Predicting Single-Cell Spatial Gene Expression from H&E using Pathology Foundation Models

Ashley P. Tsang[1,2,*] **Evan Liu**[3] **Conrad Foo**[4] **Bob Chen**[1] **Hassaan Maan**[1] **Donny Chan**[1]
Elisa Penna[4] **Veronica Ibarra Lopez**[4] **Hector Corrada Bravo**[1] **Aïcha BenTaieb**[1]
Lisa M. McGinnis[4] **Runmin Wei**[1,†] **Kenneth T. Gao**[1,†]

[1]AI for Biology & Translation, Computational Sciences Center of Excellence, Genentech
[2]Gilbert S. Omenn Department of Computational Medicine & Bioinformatics, University of Michigan
[3]Computational Biology & Medicine, Computational Sciences Center of Excellence, Genentech
[4]Digital and Spatial Pathology, Translational Genomics and Proteomics, gRED, Genentech
[*]Work conducted during an internship at Genentech. [†]Corresponding author.

Abstract

Predicting gene expression directly from hematoxylin and eosin (H&E) images offers a scalable and cost-effective approach to deriving spatial molecular insights from routine histology. Pathology foundation models, typically pretrained using tile-level objectives, may not adequately capture the fine-grained cellular features required for single-cell molecular prediction. In this work, we systematically evaluate and establish domain-informed and task-specific strategies for adapting pathology foundation models to single-cell spatial gene expression prediction. Using paired H&E and Xenium 5K data, our approach improves mean gene-wise Pearson correlation by 2.67$\times$ over the standard zero-shot baseline and achieves state-of-the-art performance on this task. We demonstrate strong gene-detection performance across the full gene panel (AUROC = 0.800) and importantly identify that predictive performance varies substantially across biological contexts, with our model achieving its highest Pearson correlation for spatially variable genes (r = 0.643). Our findings demonstrate that, despite the challenges of sparse single-cell measurements, H&E-derived representations can recover biologically meaningful spatial expression patterns, highlighting both the opportunities and limitations of morphology-based molecular inference at cellular resolution.

1 Introduction

Spatial transcriptomics (ST) enables the study of tissue biology by measuring gene expression in situ [1, 2, 3]. However, these assays are costly, experimentally demanding, and limited in the number of samples that can be profiled at scale. In contrast, hematoxylin and eosin (H&E) staining is inexpensive, routinely collected, and widely available in clinical archives. H&E images capture cellular morphology and tissue organization that reflect underlying molecular states and provide visual signals informative of spatial gene expression [4, 5, 6]. Thus, predicting gene expression from H&E offers a scalable, cost-effective approach to spatial molecular profiling.

Existing H&E-to-ST methods have been developed predominantly for spot-level ST data [4, 7, 8, 9, 10]. In these settings, gene expression is measured at spatially arrayed capture spots, each of which aggregates transcripts from multiple cells. Recent imaging-based spatial transcriptomics platforms enable single-cell resolved expression profiling, revealing how cellular states are organized across different microenvironments and how cells interact with their neighbors [11, 12, 13]. However, this increased spatial granularity makes H&E-based prediction more challenging due to sparse transcript counts and gene-specific variability in probe detection efficiency.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: AI at Scale for Clinical Impact.

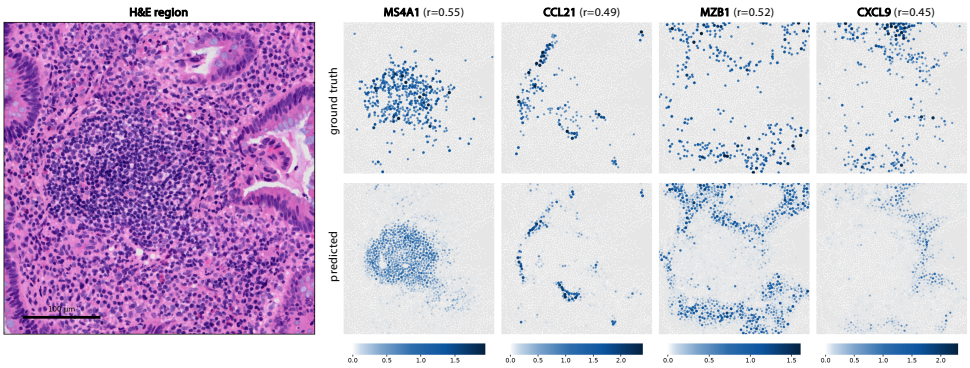


Figure 1: Single-cell predictions from our model in a tertiary lymphoid structure (TLS) region. Our model predicts expression (log1p shown) of the B-cell marker *MS4A1* within the TLS, immune-cell recruitment chemokines *CCL21* and *CXCL9*, and plasma-cell marker *MZB1* at the periphery.

Pathology foundation models (FMs) are a promising basis for this task, providing general-purpose representations for histopathology. These models are pretrained at scale on large collections of pathology images, typically using tile-level self-supervised objectives [14, 15, 16, 17]. Their embeddings capture tissue morphology, but may lack the fine-grained cellular and microenvironmental features needed for single-cell molecular prediction. Because paired H&E and single-cell spatial transcriptomics data remains scarce, training large-scale models for this task is challenging. Leveraging representations learned by existing pathology FMs and adapting them to capture cell-specific features offers a well-motivated strategy for single-cell level gene expression prediction from H&E.

In this work, we establish a practical and effective framework for single-cell gene expression prediction using pathology foundation models. Using a large cohort of paired H&E imaging and Xenium 5K spatial transcriptomics data from patients with inflammatory bowel disease (IBD)—a group of chronic inflammatory disorders associated with an increased risk of colorectal cancer—we systematically implement a series of design choices such as cell-level tokenization, input resolutions, and parameter-efficient fine-tuning to improve single-cell prediction. We show that leveraging pathology FMs as-is works well for spot-level predictions, but is insufficient at single-cell resolution, revealing a key gap from tile-level embeddings to single-cell predictions.

Each of our domain-informed adaptations substantially improves performance across the full gene panel, leading to our model, which integrates all adaptations and achieves state-of-the-art (SOTA) performance in single-cell spatial gene expression prediction. Analyses stratified by biologically informative gene sets, cell types, and pathways reveal substantial heterogeneity in predictability, with particularly strong performance for spatially variable and disease-relevant genes. Together, these results demonstrate that lightweight adaptation enables pathology foundation models to recover biologically meaningful spatial expression patterns at single-cell resolution from H&E.

2 Related Work

Pathology foundation models. Pathology FMs have emerged as powerful feature extractors for histopathology. Models such as UNI [15], Virchow [16], Prov-GigaPath [17], H-Optimus-0 [18] and others [19, 20, 21] are pretrained at scale on H&E images and transfer effectively to various downstream tasks such as tissue classification and biomarker prediction. They commonly process fixed-size image tiles and downstream applications often rely on tile-level embeddings [15, 16, 20]. These representations capture rich regional morphology but are not explicitly optimized to associate visual features with individual target cells. Our work investigates how the information encoded by these pathology FMs can be localized and adapted for cell-resolved molecular prediction.

Spot-level H&E to gene expression prediction. Prior work established the feasibility of predicting spatial gene expression from histology using paired H&E and spot-level ST data. ST-Net [4] learns a supervised mapping from H&E patches to matched expression measurements, while HisToGene [22] and Hist2ST [7] incorporate transformer-based modeling. Other approaches such as BLEEP [8]

and OmiCLIP [9] explore multi-modal representation learning. HEST-1K [23] benchmarks zero-shot pathology FMs for gene expression prediction across various ST datasets. Collectively, these studies demonstrate that histology captures expression signals at the regional level, where spot-level aggregation smooths cell-to-cell variation and noise.

Single-cell H&E to gene expression prediction. Methods such as XFuse [24] and iStar [25] bridge the resolution gap by inferring expression from spot-level ST at super-resolution. sCellST [26] also leverages lower-resolution spot-level ST to predict single-cell level expression. In contrast, our approach and recent methods such as SpatialEx [27] and GHIST [28] leverage single-cell-resolution ST for supervision. SpatialEx combines hypergraph and contrastive learning, whereas GHIST employs multi-task learning with auxiliary cell-type and neighborhood prediction tasks. Our approach achieves SOTA performance through lightweight adaptations of tile-level pathology foundation models, without requiring specialized architectures or auxiliary biological inputs.

3 Methods

3.1 Problem formulation

We formulate single-cell spatial transcriptomics prediction as a supervised learning problem that maps an H&E tile to its gene expression profile. For each cell $i \in \{1, \ldots, N\}$ with centroid c_i, the input is an RGB H&E tile $\mathbf{X}_i \in \mathbb{R}^{H \times W \times 3}$ with height H and width W centered at c_i. The target is $\mathbf{y}_i \in \mathbb{R}^G$, where G is the number of genes to predict. A pretrained pathology foundation model encoder f_{θ_0} produces a sequence of token embeddings $\mathbf{H}_i = f_{\theta_0}(\mathbf{X}_i)$. A pooling function P extracts a representation $\mathbf{z}_i = P(\mathbf{H}_i)$, which a prediction head g_ϕ maps to a G-gene expression vector.

3.2 Linear probe baseline

For the baseline, we adapt the HEST-1K [23] linear probe protocol. We keep f_{θ_0} frozen and use its final-layer class token as the representation: $\mathbf{H}_i = f_{\theta_0}(\mathbf{X}_i), P(\mathbf{H}_i) = \mathbf{h}_{i,\mathrm{cls}}$. A linear head predicts log1p-transformed ($\log(1+y)$) gene expression, $\hat{\mathbf{y}}_i = \mathrm{Linear}(\mathbf{z}_i)$, and is optimized using mean-squared error (MSE) against the log1p-transformed target. We extract an H&E patch centered on the cell centroid, spanning a physical field of view (FOV) of 112×112 μm at 20× magnification. This resolution is commonly used in pretrained pathology FMs and downstream tasks [29, 19, 15, 4]. The resulting embedding is used to predict the cell's corresponding gene expression profile.

3.3 Nonlinear prediction head

While the linear probe baseline maps the image representation to gene expression using a single linear layer, we replace this with a nonlinear prediction head. $\mathbf{z}_i$ is passed to a two-layer multilayer perceptron (MLP) prediction head, g_ϕ, with a 512-dimensional hidden layer and ReLU activation.

3.4 Core design choices for single-cell adaptation

Building on the nonlinear prediction head, we introduce a series of lightweight adaptations tailored to single-cell spatial transcriptomics and gene-expression prediction from H&E images (Figure 2).

Cell-based feature pooling. Many pathology foundation models use the class token as a global representation of an H&E tile [15, 16, 20]. For single-cell prediction, however, we use the granular patch-token embeddings associated with the cell of interest. Let $\mathbf{H}_i = \{\mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,T}\}$ denote the patch-token embeddings for image patch $\mathbf{X}_i$, and let $u_{i,t}$ denote the spatial center of token t. Given the Xenium-derived cell mask M_i, we define the token-membership indicator $m_{i,t} = \mathbf{1}\{u_{i,t} \in M_i\}$, which selects tokens whose spatial centers fall inside the mask. For cells with no token centers inside M_i, we instead select the single token whose center is nearest to c_i, so that at least one token is always selected. We then obtain the cell representation by mean pooling the selected token embeddings: $P(\mathbf{H}_i) = \frac{\sum_{t=1}^{T} m_{i,t}\mathbf{h}_{i,t}}{\sum_{t=1}^{T} m_{i,t}}$.

Input resolution. Spot-level prediction typically uses 224×224-pixel H&E tiles at 0.5 μm/pixel (20×), corresponding to a 112×112 μm FOV [4, 29, 15, 19]. For single-cell prediction, we instead



Figure 2: Framework for predicting single-cell gene expression from H&E images. We integrate domain-informed design choices including high-resolution inputs, parameter-efficient encoder fine-tuning, cell-localized features, and a Poisson-trained MLP to predict expression across 5K genes.

extract patches spanning a 47.6×47.6 μm FOV at 40× resolution. This configuration provides finer spatial detail around the target cell while retaining local tissue context.

Loss function. Because Xenium expression measurements are sparse, non-negative counts, we train on raw transcript counts using the Poisson negative log-likelihood rather than the MSE on log1p-transformed targets [30, 31]. The model predicts gene-specific rates $\boldsymbol{\lambda}_i = \mathrm{softplus}(g_\phi(\mathbf{z}_i))$, where the softplus activation ensures that all predicted rates are positive. The training objective is

$$\mathcal{L}_{\text{Poisson}} = \frac{1}{NG} \sum_{i=1}^{N} \sum_{g=1}^{G} \left(\lambda_{ig} - y_{ig} \log \lambda_{ig} + \log(y_{ig}!)\right),$$

where y_{ig} is the raw transcript count for gene g in cell i. For evaluation, predicted rates and observed counts are log1p-transformed.

Parameter-efficient fine-tuning. Since pathology FMs were not directly trained for cell-resolved transcriptomic prediction, we further augment f_{θ_0} with Low-Rank Adaptation (LoRA) [32] parameters ψ, producing the histology representation $\mathbf{H}_i = f_{\theta_0,\psi}(\mathbf{X}_i)$. Keeping θ_0 fixed, we optimize only the LoRA adapters ψ and the prediction head. We add LoRA adapters to the query, key, and value projections in every self-attention block, using rank r = 16, scaling factor α = 32, and adapter dropout 0.05. The loss is computed only from the cell-localized tokens; since self-attention operates globally over the image tile, these tokens nevertheless incorporate information from the entire tile.

4 Experiments

4.1 Experimental setup

Dataset. We used two colon tissue datasets with paired H&E imaging and 10x Genomics Xenium 5K Human Colon profiling. H&E images were acquired at 40× resolution (0.2125 μm/pixel), registered to the Xenium data, and paired with cell boundaries generated by the standard on-instrument segmentation pipeline. Of the 5,399 profiled genes, zero-variance genes were excluded, leaving 5,101 genes for training and evaluation.

The tissue microarray (TMA) cohort comprised 176 ulcerative colitis (UC) cores, spanning 117 donors, 10 TMA blocks, and totaling about 8.6 million cells. All cores from a given donor were placed on the same TMA block; thus, a block-level splitting strategy eliminates biological leakage from donor-associated effects and minimizes technical leakage from block-associated effects. Five blocks (85 cores) were used for fine-tuning the image encoder, and five (91 cores) for downstream evaluation. Evaluation used five-fold cross-validation at the block level: three blocks for training, one for validation, and one for testing. The resection cohort comprised two UC, two Crohn's disease, and two healthy control samples, totaling about 2.9 million cells, and were used for fine-tuning only.

Metrics. Performance on held-out cells was evaluated primarily using gene-wise Pearson correlation coefficient (PCC), computed between log1p-transformed predicted and measured expression across cells [8, 26, 10]. For each gene panel, we report mean PCC across its genes. We also compute per-cell PCC across all genes and report the mean across cells. Gene-detection performance was assessed using gene-wise area under the receiver operating characteristic curve (AUROC), with measured counts binarized as detected ($y_{ig} > 0$), and corresponding continuous predictions used as scores.

Table 1: Mean per-gene PCC across cumulative model adaptations. Values in parentheses indicate relative improvement over the preceding configuration. †Configuration: linear probe, class-token embeddings, 20× input, MSE loss, and frozen embeddings.

Model	SVG50	IBD	HVG50	5K
OmiCLIP [9]	0.207	0.072	0.057	0.029
BLEEP [8]	0.349	0.160	0.117	0.065
GHIST [28]	0.505	0.167	0.087	0.081
SpatialEx [27]	0.509	0.207	0.158	0.094
Baseline†	0.469	0.182	0.151	0.065
+ MLP	0.525 (+12.0%)	0.235 (+28.8%)	0.195 (+29.5%)	0.110 (+69.0%)
+ cell_pool	0.583 (+10.9%)	0.285 (+21.6%)	0.253 (+29.5%)	0.135 (+22.6%)
+ 40×	0.602 (+3.3%)	0.298 (+4.4%)	0.248 (−1.8%)	0.146 (+7.8%)
+ Poisson	0.618 (+2.7%)	0.324 (+8.8%)	0.301 (+21.0%)	0.161 (+10.6%)
+ LoRA (Ours)	**0.643** (+4.0%)	**0.356** (+9.7%)	**0.338** (+12.5%)	**0.174** (+7.9%)
Baseline → Ours	**+37.1%**	**+95.6%**	**+124.2%**	**+166.7%**

Metrics were computed separately for each test fold and summarized as mean ± standard deviation. Full metric definitions are provided in Appendix A.1.

Evaluations. We evaluate four gene panels for comprehensive evaluation: all 5,101 genes (5K), the top 50 highly variable genes (HVG50), the top 50 spatially variable genes (SVG50), and 76 literature-derived IBD-relevant genes (IBD) [33, 34, 35, 36, 37]. HVG50 and SVG50 were selected independently from each fold's training data. HVGs were ranked by normalized dispersion using Scanpy [38], and SVGs by Moran's I [39], both using log1p-transformed raw counts.

Implementation details. The following experiments use H-Optimus-0 [18], a pathology foundation model based on a ViT-g/14 [40] with a DINOv2 [14] backbone, which generates 1536-dimensional feature representations. Prediction heads were trained independently within each fold using AdamW [41]. Early stopping and model selection used mean per-gene PCC on the validation split, and the best checkpoint was evaluated once on the held-out test split. LoRA [32] adaptation used a separate cohort disjoint from all evaluation folds (Section 4.1). As with f_{θ_0}, we use $f_{\theta_0,\psi}$ as a frozen feature extractor and train the prediction head using the same cross-validation protocol. Full optimization and LoRA settings are provided in Appendix A.2.

4.2 Pathology foundation models do not natively transfer to single-cell expression prediction

We first assessed how well frozen pathology foundation models transfer to gene expression prediction at spot and single-cell resolution. Following the HEST-1K linear-probing framework [23], we evaluated both tasks using the same Xenium data. To construct Visium-like pseudo-spots, we assigned cells to a regular grid of 55×55 μm bins, approximating the diameter of a Visium spot. For each pseudo-spot, raw gene counts were summed and log1p-transformed. Each pseudo-spot was paired with a centered 112×112 μm H&E tile, resized to 224×224 pixels at 0.5 μm/pixel, to predict its corresponding pseudo-bulked expression profile. On HVG50, the mean per-gene PCC was 0.428, comparable to the spot-level performance reported by HEST-1K [23].

Linear probes on frozen pathology FM embeddings performed substantially better at the pseudo-spot level than at the single-cell level, suggesting that these embeddings capture histology-associated gene expression patterns more effectively at regional than at cellular resolution. For the full 5K panel, mean per-gene PCC decreased from 0.375 at pseudo-spot resolution to 0.065 at single-cell resolution. Similar declines were observed for the other gene panels (Figure 3a), although performance varied substantially across panels, with SVG50 achieving the highest PCCs. Together, these results suggest that native pathology FM embeddings are better suited to regional than cell-resolved expression prediction, with performance depending on both spatial resolution and gene-panel composition. We further observed that PCC declined with increasing gene expression sparsity. At the spot-level, performance decreased gradually, whereas at the single-cell level, the decline was substantially more pronounced (Figure 3b). This pattern highlights the greater challenge of inferring sparse gene expression at single-cell resolution using native pathology foundation model embeddings.

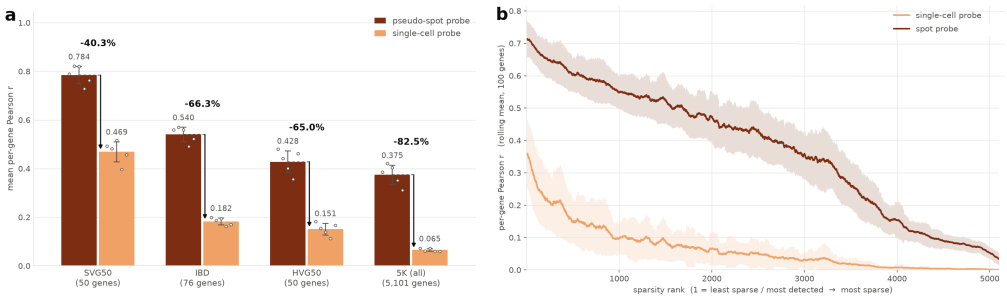


Figure 3: Baseline performance with frozen pathology foundation model embeddings at pseudo-spot and single-cell resolutions: (a) across gene panels; (b) per-gene over increasing expression sparsity.

Given these limitations, we replaced the linear probe with a two-layer MLP decoder, which improved performance across all gene panels (Table 1, +MLP). These gains suggest that frozen FM embeddings retain expression-relevant information that is better recovered by a higher-capacity nonlinear decoder, and standard linear probes may be insufficiently expressive for single-cell expression prediction.

4.3 Single-cell-specific adaptations yield substantial cumulative performance gains

Cell-localized token pooling improves single-cell representations. Pooling tokens within cell boundaries substantially improved performance relative to the class token baseline (Table 1, +cell_pool). We also evaluated centroid-based pooling strategies that do not require cell-boundary masks. For each cell, we selected tokens within a radius of 5 μm from its centroid and aggregated them using mean pooling or distance-weighted pooling (Appendix A.3, Table 3). Compared with mask-based pooling, distance-weighted and mean pooling reduced PCC for 5K by only 4.6% and 7.3%, respectively. Both approaches nevertheless substantially outperformed the class token baseline, where distance-weighted and mean pooling improved PCC by 20.2% and 16.4%, respectively.

These results show that restricting representations to the target cell and its immediate surroundings is more effective than using a global tile-level embedding, and that much of this benefit can be retained without explicit cell boundaries. In an H&E-only deployment setting, cell nuclei or boundaries would need to be inferred using a segmentation model [42, 43, 44]. To simulate this setting without the Xenium-derived cell information, we used CellViT++ [43] to detect nuclei centroids in H&E images and found high concordance with the Xenium-derived nuclei coordinates (Appendix A.4). Together with the strong performance of distance-weighted pooling, these findings support the feasibility of nuclei-based cell-localized pooling in H&E-only settings, while highlighting the potential for further gains from explicit cell boundaries.

40$\times$ inputs balance cellular detail with local tissue context. 40$\times$ inputs (47.6$\times$47.6 μm FOV) improved performance compared to 20$\times$ inputs (Table 1, +40$\times$). At this resolution, each cell spans more image pixels and spatial tokens, providing a richer representation of cellular and subcellular morphology, while also retaining local tissue context around the target cell (Appendix A.5, Figure 7). In comparison, using a broader FOV at 10$\times$ reduced PCC on 5K by 19.8% relative to 40$\times$ (Appendix A.5, Table 5), likely because the surrounding tissue dilutes target-cell-specific signals. On the other extreme, isolated single-cell crops reduced PCC by 15.5% (Appendix A.5, Table 5). Such crops remove potentially informative microenvironmental context and may also be poorly matched to pathology FM pretraining, which typically captures tiles of tissue regions rather than isolated cells.

Poisson count modeling improves single-cell expression prediction. Poisson supervision improved performance across all gene subsets (Table 1, +Poisson), with the largest gain observed for HVG50. These results indicate that modeling the raw transcript counts directly provides a more informative supervisory signal for single-cell prediction than regression on transformed expression values. Compared with a Gaussian regression, the Poisson objective better reflects the discrete nature and mean–variance relationship of count data, consistent with previous findings [30, 31].

Fine-tuning further aligns H&E features with single-cell expression. LoRA fine-tuning improved mean per-gene PCC across all gene subsets (Table 1, +LoRA). While our frozen adapted

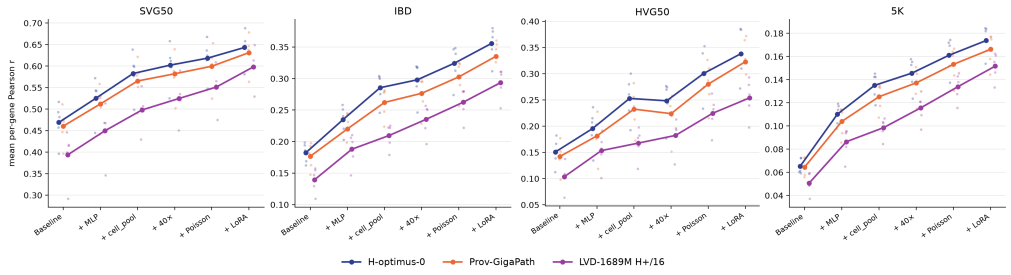


Figure 4: Our domain-driven model adaptations generalize across different foundation models.

model already translated well to the task, parameter-efficient task adaptation further aligned the features with single-cell gene expression prediction (Appendix A.6, Figure 8).

Our adapted model substantially outperformed the pre-adaptation baseline, improving performance by 166.7% on 5K, 124.2% on HVG50, 95.6% on IBD, and 37.1% on SVG50, and also achieved SOTA performance for this task (Table 1). The single-cell predictions from our model accurately recapitulated expression patterns within the tertiary lymphoid structure in colon tissue, including the spatial localization of relevant immune and signaling markers (Figure 1). Our model also demonstrated significant strengths in predicting epithelial marker genes, across both healthy and diseased samples (Appendix A.7, Figure 9).

4.4 Our adaptations generalize across foundation models

We applied these domain-driven adaptations to two additional foundation models to evaluate if these improvements generalize beyond H-Optimus-0 [18]. Prov-GigaPath [17] was selected as another pathology FM due to its comparable use of DINOv2, a 14×14 patch size, and an embedding dimension of 1536. We also evaluated the distilled DINOv3 [45] model trained on the non-histopathology LVD-1689M dataset, which uses a 16×16 patch size and a 1280 embedding dimension. We observe that our improvements successfully transfer regardless of the foundation model used (Figure 4).

4.5 Our adapted single-cell model outperforms existing H&E-to-ST methods

We benchmarked our model against several existing approaches for spatial transcriptomics prediction from H&E, including SpatialEx [27], GHIST [28], OmiCLIP [9], and BLEEP [8]. SpatialEx and GHIST predict single-cell gene expression using supervision from single-cell-resolution ST data. OmiCLIP is a visual–omics foundation model trained on Visium-scale ST and BLEEP introduced a contrastive retrieval framework for spot-level prediction from H&E. For BLEEP, we lowered the contrastive temperature to $\tau = 0.07$ to obtain a stable baseline for single-cell predictions. OmiCLIP was evaluated in a zero-shot setting, while GHIST and SpatialEx were trained end-to-end. For GHIST, its cell-type, neighborhood, and reference-expression components were disabled for a fair comparison with the intended H&E-only constraint.

Our adapted model, which combines all the task-specific adaptations, achieved the best performance across all evaluated metrics (Table 2). It outperformed SpatialEx, the strongest competing method, on all three metrics on the 5K panel: mean per-gene PCC (0.174 ± 0.014 vs. 0.094 ± 0.009), mean per-cell PCC (0.280 ± 0.033 vs. 0.242 ± 0.031), and AUROC (0.800 ± 0.009 vs. 0.642 ± 0.005). Even without LoRA fine-tuning, our frozen adapted model ranked second across all metrics, demonstrating that the single-cell architectural adaptations alone substantially improve upon existing approaches. Although precise spatial expression prediction across 5K genes at single-cell resolution remains challenging, our method consistently outperformed competing approaches, with its strong gene-detection AUROC also indicating reliable detection of expression signals.

4.6 Aggregated single-cell predictions improve spot-level expression prediction

At single-cell resolution, our model achieved higher per-gene PCCs across every gene in the 5K panel and consistently outperformed the baseline across levels of gene expression sparsity (Appendix A.8, Figure 10a). To assess whether cell-level predictions recovered coherent regional expression patterns,

Table 2: Benchmarking single-cell gene expression predictions from H&E on the 5K panel. We report mean per-gene PCC, mean per-cell PCC, and AUROC. Native granularity indicates the resolution-level the model was trained on; spot-level methods were evaluated on the single-cell prediction task.

Model	Native granularity	PCC (per-gene)	PCC (per-cell)	AUROC
OmiCLIP [9]	Spot-level	0.029 ± 0.006	0.203 ± 0.026	0.568 ± 0.016
BLEEP [8]	Spot-level	0.065 ± 0.041	0.124 ± 0.078	0.578 ± 0.062
GHIST [28]	Single-cell	0.081 ± 0.007	0.242 ± 0.027	0.605 ± 0.012
SpatialEx [27]	Single-cell	0.094 ± 0.009	0.242 ± 0.031	0.642 ± 0.005
Ours (frozen)	Single-cell	0.161 ± 0.016	0.271 ± 0.034	0.786 ± 0.005
Ours (LoRA)	Single-cell	$\mathbf{0.174 \pm 0.014}$	$\mathbf{0.280 \pm 0.033}$	$\mathbf{0.800 \pm 0.009}$

we aggregated predicted and measured single-cell expression into 55 μm and 100 μm pseudo-spots. Values were converted to counts, summed within each spot, and log1p-transformed. Aggregation improved correlations overall, and our model maintained a substantial advantage for every gene at both resolutions (Appendix A.8, Figure 10b, c). These results demonstrate that our model, optimized for single-cell prediction, also yields more accurate spatially aggregated expression profiles.

4.7 Cell-type-stratified analysis reveals populations that benefit most from our approach

We next investigated whether our adaptations preferentially benefited specific cell types. We computed the per-cell PCC across all genes and averaged these scores within each annotated cell type. Cell-type labels were obtained using CONCORD [46], which jointly embeds a reference scRNA-seq atlas with the ST data, followed by annotation with a calibrated k-nearest-neighbor classifier (Appendix A.9).

Cell-localized pooling particularly benefited stromal cells (Figure 5a). Overall, stromal cells showed the largest gains from baseline to our model, while epithelial cells showed the smallest. Epithelial cells already performed relatively well at baseline (Figure 5b), possibly because their cohesive organization and distinct architecture provide strong morphological cues in H&E images [47]. In contrast, stromal cells are heterogeneous and interact with diverse cellular and extracellular components in their microenvironment [48] (Figure 5c), potentially making them more sensitive to the inclusion of features specific to the target cell. The global class token may capture signals from the surrounding mixture of cells, whereas cell-localized pooling emphasizes target-cell features. For epithelial cells, these signals may be more concordant, limiting the additional benefit of localized pooling. T lymphocytes also showed relatively small gains, potentially because their small size and relatively uniform nuclear morphology in H&E provide limited information for molecular discrimination [49].

These results suggest that single-cell adaptations, such as cell-token pooling, are beneficial for cell types whose expression-associated features are poorly isolated by global, tile-level representations. More broadly, they emphasize the importance of evaluating performance across biologically informed populations rather than relying exclusively on aggregate metrics, which can obscure cell-type-specific improvements and persistent performance gaps.

4.8 Pathway analysis identifies biological programs associated with single-cell adaptations

To identify the biological programs that benefited most from single-cell adaptations, we ranked genes by the change in mean per-gene PCC relative to the baseline. We performed REACTOME [50] pathway enrichment analysis on the 500 most improved and 500 least improved genes using g:Profiler [51] (details in Appendix A.10).

The most improved genes from our model were enriched for pathways related to extracellular matrix organization, cell junctions, adhesion, and cell-cell communication pathways (Appendix A.10, Figure 11a). These processes are closely linked to tissue architecture and local cell–cell interactions, suggesting that the adaptations particularly improve prediction of genes with visible morphological or microenvironmental correlates. This is also consistent with our cell-type-stratified analysis, in which stromal populations exhibited large performance gains following single-cell adaptations.

The least improved genes were enriched for G-protein-coupled receptor (GPCR) signaling and neurotransmitter-related pathways (Appendix A.10, Figure 11b). These may represent signaling states that are transient, weakly reflected in H&E morphology, or dependent on molecular context

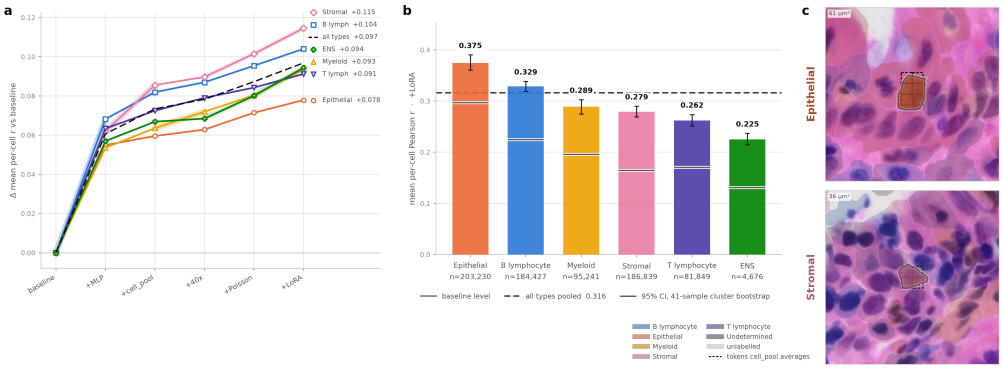


Figure 5: Cell-type-specific effects of single-cell modeling adaptations. (a) Gains in performance across adaptations. (b) Performance stratified by cell type. (c) Example H&E image regions.

that cannot be inferred from histology alone. Enrichment among the least-improved genes may also be influenced by technical and statistical factors, including sparsity, limited dynamic range, or low detection rates. We therefore interpret these enrichment results as hypothesis-generating rather than as definitive evidence that particular biological programs are or are not morphologically predictable.

5 Conclusion

In this work, we demonstrate that a set of domain-driven, task-specific design choices achieves state-of-the-art performance for predicting single-cell gene expression from H&E images. We found that frozen pathology foundation model embedding approaches transfer poorly to cell-resolved targets, motivating adaptations that better reflect the spatial resolution, sparsity, and complexity of single-cell data. Nonlinear decoding with additional parameters, cell-localized spatial-token pooling, 40$\times$ inputs that preserve both cellular detail and local tissue context, Poisson supervision, and LoRA fine-tuning together establish substantial improvements over existing methods for this task.

We further showed that aggregate performance metrics alone are insufficient to characterize this task. By evaluating performance across gene panels, expression-sparsity levels, cell types, spatial aggregation scales, and biological pathways, we identified where single-cell adaptation provides the greatest benefit and where important performance gaps remain. The modest overall correlation partly reflects the sparsity and detection variability of single-cell transcript measurements, which can limit observed correlations despite meaningful biological signals. Predictive performance varies substantially across genes, with notably higher Pearson correlations for spatially variable genes (r = 0.643) and IBD-associated genes (r = 0.356), compared with the full 5K panel (r = 0.174). Our model recovers spatial expression patterns of key immune and epithelial markers, suggesting its potential to characterize tissue microenvironments and spatially organized molecular programs.

Our work has several limitations. Our experiments relied on single-cell imaging-based ST technologies, which are susceptible to transcript spillover, imperfect cell segmentation, and cell-assignment errors. Further evaluation is needed to determine whether these predictions capture subtle cell states and within-cell-type molecular variation beyond broad cell identity. In addition, broader validation across tissues and cohorts is needed to establish generalizability. Future work should further characterize model uncertainty, particularly on a per-gene and per-cell-type basis. Although the relatively low correlation scores in the 5K predictions indicate the difficulty of predicting single-cell gene expression from H&E, our downstream cell type and pathway enrichment analyses demonstrate that the resulting predictions can provide meaningful insights into the model's behavior and capabilities.

Such analyses will be essential for determining when molecular profiles inferred from H&E are sufficiently accurate, reliable, and useful in research and translational settings. Overall, our work establishes a foundation for adapting pathology foundation models to single-cell spatial transcriptomics. We demonstrate that carefully chosen, lightweight modifications to baseline foundation model approaches—rather than extensive architectural or methodological redesign—can enable scalable and interpretable molecular profiling from routine histology.

Acknowledgments and Disclosure of Funding

We are grateful to Alma Andersson for feedback on the manuscript and valuable input, and Rushin Gindra for insightful discussions.

References

[1] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[2] 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.

[3] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C Devlin, Maayan Baron, Cristina H Hajdu, Diane M Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology*, 38(3):333–342, 2020.

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

[5] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):3877, 2020.

[6] Alona Levy-Jurgenson, Xavier Tekpli, Vessela N Kristensen, and Zohar Yakhini. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific reports*, 10(1):18802, 2020.

[7] 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):bbac297, 2022.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] Shanshan He, Ruchir Bhatt, Carl Brown, Emily A Brown, Derek L Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G Garrison, Gary Geiss, et al. High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature biotechnology*, 40(12):1794–1806, 2022.

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[15] 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.

[16] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10):2924–2935, 2024.

[17] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[18] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024.

[19] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

[20] Maximilian Alber, Timo Milbich, Alexandra Carpen-Amarie, Stephan Tietz, Jonas Dippel, Lukas Muttenthaler, Beatriz Perez Cancer, Alessandro Benetti, Panos Korfiatis, Elias Eulig, Jérôme Lüscher, Jiasen Wu, Sayed Abid Hashimi, Gabriel Dernbach, Simon Schallenberg, Neelay Shah, Moritz Krügener, Aniruddh Jammoria, Jake Matras, Patrick Duffy, Matt Redlon, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Andrew Norgan. Atlas 2 – foundation models for clinical deployment, 2026.

[21] Harshith Padigela, Shima Nofallah, Atchuth Naveen Chilaparasetti, Ryun Han, Andrew Walker, Judy Shen, Chintan Shah, Blake Martin, Aashish Sood, Elliot Miller, Ben Glass, Andy Beck, Harsha Pokkalla, and Syed Ashar Javed. Pluto-4: Frontier pathology foundation models, 2025.

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

[23] 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.

[24] Ludvig Bergenstråhle, Bryan He, Joseph Bergenstråhle, Xesús Abalo, Reza Mirzazadeh, Kim Thrane, Andrew L Ji, Alma Andersson, Ludvig Larsson, Nathalie Stakenborg, et al. Super-resolved spatial transcriptomics by deep data fusion. *Nature biotechnology*, 40(4):476–479, 2022.

[25] Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle Y. Y. Lee, Kyung S. Cho, Katalin Susztak, George X. Xu, Michael D. Feldman, Edward B. Lee, Emma E. Furth, Linghua Wang, and Mingyao Li. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 42(9):1372–1377, January 2024.

[26] 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.

[27] Yonghao Liu, Chuyao Wang, Zhikang Wang, Liang Chen, Zhi Li, Jiangning Song, Qi Zou, Rui Gao, Bin-Zhi Qian, Xiaoyue Feng, et al. High-parameter spatial multi-omics through histology-anchored integration. *Nature Methods*, 23(2):373–386, 2026.

[28] 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.

[29] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

[30] Peiyao Zhao, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biology*, 23, 2022.

[31] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, 53:770–777, 2021.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[33] Alba Garrido-Trigo, Ana M. Corraliza, Marisol Veny, Isabella Dotti, Elisa Melón-Ardanaz, Aina Rill, Helena L. Crowell, Ángel Corbí, Victoria Gudiño, Miriam Esteller, Iris Álvarez Teubel, Daniel Aguilar, M. Carme Masamunt, Emily Killingbeck, Youngmi Kim, Michael Leon, Sudha Visvanathan, Domenica Marchese, Ginevra Caratù, Albert Martin-Cardona, Maria Esteve, Ingrid Ordás, Julian Panés, Elena Ricart, Elisabetta Mereu, Holger Heyn, and Azucena Salas. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nature Communications*, 14, 2023.

[34] Grzegorz Maciag, Stine Lind Hansen, Kata Krizic, Lauge Kellermann, Maureen Joy Inventor Zøylner, Svetlana Ulyanchenko, Martti Maimets, Astrid Møller Baattrup, Lene Buhl Riis, Konstantin Khodosevich, Toshiro Sato, Raul Bardini Bressan, Ole Haagen Nielsen, and Kim B. Jensen. Jak/stat signaling promotes the emergence of unique cell states in ulcerative colitis. *Stem Cell Reports*, 19:1172–1188, 2024.

[35] Colleen McGregor, Xiao Qin, Marta Jagielowicz, Tarun Gupta, Zinan Yin, Verena Lentsch, David Fawkner-Corbett, Vy Wien Lai, Paula Gomez Castro, Esther Bridges, Chloe Hyun-Jung Lee, Huei-Wen Chuang, Lei Deng, Anna Aulicino, Renuka Teague, Sorayya Moradi, Jun Sung Park, Jeongmin Woo, Kexin Xu, Ruchi Tandon, Nicole Cianci, Jan Bornschein, Ling-Pei Ho, Paulina Siejka-Zielinska, Zoe Christoforidou, Sarah Hill, Johannes Lehmann, Rhea Kujawa, Paola Vargas Gutierrez, Carol Cheng, Maria Greco, Katherine Baker, Mark Bignell, Bruce George, Eve Fryer, Michael Vieth, Agne Antanaviciute, and Alison Simmons. Spatial fibroblast niches define crohn's fistulae. *Nature*, 649:703–712, 2025.

[36] Rasa Elmentaite, Natsuhiko Kumasaka, Kenny Roberts, Aaron Fleming, Emma Dann, Hamish W. King, Vitalii Kleshchevnikov, Monika Dabrowska, Sophie Pritchard, Liam Bolt, Sara F. Vieira, Lira Mamanova, Ni Huang, Francesca Perrone, Issac Goh Kai'En, Steven N. Lisgo, Matilda Katan, Steven Leonard, Thomas R. W. Oliver, C. Elizabeth Hook, Komal Nayak, Lia S. Campos, Cecilia Domínguez Conde, Emily Stephenson, Justin Engelbert, Rachel A. Botting, Krzysztof Polanski, Stijn van Dongen, Minal Patel, Michael D. Morgan, John C. Marioni, Omer Ali Bayraktar, Kerstin B. Meyer, Xiaoling He, Roger A. Barker, Holm H. Uhlig, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Matthias Zilbauer, Menna R. Clatworthy, Muzlifah Haniffa, Kylie R. James, and Sarah A. Teichmann. Cells of the human intestinal tract mapped across space and time. *Nature*, 597:250–255, 2021.

[37] Georgios Theocharidis, Beena E. Thomas, Debasree Sarkar, Hope L. Mumme, William J. R. Pilcher, Bhakti Dwivedi, Teresa Sandoval-Schaefer, Ruxandra F. Sîrbulescu, Antonios Kafanas, Ikram Mezghani, Peng Wang, Antonio Lobao, Ioannis S. Vlachos, Biraja Dash, Henry C. Hsia, Valerie Horsley, Swati S. Bhasin, Aristidis Veves, and Manoj Bhasin. Single cell transcriptomic landscape of diabetic foot ulcers. *Nature Communications*, 13, 2022.

[38] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[39] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[40] 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.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[42] 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.

[43] Fabian Hörst, Moritz Rempe, Helmut Becker, Lukas Heine, Julius Keyl, and Jens Kleesiek. Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models. *Computer Methods and Programs in Biomedicine*, page 109206, 2026.

[44] Markus Marks, Uriah Israel, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Ahamed Iqbal, Elora Pradhan, Ada Ates, Martin Abt, et al. Cellsam: a foundation model for cell segmentation. *Nature Methods*, 22(12):2585–2593, 2025.

[45] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

[46] Qin Zhu, Zuzhi Jiang, Binyamin Zuckerman, Leor Weinberger, Matt Thomson, and Zev J. Gartner. Revealing a coherent cell-state landscape across single-cell datasets with concord. *Nature Biotechnology*, 2026.

[47] Vincenzo Villanacci, Luca Reggiani-Bonetti, Tiziana Salviato, Giuseppe Leoncini, Moris Cadei, Luca Albarello, Alessandro Caputo, Maria Costanza Aquilano, Serena Battista, and Paola Parente. Histopathology of ibd colitis. a practical approach from the pathologists of the italian group for the study of the gastrointestinal tract (gipad). *Pathologica*, 113(1):39, 2021.

[48] MC Barnhoorn, SK Hakuno, RS Bruckner, G Rogler, LJAC Hawinkels, and M Scharl. Stromal cells in the pathogenesis of inflammatory bowel disease. *Journal of Crohn's and Colitis*, 14(7):995–1009, 2020.

[49] R Luz Elena Cano and H Damaris E Lopera. Introduction to t and b lymphocytes. In *Autoimmunity: from bench to bedside [Internet]*. El Rosario University Press, 2013.

[50] Eliot Ragueneau, Chuqiao Gong, Pierre Sinquin, Cristoffer Sevilla, Deidre Beavers, Alexander Grentner, Johannes Griss, Gregory F J Hogue, Nancy T Li, Lisa Matthews, Bruce May, Marija Milacic, Helia Mohammadi, Robert Petryszak, Karen Rothfels, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Joel Weiser, Adam Wright, Marc Gillespie, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome knowledgebase 2026. *Nucleic Acids Research*, 54:D673–D681, 2025.

[51] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic acids research*, 51(W1):W207–W212, 2023.

A Appendix

A.1 Formal metric definitions

For each gene g, we compute the Pearson correlation (PCC) across cells in the evaluation set:

$$r_g = \frac{\sum_{i=1}^{N_{\text{test}}} \left(\hat{y}_{ig} - \bar{\hat{y}}_g\right)\left(y_{ig} - \bar{y}_g\right)}{\sqrt{\sum_{i=1}^{N_{\text{test}}} \left(\hat{y}_{ig} - \bar{\hat{y}}_g\right)^2}\sqrt{\sum_{i=1}^{N_{\text{test}}} \left(y_{ig} - \bar{y}_g\right)^2}},$$

where y_{ig} and $\hat{y}_{ig}$ denote the measured and predicted expression values for gene g in cell i, respectively, and $\bar{y}_g$ and $\bar{\hat{y}}_g$ denote the corresponding means across the evaluation cells. Overall performance for a gene set $\mathcal{G}$ is summarized as the mean PCC across genes,

$$\text{PCC}(\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} r_g.$$

To inform correlation in a cell-centric manner, we also compute PCC at the cell level. For cell i, the per-cell PCC is

$$r_i = \frac{\sum_{g\in\mathcal{G}} \left(\hat{s}_{ig} - \bar{\hat{s}}_i\right)\left(s_{ig} - \bar{s}_i\right)}{\sqrt{\sum_{g\in\mathcal{G}} \left(\hat{s}_{ig} - \bar{\hat{s}}_i\right)^2}\sqrt{\sum_{g\in\mathcal{G}} \left(s_{ig} - \bar{s}_i\right)^2}},$$

where $\bar{s}_i$ and $\bar{\hat{s}}_i$ denote the mean measured and predicted expression across genes in $\mathcal{G}$ for cell i.

To assess whether models correctly identify gene detection events, we compute AUROC for each gene by binarizing measured expression as

$$a_{ig} = \mathbb{I}(y_{ig} > 0),$$

where $a_{ig} = 1$ indicates that gene g was detected in cell i. The continuous prediction score was the predicted expression value $\hat{s}_{ig}$. For each gene g, let $\mathcal{P}_g = \{i : a_{ig} = 1\}$ and $\mathcal{N}_g = \{i : a_{ig} = 0\}$. The gene-wise AUROC is

$$\text{AUROC}_g = \frac{1}{|\mathcal{P}_g||\mathcal{N}_g|}\sum_{i\in\mathcal{P}_g}\sum_{j\in\mathcal{N}_g}\left[\mathbb{I}(\hat{s}_{ig} > \hat{s}_{jg}) + \frac{1}{2}\mathbb{I}(\hat{s}_{ig} = \hat{s}_{jg})\right].$$

AUROC for a gene set $\mathcal{G}$ is summarized as

$$\text{AUROC}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \text{AUROC}_g.$$

Genes with only one detected class in an evaluation fold, i.e. all detected or all undetected, were excluded from AUROC calculation for that fold.

A.2 Implementation full details

Heads are optimized with AdamW [41] at a learning rate of 10^{-3}, batch size 8192, for at most 30 epochs with gradient clipping at 1.0 and mixed precision, on a single A100 GPU. Model selection and early stopping use mean per-gene PCC on the fold's validation split (patience 5 epochs); the best-scoring checkpoint is restored before the held-out test split is scored once.

LoRA [32] is run on the fine-tuning cohort, disjoint from the evaluation folds. The encoder and the head use separate learning rates (10^{-4} and 10^{-3}), with 500 steps of linear warm-up followed by cosine decay to 10^{-6}, AdamW, batch size 128 with gradient accumulation over 2 steps (effective batch 256), gradient clipping at 1.0, and a cap of 30,000 optimizer steps. We early-stop on validation mean per-gene PCC with patience 4 and a minimum improvement of 0.001. The adapted encoder is then frozen and used as a feature extractor: all cells are re-embedded, and the readout is refit under exactly the same k-fold protocol as the frozen encoders, so the adaptation and frozen rungs differ only in the encoder weights.

A.3 Cell-token pooling experiments

Tokens within a fixed $5\mu\mathrm{m}$ radius of the centroid (12 of 256 tokens at $40\times$) were pooled either uniformly or using Gaussian distance weights, $w_k = \exp(-\frac{1}{2}(d_k/\sigma)^2)$, with $\sigma = 2.5\mu\mathrm{m}$. Distance-weighted pooling performed slightly worse than mask-guided pooling (PCC = 0.154 vs. 0.161) but better than uniform radius pooling (PCC = 0.149) and the class-token baseline (PCC = 0.128).

Table 3: Comparison of cell-token pooling strategies within $40\times$ tiles. Performance is reported as mean per-gene PCC across the 5K panel; Δ denotes the absolute and relative change from mask-guided pooling.

Configuration	PCC (5K)	Δ
Cell mask (reference)	**0.161 ± 0.016**	—
Disc r=5 µm, Gaussian σ=2.5 µm	0.154 ± 0.016	−0.007 (−4.6%)
Disc r=5 µm, unweighted mean	0.149 ± 0.015	−0.012 (−7.3%)
Class-token (baseline)	0.128 ± 0.013	−0.035 (−20.2%)

A.4 H&E-derived centroids

To assess the concordance between cell coordinates derived from the Xenium platform and those derived from H&E only, we generated nuclei centroids for all samples in the dataset using CellViT++ [43] and the CellViT-SAM-H-x40-AMP backbone at $40\times$ and 0.2125 μm/pixel. This gave 2,373,383 nuclei across the 91 samples as compared to the 2,451,764 total cells from Xenium. Next, a one-to-one matching was performed in nested distance tiers of 2, 3, and 5 μm, where at each tier, the optimal assignment among unmatched cells was computed with the Hungarian algorithm to maximize the number of pairs while minimizing the total distance between pairs. Matching resulted in a precision of 0.901 and recall of 0.876 using the conservative distance of between cell and nuclei centroids (Table 4, Figure 6).

Table 4: Agreement between CellViT++ nuclei centroids and Xenium platform-derived cell centroids. Values are median (interquartile range). Thresholds are cumulative: a pair counts if its distance is within the threshold in the nested one-to-one matching.

Distance threshold	Precision	Recall	F1
$\leq 2\ \mu$m	0.758 (0.742–0.776)	0.739 (0.706–0.764)	0.748 (0.722–0.766)
$\leq 3\ \mu$m	0.853 (0.844–0.865)	0.832 (0.809–0.851)	0.841 (0.827–0.854)
$\leq 5\ \mu$m	0.901 (0.892–0.909)	0.876 (0.855–0.890)	0.887 (0.876–0.896)

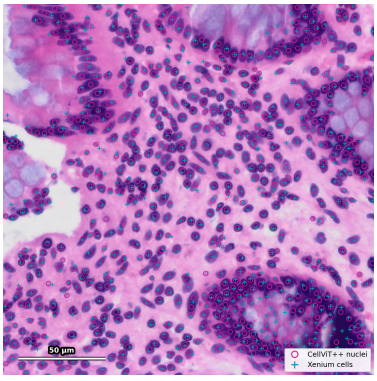


Figure 6: CellViT++ nuclei centroids and Xenium cell centroids overlaid on a representative H&E image.

A.5 Input resolution ablation

We compared inputs spanning $10\times$, $20\times$, and $40\times$ resolution, as well as $14\times14\mu\mathrm{m}$ single-cell crops (Figure 7, Table 5). The $40\times$, $47.6\mu\mathrm{m}\times47.6\mu\mathrm{m}$ input performed best (PCC = 0.161). Lower-resolution inputs reduced performance by 5.6% at $20\times$ and 19.8% at $10\times$. Single-cell crops also reduced performance by 15.5%, suggesting that prediction benefits from both fine cellular detail and surrounding tissue context.

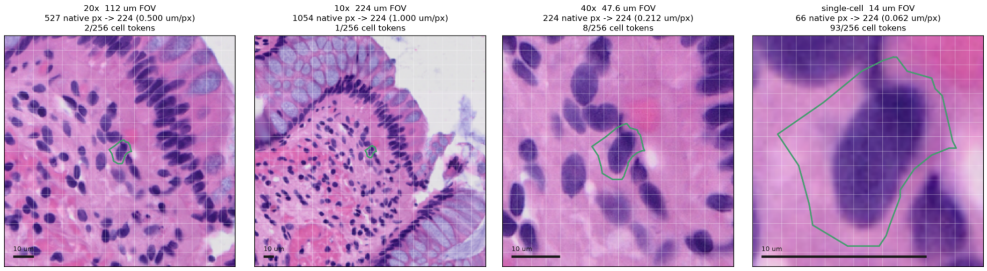


Figure 7: H&E input configurations evaluated across spatial resolutions and fields of view. Each image is centered on the target cell, with its Xenium-derived boundary shown in green and patch-token boundaries overlaid as a grid.

Table 5: Ablation of H&E input resolution and field of view. Performance is reported as mean gene-wise PCC across the 5K panel; Δ denotes the absolute and relative change from $40\times$.

Input	Pixel size (μm/px)	FOV (μm)	Token width (μm)	Tokens/cell median (mean)	PCC (5K)	Δ
$40\times$	0.2125	47.6	2.98	4 (5.1)	$\mathbf{0.1611 \pm 0.0158}$	—
$20\times$	0.5	112	7.00	1 (1.2)	0.1521 ± 0.0125	-5.6%
$10\times$	1.0	224	14.0	1 (1.0)	0.1292 ± 0.0107	-19.8%
Single-cell crop	0.2125	14.0	0.88	50 (56.8)	0.1361 ± 0.0186	-15.5%

A.6 UMAP embedding representations

We sample 150,000 cells from one of the held out test folds, and plot a UMAP of the embeddings which were used for input to the prediction head. From left to right of Figure 8, (a) shows the standard class-token embedding with $20\times$ inputs, (b) shows the cell-token pooled embeddings also with $20\times$ inputs, (c) shows the cell-token pooled embeddings with $40\times$ inputs, and (d) shows the LoRA fine-tuned cell-token pooled embeddings.

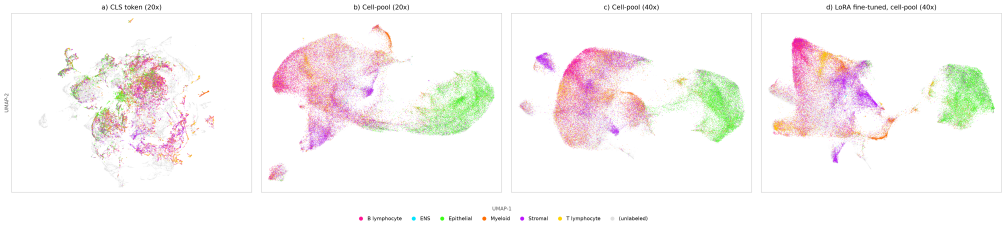


Figure 8: UMAP embeddings across adaptations, colored by cell type annotations.

A.7 Epithelial gene predictions

Our model accurately predicts single-cell spatial gene expression of epithelial markers in both healthy and inflamed epithelial regions (Figure 9).

A.8 Evaluation of spatially aggregated predictions

Our model maintained its advantage over the baseline across levels of gene expression sparsity, and the decline in performance for sparse genes became less pronounced after aggregation (Figure 10). This suggests that the adapted model captures morphology-associated transcriptional signals that are spatially coherent at both cellular and regional scales.

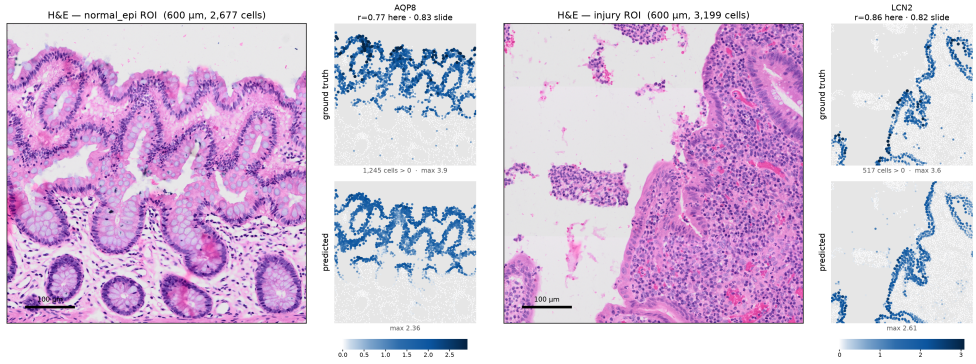


Figure 9: Predicted and observed single-cell expression of representative epithelial genes in a healthy region (left), and a diseased region (right).

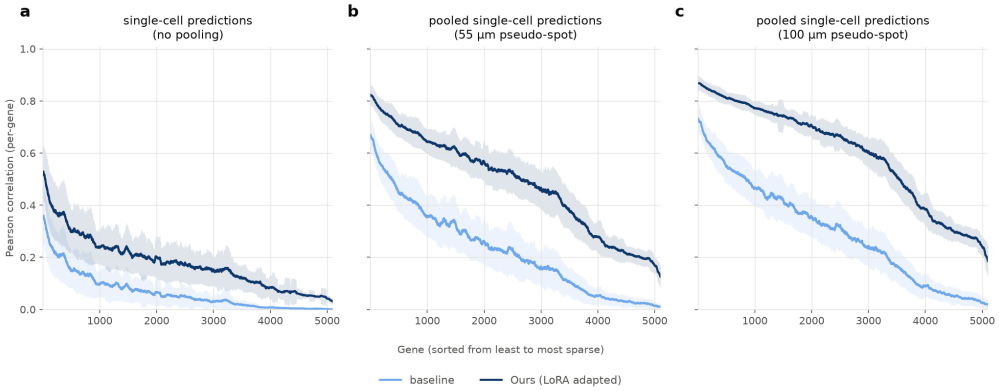


Figure 10: Per-gene PCC as a function of gene expression sparsity for the baseline linear probe vs. our adapted model. (a) Performance is evaluated at single-cell resolution and after aggregating predicted and measured single-cell expression into (b) 55 μm and (c) 100 μm pseudo-spots.

A.9 Cell type annotation coverage

Cell type annotations were available for 41 of the 91 TMA cores, covering approximately 78% of cells within those cores. Approximately 2.4 million TMA cells were scored for annotations, including cells with undetermined cell types. Of these cells, 31.4% were assigned to one of the six classes: epithelial, stromal, B lymphocyte, myeloid, T lymphocyte, or enteric nervous system (ENS).

A.10 Pathway enrichment analysis

Using the top 500 most significantly improved genes, and the bottom 500 least improved genes through the adaptation changes to pathology FMs, we sought to determine if the genes in both of these groups were enriched for specific biological programs and pathways.

To this end, we used the g:Profiler [51] web tool to test both sets of genes for enrichment of specific pathways. For the top 500 most significantly improved genes (S1), we performed an ordered query with the default g:Profiler significance thresholds, and the REACTOME [50] database of pathways (Hsapiens (Human) - version: GRCh38.p14 - REAC – annotations: BioMart, classes: 2026-3-20). For the 500 least improved genes (S2), since order is not as straightforward to consider here, we performed an unordered query using the same REACTOME [50] pathway database, with the same default parameters for g:Profiler. Multiple testing correction was done using the g:SCS method [51] developed specifically for pathway analysis multiple hypothesis testing correction.

All enriched pathways were plotted, sorted based on the adjusted p-values (significance), and overlap of genes in S1 and S2 in specific pathways/gene sets (Figure 11).

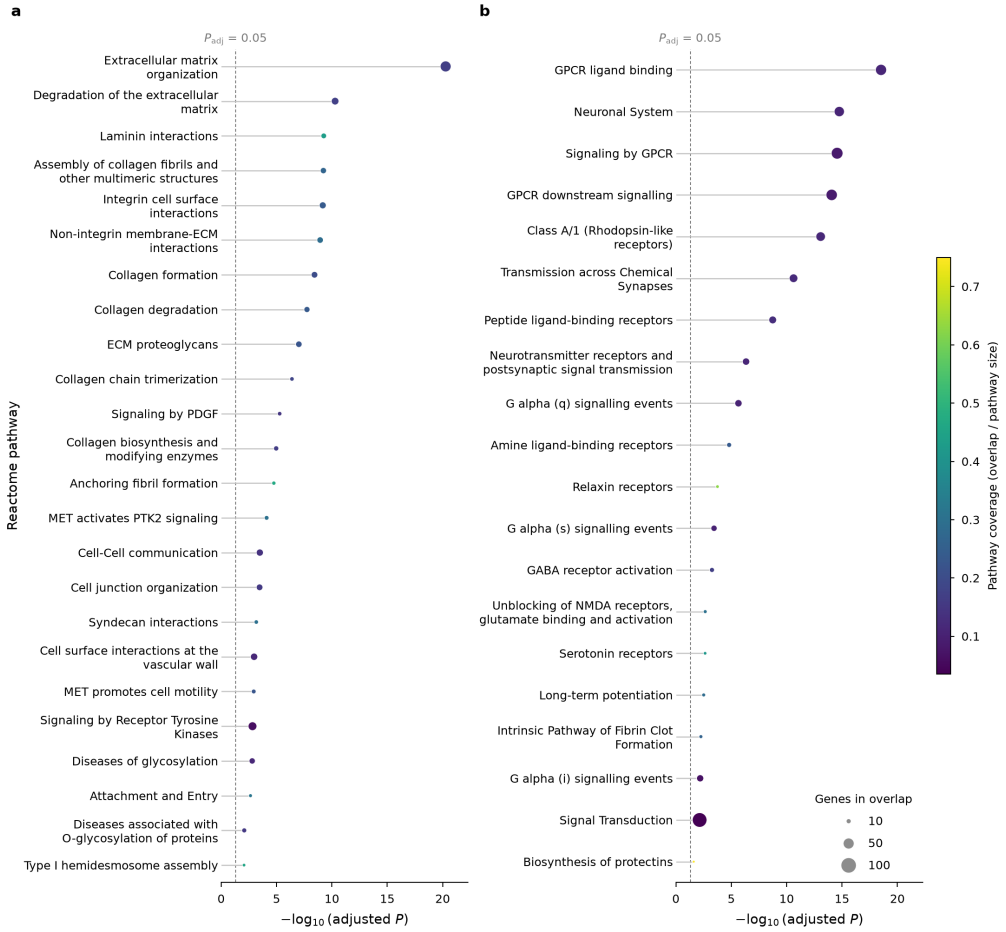


Figure 11: Pathway enrichment analysis was done using g:Profiler and the REACTOME database on the (a) 500 most improved genes and (b) 500 least improved genes.