **Table A3**). Relative to the strongest baseline, GHIST, TIGER improves PCC by 31.1% (0.295 vs. 0.225) and PCC_{50} by 11% (0.617 vs. 0.556), while also attaining the lowest bMAE (0.427) and bRMSE (0.493). On sparsity-aware metrics, TIGER achieves the highest PCC+ (0.242), AUC-PR (0.324), and Recall (0.333), demonstrating also superior ability to detect and localize expressed genes despite zero-inflation. This advantage is not driven by a single cohort: TIGER ranks first on all seven metrics in each of the lung, colorectal, and breast cohorts (**Table A4**). For Img2ST-Net and iStar, we verify that the gap is not introduced by conversion to cell-level predictions (**Table A2**).

TIGER is also markedly more efficient: it trains $15\times$ faster per epoch than GHIST with $5.6\times$ less GPU memory (**Table A5**), as token embeddings are extracted once on a fixed grid and shared across cells. In addition, TIGER shows higher per-gene concordance with ground-truth expression *variance* than all baselines (**Figure A2**), better capturing the biological heterogeneity across cells.

Table 2: ***Single-cell* ST prediction evaluation.** Performance averaged across the three internal cohorts ($N_{samples} = 43$). The training strategy (Full- vs. weakly-supervised) and resolution (bin vs. cell) are specified under each model. All predictions are performed at cell level. Best performance in **bold**, second best underlined; standard deviation across folds in parentheses.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Img2ST-Net (Full, bin)	0.008 (±0.011)	0.071 (±0.027)	0.000 (±0.009)	0.531 (±0.024)	0.614 (±0.020)	0.190 (±0.030)	0.191 (±0.030)
iStar (Weak, bin)	0.123 (±0.021)	0.356 (±0.045)	0.059 (±0.021)	0.485 (±0.019)	0.524 (±0.023)	0.246 (±0.022)	0.239 (±0.024)
sCellST (Weak, cell)	0.129 (±0.017)	0.351 (±0.042)	0.091 (±0.015)	0.464 (±0.011)	0.552 (±0.014)	0.250 (±0.020)	0.244 (±0.023)
DeepSpot2Cell (Weak, cell)	0.208 (±0.023)	0.445 (±0.054)	0.157 (±0.021)	0.489 (±0.014)	0.547 (±0.020)	0.291 (±0.023)	0.281 (±0.024)
GHIST (Full, cell)	0.225 (±0.020)	0.556 (±0.048)	0.170 (±0.014)	0.445 (±0.014)	0.521 (±0.016)	0.292 (±0.019)	0.285 (±0.021)
TIGER (ours) (Full, cell)	**0.295** (±0.024)	**0.617** (±0.038)	**0.242** (±0.022)	**0.427** (±0.014)	**0.493** (±0.017)	**0.333** (±0.022)	**0.324** (±0.025)

***External evaluation*.** In addition to the cross-validation within the internal cohorts, we assess generalization by applying the models trained on the internal lung and breast cohorts to two independent external Xenium datasets ($N_{samples} = 36$; $N_{cells} = 4.6 \times 10^6$) without further fine-tuning. TIGER transfers markedly better than the other baselines and Phoenix, a large zero-shot pan-cancer model (**Table 3**; with additional metrics stratified per cancer cohort in **Tables A6, A7**). Relative to DeepSpot2Cell, the strongest baseline in this task, TIGER improves PCC by +112% (0.091 *vs.* 0.193) and PCC_{50} by +108% (0.103 *vs.* 0.214) while also yielding substantial improvements across error and sparsity-aware metrics. Slide-level paired tests confirm these gains: on the breast cohort, TIGER is significantly better than GHIST on all seven metrics ($q < 0.05$), and on the lung cohort on all correlation and sparsity-aware detection metrics ($q < 10^{-5}$). This gain also holds across spatial scales, with TIGER performing consistently best as the predictions are aggregated to progressively coarser resolutions, even outperforming a wider range of baselines at ST spot-level resolution (**Appendix A.7.4**).

Table 3: ***Single-cell* ST prediction performance in the external cancer cohorts.** Best in **bold**, second best underlined; standard deviation in parentheses.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	Recall (↑)
Img2ST-Net	0.006 (±0.008)	0.011 (±0.009)	0.003 (±0.014)	0.493 (±0.026)	0.140 (±0.042)
iStar	0.045 (±0.032)	0.054 (±0.035)	0.017 (±0.027)	0.459 (±0.028)	0.165 (±0.050)
sCellST	0.050 (±0.036)	0.058 (±0.037)	0.040 (±0.028)	0.514 (±0.030)	0.164 (±0.053)
DeepSpot2Cell	0.091 (±0.051)	0.103 (±0.055)	0.085 (±0.050)	0.475 (±0.026)	0.188 (±0.058)
GHIST	0.085 (±0.059)	0.092 (±0.060)	0.069 (±0.046)	0.449 (±0.027)	0.181 (±0.059)
Phoenix	0.074 (±0.044)	0.084 (±0.050)	0.077 (±0.048)	0.450 (±0.027)	0.174 (±0.048)
TIGER (ours)	**0.193** (±0.104)	**0.214** (±0.113)	**0.169** (±0.098)	**0.437** (±0.034)	**0.244** (±0.077)

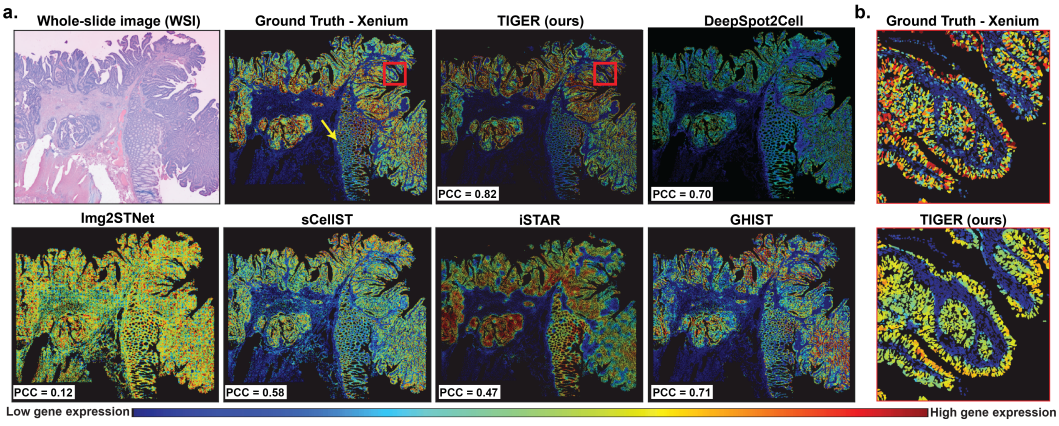


Figure 3: **Visualization of TIGER prediction. a.** Visualization of *CEACAM5* expression, a colorectal cancer cell-associated marker, that shows tumor-enriched expression. The yellow arrow indicates adjacent normal mucosa with reduced expression relative to the tumor epithelium. **b.** Zoomed red-box region for Xenium and TIGER single-cell expression highlighting dysplastic epithelium.

Visualization example. An example visualization of *CEACAM5* gene, a colorectal cancer cell-associated marker, shows that TIGER recovers the elevated expression in superficial dysplastic as well as invasive malignant epithelial cells while maintaining lower expression in adjacent normal mucosa (**Figure 3a**, yellow arrow). In contrast, other methods capture a general increase in epithelial expression but fail to preserve this fine-grained distinction. TIGER maintains this specificity at the single-cell level (**Figure 3b**). Beyond highly expressed targets, TIGER also better captures spatial patterns for sparse genes expressed in a small fraction of cells such as *CD163*, spatially confined to M2-macrophage aggregates (**Figure A4b**). Additional examples are depicted in **Figures A4, A5**.

5.2 Design choices for TIGER

Pathology foundation model ablation. We evaluate how the performance changes with different PFMs, using Kaiko (Aben et al. (2024)), Virchow2 (Zimmermann et al. (2024)), and UNIv2 (Chen et al. (2024)) (**Table 4**, **Table A8**). These are ViT-L or ViT-H backbones that differ in pretraining strategy and data. We also benchmark against EVA-02 (Fang et al. (2023)), a ViT model pretrained on ImageNet, as an out-of-domain encoder. We observe that cell embeddings built from PFM features consistently outperform EVA-02, underscoring the benefits of in-domain PFM representations. TIGER achieves the strongest performance with UNIv2, followed by Virchow2.

Table 4: **Pathology foundation model ablations.** TIGER macro-averaged across internal cohorts. Best in **bold**, second best underlined.

Model	PCC	PCC_{50}	PCC+	Recall	AUC-PR
EVA-02	0.200 (± 0.022)	0.474 (± 0.046)	0.135 (± 0.016)	0.278 (± 0.019)	0.271 (± 0.021)
Kaiko	0.286 (± 0.026)	0.601 (± 0.042)	0.223 (± 0.025)	0.324 (± 0.021)	0.315 (± 0.025)
Virchow2	<u>0.295</u> (± 0.027)	<u>0.611</u> (± 0.039)	<u>0.234</u> (± 0.022)	<u>0.329</u> (± 0.021)	<u>0.320</u> (± 0.024)
UNIv2	**0.295** (± 0.024)	**0.617** (± 0.036)	**0.242** (± 0.023)	**0.333** (± 0.021)	**0.324** (± 0.024)

Expert ablation. We assess the performance behavior of each expert, $\text{TIGER}_{\text{focal}}$ and $\text{TIGER}_{\text{context}}$ (**Table 5**). We observe that the two experts are complementary. $\text{TIGER}_{\text{focal}}$ alone already outperforms every baseline (**Table 2**), which we attribute to the PFM-derived cell embedding, and it clearly dominates the *sparsity-aware* metrics, improving PCC+, Recall and AUC-PR by 19.5%, 8.0% and 5.1% respectively over $\text{TIGER}_{\text{context}}$. In contrast, $\text{TIGER}_{\text{context}}$ attains a higher overall PCC, surpassing $\text{TIGER}_{\text{focal}}$ by 3.5%. These overall trends are preserved across cohorts (**Table A9**). This confirms that $\text{TIGER}_{\text{focal}}$ is sensitive to localized expression signals without exploiting tissue context. Conversely, $\text{TIGER}_{\text{context}}$ leverages information from the cell and tissue microenvironment to produce spatially smoother predictions albeit with less sensitivity towards sparse expression. Within $\text{TIGER}_{\text{context}}$, we observe that the integration of *global* and *local* context consistently outperforms either of the components alone (**Table A10**).

Table 5: **Model design ablations.** TIGER experts and gating strategies, macro-averaged across the three internal cohorts. Best in **bold**, second best underlined.

Model	PCC	PCC_{50}	PCC+	Recall	AUC-PR
$\text{TIGER}_{\text{focal}}$	0.286 (± 0.027)	0.610 (± 0.043)	<u>0.239</u> (± 0.026)	<u>0.330</u> (± 0.023)	<u>0.320</u> (± 0.026)
$\text{TIGER}_{\text{context}}$	**0.296** (± 0.025)	0.609 (± 0.038)	0.200 (± 0.020)	0.306 (± 0.019)	0.305 (± 0.022)
TIGER (*Average*)	0.280 (± 0.026)	0.608 (± 0.040)	0.236 (± 0.023)	0.328 (± 0.022)	0.319 (± 0.025)
TIGER (*Late fusion*)	0.280 (± 0.025)	<u>0.614</u> (± 0.042)	0.236 (± 0.022)	0.330 (± 0.021)	0.320 (± 0.024)
TIGER (*Gene adaptive*)	0.287 (± 0.026)	0.608 (± 0.042)	0.236 (± 0.024)	0.328 (± 0.021)	0.319 (± 0.025)
TIGER (*Fully adaptive*)	<u>0.295</u> (± 0.024)	**0.617** (± 0.038)	**0.242** (± 0.022)	**0.333** (± 0.022)	**0.324** (± 0.025)

Gating ablation. While the previous results naturally motivate combining both experts to leverage complementary strengths, a naive combination would risk over-smoothing localized expression signals. To determine the optimal strategy for combining the two experts, we evaluated three strategies in addition to the *fully adaptive* scheme that TIGER employs,

$$\begin{aligned} (\textit{averaging}) \quad & \widehat{\mathbf{y}}_c = 0.5\widehat{\mathbf{y}}_c^{\text{Focal}} + 0.5\widehat{\mathbf{y}}_c^{\text{Context}} \quad (\textit{late fusion}) \quad \widehat{\mathbf{y}}_c = f_{\text{MLP}}([\mathbf{h}_c^{\text{Focal}}; \mathbf{h}_c^{\text{Context}}]) \\ (\textit{gene adaptive}) \quad & \widehat{\mathbf{y}}_c = \boldsymbol{\alpha} \odot \widehat{\mathbf{y}}_c^{\text{Focal}} + (1-\boldsymbol{\alpha}) \odot \widehat{\mathbf{y}}_c^{\text{Context}}, \ \boldsymbol{\alpha}_c = \boldsymbol{\alpha} \text{ for all } c. \end{aligned} \tag{6}$$

As shown in **Table 5**, the three strategies often fall short of individual expert performance instead of synergistically combining them. This can be attributed to *average* and *gene adaptive* sharing the same weight across cells, limiting adaptation to each cell's local environment. *Fully adaptive* achieves the strongest performance across all metrics, with significant improvements in all seven metrics against *late fusion* and in six metrics against *average* and *gene adaptive* (**Table A11**). **Figure 4** shows an example, where *gene adaptive* fails to integrate both experts, whereas the *fully adaptive* mechanism keeps *SCUBE2* spatially coherent within the expressing tumor cells while suppressing it in the surrounding stroma. Additional details can be found in **Appendix A.8.2**.

Based on these results, TIGER adopts the *fully adaptive* gate to flexibly combine the two experts at each gene and cell. TIGER nearly matches or exceeds the two experts (**Table 5**). This advantage holds gene by gene: TIGER matches or beats the stronger expert for that gene on all six per-gene metrics at once in at least 64% of genes in every cohort, and in over 80% when scored on PCC alone. The spatially adaptive formulation lets the model borrow from each expert where that expert is reliable: for example, it suppresses $\text{TIGER}_{\text{focal}}$'s false positives while keeping its sensitivity to spatially sparse expression, recovering *CXCL13* in tertiary lymphoid structure (TLS) immune niches (**Figure A6**), and it can sharpen $\text{TIGER}_{\text{context}}$'s predictions without losing their spatial coherence (**Figure A7**).

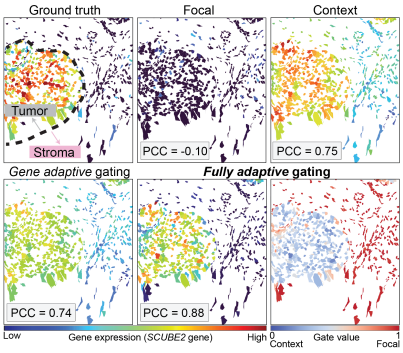


Figure 4: ***Fully adaptive gating* example.** Tumor/stroma annotations included in ground truth inset.

5.3 TIGER ENHANCES CELL PHENOTYPING PERFORMANCE

Cell phenotyping is a central requirement for answering numerous biological questions in single-cell ST data, by enabling the study of cell-to-cell interactions and tumor niches (Barrera et al. (2023); Kiemen et al. (2026)). While ML-based cell phenotyping can be achieved directly from H&E images (Hörst et al. (2024); Baumann et al. (2024)), virtual single-cell ST can provide an additional supervisory signal for cell classification. Cell phenotyping is performed in the colorectal cohort using SINGLER (Aran et al. (2019)), which assigns cell types from ST data based on a reference single-cell RNA-seq dataset. **Appendix A.9** includes a detailed explanation. Overall, we identified seven cell types that can be reliably identified based on Xenium: epithelial, tumor, stroma, endothelial, smooth muscle, myeloid, and lymphoid immune cells (**Figures 5**, **A8**, **A9**).

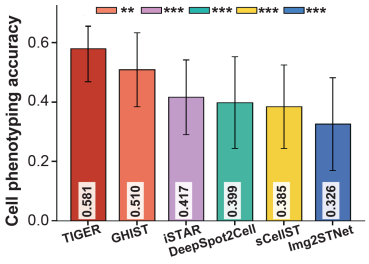


Figure 5: **Cell phenotyping from virtual ST.** Pooled accuracy, mean and std across samples. Paired t-test, TIGER *vs.* each baseline: $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

TIGER performs cell phenotyping with an accuracy of 0.581 (**Figure 5**, **Figure A10**). We observe that virtual ST prediction and cell phenotyping performances are correlated. TIGER significantly surpasses all baselines, including the second-best model GHIST, by +13.9% (0.581 vs. 0.510). It achieves strong performance for tumor epithelial cells (recall 96.4%), the most clinically relevant class for colorectal cancer. Most misclassifications concentrate between related classes, such as tumor versus non-malignant epithelium, which the targeted panel lacks the markers to separate.

5.4 TIGER PREDICTION ENHANCES CLINICAL DOWNSTREAM TASKS

We evaluated TIGER across two breast cancer downstream settings: recovery of established molecular phenotypes and added value beyond histology for treatment response prediction. As a direct test of molecular fidelity, we predicted HER2 and ER status in 166 patients (Sammut et al. (2022)) using established breast cancer gene-expression signatures (Zhao et al. (2019)). TIGER achieves AUROCs of 0.712 for HER2 and 0.747 for ER, versus 0.655 and 0.712 for image-based prediction ($p < 0.001$ and $p = 0.008$, respectively), while outperforming all virtual ST baselines (**Table A12**).

We further evaluated trastuzumab response prediction in 62 HER2-positive patients (Sammut et al. (2022)) using a multimodal model combining H&E and virtual ST features. Incorporating TIGER improves AUROC from 0.620 to 0.658, and surpassing all virtual ST baselines (**Table A12**), with a significant gain over image-only prediction ($p = 0.006$). Further details are in **Appendix A.10**.

6 DISCUSSION

We introduced TIGER, a *cell-level* virtual ST framework that flexibly combines fine-grained morphology with local and global spatial context. TIGER significantly outperforms six virtual ST approaches across a comprehensive set of metrics and on downstream tasks such as cell phenotyping and treatment response prediction. Future work includes extending TIGER to recent single-cell assays with broader transcriptomic coverage, and evaluating transfer across additional cancer types.

AI USE STATEMENT

In this work, we used generative AI tools for results interpretation. We have not used generative AI tools to autonomously conduct experiments or generate experimental results, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for language editing and software coding assistance. AI-assisted interpretations were independently verified against the underlying experimental results, while AI-assisted code and text were manually reviewed and validated for correctness, scientific accuracy, and consistency with the reported analyses. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses human-derived histopathology and spatial transcriptomics (ST) data. Breast and lung cancer samples with corresponding ST datasets were obtained from previously published or publicly released sources and were used in accordance with their respective data-use terms and original ethical approvals. Preprocessed data derived from these publicly available cohorts is available at https://anonymous-hf.com/a/xgn0n1j4ylek/. The colorectal cancer cohort comprises internal and external datasets. The internal colorectal cancer samples were acquired under appropriate IRB approval, details of which will be disclosed upon acceptance, and the corresponding data will be made publicly available. The downstream treatment-response experiments constitute secondary analyses of previously collected cohorts, for which ethical approval was obtained in the original studies.

TIGER is intended as a research tool for virtual ST rather than for clinical diagnosis or treatment decision-making. In particular, the treatment-response analyses are retrospective and do not establish prospective clinical utility. The cohorts used in this study may not capture the full diversity of patient populations, tissue preparation protocols, or imaging conditions, and these factors should be considered when interpreting or extending the model.

REPRODUCIBILITY STATEMENT

Details necessary for the reproducibility of our experiments are provided in the Appendix. **Appendix A.2** describes H&E–Xenium registration, cell segmentation and matching, cell-level embedding construction, gene selection, and expression preprocessing, while **Appendix A.3** provides detailed descriptions of the TIGER architecture, the complete training procedure, optimization settings, random seed, checkpoint selection, and inference configuration. Dataset composition and cohort construction are documented in **Appendix A.4**, and implementation and training details for all comparison methods are provided in **Appendix A.5**. Definitions of all evaluation metrics and statistical procedures are given in **Appendix A.6**, with additional experiments, architectural ablations, cell-phenotyping analyses, and the downstream treatment-response protocol described in **Appendix A.7–A.10**. All cross-validation splits are performed at the patient level to prevent overlap between training and test samples. An anonymized implementation of TIGER, together with data-processing, training, and evaluation scripts and the experimental configurations used in this work, is provided in the manuscript with the following link: https://anonymous-hf.com/a/xgn0n1j4ylek/.

REFERENCES

Nanne Aben, Edwin D de Jong, Ioannis Gatopoulos, Nicolas Känzig, Mikhail Karasikov, Axel Lagré, Roman Moser, Joost van Doorn, Fei Tang, et al. Towards large-scale training of pathology foundation models. *arXiv preprint arXiv:2404.15217*, 2024.

Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

Cristian Barrera, Germán Corredor, Vidya Sankar Viswanathan, Ruiwen Ding, Paula Toro, Pingfu Fu, Christina Buzzy, Cheng Lu, Priya Velu, Philipp Zens, et al. Deep computational image analysis of immune cell niches reveals treatment-specific outcome associations in lung cancer. *NPJ precision oncology*, 7(1):52, 2023.

Elias Baumann, Bastian Dislich, Josef Lorenz Rumberger, Iris D Nagtegaal, Maria Rodriguez Martinez, and Inti Zlobec. Hover-next: A fast nuclei segmentation and classification pipeline for next generation histopathology. In *Medical Imaging with Deep Learning*, 2024.

Mariia Bilous, Daria Buszta, Jonathan Bac, Senbai Kang, Yixing Dong, Stephanie Tissot, Sylvie Andre, Marina Alexandre Gaveta, Christel Voize, Solange Peters, et al. Resolving sensitivity, specificity and signal contamination in xenium spatial transcriptomics. *Nature Methods*, pp. 1–11, 2026.

Quentin Blampey, Hakim Benkirane, Nadège Bercovici, Kevin Mulder, Grégoire Gessain, Florent Ginhoux, Fabrice André, and Paul-Henry Cournède. Novae: a graph-based foundation model for spatial transcriptomics data. *Nature Methods*, 22(12):2539–2550, 2025.

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

Gabriele Campanella, Neeraj Kumar, Swaraj Nanda, Siddharth Singi, Eugene Fluder, Ricky Kwan, Silke Muehlstedt, Nicole Pfarr, Peter J Schüffler, Ida Häggström, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine*, 31(9):3002–3010, 2025.

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.

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.

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

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.

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

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pp. 1–13, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

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.

Chandler D Gatenbee, Ann-Marie Baker, Sandhya Prabhakaran, Ottilie Swinyard, Robbert JC Slebos, Gunjan Mandal, Eoghan Mulholland, Noemi Andor, Andriy Marusyk, Simon Leedham, et al. Virtual alignment of pathology image series for multi-gigapixel whole slide images. *Nature communications*, 14(1):4502, 2023.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

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, 2020a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020b.

Konstantin Hemker, Andrew H Song, Cristina Almagro-Pérez, Guillaume Jaume, Sophia J Wagner, Anurag Vaidya, Nikola Simidjievski, Mateja Jamnik, and Faisal Mahmood. Towards spatial transcriptomics-driven pathology foundation models. *arXiv preprint arXiv:2602.14177*, 2026.

Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical image analysis*, 94:103143, 2024.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1055–1059. Ieee, 2020.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *arXiv preprint arXiv:2506.05361*, 2025.

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

Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications*, 14(1):8353, 2023.

Amanda Janesick, Stephanie N Kravitz, Weston Stauffer, Miriam Valencia, and Sarah EB Taylor. Biomarker quantification in breast cancer using xenium in situ. *bioRxiv*, pp. 2025–12, 2025.

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew F Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024.

Ashley L Kiemen, Valentina Matos, Cristina Almagro-Pérez, André Forjaz, Alicia M Braxton Hickman, Lucie Dequiedt, Jeeun Parksong, Courtney D Cannon, Xuan Yuan, Sarah M Shin, et al. Quantitative 3d histology reveals localized immune remodeling during early pancreatic cancer progression. *Cell Press Blue*, 2026.

Antti Kiviaho, Sini K Eerola, Heini ML Kallio, Maria K Andersen, Miina Hoikka, Aliisa M Tiihonen, Iida Salonen, Xander Spotbeen, Alexander Giesen, Charles TA Parker, et al. Single cell and spatial transcriptomics highlight the interaction of club-like cells with immunosuppressive myeloid cells in prostate cancer. *Nature communications*, 15(1):9949, 2024.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pp. 255–268, 1989.

Yunhe Liu, Yibo Dai, and Linghua Wang. Spatial omics at the forefront: emerging technologies, analytical innovations, and clinical applications. *Cancer cell*, 2025.

Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

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

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

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

Kalin Nonchev, Sebastian Dawo, Karina Silina, Holger Moch, Sonali Andani, Tumor Profiler Consortium, Viktor H Koelzer, and Gunnar Rätsch. Deepspot: Leveraging spatial context for enhanced spatial transcriptomics prediction from h&e images. *MedRxiv*, pp. 2025–02, 2025a.

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, 2025b.

Michelli Faria de Oliveira, Juan Pablo Romero, Meii Chung, Stephen R Williams, Andrew D Gottscho, Anushka Gupta, Susan E Pilipauskas, Seayar Mohabbat, Nandhini Raman, David J Sukovich, et al. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics*, 57(6):1512–1523, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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.

Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A Bardwell, Wenxin Ma, Wei Cope, Ali Dariush, Sarah-Jane Dawson, Jean E Abraham, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022.

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.

Yuma Takano, Jun Suzuki, Kotaro Nomura, Gento Fujii, Junko Zenkoh, Hitomi Kawai, Yuta Kuze, Yukie Kashima, Satoi Nagasawa, Yuka Nakamura, et al. Spatially resolved gene expression profiling of tumor microenvironment reveals key steps of lung adenocarcinoma development. *Nature Communications*, 15(1):10637, 2024.

Manuel Tran, Rushin H. Gindra, Philipp Putze, Kang Senbai, Giovanni Palla, Tina Kos, Chiara Falcomatà, Chen Wang, Ruifeng (Ray) Guo, Melanie Boxberg, Luc M. Berclaz, Lars H. Lindner, Linda Bergmayr, Thomas Knösel, Philipp Jurmeister, Frederick Klauschen, Krisztian Homicsko, Raphael Gottardo, Markus Eckstein, Christian Matek, Andreas Mock, Fabian J. Theis, Dieter Saur, and Tingying Peng. Pan-cancer virtual spatial transcriptomics from routine histology with phoenix. *bioRxiv*, 2026. doi: 10.64898/2026.04.25.720812. URL `https://www.biorxiv.org/content/early/2026/04/29/2026.04.25.720812`.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

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

Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5039–5048, 2023.

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.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5), 2022.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

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.

Ruikun Zhang, Yan Yang, and Liyuan Pan. From spots to pixels: Dense spatial gene expression prediction from histology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19791–19800, 2026.

Shuang G Zhao, William S Chen, Rajdeep Das, S Laura Chang, Scott A Tomlins, Jonathan Chou, David A Quigley, Ha X Dang, Travis J Barnard, Brandon A Mahal, et al. Clinical and genomic implications of luminal and basal subtypes across carcinomas. *Clinical cancer research*, 25(8): 2450–2457, 2019.

Junchao Zhu, Ruining Deng, Junlin Guo, Tianyuan Yao, Siqi Lu, Chongyu Qu, Juming Xiong, Yanfan Zhu, Zhengyi Lu, Yuechen Yang, et al. Computer vision methods for spatial transcriptomics: A survey. *bioRxiv*, pp. 2025–10, 2025a.

Junchao Zhu, Ruining Deng, Junlin Guo, Tianyuan Yao, Juming Xiong, Chongyu Qu, Mengmeng Yin, Yu Wang, Shilin Zhao, Haichun Yang, et al. Img2st-net: efficient high-resolution spatial omics prediction from whole-slide histology images via fully convolutional image-to-image learning. *Journal of Medical Imaging*, 12(6):061410–061410, 2025b.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Appendix

Appendix Content

A.1 Sparsity analysis . . . 16
A.2 Data preprocessing . . . 16
A.3 Architectural and Training details . . . 17
A.3.1 Context module: Global context branch embedding . . . 17
A.3.2 Training details . . . 18
A.4 Datasets . . . 18
A.4.1 Internal datasets . . . 18
A.4.2 External datasets . . . 19
A.5 Baselines . . . 19
A.6 Evaluation details . . . 21
A.7 Additional experiments . . . 22
A.7.1 TIGER's performance on internal cancer cohorts . . . 22
A.7.2 Preservation of gene expression variability in inferred single-cell ST . . . 24
A.7.3 Computational considerations . . . 24
A.7.4 TIGER's performance on external cancer cohorts . . . 25
A.7.5 Additional visualization examples . . . 25
A.8 Design choices for TIGER architecture . . . 30
A.8.1 Visualizations and interpretability analysis . . . 30
A.8.2 TIGER components and gating mechanism ablations . . . 30
A.9 Cell phenotyping considerations . . . 34
A.10 Downstream clinical application . . . 35
A.11 Broader impacts . . . 38

A.1 SPARSITY ANALYSIS

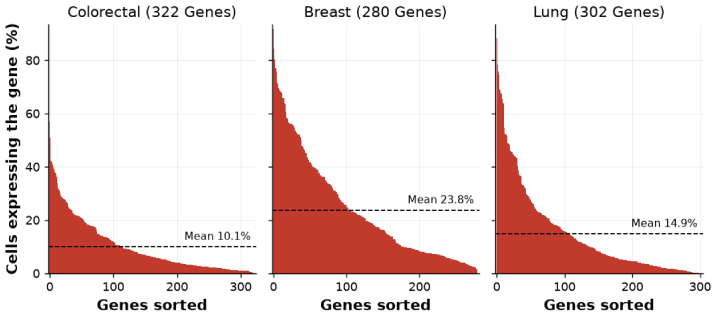


Figure A1: **ST at single-cell resolution suffers from extreme sparsity.** Percentage of cells expressing each gene, averaged over the slides of each internal dataset (refer to **Section A.4.1**), with genes sorted from most to least detected and the dashed line at the cohort mean (10.1%, 23.8% and 14.9%). Analysis on Xenium (ground truth) data.

A.2 DATA PREPROCESSING

H&E and Xenium Alignment. Raw Xenium outputs include DAPI-stained fluorescence images alongside transcript coordinates and cell segmentation boundaries that are natively registered to the DAPI coordinate system. To align the companion H&E whole-slide image (WSI) with the Xenium coordinate space, we employ VALIS (Gatenbee et al. (2023)), an image registration framework for histopathological images. VALIS performs deformable registration between the H&E bright-field image and the Xenium DAPI image, producing a warp field that is subsequently applied to H&E pixel coordinates. Following alignment, all spatial objects in the Xenium coordinate space, including transcript positions, cell segmentation polygons, and nucleus segmentation polygons, are transformed into the registered H&E coordinate system. The resulting multi-modal data are stored in HEST-compatible format (Jaume et al. (2024)) for downstream processing.

Cell Segmentation and H&E–Xenium Cell Matching. TIGER is agnostic to the choice of cell segmentation method used to derive histology-based cell boundaries. In our implementation, we use CellViT (Hörst et al. (2024)) on the aligned H&E WSI to detect and delineate individual nuclei. Detected nuclei are matched to Xenium native cell identifiers based on spatial overlap; specifically, a match is established if at least 50% of the area of a CellViT nucleus mask overlaps with a corresponding Xenium cell boundary. This establishes a shared `id_histology` mapping stored in per-sample cell AnnData files. We further evaluate the effect of the boundary source by replacing the histology-derived CellViT masks with the cell boundaries provided directly by Xenium (cell contours derived from fluorescence morphology images). Using CellViT-derived boundaries results in only a slight decrease in performance compared with the native Xenium boundaries, despite relying solely on information available from the H&E WSI at inference time (**Table A1**).

Table A1: **Ablation study of the cell boundary source.** Performance of TIGER using either Xenium-native cell boundaries, derived from fluorescence-based Xenium morphology images, or H&E-derived cell boundaries obtained with CellViT (Hörst et al. (2024)). Values are macro-averaged over the lung, breast and colorectal cohorts. While the Xenium-native boundaries yield slightly higher performance, they rely on fluorescence imaging that is not available in standard H&E WSI-only settings; the CellViT-based setting therefore reflects the practical deployment scenario of TIGER. The best result for each metric in **bold**.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
TIGER (Xenium boundary)	**0.301** (± 0.025)	**0.625** (± 0.040)	**0.251** (± 0.023)	**0.426** (± 0.013)	**0.492** (± 0.014)	**0.336** (± 0.022)	**0.328** (± 0.025)
TIGER (CellViT boundary)	0.295 (± 0.024)	0.617 (± 0.038)	0.242 (± 0.022)	0.427 (± 0.013)	0.493 (± 0.015)	0.333 (± 0.022)	0.324 (± 0.025)

Cell Embedding Extraction. To efficiently extract *cell-level* histology embeddings, we construct a spatial grid over the H&E image and extract pathology foundation model (PFM) token features, which we then assign to individual cells based on token-cell overlap, rather than independently cropping and encoding the H&E image for each cell. We first tessellate the H&E image into non-overlapping $128\,\mu\text{m} \times 128\,\mu\text{m}$ regions, referred to as *large bins*, each corresponding to a 256×256-pixel patch at $0.5\,\mu\text{m/px}$ (equivalent to $20\times$ magnification). Each patch is rescaled to 224×224 pixels and encoded once using UNIv2 (Chen et al. (2024)), a ViT-H/14 pathology foundation model pre-trained on large-scale histopathology data. UNIv2 produces a 16×16 grid of 256 patch-token embeddings with feature dimension $d = 1536$. Each element of this 16×16 grid corresponds to $8\,\mu\text{m} \times 8\,\mu\text{m}$ regions, referred to as *small bins*. For each segmented cell, we identify the small bins overlapping its CellViT segmentation mask and compute its histology embedding as the normalized weighted average of the corresponding UNIv2 token representations, with weights determined by their spatial overlap with the cell mask. On average, each cell was represented by a weighted combination of 3 – 4 tokens.

Gene Selection. We utilize all genes available in the Xenium gene panel for each internal cohort. Negative and quality control probes (`NegControlCodeword`, `NegControlProbe`, `UnassignedCodeword`) are excluded. Genes are selected as the intersection of panel genes across all samples in each cohort to ensure consistent gene vocabularies. This results in $G = 302$ genes for the lung cancer cohort, $G = 280$ genes for the breast cancer cohort, and $G = 322$ genes for the colorectal cancer cohort. For the external lung and breast cohorts, we do not use their full native gene panels; instead, we restrict evaluation to genes shared with the corresponding internal training panel, yielding an intersection of $G = 93$ genes for the lung external cohort and $G = 61$ genes for the breast external cohort.

Gene Expression Transformation. Raw per-cell transcript counts are transformed element-wise using $\log(1 + x)$, without library-size normalization (CPM/total-count scaling):

$$y_{c,g} = \log(1 + x_{c,g})\,, \tag{7}$$

where $x_{c,g}$ denotes the raw transcript count of gene g in cell c. All models are trained to predict the log-transformed counts, and evaluation is performed in the same space.

Baseline-specific Spot Assignment For baselines (GHIST, sCellST, DeepSpot2Cell) operating on spot-level inputs, a cell-to-spot assignment map is pre-computed for each sample by assigning cells whose nucleus centroids fall within a given spot boundary to that spot.

A.3 Architectural and Training details

A.3.1 Context module: Global context branch embedding

The weights $w_{c,n}$ of **Eq. 4** are bilinear interpolation weights, which sample the patch-level context field continuously at the cell centroid rather than assigning every cell of patch x_n the same vector $\mathbf{h}_n^{\text{global}}$. Let $\mathbf{q}_c = (q_{c,1}, q_{c,2})$ be the centroid of cell c and let n_0 be the patch whose center $\mathbf{p}_{n_0}$ is closest to it. We express the offset of the centroid from that center as a fraction of the patch width W along each axis, $\omega_i = \min\left(|q_{c,i} - p_{n_0,i}|/W,\, 1\right)$ for $i \in \{1, 2\}$, and let $\delta_i = \text{sign}(q_{c,i} - p_{n_0,i})$. The set $\mathcal{N}_4^{\text{patch}}(c)$ is then the quad enclosing $\mathbf{q}_c$, formed by x_{n_0} and its lattice neighbors displaced by δ_1, δ_2, and both, with the separable weights $(1-\omega_1)(1-\omega_2)$, $\omega_1(1-\omega_2)$, $(1-\omega_1)\omega_2$ and $\omega_1\omega_2$ respectively. Patches of the quad that fall outside the tissue, at a tissue or slide border, are given zero weight and the remaining weights are renormalized to sum to one, so that border cells degrade gracefully to a three-, two- or one-patch average.

These weights vary smoothly and monotonically with the position of the centroid within the quad: a cell lying on a patch center recovers that patch alone, a cell lying on a patch center recovers that patch alone, a cell at the midpoint between two patch centers averages those two equally, and a cell at the center of the quad averages all four equally. Cells within one patch therefore no longer share an identical context vector, spatially adjacent cells receive similar contexts even across a patch border, and the gradient from a single cell reaches four patch embeddings instead of one, which densifies the training signal reaching the aggregator $f_{\text{Trans.}}^{\text{global}}$.

A.3.2 TRAINING DETAILS

We train TIGER using a 5-fold cross-validation. All experiments use a fixed random seed of 2026 for reproducibility. Mixed-precision training (FP16) is enabled throughout via PyTorch automatic mixed precision (AMP).

End-to-end training. $\text{TIGER}_{\text{focal}}$, $\text{TIGER}_{\text{context}}$ and the gating function f_{gate} are optimized jointly, in a single stage. The image encoder f_{enc} (UNIv2) remains frozen throughout, and all inputs are derived from its cached token grid, so the only trainable parameters are those of the two experts and the gate. The gate is made non-competitive with the experts by detaching its inputs – the cell's own pooled input feature, $\text{TIGER}_{\text{context}}$'s internal representation, and the two experts' logit disagreement –, so that the gate cannot shape the experts' representations through its inputs. The final layer of the router is zero-initialized, so all gates are exactly 0.5 at initialization and optimization begins from the average of the two experts.

Objective. All parameters are optimized with a gene-wise scaled mean squared error on log-transformed cell-level expression, which prevents the few high-variance genes of the gene panel from dominating the gradient:

$$\mathcal{L} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{\sigma_g}\cdot\frac{1}{C}\sum_{c=1}^{C}\left(\widehat{y}_{c,g} - y_{c,g}\right)^2, \tag{8}$$

where C is the number of cells in a mini-batch, $\widehat{y}_{c,g}$ and $y_{c,g}$ are the predicted and ground-truth log-transformed expression of gene g in cell c, and σ_g is the standard deviation of gene g computed from the *training* cells of that fold only (recomputed independently per fold, so no validation or test statistics leak into the loss).

Optimization. All models are trained with the AdamW optimizer (lr $= 10^{-4}$) with a learning rate schedule composed of (i) a 5-epoch linear warmup from $0.01 \times$ lr to lr, followed by (ii) cosine annealing decay to a minimum of 10^{-6}. Training runs for up to 100 epochs with early stopping (patience = 10 epochs) monitored on validation *cell-level* Pearson Correlation Coefficient (PCC). Training batches contain 32 contiguous groups of 256 cells (8,192 cells per optimization step), with 4 data loader workers; at inference, cells are processed in chunks of 4,096 while the whole-slide context is computed once per slide.

A.4 DATASETS

All datasets used in this study were generated using the 10x Genomics Xenium In Situ platform, which provides subcellular-resolution transcript coordinates along with cell segmentation boundaries. Each tissue section is accompanied by a paired post-Xenium H&E image. The data can be accessed at our anonymous database and code repository: https://anonymous-hf.com/a/xgn0n1j4ylek/ and will be made publicly available upon acceptance.

A.4.1 INTERNAL DATASETS

Colorectal Cancer We collected an internal Xenium colorectal adenocarcinoma dataset, consisting of 11 formalin-fixed, paraffin-embedded (FFPE) CRCs sampled from different locations throughout the colorectum. All tumours were locally advanced (pT3–pT4) with nodal stages ranging from pN0 to pN2. The cohort was deliberately composed to capture major molecular subtypes of CRC, in particular microsatellite-stable (MSS) and microsatellite-unstable (MSI) tumors. Spatial transcriptomic profiling was performed using the 10x Genomics Xenium platform with the Human Colon Gene Expression Panel supplemented by custom add-on genes (322 pre-designed[1] + 100 custom genes, 422 genes total). Paired post-Xenium H&E images were used for morphological correlation. This internal cohort of 11 sections was complemented with five publicly available colorectal sections from 10x Genomics, for a total of 16.

[1] https://www.10xgenomics.com/support/software/xenium-panel-designer/latest/tutorials/pre-designed-xenium-v1

Lung Cancer We use the Xenium lung adenocarcinoma (LUAD) dataset from Takano et al. (Takano et al. (2024)), which profiled tissue sections from lung adenocarcinoma patients spanning non-invasive to invasive stages using a custom 302-gene panel (202 pre-designed and 100 custom genes). The cohort includes a mixture of FFPE and fresh-frozen (FF) preserved samples with paired H&E images. The original study includes 17 Xenium-profiled samples; we use 16 of them, as one sample was inaccessible due to a data retrieval issue.

Breast Cancer We use a publicly available Xenium breast cancer dataset released by 10x Genomics (Janesick et al. (2025)), comprising 12 FFPE human breast tissue sections—8 ductal carcinoma in situ (DCIS), 3 invasive ductal carcinoma (IDC), and 1 normal—profiled with a custom 280-gene panel[2]. The gene panel covers breast tissue-relevant markers, basement membrane genes, established tumor biomarkers (e.g., *ERBB2*, *PGR*, *ESR1*), and 45 housekeeping gene candidates. We exclude the normal breast tissue section, using the remaining 11 cancer sections for all experiments.

A.4.2 External datasets

Lung Cancer To test generalization beyond the training/validation cohort without any fine-tuning, we evaluate our lung cancer-trained models on an independent, publicly available non-small-cell lung cancer (NSCLC) Xenium cohort from Lausanne University Hospital (CHUV)[3], comprising 22 Xenium samples from 10 patients profiled with three different Xenium gene panels (Xenium Prime 5K, the standard Xenium human lung panel, and a custom immuno-oncology panel). We exclude 2 patients whose only available section lacks a paired post-Xenium H&E image, and 3 further patients already represented by a Xenium Prime 5K section, a duplicate standard-panel section of the same tissue, yielding 17 Xenium samples from 8 patients for evaluation. We evaluate model predictions on the intersection between the 302-gene lung cancer training panel and the cohort's native gene panel—93 genes in total at the panel level—the set of genes for which the models were trained to make predictions.

Breast Cancer Analogously, we evaluate our breast cancer-trained models, again without fine-tuning, on an independent breast cancer Xenium cohort from the same Lausanne (CHUV) series, comprising 19 Xenium samples from 17 patients (one patient contributed 3 sections), uniformly profiled with the standard Xenium human breast gene panel. As with the lung external cohort, we restrict evaluation to the intersection of the 280-gene breast cancer training panel with the cohort's native gene panel (61 genes).

A.5 Baselines

We compare TIGER against six recent methods for high-resolution spatial gene expression prediction from histology:

GHIST: GHIST (Fu et al. (2025)) is a multitask deep learning framework that predicts single-cell gene expression from H&E patches by jointly learning four interrelated tasks: nuclei segmentation and morphology, cell-type classification, neighborhood composition estimation, and gene expression prediction. The backbone is a Unet 3+ architecture (Huang et al. (2020)) that performs pixel-level nuclei segmentation. Nucleus-level features are extracted by masking the first and last convolutional feature maps with each nucleus's segmentation mask, and are concatenated with patch-level features to form a 768-dimensional vector per cell, which is then projected to a 128-dimensional embedding. A separate MLP head classifies each cell's type from this embedding. For gene expression prediction, the per-cell embedding is passed through two fully connected layers to produce initial per-gene expression values. These values are then refined by a multi-head cross-attention module (eight heads), where the estimated neighborhood cell-type composition serves as the query and the expression features serve as the key and value; the final per-cell expression is obtained by applying ReLU to the cross-attention output. Separate prediction heads with the same architecture are maintained for immune and invasive cell types. An auxiliary module further ensures biological consistency by predicting cell type from the predicted expression, enforcing that predicted profiles correspond to meaningful cell-type signatures. During training, all losses—including segmentation

[2] https://www.10xgenomics.com/datasets/xenium-ffpe-human-breast-biomarkers

[3] GEO accession GSE311609

(cross-entropy), cell-type classification (cross-entropy), cell-type embedding consistency (cosine embedding loss scaled by 100), cell-type logit consistency (MSE), neighborhood composition (KL divergence), and gene expression (MSE)—are jointly minimized end-to-end. In our implementation, we used the cell-type classification and neighborhood composition modules but did not provide the optional averaged single-cell reference expression profiles. The basic model architecture followed the descriptions provided in the original publication, while all detailed implementation aspects followed the official repository.

sCellST: sCellST (Chadoutaud et al. (2026)) adopts a Multiple Instance Learning (MIL) formulation, treating each ST spot as a bag and the individual cells detected within it as instances. A pretrained nucleus detection model (i.e. CellViT (Hörst et al. (2024)), HoverNet (Graham et al. (2019))) first segments cells from the H&E image, and cell-centered crops are extracted and encoded into embedding vectors. A feed-forward network then predicts per-cell gene expression scores, which are aggregated via mean pooling to produce spot-level predictions supervised against measured spot-level expression. In our implementation, spot-level expression is defined as the average expression of cells located within a 128μm region. For cell image embedding extraction, we extract 48×48-pixel cell-centered tiles at $0.25\ \mu\text{m}/\text{px}$ and resize them to 224×224 pixels before encoding. Because the MoCo v3 (He et al. (2020b))-pretrained ResNet-50 checkpoint used in the original study was not publicly available, we replace it with the frozen UNIv2 encoder, generating per-cell embeddings of dimension $d = 1536$. At inference, per-cell predictions are reported at the instance level.

DeepSpot2Cell: DeepSpot2Cell (Nonchev et al. (2025b)) extends DeepSpot (Nonchev et al. (2025a)) by integrating pathology foundation model embeddings at three spatial scales (cell, spot, and neighboring spot tiles), and aggregates per-cell predictions via summation to match spot-level supervision, enabling *single-cell* inference at test time. For each cell, the smallest axis-aligned square patch fully enclosing the segmented nucleus boundary is extracted from the aligned H&E WSI, resized to 224×224 pixels using bilinear interpolation with antialiasing, and encoded with the frozen UNIv2 encoder to produce a per-cell image embedding. A pseudospot is then constructed for each $128\ \mu$m bin by assembling the cell embeddings of all assigned cells together with the UNIv2 CLS-token embeddings of the bin's own H&E patch and its up to 8 spatial neighbors in a 3×3 grid (zero-padded for boundary bins). During training, cell-level predictions are summed to the spot level and supervised with spot-level expression. During evaluation, per-cell predictions are reported.

Img2ST-Net: *Img2ST-Net* (Zhu et al. (2025b)) reformulates ST prediction as an image-to-image translation task. It adopts a UNet-style fully convolutional architecture with an ImageNet-pretrained ResNet-50 encoder (He et al. (2016)), where the third and fourth residual blocks produce feature maps at 1/16 and 1/32 of the input resolution, respectively, and a single-layer upsampling decoder with skip connections from the third block recovers spatial resolution. Since the U-Net decoder consumes intermediate ResNet feature maps directly via skip connections, this architecture cannot be adapted to incorporate UNIv2 vision transformer (ViT). We extract 256×256-pixel patches from the WSI and resize them to 224×224 as model input, with random horizontal/vertical flips and 90-degree rotations applied during training. The model jointly predicts gene expression for all bins within each region in a single forward pass, with the decoded feature maps adaptively pooled to the target bin grid. A frozen MLP-based gene expression encoder provides latent reference embeddings, which are aligned with intermediate 512-dimensional image features via an InfoNCE loss Oord et al. (2018). We followed the author's implementation [4]. The model generates dense gene-expression maps at an 8μm bin resolution. We assigned these predictions to individual cells based on their area overlap with CellViT masks, enabling cell-level evaluation, as is done for *iStar*.

iStar: *iStar* (Zhang et al. (2024)) divides the whole-slide image into 4096×4096-pixel regions and extracts hierarchical histology features using HIPT (Chen et al. (2022)), a self-supervised hierarchical vision transformer. Within each region, a local vision transformer first encodes 16×16-pixel tiles into fine-grained cellular features, then aggregates them at the 256×256-pixel scale to capture broader tissue context; a global ViT subsequently models long-range dependencies across all 256×256 tiles to produce image-wide structural features. These multi-scale feature representations are concatenated with the original RGB values and fed into a feed-forward neural network, which is trained through weakly supervised learning to distribute spot-level expression across superpixels. It predicts gene expression at a fixed grid resolution with cell-level gene expression obtained by a weighted sum of the superpixel-level gene expressions, with the weight equal to the proportion of

[4] https://github.com/hrlblab/Img2ST-Net

the superpixel that overlaps with the cell mask. We followed the author's implementation[5] and use CellViT (Hörst et al. (2024)) masks as the cell segmenter that matches TIGER and baselines.

Phoenix: Phoenix (Tran et al. (2026)) is a pan-cancer foundation model that predicts single-cell gene expression directly from H&E histology via latent flow matching: a frozen DINOv2 ViT-Giant encoder conditions a cross-attention flow transformer, integrated by an ODE solver into per-cell expression profiles. Unlike every other baseline here, it is evaluated strictly zero-shot, using the publicly released `Nest` checkpoint with no fine-tuning on our training cohorts. We crop cell-centered 224×224 patches using the same segmentation pipeline as our other cell-level baselines, and evaluate Phoenix on the gene panels (93 genes for the external lung dataset and 61 genes for the external breast dataset) used for external evaluation. Phoenix is uniquely compared against the *zero-shot* version (out-of-the-box; using the pretrained checkpoints on internal breast and lung cancer without fine-tuning) of TIGER in the external lung and breast cohorts.

For any remaining implementation details, we followed the descriptions provided in the original publication. *iStar*, *sCellST*, and *DeepSpot2Cell* are *weakly supervised*, learning from spot-level aggregates without access to per-cell expression during training, while *GHIST* and *Img2ST-Net* are *fully supervised*, requiring matched high-resolution ground-truth expression for training.

A.6 Evaluation details

All metrics considered in TIGER are computed *per-gene* and then macro-averaged over the G genes of the panel, giving one value per slide. For gene g, we write $y_{c,g}$ and $\widehat{y}_{c,g}$ for the ground-truth and predicted expression in cell c, and split the C cells of a sample into the *expressing* and *non-expressing* populations of the *ground truth*:

$$P_g = \{\, c : y_{c,g} > 0 \,\}, \qquad N_g = \{\, c : y_{c,g} = 0 \,\}, \tag{9}$$

noting that $|N_g|/C \approx 0.9$ at cell level, highlighting the sparsity of single-cell ST data.

(1) PCC. The Pearson Correlation Coefficient between predicted and ground-truth expression, computed across all cells of a sample for each gene and averaged over all genes.

(2) Top-50 PCC. The same quantity restricted to the 50 most highly predictive genes, identified independently for each cross-validation fold (so that gene selection never uses test-fold information).

(3) PCC+. The Pearson Correlation Coefficient restricted to the cells in P_g, i.e. computed only where the gene is actually expressed, which prevents the dominant zero population from inflating the correlation.

(4) bMAE. Plain Mean Absolute Error (MAE) can be misleading at *cell-level*: since roughly 90% of the *(cell, gene)* entries are exact zeros, a model predicting approximately zero everywhere obtains an excellent MAE while not closely recapitulating the spatial expression patterns. We therefore weight the two populations equally, in the same spirit as balanced accuracy for imbalanced classification. Defining the error on the expressing and non-expressing cells separately,

$$\mathrm{MAE}_g^{+} = \frac{1}{|P_g|} \sum_{c \in P_g} \left| \widehat{y}_{c,g} - y_{c,g} \right|, \qquad \mathrm{MAE}_g^{-} = \frac{1}{|N_g|} \sum_{c \in N_g} \left| \widehat{y}_{c,g} \right|, \tag{10}$$

the *balanced* mean absolute error is defined as their average, macro-averaged over genes:

$$\mathrm{bMAE} = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{2} \left(\mathrm{MAE}_g^{+} + \mathrm{MAE}_g^{-} \right). \tag{11}$$

Neither degenerate strategy can win: predicting zero everywhere makes MAE_g^{-} vanish but leaves $\mathrm{MAE}_g^{+} = \frac{1}{|P_g|} \sum_{c \in P_g} y_{c,g}$, the mean expression of gene g over the cells that express it, while globally inflating the predictions lowers MAE_g^{+} at the cost of MAE_g^{-}. Note that conventional MAE is exactly the sparsity-weighted mixture of the two terms, $\mathrm{MAE}_g = (1 - \pi_g)\, \mathrm{MAE}_g^{-} + \pi_g\, \mathrm{MAE}_g^{+}$ with $\pi_g = |P_g|/C$, which is why it is dominated by the zeros. Unlike PCC+, bMAE is deliberately scale-sensitive.

[5] `https://github.com/daviddaiweizhang/istar`

(5) bRMSE. We define the *balanced* root mean squared error in the same spirit as bMAE, computing the error on the expressing and non-expressing cells separately,

$$\mathrm{RMSE}_g^+ = \Big(\tfrac{1}{|P_g|}\sum_{c\in P_g}(\widehat{y}_{c,g}-y_{c,g})^2\Big)^{1/2}, \qquad \mathrm{RMSE}_g^- = \Big(\tfrac{1}{|N_g|}\sum_{c\in N_g}(\widehat{y}_{c,g}-y_{c,g})^2\Big)^{1/2}, \tag{12}$$

and averaging the two root mean squared errors, macro-averaged over genes:

$$\mathrm{bRMSE} = \tfrac{1}{2G}\sum_{g=1}^{G}\big(\mathrm{RMSE}_g^+ + \mathrm{RMSE}_g^-\big). \tag{13}$$

Like bMAE it is scale-sensitive and removes the domination of the much larger zero-expression population, but it penalises large deviations more strongly.

(6) Recall. This metric measures whether the model assigns high predicted expression to the cells that truly express the gene. For gene g, let T_g denote the set of the $K_g = |P_g|$ cells with the largest predicted expression $\widehat{y}_{c,g}$, i.e. we retain exactly the same number of cells that express a given gene considering the ground truth gene expression. Recall is then defined as the fraction of the truly expressing cells that are recovered in that top-K_g set, macro-averaged over genes:

$$\mathrm{Recall} = \frac{1}{G}\sum_{g=1}^{G}\frac{|T_g \cap P_g|}{|P_g|}. \tag{14}$$

The metric depends only on the ranking of cells induced by the prediction, not on its magnitude. Since $K_g = |P_g|$, recall equals precision at this cutoff (also known as R-precision), so missed expressing cells and false positives are penalized equally, and a random ranking attains $|P_g|/N$ in expectation.

(7) AUC-PR. Similar to *Recall*, gene detection is formulated as binary classification: for gene g the labels are $\mathbf{1}[y_{c,g} > 0]$ and the scores are the predicted expressions $\widehat{y}_{c,g}$. The area under the precision-recall curve (AUC-PR) is invariant to any monotone rescaling of the predictions and only assesses whether expressing cells are ranked above non-expressing ones. Genes with no positive or no negative cell are excluded.

A.7 Additional experiments

A.7.1 TIGER's performance on internal cancer cohorts

We further analyze TIGER on the internal cancer cohorts through additional evaluations. First, we compare bin-level and cell-level prediction performance for the bin-level baseline models (**Table A2**). We then assess the statistical significance of TIGER's improvements over each baseline at the patient level (**Table A3**), followed by cohort-specific cell-level performance across lung, colorectal, and breast cancer cohorts (**Table A4**).

Table A2: **Bin-level prediction performance, macro-averaged across the three internal cohorts.** Bin-level and cell-level performance are comparable for these models, and both remain consistently lower than TIGER across metrics. Best performance in **bold**, and standard deviation across the five-fold cross-validation in parentheses.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Img2ST-Net (**Bin**)	0.038 (±0.024)	0.126 (±0.048)	0.007 (±0.008)	0.487 (±0.021)	0.563 (±0.015)	0.112 (±0.031)	0.110 (±0.031)
Img2ST-Net (**Cell**)	0.008 (±0.011)	0.071 (±0.027)	0.000 (±0.009)	0.531 (±0.024)	0.614 (±0.020)	0.190 (±0.030)	0.191 (±0.030)
iStar (**bin**)	0.127 (±0.020)	0.346 (±0.039)	0.065 (±0.022)	0.474 (±0.019)	0.515 (±0.022)	0.238 (±0.029)	0.230 (±0.030)
iStar (**Cell**)	0.123 (±0.021)	0.356 (±0.045)	0.059 (±0.021)	0.485 (±0.019)	0.524 (±0.023)	0.246 (±0.022)	0.239 (±0.024)
TIGER (ours)	**0.295** (±0.024)	**0.617** (±0.038)	**0.242** (±0.022)	**0.427** (±0.014)	**0.493** (±0.017)	**0.333** (±0.022)	**0.324** (±0.025)

Table A3: **Patient-level significance of TIGER over each baseline.** Entries are metric values averaged over the 43 internal-cohort patients (one tissue section each), so every patient is weighted equally. No distinction is made based on cancer type. Stars in a baseline row mark a statistically significant superiority of TIGER over that baseline on the corresponding metric, from a one-sided paired t-test on the per-patient differences: $^{***}q < 0.001$, $^{**}q < 0.01$, $^{*}q < 0.05$, with q-values Benjamini–Hochberg-adjusted across all 35 baseline $\times$ metric comparisons. **Bold** indicates the best value and underlining the second best, independently of significance.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Img2ST-Net	0.008^{***}	0.066^{***}	0.000^{***}	0.515^{***}	0.595^{***}	0.178^{***}	0.179^{***}
iStar	0.116^{***}	0.341^{***}	0.057^{***}	0.471^{***}	0.507^{***}	0.230^{***}	0.221^{***}
DeepSpot2Cell	0.207^{***}	0.482^{***}	0.159^{***}	0.478^{***}	0.531^{***}	$\underline{0.283}^{***}$	$\underline{0.272}^{***}$
sCellST	0.117^{***}	0.323^{***}	0.079^{***}	0.518^{***}	0.619^{***}	0.235^{***}	0.228^{***}
GHIST	$\underline{0.213}^{***}$	$\underline{0.544}^{***}$	$\underline{0.162}^{***}$	$\underline{0.437}^{***}$	$\underline{0.507}^{***}$	0.277^{***}	0.270^{***}
TIGER (ours)	**0.283**	**0.608**	**0.232**	**0.424**	**0.487**	**0.318**	**0.308**

Table A4: **Single-cell ST prediction in the internal cohorts.** Results of TIGER and baselines, all evaluated at the *cell level*. The training strategy (Full- vs. weakly-supervised) and resolution (bin vs. cell) are specified under each model. Best performance in **bold**, second best underlined; standard deviation across the five-fold cross-validation in parentheses.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Lung cancer							
Img2ST-Net (Full, bin)	0.003 (±0.006)	0.044 (±0.007)	0.002 (±0.004)	0.481 (±0.011)	0.576 (±0.012)	0.180 (±0.018)	0.181 (±0.018)
iStar (Weak, bin)	0.081 (±0.009)	0.262 (±0.020)	0.060 (±0.003)	0.466 (±0.012)	0.500 (±0.014)	0.219 (±0.019)	0.214 (±0.019)
DeepSpot2Cell (Weak, cell)	0.140 (±0.016)	0.372 (±0.033)	0.127 (±0.015)	0.499 (±0.011)	0.544 (±0.014)	0.246 (±0.022)	0.237 (±0.022)
sCellST (Weak, cell)	0.046 (±0.007)	0.164 (±0.019)	0.040 (±0.007)	0.476 (±0.009)	0.556 (±0.010)	0.200 (±0.016)	0.197 (±0.016)
GHIST (Full, cell)	0.162 (±0.010)	0.468 (±0.037)	0.136 (±0.005)	0.444 (±0.013)	0.506 (±0.014)	0.250 (±0.016)	0.246 (±0.018)
TIGER (ours) (Full, cell)	**0.230** (±0.023)	**0.540** (±0.037)	**0.211** (±0.020)	**0.434** (±0.012)	**0.488** (±0.014)	**0.290** (±0.024)	**0.280** (±0.026)
Colorectal cancer							
Img2ST-Net (Full, bin)	0.012 (±0.012)	0.085 (±0.040)	0.002 (±0.010)	0.498 (±0.039)	0.578 (±0.036)	0.111 (±0.022)	0.110 (±0.023)
iStar (Weak, bin)	0.127 (±0.024)	0.398 (±0.049)	0.047 (±0.031)	0.456 (±0.024)	0.491 (±0.031)	0.176 (±0.017)	0.163 (±0.020)
DeepSpot2Cell (Weak, cell)	0.202 (±0.012)	0.415 (±0.066)	0.125 (±0.018)	0.420 (±0.020)	0.472 (±0.027)	0.220 (±0.018)	0.204 (±0.022)
sCellST (Weak, cell)	0.120 (±0.012)	0.400 (±0.042)	0.070 (±0.011)	0.415 (±0.016)	0.479 (±0.016)	0.175 (±0.017)	0.163 (±0.021)
GHIST (Full, cell)	0.192 (±0.011)	0.561 (±0.043)	0.126 (±0.007)	0.412 (±0.016)	0.479 (±0.024)	0.214 (±0.019)	0.198 (±0.022)
TIGER (ours) (Full, cell)	**0.264** (±0.007)	**0.621** (±0.025)	**0.191** (±0.015)	**0.391** (±0.020)	**0.444** (±0.025)	**0.253** (±0.016)	**0.234** (±0.021)
Breast cancer							
Img2ST-Net (Full, bin)	0.010 (±0.016)	0.085 (±0.034)	-0.003 (±0.013)	0.615 (±0.021)	0.687 (±0.011)	0.280 (±0.049)	0.283 (±0.049)
iStar (Weak, bin)	0.161 (±0.031)	0.408 (±0.065)	0.069 (±0.029)	0.533 (±0.020)	0.582 (±0.023)	0.343 (±0.030)	0.340 (±0.033)
DeepSpot2Cell (Weak, cell)	0.283 (±0.041)	0.547 (±0.062)	0.219 (±0.031)	0.548 (±0.011)	0.624 (±0.018)	0.406 (±0.029)	0.403 (±0.029)
sCellST (Weak, cell)	0.222 (±0.033)	0.489 (±0.066)	0.163 (±0.027)	0.500 (±0.007)	0.622 (±0.017)	0.375 (±0.028)	0.373 (±0.032)
GHIST (Full, cell)	0.322 (±0.040)	0.643 (±0.064)	0.249 (±0.029)	0.476 (±0.012)	0.577 (±0.010)	0.419 (±0.022)	0.419 (±0.024)
TIGER (ours) (Full, cell)	**0.390** (±0.043)	**0.690** (±0.052)	**0.324** (±0.032)	**0.456** (±0.010)	**0.548** (±0.011)	**0.456** (±0.025)	**0.457** (±0.028)

A.7.2 PRESERVATION OF GENE EXPRESSION VARIABILITY IN INFERRED SINGLE-CELL ST

Pearson correlation alone does not capture whether predicted expression preserves the true cell-to-cell scale of variation, as it is invariant to affine rescaling. We therefore evaluate concordance between predicted and ground-truth expression using Lin's Concordance Correlation Coefficient (CCC Lawrence & Lin (1989)).

For each gene g, we compute:

$$\mathrm{CCC}_g = \frac{2\,\mathrm{Cov}(\widehat{y}_g, y_g)}{\mathrm{Var}(\widehat{y}_g) + \mathrm{Var}(y_g) + (\mu_{\widehat{y}_g} - \mu_{y_g})^2},$$

where $\widehat{y}_g$ and y_g denote predicted and ground-truth $\log(1+x)$ expression across cells. CCC can be written as $\mathrm{CCC}_g = \rho_g C_{b,g}$, combining Pearson correlation ρ_g with a bias-correction term $C_{b,g}$ that penalizes mismatch in scale and location. Thus, high CCC requires both correct cell-level correspondence and agreement in expression scale. TIGER achieves the highest mean CCC in all three cohorts (LUAD: 0.18, IDC: 0.34, CRC: 0.21; **Figure A2**).

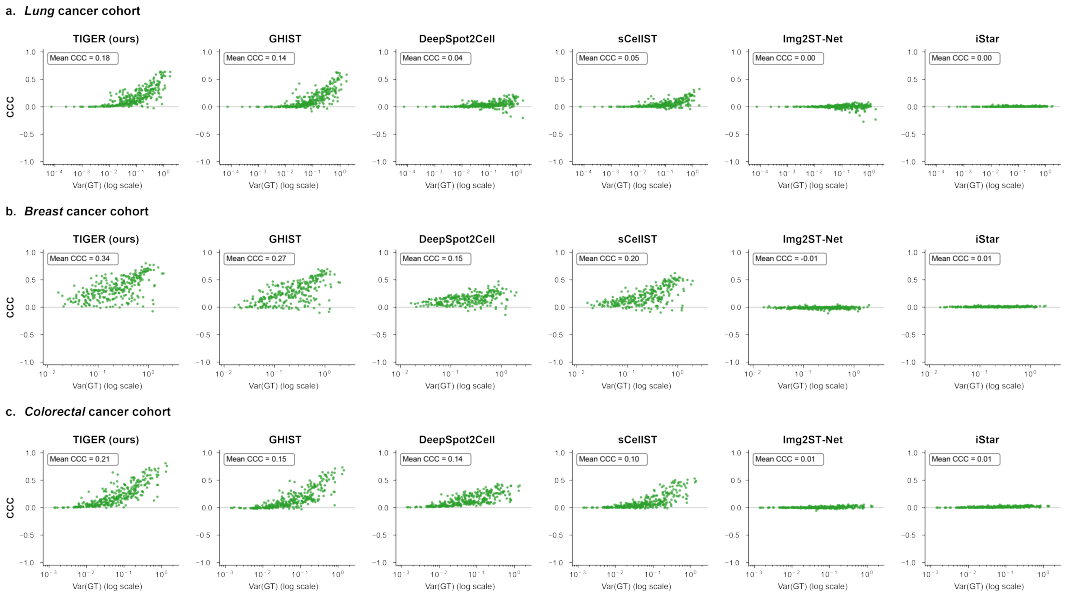


Figure A2: **Per-gene concordance across expression variability in TIGER and baseline models.** For each gene, we plot the Concordance Correlation Coefficient (CCC) between predicted and ground-truth cell-level expression against the gene's ground-truth variance, $\mathrm{Var}(y_g)$ (x-axis, log scale). **a.** Lung, **b.** breast, and **c.** colorectal cancer cohorts. Insets report the mean CCC across genes, $\frac{1}{G}\sum_{g=1}^{G}\mathrm{CCC}_g$. CCC jointly captures cell-level correlation and agreement in expression scale and location; thus, high values require both correct cell-level correspondence and well-calibrated predictions.

A.7.3 COMPUTATIONAL CONSIDERATIONS

Hardware and Software. All models were trained on a single NVIDIA RTX A5000 or RTX 3090 GPU (each with 24GB of memory). The framework was implemented in `PyTorch` and used `Automatic Mixed Precision (AMP)` to reduce memory usage and accelerate training.

Training Efficiency. TIGER is trained end-to-end in a *single* stage using 5-fold cross-validation: the focal and context branches and the per-gene gate are optimized jointly, with no pretraining or freezing of the individual experts. We use a batch size of 32 cell chunks of 256 cells each, corresponding to $32 \times 256 = 8{,}192$ cells with associated gene expression per optimization step. On a single NVIDIA RTX 3090 (24 GB), one fold converges in approximately 1 hour, giving roughly 5 hours for the complete 5-fold cross-validation. Table A5 compares TIGER to GHIST, the strongest baseline: TIGER trains $15\times$ faster per epoch while using less than a fifth of the peak GPU memory.

Table A5: **Training efficiency.** Wall-clock time per training epoch (averaged over 8 epochs) and peak GPU memory for TIGER and GHIST, the strongest-performing baseline (Table 2). Best performance in **bold**. TIGER outperforms GHIST across all metrics while training $15\times$ faster per epoch on $5.6\times$ less GPU memory.

Model	Time / epoch (↓)	Peak GPU mem. (↓)
GHIST (Fu et al. (2025))	1769.7 s	19,921 MiB
TIGER	**117.7 s**	**3,559 MiB**

A.7.4 TIGER'S PERFORMANCE ON EXTERNAL CANCER COHORTS

We report detailed performance on the external breast and lung cancer cohorts in **Tables A6, A7**, respectively. Across both cohorts, TIGER achieves the strongest overall performance on correlation- and sparsity-aware metrics, with statistically significant improvements over most baselines, supporting its generalization to independent datasets.

Performance across multiple spatial scales on external cohorts. To assess TIGER's generalization across spatial scales, we pooled predictions into progressively coarser bins (16, 32, 64, 96, and 128 μm edge length) and compared against five bin- and cell-level baselines (DeepSpot2Cell, iStar, Img2ST-Net, sCellST, GHIST) and four spot-level methods (ST-Net, BLEEP, DeepSpot, ST-Flow) trained at the 128 μm patch scale. Because all models predict log-transformed expression, predictions were first mapped back to count space using $\mathrm{expm1}(\cdot)$. Per-cell predictions from TIGER, GHIST, DeepSpot2Cell, and sCellST were sum-pooled within each bin, with boundary-crossing cells weighted by their segmented area fraction, whereas Img2ST-Net and iStar predictions on the 8 μm grid were directly sum-pooled. Ground truth was obtained by summing raw Xenium counts within each bin. Both predictions and ground truth were then transformed with $\log(1 + x)$ before evaluation, so that both represent total expression per spatial bin.

Pearson correlation (PCC) between pooled prediction and pooled ground truth was computed per gene per slide, then averaged over genes and slides within each cohort separately; the reported PCC and its standard error of the mean (SEM) are each the simple average of the two cohorts' own per-cohort statistics (external breast and lung cohorts). All models were evaluated zero-shot on the two external cohorts, using checkpoints trained only on the matched-organ internal cohort, with no fine-tuning on external data.

Aggregating cell-level predictions to coarser bin sizes, from 16 μm to 128 μm (patch size), progressively improves PCC of TIGER by 145% (**Figure A3**). This implies that spatial averaging reduces noise and sparsity in gene expression, facilitating the learning of the correlative signature. Notably, even at the patch level, TIGER outperforms four spot-level baselines trained for 128 μm-bin expression prediction (i.e., BLEEP (Xie et al. (2023)), STFlow (Huang et al. (2025)), ST-Net (He et al. (2020a)), and DeepSpot (Nonchev et al. (2025a))), and surpasses other single-cell baselines when extended to larger scales. This suggests that TIGER can reliably operate across different scales, with the model incorporating cell, microenvironment, and tissue-level context by design.

A.7.5 ADDITIONAL VISUALIZATION EXAMPLES

Figures A5 and A4 show additional examples of *single-cell* virtual ST for TIGER and the baselines. In the breast cancer cohort, TIGER accurately captures tumor-specific expression of *ERBB2*, an oncogene and clinically actionable marker. In **Figure A5**, the sample contains high-grade *ductal carcinoma in situ* (DCIS) with comedonecrosis (yellow arrows) showing high *ERBB2* expression, a well-established pattern, as well as more heterogeneous *ERBB2* expression in invasive tumor cells. TIGER successfully recapitulates this spatial expression pattern. In addition to highly expressed genes, TIGER also captures *sparser* spatial patterns. One example is depicted in **Figure A5b**, where the endothelial marker of blood vasculature *VWF* is only expressed in a few cells within the slide and TIGER successfully predicts this expression pattern. In the lung cancer cohort, we observe similar trends for pneumocyte markers such as *NAPSA*, where TIGER accurately highlights regular lung epithelium as well as adenocarcinoma (PCC = 0.78), and sparser genes such as *CD163*, a macrophage-associated marker, where TIGER delineates localized immune-enriched aggregates

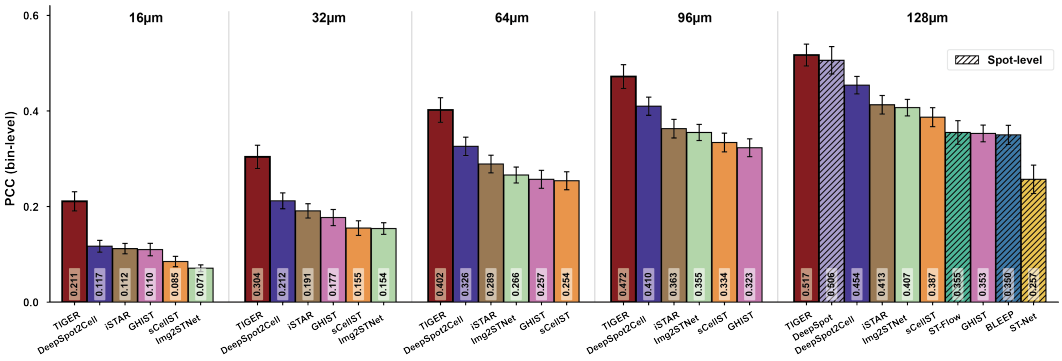


Figure A3: **Prediction across spatial scales in the external cohorts.** Cell-level expression of TIGER and five cell-/bin-level baselines (DeepSpot2Cell, iStar, Img2ST-Net, sCellST, GHIST), pooled to 16, 32, 64, 96, and 128 μm spot bins, compared against four spot-level baselines (DeepSpot, ST-Flow, BLEEP, ST-Net) natively trained at 128 μm resolution (hatched bars). Values are averaged over the two external cancer Xenium cohorts (breast, 19 slides; lung, 17 slides). Within each resolution, models are sorted by descending PCC (TIGER sits always leftmost); error bars denote standard error of the mean (SEM), averaged across the two cohorts' own per-cohort SEM. PCC: Pearson Correlation Coefficient.

Table A6: **Single-cell ST prediction in the *external breast cancer* cohort.** Best performance in **bold**, second best underlined; standard deviation across the 19 slides in parentheses. Stars in a baseline row mark a statistically significant superiority of TIGER over that baseline, from a one-sided paired t-test on the per-slide differences: $^{***}q < 0.001$, $^{**}q < 0.01$, $^{*}q < 0.05$, with q-values Benjamini–Hochberg-adjusted across all 42 (baseline $\times$ metric) comparisons.

Model	**PCC** (↑)	**PCC$_{50}$** (↑)	**PCC+** (↑)	**bMAE** (↓)	**bRMSE** (↓)	**Recall** (↑)	**AUC-PR** (↑)
Img2ST-Net	0.009^{***} ($\pm$0.009)	0.011^{***} ($\pm$0.009)	0.004^{***} ($\pm$0.013)	0.540^{***} ($\pm$0.032)	0.658^{***} ($\pm$0.039)	0.193^{***} ($\pm$0.045)	0.193^{***} ($\pm$0.046)
iStar	0.065^{***} ($\pm$0.047)	0.071^{***} ($\pm$0.049)	0.028^{***} ($\pm$0.034)	0.494^{***} ($\pm$0.031)	0.546 ($\pm$0.037)	0.228^{***} ($\pm$0.057)	0.224^{***} ($\pm$0.056)
DeepSpot2Cell	0.139^{***} ($\pm$0.081)	0.147^{***} ($\pm$0.083)	0.131^{***} ($\pm$0.077)	0.494^{***} ($\pm$0.032)	0.558^{**} ($\pm$0.039)	0.267^{***} ($\pm$0.072)	0.256^{***} ($\pm$0.071)
sCellST	0.073^{***} ($\pm$0.055)	0.078^{***} ($\pm$0.057)	0.062^{***} ($\pm$0.044)	0.555^{***} ($\pm$0.034)	0.649^{***} ($\pm$0.034)	0.228^{***} ($\pm$0.063)	0.224^{***} ($\pm$0.063)
GHIST	0.114^{***} ($\pm$0.080)	0.114^{***} ($\pm$0.080)	0.090^{***} ($\pm$0.062)	0.480^{***} ($\pm$0.032)	0.555^{*} ($\pm$0.038)	0.246^{***} ($\pm$0.067)	0.239^{***} ($\pm$0.068)
Phoenix	0.077^{***} ($\pm$0.052)	0.080^{***} ($\pm$0.054)	0.079^{***} ($\pm$0.051)	0.489^{***} ($\pm$0.031)	0.559^{*} ($\pm$0.039)	0.224^{***} ($\pm$0.054)	0.220^{***} ($\pm$0.057)
TIGER (ours)	**0.238** ($\pm$0.138)	**0.248** ($\pm$0.143)	**0.210** ($\pm$0.127)	**0.460** ($\pm$0.042)	**0.540** ($\pm$0.054)	**0.318** ($\pm$0.092)	**0.308** ($\pm$0.093)

consistent with tumor-associated macrophages (PCC = 0.68). Most baseline methods struggle to recover this spatially confined pattern.

Table A7: **Single-cell ST prediction in the *external lung cancer* cohort.** Best performance in **bold**, second best underlined; standard deviation across the 17 slides in parentheses. Stars in a baseline row mark a statistically significant superiority of TIGER over that baseline, from a one-sided paired t-test on the per-slide differences: $^{***}q < 0.001$, $^{**}q < 0.01$, $^{*}q < 0.05$, with q-values Benjamini–Hochberg-adjusted across all 42 (baseline $\times$ metric) comparisons.

Model	**PCC** (↑)	**PCC$_{50}$** (↑)	**PCC+** (↑)	**bMAE** (↓)	**bRMSE** (↓)	**Recall** (↑)	**AUC-PR** (↑)
Img2ST-Net	0.004^{***} (±0.007)	0.011^{***} (±0.009)	0.002^{***} (±0.015)	0.445^{***} (±0.019)	0.536^{***} (±0.027)	0.087^{***} (±0.039)	0.087^{***} (±0.039)
iStar	0.025^{***} (±0.017)	0.038^{***} (±0.020)	0.005^{***} (±0.019)	0.423^{**} (±0.024)	**0.453** (±0.031)	0.101^{***} (±0.042)	0.097^{***} (±0.042)
DeepSpot2Cell	0.044^{***} (±0.020)	0.060^{***} (±0.026)	0.039^{***} (±0.022)	0.456^{***} (±0.020)	0.492^{***} (±0.027)	0.109^{***} (±0.043)	0.102^{***} (±0.044)
sCellST	0.026^{***} (±0.016)	0.038^{***} (±0.017)	0.018^{***} (±0.012)	0.473^{***} (±0.026)	0.553^{***} (±0.034)	0.099^{***} (±0.042)	0.096^{***} (±0.042)
GHIST	0.055^{***} (±0.037)	0.069^{***} (±0.039)	0.047^{***} (±0.029)	0.418^{*} (±0.022)	0.462 (±0.030)	0.115^{***} (±0.050)	0.108^{***} (±0.049)
Phoenix	$\underline{0.070}^{***}$ (±0.036)	$\underline{0.088}^{***}$ (±0.045)	$\underline{0.074}^{***}$ (±0.044)	**0.410** (±0.023)	0.466^{*} (±0.028)	$\underline{0.124}^{***}$ (±0.042)	$\underline{0.112}^{***}$ (±0.046)
TIGER (ours)	**0.147** (±0.069)	**0.180** (±0.083)	**0.128** (±0.069)	<u>0.413</u> (±0.025)	<u>0.460</u> (±0.030)	**0.169** (±0.061)	**0.153** (±0.061)

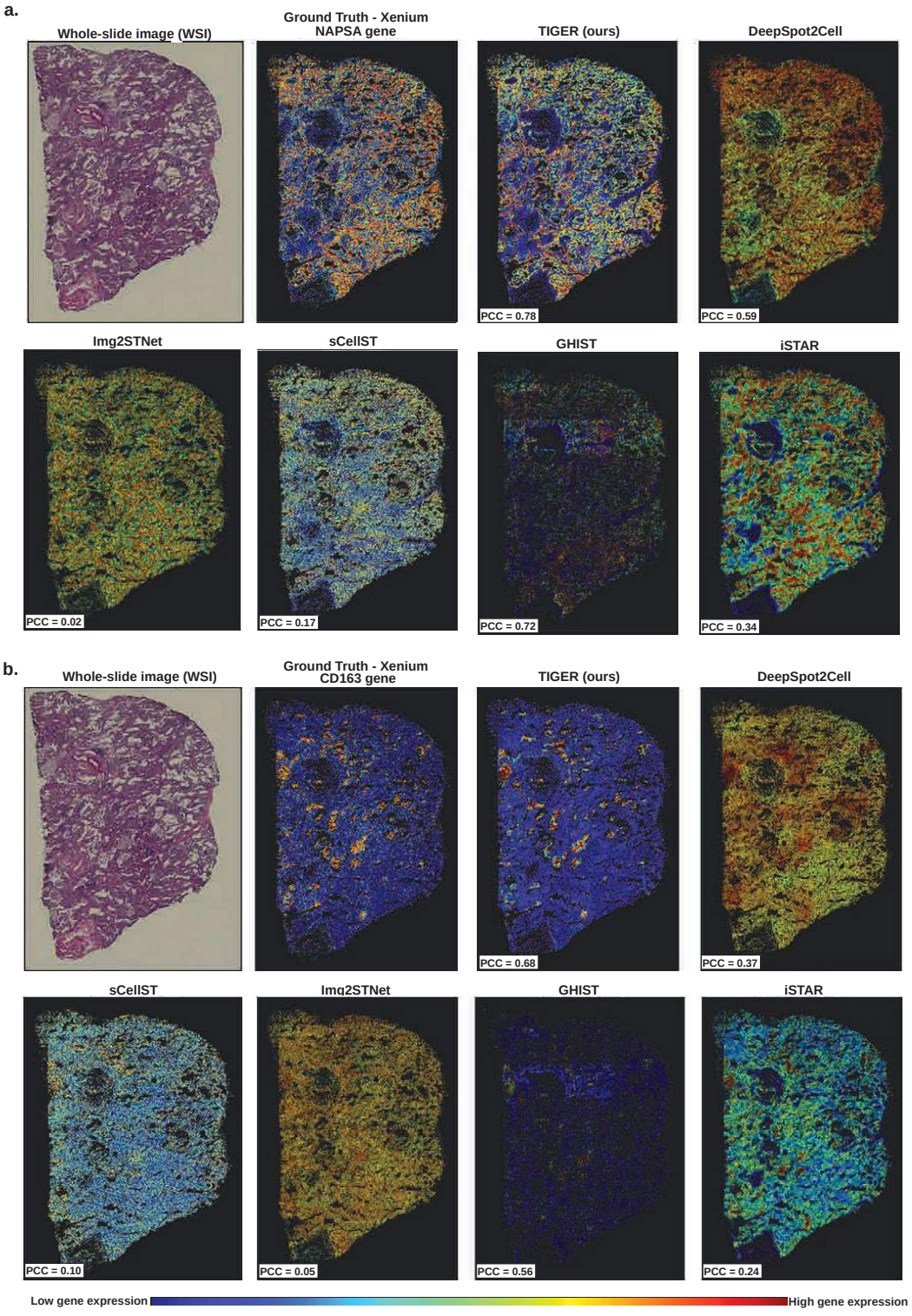


Figure A4: **Visualization of *single-cell* ST expression in lung cancer tissue. a.** Visualization of *NAPSA* expression, a lung adenocarcinoma–associated epithelial marker (typically expressed in pneumocytes but also in malignant epithelial cells in lung adenocarcinomas), highlighting tumor-cell–enriched patterns. **b.** Visualization of *CD163* expression, a macrophage-associated marker, highlighting immune-enriched spatial patterns (tumor-associated macrophages).

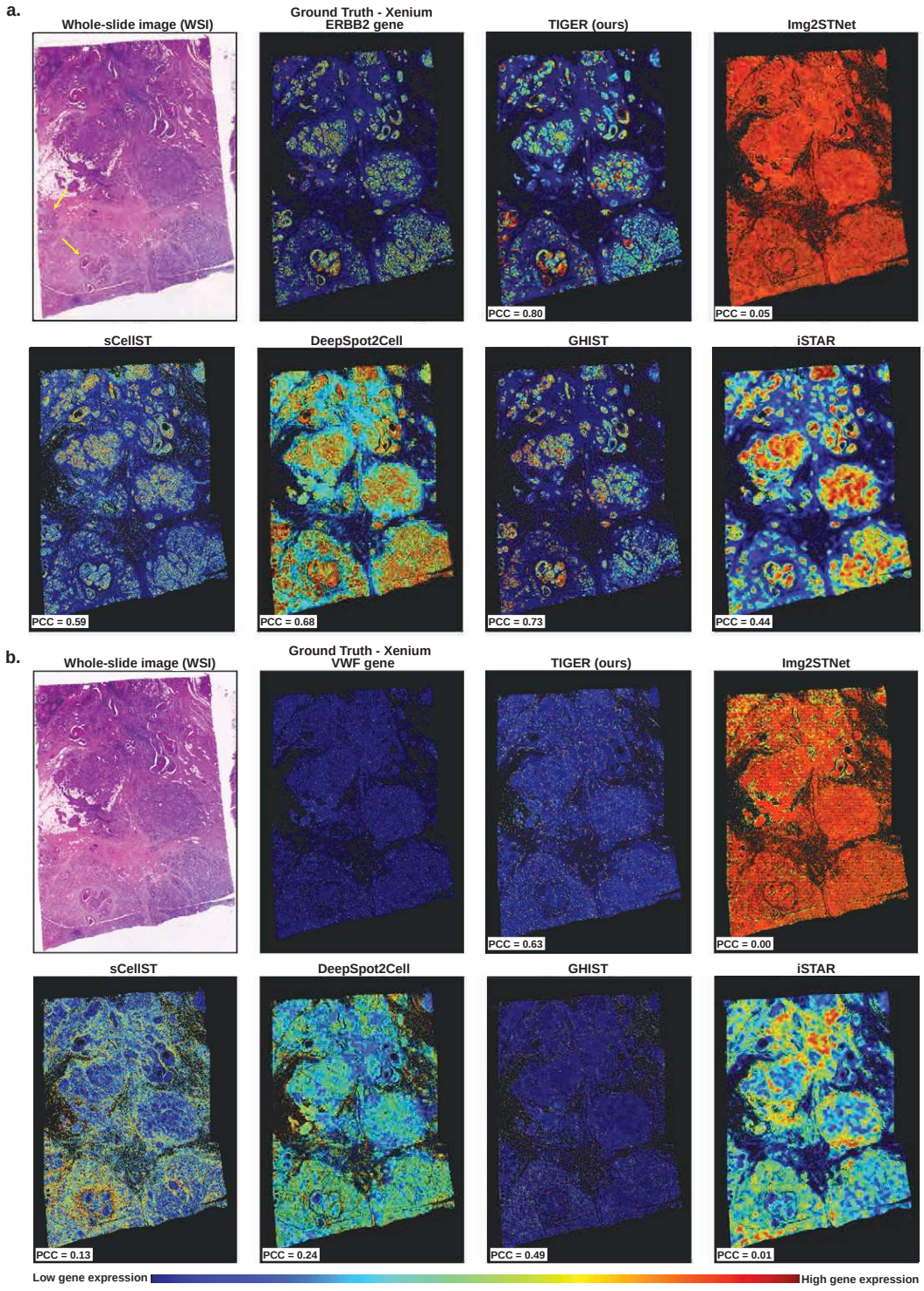


Figure A5: **Visualization of single-cell ST expression in breast cancer tissue. a.** Visualization of *ERBB2* (*HER2*) expression, a breast cancer–associated oncogene and clinically actionable marker, enriched in both invasive carcinoma and especially ductal carcinoma in situ (DCIS) across TIGER and baseline methods. Yellow arrows highlight DCIS with elevated ERBB2 expression. **b.** Visualization of *VWF* expression, an endothelial cell–associated marker of blood vasculature, highlighting vessel-enriched spatial patterns. TIGER better captures the expression of this sparse gene.

A.8 Design choices for TIGER architecture

We further examine the architectural design choices of TIGER through a series of interpretability analyses and ablation studies. We first visualize the learned gating behavior and analyze how the two experts contribute across spatial locations and gene-expression patterns (**Appendix A.8.1**). We then evaluate alternative fusion strategies, the contribution of local and global context, and the choice of encoder (**Appendix A.8.2**).

A.8.1 Visualizations and interpretability analysis

TIGER integrates $\text{TIGER}_{\text{context}}$ and $\text{TIGER}_{\text{focal}}$ to leverage the benefits of *local* and *contextualized* morphological understanding to accurately predict the expression of genes with distinct support, from sparse to highly expressed genes. **Figure A6** illustrates how different regions of a whole-slide image are routed to the model branches, enabling the capture of sparse expression while limiting false positives in areas lacking expression of the gene. **Figure A7** showcases how the *fully adaptive* gating preserves spatially coherent patterns while sharpening its predictions.

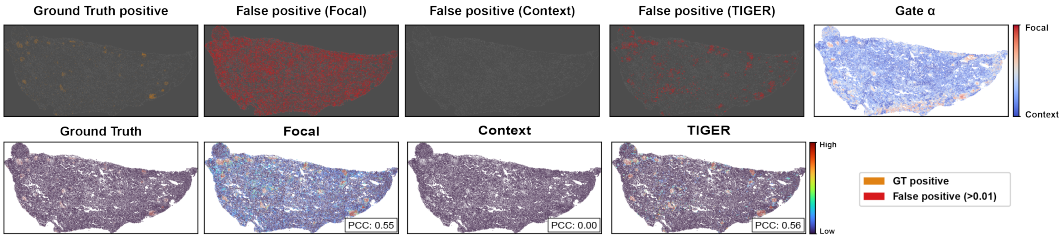


Figure A6: **Analysis of TIGER's gating behavior for *CXCL13* expression in a lung-cancer sample. Top:** From left to right, ground-truth-positive locations, false-positive locations for $\text{TIGER}_{\text{focal}}$, $\text{TIGER}_{\text{context}}$, and TIGER, and the learned gating values α. Ground-truth-positive locations denote spatial locations where *CXCL13* is expressed, often corresponding to tertiary-lymphoid-structure-like immune niches. False-positive locations are locations with zero ground-truth expression but predicted expression above 0.01. **Bottom:** Predicted *CXCL13* expression. $\text{TIGER}_{\text{focal}}$ produces spatially precise but noisy predictions, yielding high sensitivity but reduced specificity, whereas $\text{TIGER}_{\text{context}}$ oversmooths expression and misses sparse signals. *The gating mechanism adaptively combines the two, reducing false positives while preserving sensitivity to spatially sparse expression.*

A.8.2 TIGER components and gating mechanism ablations

TIGER predicts single-cell expression as a convex combination of two experts: $\text{TIGER}_{\text{focal}}$, which sees only the cell's own embedding $\mathbf{z}_c$, and $\text{TIGER}_{\text{context}}$, which adds the local cellular neighbourhood and the slide-level tissue context (**Eq. 5**). The combination is controlled by the gate $\boldsymbol{\alpha}_c \in [0,1]^G$, which decides, separately for every gene, how much of the prediction comes from cell morphology alone and how much from its surroundings. This subsection isolates two questions: how much each expert contributes on its own, and how much is gained by *learning* the gate rather than fixing it. We therefore compare the two experts in isolation against four ways of combining them, all trained under an identical setup and evaluated with the same five-fold cross-validation. Table A9 then reports all six variants in the three internal cohorts.

- **$\text{TIGER}_{\text{focal}}$** and **$\text{TIGER}_{\text{context}}$**: each expert trained and evaluated on its own, with no gate. These bound what fusion has to improve upon.
- **TIGER (*Average*)**: the gate is fixed to $\boldsymbol{\alpha}_c = 0.5$ for every gene and every cell, i.e. a plain average of the two experts. No gating parameters are learned, so the gap to the other variants measures what any form of gating contributes.
- **TIGER (*Late fusion*)**: no gate. The two experts' hidden representations are concatenated and decoded jointly,

$$\widehat{\mathbf{y}}_c = \boldsymbol{f}_{\text{MLP}}\left(\left[\mathbf{h}_c^{\text{Focal}}\,;\,\mathbf{h}_c^{\text{Context}}\right]\right), \tag{15}$$

  where $\mathbf{h}_c^{\text{Focal}} \in \mathbb{R}^{512}$ is $\text{TIGER}_{\text{focal}}$'s penultimate activation, that is its MLP without the final gene-wise layer, and $\mathbf{h}_c^{\text{Context}}$ is the context embedding of Eq. 1. The fusion MLP

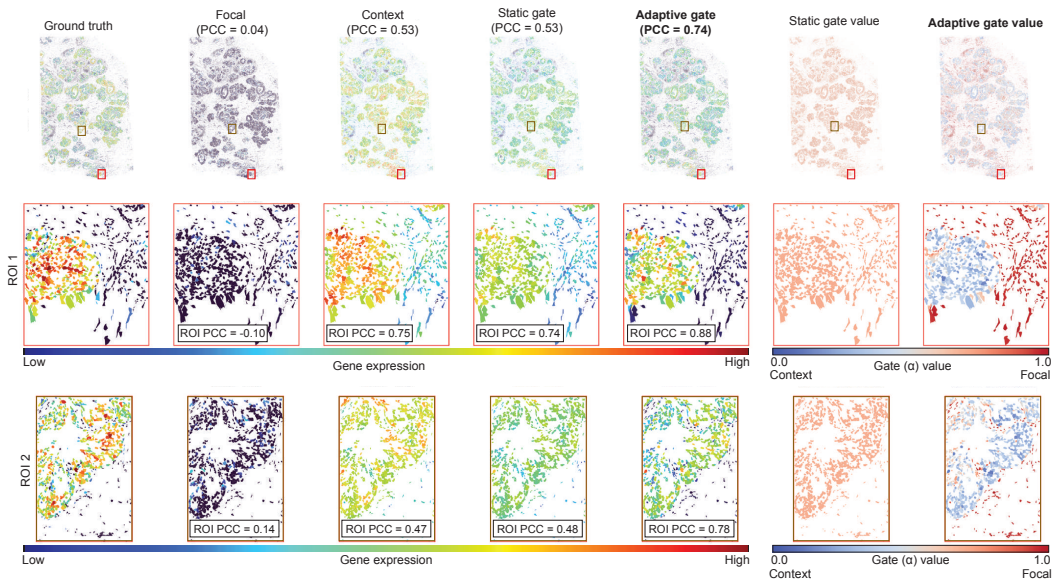


Figure A7: **Analysis of TIGER's gating behavior for *SCUBE2* gene in a breast cancer sample (TENX169). Columns:** ground-truth Xenium expression, $\text{TIGER}_{\text{focal}}$, $\text{TIGER}_{\text{context}}$, the *gene adaptive gate* ablation (locally *static* for a gene), and TIGER with the *fully adaptive* gate, followed by the per-cell gate value α_c for the gene and fully adaptive variants. **Rows:** whole slide image (top) and two magnified regions of interest (ROI 1, red; ROI 2, brown). $\text{TIGER}_{\text{focal}}$ misses the pattern almost entirely and $\text{TIGER}_{\text{context}}$ recovers the broad tumor structure but oversmooths it. The gene gate applies a constant gating and it is not able to recover the spatial gene profile whereas the fully adaptive gate varies per cell, deferring to the context branch inside glandular regions and to the focal branch in the surrounding non-expressing cells, which sharpens the prediction.

matches the gate in width and depth, so the output is not constrained to lie between the two experts' predictions, at the cost of losing the interpretable mixing weight.

- **TIGER (*gene adaptive* gate)**: the gate is a single learned parameter per gene, shared by all cells,

$$\boldsymbol{\alpha}_c = \boldsymbol{\alpha} = \sigma(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \mathbb{R}^G. \tag{16}$$

  with σ the element-wise logistic function and $\boldsymbol{\theta}$ optimised jointly with the two experts and initialised at $\mathbf{0}$, so training starts from the average ensemble. Since $\boldsymbol{\alpha}$ does not depend on $\mathbf{z}_c$, the model can learn that some genes are better predicted with $\text{TIGER}_{\text{focal}}$ and some genes with $\text{TIGER}_{\text{context}}$ but it cannot adapt locally to different cell microenvironments.

- **TIGER (*Adaptive* gate)**: the mechanism used in the final model (**Eq. 5**). A lightweight router MLP predicts $\boldsymbol{\alpha}_c$ from the cell's own embedding, its context representation and the disagreement between the two experts, so *the mixing weight varies per-gene and per-cell*. This behavior is illustrated in **Figure A7**, where the fully adaptive gate varies spatially across cells and selectively shifts between $\text{TIGER}_{\text{focal}}$ and $\text{TIGER}_{\text{context}}$ according to the local expression pattern. In contrast, the gene adaptive gate applies the same gene-specific weight throughout the tissue and therefore cannot recover such spatially heterogeneous routing.

The following tables include ablations of TIGER architectural components. **Table A8** shows the vision encoder ablation, stratified by cancer cohort. **Table A9** depicts the two TIGER experts and the distinct gating mechanisms, **Table A10** considers the contribution of the two context components and **Table A11** includes the per-patient statistical analysis between gating strategies.

Table A8: **Ablation of ViT-based patch encoders in TIGER, stratified by cancer cohort.** Kaiko (Embed. dim = 1024), Virchow2 (Embed. dim = 1280) and UNIv2 (Embed. dim = 1536) are pathology foundation models (PFMs), while EVA-02 (Embed. dim = 1024) was pretrained on ImageNet-21K. PFM token embeddings enhance *single-cell* ST prediction in all three cohorts. Best performance in **bold**, second best underlined; standard deviation across the five-fold cross-validation in parentheses.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Lung cancer							
EVA-02 (Fang et al. (2023))	0.127 (±0.010)	0.342 (±0.029)	0.100 (±0.006)	0.461 (±0.012)	0.509 (±0.014)	0.232 (±0.017)	0.226 (±0.018)
Kaiko (Aben et al. (2024))	0.229 (±0.027)	0.522 (±0.038)	0.195 (±0.019)	0.438 (±0.012)	0.490 (±0.015)	0.281 (±0.024)	0.272 (±0.026)
Virchow2 (Zimmermann et al. (2024))	**0.236** (±0.028)	0.535 (±0.039)	0.204 (±0.022)	0.437 (±0.013)	0.489 (±0.014)	0.286 (±0.024)	0.277 (±0.025)
UNIv2 (Chen et al. (2024))	0.230 (±0.023)	**0.540** (±0.037)	**0.211** (±0.020)	**0.434** (±0.012)	**0.488** (±0.014)	**0.290** (±0.024)	**0.280** (±0.026)
Colorectal cancer							
EVA-02 (Fang et al. (2023))	0.186 (±0.009)	0.494 (±0.030)	0.098 (±0.009)	0.414 (±0.018)	0.461 (±0.019)	0.200 (±0.014)	0.186 (±0.019)
Kaiko (Aben et al. (2024))	0.258 (±0.008)	0.608 (±0.028)	0.172 (±0.021)	0.396 (±0.018)	0.447 (±0.020)	0.245 (±0.015)	0.227 (±0.020)
Virchow2 (Zimmermann et al. (2024))	0.263 (±0.010)	0.614 (±0.025)	0.183 (±0.014)	0.393 (±0.016)	0.445 (±0.018)	0.249 (±0.015)	0.231 (±0.020)
UNIv2 (Chen et al. (2024))	**0.264** (±0.007)	**0.621** (±0.019)	**0.190** (±0.017)	**0.391** (±0.018)	**0.444** (±0.019)	**0.254** (±0.015)	**0.234** (±0.020)
Breast cancer							
EVA-02 (Fang et al. (2023))	0.288 (±0.047)	0.585 (±0.080)	0.207 (±0.034)	0.481 (±0.011)	0.571 (±0.013)	0.403 (±0.025)	0.401 (±0.027)
Kaiko (Aben et al. (2024))	0.372 (±0.044)	0.672 (±0.060)	0.301 (±0.035)	0.462 (±0.014)	0.552 (±0.011)	0.445 (±0.025)	0.446 (±0.028)
Virchow2 (Zimmermann et al. (2024))	0.385 (±0.042)	0.684 (±0.053)	0.314 (±0.031)	0.462 (±0.015)	0.551 (±0.012)	0.452 (±0.025)	0.453 (±0.027)
UNIv2 (Chen et al. (2024))	**0.393** (±0.042)	**0.692** (±0.052)	**0.325** (±0.031)	**0.458** (±0.014)	**0.549** (±0.014)	**0.456** (±0.027)	**0.458** (±0.027)
Macro-average (Three cohorts)							
EVA-02 (Fang et al. (2023))	0.200 (±0.022)	0.474 (±0.046)	0.135 (±0.016)	0.452 (±0.014)	0.514 (±0.015)	0.278 (±0.019)	0.271 (±0.021)
Kaiko (Aben et al. (2024))	0.286 (±0.026)	0.601 (±0.042)	0.223 (±0.025)	0.432 (±0.015)	0.496 (±0.015)	0.324 (±0.021)	0.315 (±0.025)
Virchow2 (Zimmermann et al. (2024))	0.295 (±0.027)	0.611 (±0.039)	0.234 (±0.022)	0.431 (±0.015)	0.495 (±0.015)	0.329 (±0.021)	0.320 (±0.024)
UNIv2 (Chen et al. (2024))	**0.295** (±0.024)	**0.617** (±0.036)	**0.242** (±0.023)	**0.429** (±0.015)	**0.494** (±0.016)	**0.333** (±0.021)	**0.324** (±0.024)

Table A9: **Performance of TIGER and its architectural components along with ablations on the gating mechanism stratified by cancer cohort.** Best performance in **bold**, second best underlined; standard deviation across the five-fold cross-validation in parentheses. The fully adaptive gating mechanism of TIGER matches or surpasses the better of its two experts ($\text{TIGER}_{\text{focal}}$ or $\text{TIGER}_{\text{context}}$) on PCC_{50}, Recall and AUC-PR in all three cohorts, and on six of the seven metrics in lung and breast cancer.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Lung cancer							
$\text{TIGER}_{\text{focal}}$	0.220 (± 0.026)	0.528 (± 0.044)	0.205 (± 0.025)	0.435 (± 0.013)	0.490 (± 0.014)	0.287 (± 0.026)	0.276 (± 0.027)
$\text{TIGER}_{\text{context}}$	**0.232** (± 0.013)	0.528 (± 0.039)	0.162 (± 0.011)	0.439 (± 0.011)	0.494 (± 0.013)	0.259 (± 0.018)	0.258 (± 0.020)
TIGER (*Average*)	0.218 (± 0.024)	0.530 (± 0.045)	0.204 (± 0.021)	0.436 (± 0.014)	0.492 (± 0.015)	0.284 (± 0.025)	0.275 (± 0.026)
TIGER (*Late fusion*)	0.219 (± 0.022)	0.537 (± 0.041)	0.203 (± 0.019)	0.436 (± 0.013)	0.491 (± 0.014)	0.287 (± 0.024)	0.277 (± 0.025)
TIGER (*Gene adaptive* gate)	0.219 (± 0.024)	0.531 (± 0.044)	0.206 (± 0.022)	0.435 (± 0.012)	0.492 (± 0.014)	0.285 (± 0.024)	0.275 (± 0.026)
TIGER (*Adaptive* gate)	0.230 (± 0.023)	**0.540** (± 0.037)	**0.211** (± 0.020)	**0.434** (± 0.012)	**0.488** (± 0.014)	**0.290** (± 0.024)	**0.280** (± 0.026)
Colorectal cancer							
$\text{TIGER}_{\text{focal}}$	0.253 (± 0.010)	0.617 (± 0.030)	**0.191** (± 0.018)	**0.389** (± 0.018)	**0.441** (± 0.020)	0.250 (± 0.016)	0.230 (± 0.021)
$\text{TIGER}_{\text{context}}$	**0.280** (± 0.015)	0.609 (± 0.024)	0.131 (± 0.013)	0.396 (± 0.017)	0.449 (± 0.019)	0.215 (± 0.014)	0.208 (± 0.020)
TIGER (*Average*)	0.243 (± 0.009)	0.610 (± 0.022)	0.186 (± 0.016)	0.392 (± 0.018)	0.447 (± 0.020)	0.247 (± 0.015)	0.229 (± 0.020)
TIGER (*Late fusion*)	0.244 (± 0.008)	0.620 (± 0.031)	0.188 (± 0.016)	0.392 (± 0.015)	0.444 (± 0.020)	0.250 (± 0.015)	0.229 (± 0.021)
TIGER (*Gene adaptive* gate)	0.261 (± 0.010)	0.611 (± 0.028)	0.185 (± 0.016)	0.392 (± 0.017)	0.447 (± 0.020)	0.247 (± 0.015)	0.228 (± 0.021)
TIGER (*Adaptive* gate)	0.264 (± 0.007)	**0.621** (± 0.025)	**0.191** (± 0.015)	0.391 (± 0.018)	0.444 (± 0.019)	**0.253** (± 0.016)	**0.234** (± 0.021)
Breast cancer							
$\text{TIGER}_{\text{focal}}$	0.385 (± 0.045)	0.686 (± 0.054)	0.321 (± 0.035)	0.456 (± 0.013)	**0.546** (± 0.011)	0.454 (± 0.026)	0.455 (± 0.029)
$\text{TIGER}_{\text{context}}$	0.377 (± 0.046)	0.690 (± 0.052)	0.307 (± 0.035)	0.457 (± 0.010)	0.554 (± 0.005)	0.444 (± 0.024)	0.448 (± 0.026)
TIGER (*Average*)	0.380 (± 0.044)	0.683 (± 0.052)	0.317 (± 0.033)	0.457 (± 0.014)	0.552 (± 0.015)	0.453 (± 0.025)	0.454 (± 0.028)
TIGER (*Late fusion*)	0.378 (± 0.044)	0.684 (± 0.053)	0.318 (± 0.031)	0.460 (± 0.012)	0.553 (± 0.013)	0.452 (± 0.024)	0.453 (± 0.027)
TIGER (*Gene adaptive* gate)	0.380 (± 0.045)	0.683 (± 0.053)	0.317 (± 0.033)	0.457 (± 0.012)	0.551 (± 0.011)	0.452 (± 0.025)	0.454 (± 0.028)
TIGER (*Adaptive* gate)	**0.390** (± 0.043)	**0.690** (± 0.052)	**0.324** (± 0.032)	**0.456** (± 0.010)	0.548 (± 0.011)	**0.456** (± 0.025)	**0.457** (± 0.028)

Table A10: **Ablation study of the Global and Local Context Branches (GCB and LCB).** Performance comparison of $\text{TIGER}_{\text{context}}$ with and without the GCB and LCB across three cancer types (Lung, Colorectal, and Breast) in the internal cohorts. Overall, integrating both branches achieves the best performance across all cancer types, demonstrating their complementary roles. The best result for each metric is shown in **bold**.

Model	PCC (↑)	PCC_{50} (↑)	PCC+ (↑)	bMAE (↓)	bRMSE (↓)	Recall (↑)	AUC-PR (↑)
Lung cancer							
$\text{TIGER}_{\text{context}}$ (w/o GCB)	0.217 (± 0.010)	0.514 (± 0.045)	0.152 (± 0.008)	**0.439** (± 0.012)	0.495 (± 0.013)	0.254 (± 0.017)	0.254 (± 0.019)
$\text{TIGER}_{\text{context}}$ (w/o LCB)	0.151 (± 0.013)	0.360 (± 0.043)	0.090 (± 0.005)	0.461 (± 0.013)	0.510 (± 0.015)	0.228 (± 0.017)	0.226 (± 0.018)
$\text{TIGER}_{\text{context}}$	**0.232** (± 0.013)	**0.528** (± 0.039)	**0.162** (± 0.011)	**0.439** (± 0.011)	**0.494** (± 0.013)	**0.259** (± 0.018)	**0.258** (± 0.020)
Colorectal cancer							
$\text{TIGER}_{\text{context}}$ (w/o GCB)	0.241 (± 0.009)	0.603 (± 0.027)	**0.138** (± 0.013)	0.395 (± 0.017)	0.448 (± 0.020)	0.217 (± 0.017)	0.211 (± 0.022)
$\text{TIGER}_{\text{context}}$ (w/o LCB)	0.195 (± 0.023)	0.481 (± 0.024)	0.078 (± 0.009)	0.414 (± 0.019)	0.463 (± 0.021)	0.180 (± 0.012)	0.176 (± 0.018)
$\text{TIGER}_{\text{context}}$	**0.280** (± 0.015)	**0.609** (± 0.024)	0.131 (± 0.013)	**0.396** (± 0.017)	**0.449** (± 0.019)	**0.222** (± 0.014)	**0.213** (± 0.020)
Breast cancer							
$\text{TIGER}_{\text{context}}$ (w/o GCB)	0.359 (± 0.046)	0.675 (± 0.058)	0.292 (± 0.034)	0.458 (± 0.012)	0.559 (± 0.011)	0.438 (± 0.024)	0.442 (± 0.027)
$\text{TIGER}_{\text{context}}$ (w/o LCB)	0.233 (± 0.050)	0.496 (± 0.095)	0.151 (± 0.040)	0.514 (± 0.016)	0.602 (± 0.007)	0.373 (± 0.030)	0.372 (± 0.029)
$\text{TIGER}_{\text{context}}$	**0.377** (± 0.046)	**0.690** (± 0.052)	**0.307** (± 0.035)	**0.457** (± 0.010)	**0.554** (± 0.005)	**0.444** (± 0.024)	**0.448** (± 0.026)

Table A11: ***Fully adaptive gating* vs. alternative fusion strategies.** Cell-level metrics averaged over all 43 held-out patient slides across the internal colorectal, breast and lung cancer cohorts. Wins denotes the number of slides on which *fully adaptive* fusion outperforms the corresponding baseline. q-values are from a one-sided paired t-test on the per-slide differences, Benjamini–Hochberg-adjusted across all 21 baseline×metric comparisons; significant values ($q < 0.05$) in bold.

	Local	Average			Late fusion			Gene adaptive		
Metric	**adaptive**	mean	wins	q	mean	wins	q	mean	wins	q
PCC (↑)	0.283	0.269	43/43	**4.6e–15**	0.269	41/43	**4.6e–15**	0.276	36/43	**2.4e–09**
PCC50 (↑)	0.608	0.598	41/43	**6.0e–09**	0.604	31/43	**9.0e–03**	0.599	38/43	**7.7e–08**
PCC+ (↑)	0.232	0.226	38/43	**3.8e–09**	0.227	34/43	**1.0e–06**	0.227	35/43	**8.8e–07**
bMAE (↓)	0.4235	0.4245	30/43	7.3e–02	0.4254	28/43	**3.9e–03**	0.4243	27/43	5.6e–02
bRMSE (↓)	0.4867	0.4902	33/43	**3.3e–05**	0.4888	30/43	**3.9e–03**	0.4901	34/43	**2.3e–06**
Recall (↑)	0.318	0.313	42/43	**3.4e–11**	0.315	38/43	**3.8e–09**	0.313	41/43	**2.2e–11**
AUC-PR (↑)	0.308	0.303	42/43	**1.6e–12**	0.304	39/43	**1.5e–11**	0.303	42/43	**7.0e–13**

A.9 Cell phenotyping considerations

SingleR (Aran et al. (2019)) was applied to both ground-truth Xenium and TIGER single-cell predicted expression using CRC_{c295} from the Human Cell Atlas as the reference scRNA-seq dataset (https://singlecell.broadinstitute.org/single_cell/study/SCP1162/human-colon-cancer-atlas-c295). SingleR assigns cell type labels by computing Spearman rank correlations between each query cell's expression profile and reference cell-type profiles. Because the Xenium panel comprises a targeted set of 422 genes (322 overlapping with the TIGER prediction space), reference-based annotation with SingleR is constrained by limited marker coverage for some populations.

Several canonical lineage-defining genes present in scRNA-seq atlases are absent from the panel/predictions, including *EPCAM*, *KRT8*, and *KRT20* for epithelial identity, *COL1A1* and *DCN*

for fibroblast identity, or *PECAM1* and *VWF* for endothelial cells. Hence, ground-truth labels derived from Xenium expression should be interpreted as approximate reference-based annotations rather than definitive cell identities. In addition, CRC_{c295} is based on scRNA-seq rather than targeted spatial profiling, introducing modality differences in gene detection sensitivity and transcript representation. Despite these limitations, we applied the same annotation pipeline to both Xenium-measured and TIGER-predicted expression to ensure a fair comparison and to assess whether TIGER preserves biologically meaningful cell type structure and spatial heterogeneity at the level supported by the available gene panel. More refined subtype resolution would likely require manual marker-based curation and/or expanded gene panels.

To improve robustness of cross-modality cell type comparisons, we performed annotation using intermediate-resolution reference labels from the CRC_{c295} atlas and subsequently mapped these to broader lineage-level categories. Specifically, related subtypes were consolidated into major compartments including epithelial (Epi), tumor epithelial (EpiT), fibroblastic stroma (Fibro, Schwann), endothelial (Endo), smooth muscle/perivascular (SmoothMuscle, Peri), myeloid immune (Macro, Mono, Granulo, DC, Mast), and lymphoid immune populations (TCD4, TCD8, Tgd, TZBTB16, NK, ILC, B, Plasma). This hierarchical mapping reduces sensitivity to missing subtype-specific markers in the targeted Xenium gene panel and enables more reliable comparison between ground-truth and TIGER-predicted annotations at the lineage level. Visual examples of the cell phenotyping obtained with TIGER are included in **Figures A8** and **Figures A9**. Confusion matrices for TIGER and all baselines are depicted in **Figure A10**.

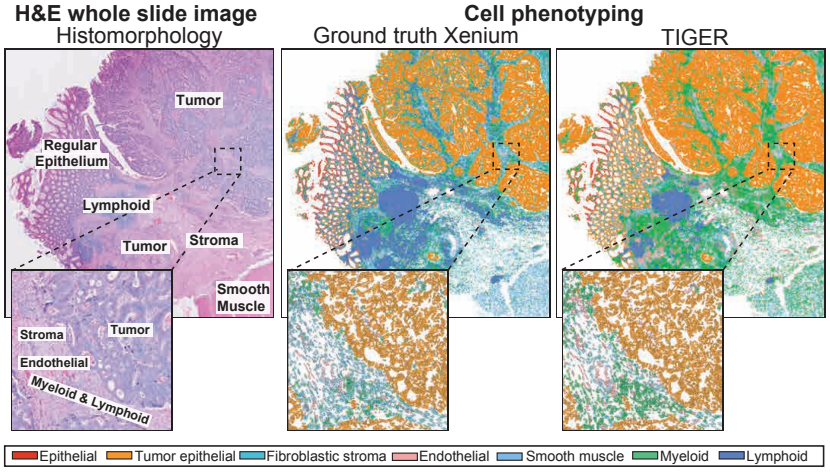


Figure A8: **Example of single-cell type annotation from predicted spatial transcriptomics in colorectal adenocarcinoma.** Left: H&E-stained histology with pathologist-annotated major cell classes. Middle: SINGLER cell type annotations based on ground truth Xenium data. Right: SINGLER cell type annotations based on TIGER virtual ST. Insets show magnified regions at the tumor-stroma interface.

A.10 DOWNSTREAM CLINICAL APPLICATION

Datasets. We evaluate two downstream tasks using different patient subsets from a previously published breast cancer cohort (Sammut et al. (2022)) with pre-treatment H&E-stained whole slide images (WSIs). For molecular phenotyping, we use 166 patients with clinically annotated *HER2* and *ER receptor status* (*HER2*/*ER*) to test whether TIGER recovers established molecular phenotypes directly from predicted single-cell expression. For treatment response prediction, we use a separate subset of 62 HER2-positive patients treated with trastuzumab in combination with neoadjuvant chemotherapy, with response assessed as pathological *complete response* (*pCR*) versus *residual disease* (*RD*).

Shared evaluation setup. For image-based prediction in both tasks, H&E patch features are extracted using the frozen UNIv2 encoder (embedding dim: 1536) and aggregated into a slide-level

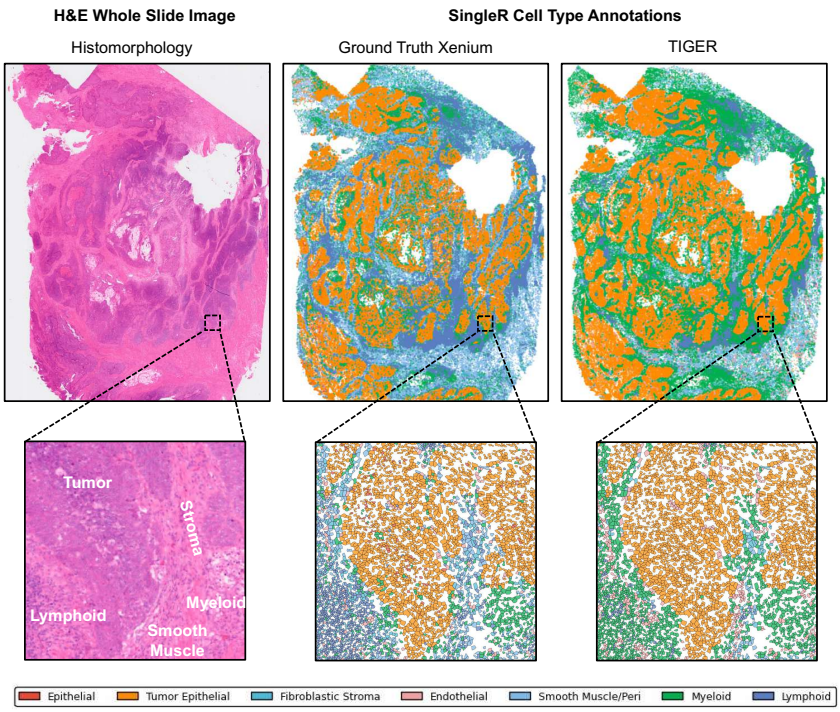


Figure A9: **Single-cell type annotation from predicted spatial transcriptomics in colorectal adenocarcinoma.** Left: H&E histopathology of a poorly differentiated colorectal adenocarcinoma. Middle: SINGLER cell type annotations based on ground truth Xenium data. Right: SINGLER cell type annotations based on TIGER virtual ST. Insets show magnified regions at the tumor-stroma interface. Tumor cell phenotyping is attained successfully while fine-grained immune cell subtyping between myeloid and lymphoid populations remains challenging due to both partial morphological similarity in H&E and transcriptomic overlap between these lineages in the underlying Xenium panel.

representation using attention-based multiple instance learning (ABMIL; Ilse et al. (2018)). All image-based classifiers are trained using Patho-Bench (Zhang et al. (2025)), which standardizes data splitting, model training, evaluation, and metric aggregation.

Virtual single-cell ST profiles for all baselines (GHIST, sCellST, DeepSpot2Cell, Img2ST-Net, iStar) and TIGER are generated using the same preprocessing protocol described in Appendix A.2. Each model is trained once on the internal breast cohort and applied directly to the TransNEO WSIs. For both downstream tasks, we use 50 Monte-Carlo cross-validation splits with stratified 80%/20% train/test partitions; AUROC and balanced accuracy are reported as mean $\pm$ standard error across test folds.

Marker-Based Molecular Phenotyping. To test whether predicted single-cell expression alone recovers established receptor status, independent of image features or ABMIL aggregation, we construct per-slide marker feature vectors directly from each model's predicted expression and classify receptor status using logistic regression. Following established breast cancer gene-expression signatures (Zhao et al. (2019)), HER2 status is represented by the 99th-percentile per-cell expression of *FGFR4*, *ERBB2*, and *GRB7*, while ER status uses the 99th-percentile expression of 15 ER-associated genes (*BLVRA*, *BAG1*, *BCL2*, *CXXC5*, *ESR1*, *GPR160*, *FOXA1*, *MLPH*, *NAT1*, *SLC39A6*, *MAPT*, *PGR*, *MDM2*, *TMEM45B*, *MMP11*). We use the 99th percentile to capture strong marker-expression signals while reducing sensitivity to isolated extreme values. Features are standardized using training-set statistics and evaluated with class-balanced logistic regression over 50 stratified 80%/20% Monte-Carlo train/test splits, matching the image-based evaluation. Evaluation uses 50 stratified Monte-Carlo splits with 80%/20% train/test partitions, matching the splits used for the image-based evaluation.

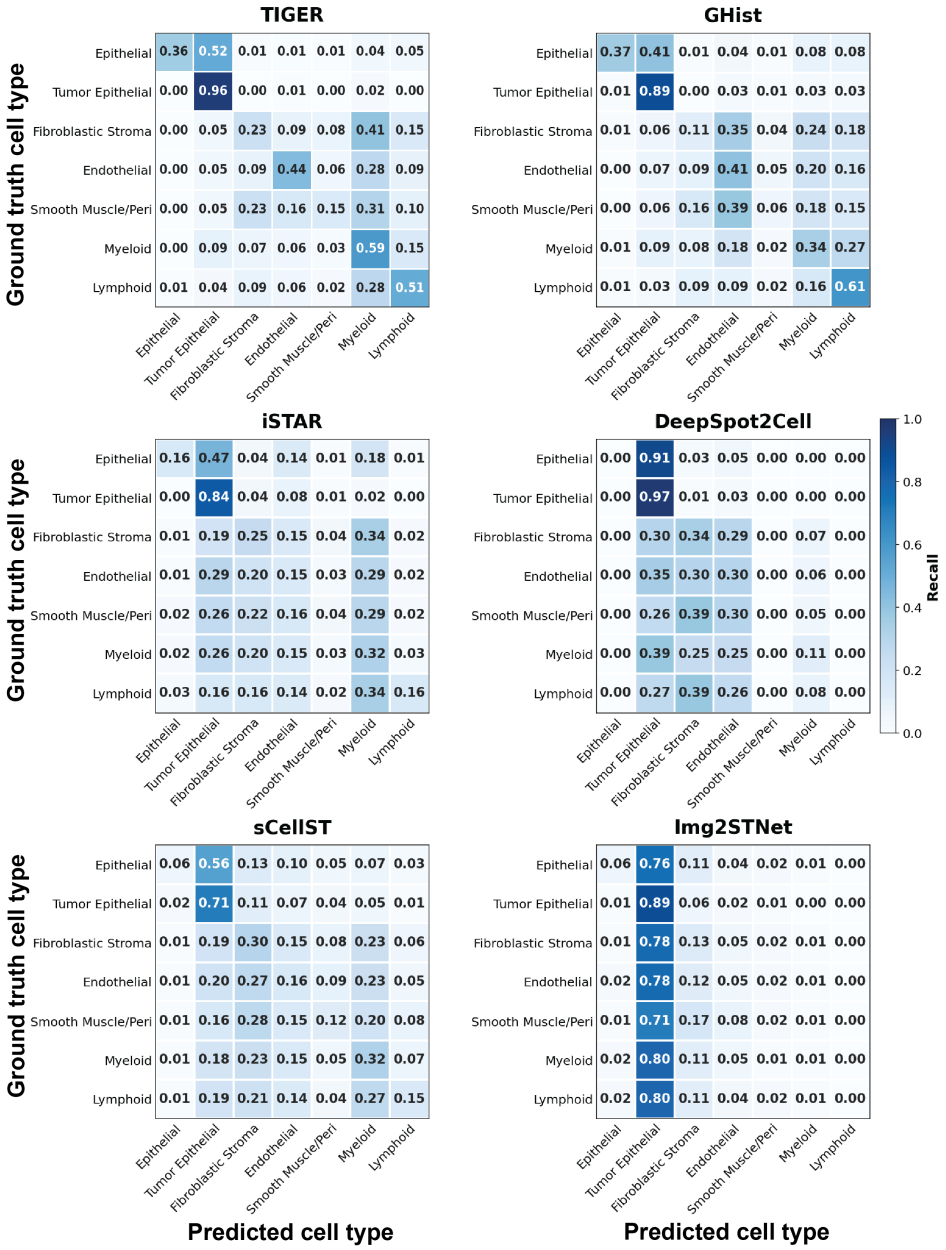


Figure A10: **Confusion matrix of cell type predictions against Xenium ground truth for TIGER and baselines.** Values represent per-class recall (row-normalized). The analysis was performed in the colorectal cancer cohort (N=16). TIGER shows strong overall recovery of major cell phenotypes compared with alternative spatial transcriptomics prediction methods.

Multimodal Treatment Response Prediction. For treatment response prediction, the shared image representation is complemented with virtual single-cell ST features. Predicted single-cell ST profiles are encoded using Novae (Blampey et al. (2025)) (embedding dim: 64), a foundation model pretrained on large-scale single-cell ST data. Image and virtual ST features are independently projected to a shared dimension and aggregated with ABMIL, after which the resulting slide-level representations are combined by element-wise late fusion for binary response prediction using binary cross-entropy loss.

As shown in **Table A12**, TIGER achieves the best performance among all virtual ST baselines on every metric across all three tasks, with a statistically significant improvement over the image-only baseline in each case.

Table A12: **Molecular phenotyping and treatment-response prediction from TIGER-predicted single-cell spatial transcriptomics.** *HER2 status* and *ER status*: receptor status of 166 patients predicted by logistic regression using gene markers derived solely from each model's predicted cell-level expression; *Drug Response*: response to neoadjuvant therapy in 62 HER2-positive patients, predicted using a multimodal ABMIL model combining H&E with profiles predicted by the corresponding virtual ST model; *Image only* uses H&E features alone with an ABMIL classifier. Best performance in **bold**. * indicates significant improvement over *Image only* (one-sided paired t-test over 50 folds, $p < 0.05$).

	HER2 status Virtual ST		**ER status** Virtual ST		**Drug Response** Image + Virtual ST	
Model	**AUROC** (↑)	**Bal. Acc.** (↑)	**AUROC** (↑)	**Bal. Acc.** (↑)	**AUROC** (↑)	**Bal. Acc.** (↑)
Image only	0.655 (± 0.013)	0.626 (± 0.008)	0.712 (± 0.014)	0.633 (± 0.012)	0.620 (± 0.018)	0.570 (± 0.015)
Img2ST-Net	0.451 (± 0.013)	0.463 (± 0.013)	0.583 (± 0.014)	0.572 (± 0.013)	0.610 (± 0.018)	0.573 (± 0.015)
iStar	0.573 (± 0.013)	0.551 (± 0.011)	0.662 (± 0.012)	0.608 (± 0.012)	0.621 (± 0.018)	0.567 (± 0.017)
DeepSpot2Cell	0.654 (± 0.012)	0.613 (± 0.010)	0.648 (± 0.014)	0.606 (± 0.013)	0.619 (± 0.018)	0.572 (± 0.017)
sCellST	0.648 (± 0.014)	0.619 (± 0.012)	0.585 (± 0.013)	0.563 (± 0.014)	0.624 (± 0.019)	0.571 (± 0.016)
GHIST	0.539 (± 0.015)	0.520 (± 0.013)	0.455 (± 0.014)	0.459 (± 0.013)	0.632 (± 0.020)	0.571 (± 0.017)
TIGER	**0.712*** (± 0.010)	**0.653*** (± 0.010)	**0.747*** (± 0.012)	**0.703*** (± 0.012)	**0.658*** (± 0.018)	**0.590*** (± 0.018)

A.11 BROADER IMPACTS

By judiciously combining the context and focal modules, TIGER accurately captures spatial expression patterns in both highly expressed genes and sparse, low-support genes, leading to improved prediction performance over prior work. In this study, we also show the utility of single-cell virtual ST by performing fine-grained cell phenotyping and improving clinical outcome prediction compared to pathology images alone. We anticipate releasing TIGER as an open-source, user-friendly tool to enable broad application in spatial biology studies.