CELL-LEVEL REPRESENTATIONS IN PATHOLOGY FOUNDATION MODELS ARE A QUESTION OF READ-OUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology foundation models (PFMs) are trained and evaluated on tiles or slides, yet many biologically relevant tasks are defined at the level of individual cells. A cell's identity depends both on its own morphology and on its surroundings, so a cell embedding should be specific to the cell while integrating its context. Current read-outs trade one for the other. One approach crops a small window around the cell, resizes it to the encoder's input size, and keeps the CLS token. This isolates the cell, but it discards the surrounding tissue and presents the image at a magnification far from the one used in pretraining. Keeping the CLS token of a wider tile retains context, but the embedding summarizes the whole tile and is shared by many, often heterogeneous, cells. We study *spatial indexing*, which removes this trade-off: a tile is centered on the cell, and the patch tokens at its location are read out. The tile size sets how much context the encoder integrates, the token position sets which cell is represented, and DINOv2's token-level iBOT objective makes these tokens informative. To evaluate read-outs at scale, we curate 111 public Xenium samples with registered H&E, comprising 20.75 million cells with gene expression and expression-derived cell-type labels, the largest cell-level annotated pathology dataset to date. Across four PFMs, spatially indexed tokens outperform the CLS token of the same tile by 0.10 macro-F1 on cell-type classification, crop-and-resize is the weakest read-out, and the gains extend to single-cell gene-expression prediction. Linear probes and iBOT-token masking confirm that a single token holds both sides of the trade-off and that both pieces of information are complementary for classification. Spatial indexing is also cheaper: reading out every cell of a tile from a single forward pass needs 1.4% of the forward passes at a small cost in accuracy, and because the tokens already carry context, the neighbor-aggregation module of a gene-expression model can be removed, cutting its compute 26×. We release the resource to support the development of cell-level pathology models.

1 INTRODUCTION

Pathology foundation models (PFMs) such as UNI2 (Chen et al., 2024), Virchow2 (Zimmermann et al., 2024), H-Optimus-1 (Scalbert et al., 2026) and Phikon-v2 (Filiot et al., 2024) are ViTs (Dosovitskiy et al., 2021) pretrained with DINOv2 (Oquab et al., 2024) on tiles of roughly 224 px, yet they are increasingly applied to cell-level tasks (Figure 1). This requires a *read-out* that turns the encoder output into one vector per cell. Because a cell's identity depends both on its own morphology and on its surroundings, a good read-out should be specific to the cell while integrating its context. Current read-outs trade one for the other. Recent efforts (Glettig et al., 2025; Nonchev et al., 2025a) used to crop a small window around the cell, resize it, and keep the CLS token: the crop isolates the cell but discards its neighborhood and is upsampled far beyond the training scale. Keeping the CLS token of a wider tile retains context, but the embedding summarizes the whole tile and is shared by many, often heterogeneous, cells.

This trade-off is avoidable. DINOv2's iBOT objective (Zhou et al., 2022) supervises individual patch tokens, so a tile's token grid is already a spatially resolved map of the tissue. We therefore center a tile on the cell and read out the tokens at its location, which we call *spatial indexing* (Figure 2a).

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.

The tile size sets how much context the encoder integrates and the token position sets which cell is represented, so cell specificity and context no longer compete.

Token-level extraction has appeared as an implementation choice in single cell pipelines (Hörst et al., 2025; Acosta et al., 2025) and benchmarks (Lüscher et al., 2025). Concurrent work (Zhao et al., 2026) shows across 30 PFMs that sampling tiled feature maps at nuclear coordinates beats the CLS token and mean pooling of the tile. However, none of these works compares against crop-and-resize, the common practice for single-cell tasks. They also stop at cell-type classification and do not consider the trade-off between accuracy and computational cost across read-outs. Most importantly, they report classification gains without asking where they come from: the cell itself, better integration of its context, or both.

We address these questions with imaging-based spatial transcriptomics (Janesick et al., 2023), which yields expression-derived single-cell labels at a scale and diversity that manually annotated datasets such as PanNuke (Gamper et al., 2019) and Lizard (Graham et al., 2021) cannot match. We curate a resource of 111 public Xenium samples with registered H&E and, for controlled evaluation, derive three benchmarks from it: pan-cancer, colon cancer and lung fibrosis. Our contributions are:

- **Resource.** 111 Xenium samples (20.75M cells) with registered H&E, expression and high-confidence cell types (to our knowledge, the largest cell-level annotated pathology dataset to date), from which we derive three benchmarks.
- **Read-out comparison.** Across four PFMs, central tokens beat the CLS token of the same tile by 0.10 macro-F1, crop-and-resize is weakest, and the ranking holds for expression prediction.
- **Mechanism.** Probes and iBOT masking show that a single token encodes both neighborhood composition and cell-intrinsic features, and that both contribute.
- **Efficiency.** Multi-cell extraction needs 70$\times$ fewer forward passes, and DeepCell's neighbor module can be dropped (26$\times$ less compute).

2 Related Work

Evaluating pathology foundation models. PFMs are mostly benchmarked on tile and slide level tasks such as tissue classification, biomarker prediction and survival analysis (Campanella et al., 2025; Neidlinger et al., 2026). Cell-level evaluation has relied on manually annotated datasets such as PanNuke (Gamper et al., 2019) and Lizard (Graham et al., 2021), which are small and cover few cell types. PathoCellBench (Lüscher et al., 2025) labels cells through multiplexed immunofluorescence and trains dense decoders on frozen token maps. CellPath-Bench (Zhao et al., 2026) uses Xenium derived labels across 11 organs and shows that sampling tiled feature maps at nuclear coordinates beats patch-level pooling for 30 PFMs. Both works rank encoders. We instead hold the encoder fixed and study how the read-out, the amount of context, and the position of the cell within it shape the representation, and why.

Cell-level representations from histology. Cell detection and classification pipelines such as HoVer-Net (Graham et al., 2019), CellViT (Hörst et al., 2024) and CellViT++ (Hörst et al., 2025) read cell information from dense ViT features, with the latter extracting per-cell embeddings from the token grid of a foundation model encoder. CellSymphony (Acosta et al., 2025) is likewise based on token-level features. In these works, token extraction is an implementation detail rather than an object of study. For tasks that require a standalone embedding per cell, another common practice remains to crop and resize a window around the cell (Glettig et al., 2025; Nonchev et al., 2025a).

Predicting expression from histology. Early work predicted spot-level expression (Ståhl et al., 2016) from H&E on Visium-type data, and HEST-1k (Jaume et al., 2024) standardized this task as a PFM benchmark. Imaging-based platforms such as Xenium (Janesick et al., 2023) now resolve expression at single-cell resolution, allowing single-cell prediction models (Nonchev et al., 2025a). Because a cropped cell lacks context, this model adds explicit neighborhood modules that aggregate embeddings of the $K = 30$ nearest cells (Nonchev et al., 2025b). We show that this module becomes redundant once the read-out itself sees context.

Table 1: Benchmarks derived from the full 111-sample resource. *Mag.*: H&E magnification; *Classif.*: cell-type classification; *Expr.*: gene-expression prediction; *Probes*: linear probing of nucleus morphology and neighborhood composition.

Dataset	Tissue	#Slides	#Cells	Mag.	Classif.	Expr.	Probes
Pan-cancer	Multiple	7	1,106,713	40×	✓		
Colon cancer	Colon	5	1,286,543	40×	✓	✓*	✓
Lung fibrosis	Lung	15	227,623	20×		✓	

* Only in Section 4.3.

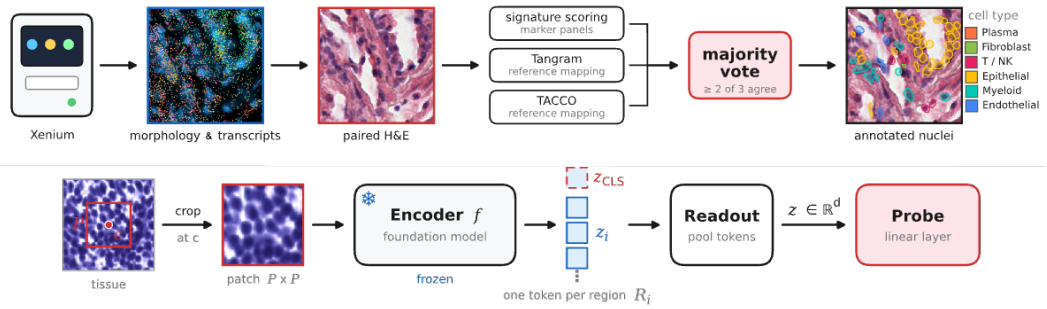


Figure 1: **Overview. Top: annotation pipeline.** Each Xenium sample is paired with a co-registered H&E image of the same section. Every cell is annotated by reference mapping (Tangram), label transfer (TACCO) and marker-based cluster annotation; the three labels are combined by majority vote, and cells with inconsistent labels or low-quality H&E are removed. **Bottom: embedding pipeline.** A square patch of side P is cropped around the nucleus centroid c and passed through the encoder f, which returns a CLS token z_{CLS} and one output token z_i per $p \times p$ region R_i. A read-out then reduces these outputs to a single cell embedding in $\mathbb{R}^d$, which is fed to the downstream task.

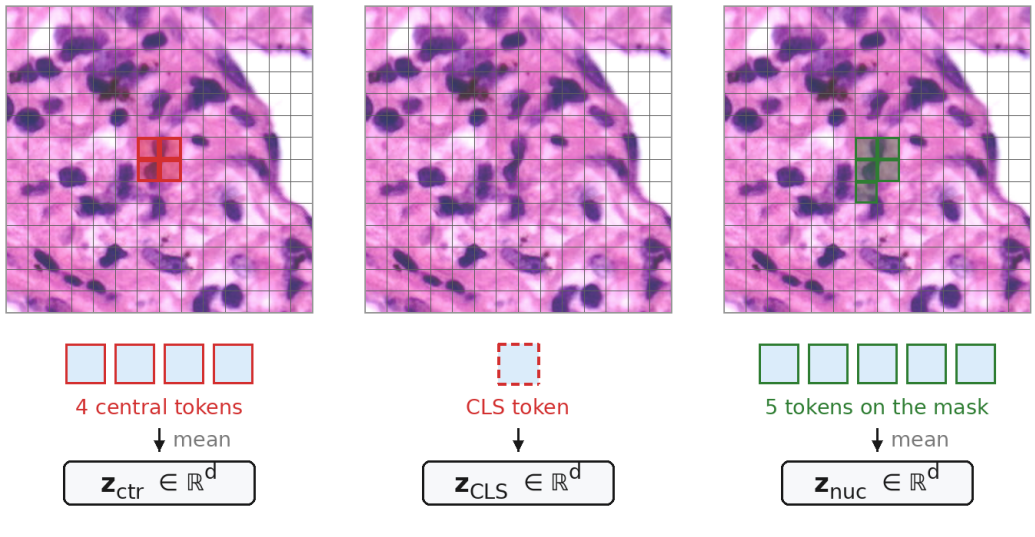
3 Methods

3.1 Datasets

We curate the to our knowledge to date largest cell-level annotated pathology dataset. Each Xenium sample is paired with a co-registered H&E image of the same section. Cell types are derived from expression: each cell is annotated independently by reference mapping with Tangram (Biancalani et al., 2021), label transfer with TACCO (Mages et al., 2023) and marker-based cluster annotation, and the three labels are combined by majority vote. Cells with inconsistent annotations or low-quality H&E are removed (Appendix A). Fine-grained labels are then collapsed into five coarse classes shared across tissues: lymphoid, stromal, epithelial, malignant and myeloid. From the full resource we derive three benchmarks (Table 1): a pan-cancer benchmark of seven slides from lung, colon, ovarian, liver and skin cancer; a colon cancer benchmark of four tumor slides and one non-diseased slide; and a lung fibrosis benchmark of 15 slides from Vannan et al. (2025).

3.2 Embedding extraction

All embeddings in this work are extracted with the same pipeline (Figure 1). We compare four read-outs, which differ in which encoder outputs they pool and therefore in how they balance cell specificity and context. Each returns a single vector in $\mathbb{R}^d$, so the resulting embeddings are directly comparable. The extraction settings (Section 3.2.2) differ in how crops are placed and which cells are read out from each forward pass. We use three extraction settings, illustrated in Appendix B.



(a) Central tokens, $\rho^{\text{P}}_{\text{tok}}$ (b) CLS token, $\rho^{\text{P}}_{\text{CLS}}$ (c) Nucleus token, $\rho^{\text{P}}_{\text{nuc}}$

Figure 2: **Token read-outs.** Each read-out maps the encoder outputs to a single cell embedding in $\mathbb{R}^d$. **(a)** Central $\rho^{\text{P}}_{\text{tok}}$: mean of the 2×2 tokens at the center of the patch. **(b)** CLS $\rho^{\text{P}}_{\text{CLS}}$: the class token, a summary of the whole crop. **(c)** Nucleus $\rho^{\text{P}}_{\text{nuc}}$: mean of the tokens overlapping the nucleus mask M. The mixed read-out $\rho^{\text{P}}_{\text{mix}}$, not shown, averages the four central tokens with the CLS token.

3.2.1 READ-OUTS

CLS $\rho^{\text{P}}_{\text{CLS}}$ We take the CLS token, z_{CLS}, as the cell embedding. This is the read-out the encoder was pretrained to make informative. It is, however, a summary of the *whole* crop: the target cell is only one of the objects it describes, and its relative contribution shrinks as the patch size P grows and more neighboring cells enter the field of view. With a small crop resized to the encoder's input size, it implements the common *crop-and-resize* practice; with a cell-centered patch at the training scale, it serves as a context-aware baseline.

Nucleus $\rho^{\text{P}}_{\text{nuc}}$ We mean pool the output tokens that overlap the nucleus region. This kind of read-out is enabled by the iBOT component (Zhou et al., 2022) of the DINOv2 loss (Oquab et al., 2024) which pushes individual tokens (not just the CLS one) to carry localized semantic information. Its main appeal is that it breaks the trade-off described above rather than arbitrating it: the patch size P controls how much context the encoder attends to, while the mask M controls which of the resulting tokens are read out. Cell specificity and context awareness are thus set independently, and a large crop can be used without diluting the target cell. The pooling support also adapts to each cell, so a small lymphocyte and a large tumor nucleus are each summarized over their own extent instead of a fixed footprint.

Central $\rho^{\text{P}}_{\text{tok}}$ Because the crop is centered on c and the token grid has an even side, the centroid lies at the shared corner of the four central tokens C, and $\rho^{\text{P}}_{\text{tok}} = \frac{1}{4}\sum_{i \in C} z_i$. Like $\rho^{\text{P}}_{\text{nuc}}$, this read-out is cell-specific and context-aware, but its support is fixed rather than adapted to the cell, so it may include parts of neighboring cells or background. In exchange, it requires no segmentation, whose quality varies across slides, and we therefore use it for our main evaluations. At both 100 and 224 px, $\rho^{\text{P}}_{\text{tok}}$ and $\rho^{\text{P}}_{\text{nuc}}$ reach the same classification performance (Appendix C).

Mixed CLS and central $\rho^{\text{P}}_{\text{mix}}$ $\rho^{\text{P}}_{\text{mix}} = \frac{1}{5}\left(z_{\text{CLS}} + \sum_{i \in C} z_i\right)$, the mean of the CLS token and the four central tokens. It tests whether the two carry complementary information (Section 4.1 and Appendix C).

3.2.2 SETTINGS

One cell extraction Each cell is embedded in its own forward pass, with the crop centered on its nucleus centroid (Figure 5a). Every cell thus has a symmetric, complete context window, at the cost of one forward pass per cell. This is the default setting in all experiments.

Multi-cell extraction All cells whose centroid lies inside a central square W of side $S \leq P$ are read out from a single forward pass, each from the tokens overlapping its nucleus. The cost per cell drops from one forward pass to $1/n_S$, where n_S is the number of cells in the square. A larger square extracts more cells per pass but moves them towards the crop boundary: a cell at the edge of the square sees only $(P - S)/2$ pixels of tissue on that side instead of $P/2$, and its context is no longer centered on it. $S = 0$ recovers one-cell extraction. We use $S \in \{448, 224, 112\}$ and a field of view of 448 pixels in Section 4.3.

Masked cell extraction As in one-cell extraction, but the central 4×4 tokens are replaced by the mask token learned during iBOT pretraining (Zhou et al., 2022), so that the encoder receives an input seen during training rather than a hand-crafted mask. The encoder then sees the surroundings but not the cell. At 448 px, the masked region spans about 128×128 slide pixels and also hides the cell's immediate neighbors which could leak cell level information. Comparing masked and unmasked embeddings isolates what the cell itself contributes (Section 4.2). Phikon-v2 is the only model whose mask token is publicly available, so this setting is restricted to it.

3.3 TASKS

Cell-type classification. We predict the five expression-derived classes from the frozen embedding with a linear classifier and report macro-F1. We use leave-one-slide-out (LOSO) cross-validation: each split holds out one test slide and one validation slide for hyperparameter selection and trains on the rest, giving 14 splits on pan-cancer and 8 on colon (Appendix A). Because difficulty varies strongly across slides, we also report differences to a reference read-out, paired within each split.

Gene-expression prediction. We use DeepCell, the single-cell variant of DeepSpot (Nonchev et al., 2025a): the embeddings of the target cell and of its K nearest neighbors ($K = 30$ by default) pass through a shared MLP, the pooled neighborhood representation is concatenated with the cell representation, and a second MLP predicts expression. Beyond asking which read-out predicts better, this lets us test whether the neighborhood module is still needed once the read-out sees context. Because Xenium gene panels differ across slides, we predict the 300 most variable genes shared by all panels of a benchmark and report the mean Pearson correlation across genes. The main evaluation uses the lung benchmark, with 15 LOSO splits (14 training slides, one test slide) and fixed hyperparameters. For the K sweep in Section 4.3 we use colon, because all the slides are at $40\times$ magnification which lets us obtain a patch with very few neighbors at a field of view of 56 pixels while this would correspond to 28 pixels for a $20\times$ lung slide.

Context and cell probes. Classification and expression prediction both depend on the cell and its context. To understand why some read-outs outperform others, we fit ridge-regression probes on the frozen embeddings on colon. Two probes target the context: the cell-type composition and the number of cells in a 224 px square centered on the cell. One probe targets the cell itself: the area of its nucleus. We report results using the R^2 metric.

4 EXPERIMENTS AND RESULTS

4.1 CELL-TYPE CLASSIFICATION AND GENE-EXPRESSION PREDICTION

Cell-type classification. We evaluate four PFMs with two read-outs and two fields of view in the one-cell setting (Table 2). Performance varies strongly across held-out slides (per-split macro-F1 ranges from 0.53 to 0.78 for UNI2 on colon), which shows that single-cell evaluation of PFMs requires diverse test slides. We therefore focus on differences to the ρ^{224}_{CLS} baseline, paired within each split.

Table 2: Cell-type classification (macro-F1) across read-outs on the pan-cancer and colon benchmarks. Each setting reports the raw score and the difference to the ρ^{224}_{CLS} baseline, computed per split and then averaged. Values are mean ± standard deviation over LOSO splits ($n = 14$ pan-cancer, $n = 8$ colon). Best read-out per model in bold.

		Pan-cancer		Colon	
Model	**Read-out**	Macro-F1	Δ vs. ρ^{224}_{CLS}	Macro-F1	Δ vs. ρ^{224}_{CLS}
UNI2	ρ^{224}_{CLS}	0.437 ± 0.061	ref.	0.582 ± 0.108	ref.
	ρ^{100}_{CLS}	0.406 ± 0.102	-0.030 ± 0.076	0.542 ± 0.106	-0.040 ± 0.027
	ρ^{224}_{tok}	$\mathbf{0.555 \pm 0.096}$	$\mathbf{0.118 \pm 0.075}$	$\mathbf{0.684 \pm 0.099}$	$\mathbf{0.101 \pm 0.038}$
	ρ^{100}_{tok}	0.477 ± 0.093	0.040 ± 0.068	0.608 ± 0.090	0.026 ± 0.052
H-Optimus-1	ρ^{224}_{CLS}	0.420 ± 0.075	ref.	0.588 ± 0.108	ref.
	ρ^{100}_{CLS}	0.394 ± 0.109	-0.026 ± 0.045	0.530 ± 0.099	-0.058 ± 0.047
	ρ^{224}_{tok}	$\mathbf{0.516 \pm 0.115}$	$\mathbf{0.096 \pm 0.064}$	$\mathbf{0.686 \pm 0.101}$	$\mathbf{0.098 \pm 0.050}$
	ρ^{100}_{tok}	0.444 ± 0.107	0.024 ± 0.044	0.600 ± 0.087	0.013 ± 0.070
Phikon-v2	ρ^{224}_{CLS}	0.381 ± 0.096	ref.	0.542 ± 0.119	ref.
	ρ^{100}_{CLS}	0.351 ± 0.112	-0.029 ± 0.050	0.485 ± 0.138	-0.057 ± 0.037
	ρ^{224}_{tok}	$\mathbf{0.492 \pm 0.115}$	$\mathbf{0.111 \pm 0.066}$	$\mathbf{0.639 \pm 0.096}$	$\mathbf{0.097 \pm 0.050}$
	ρ^{100}_{tok}	0.420 ± 0.108	0.039 ± 0.053	0.555 ± 0.096	0.013 ± 0.060
Virchow2	ρ^{224}_{CLS}	0.391 ± 0.096	ref.	0.571 ± 0.078	ref.
	ρ^{100}_{CLS}	0.343 ± 0.112	-0.048 ± 0.046	0.430 ± 0.155	-0.140 ± 0.084
	ρ^{224}_{tok}	$\mathbf{0.514 \pm 0.110}$	$\mathbf{0.123 \pm 0.071}$	$\mathbf{0.661 \pm 0.100}$	$\mathbf{0.091 \pm 0.046}$
	ρ^{100}_{tok}	0.387 ± 0.109	-0.004 ± 0.057	0.485 ± 0.123	-0.085 ± 0.064

For every model and both benchmarks, ρ^{224}_{tok} performs best. It improves over ρ^{224}_{CLS} by 0.10 macro-F1 on average, and by a similar margin over ρ^{100}_{tok}, which sees a smaller field of view. Spatially indexed tokens thus carry more cell-relevant information than the CLS token, the output PFMs are designed to expose. Both sides of the trade-off contribute. Cell specificity alone already helps: even from an upsampled 100 px crop, central tokens beat the CLS token of the full 224 px tile for three of the four models. Enlarging the field of view to 224 px then adds a further gain. Whether this second gain comes from context or from matching the pretraining scale is examined in Section 4.2. Crop-and-resize (ρ^{100}_{CLS}), the common practice, is the weakest read-out in every setting, 0.05 macro-F1 below ρ^{224}_{CLS} on average. Spatially indexed tokens of a cell-centered tile should therefore be preferred for single-cell tasks.

Comparison with other public encoders. To place these results among other public encoders, Table 11 reports the macro-F1 gain on pan-cancer over a pixel baseline: a linear classifier on the first 50 principal components of the raw RGB values of the 224 px crop (details in Appendix H). With spatial indexing, the four PFMs gain 0.32–0.38 over this baseline, ahead of CellViT (0.22), whose SAM encoder is fine-tuned for cell segmentation and classification, and CONCH (0.14). This comparison is not controlled: CellViT is fine-tuned on far less histology data, and CONCH was not designed for cell-level tasks. It nevertheless shows that, among the public encoders we tested, general-purpose PFMs read out with spatial indexing give the strongest cell embeddings.

UNI2 is the strongest model on pan-cancer and on par with H-Optimus-1 on colon (0.684 vs. 0.686), and the ranking of read-outs is the same for all four models. We therefore use UNI2 in the remaining analyses, except for the masking experiments, which require Phikon-v2 (Section 4.2).

Gene-expression prediction. Spatially indexed tokens also improve single-cell gene-expression prediction on lung (Table 3): ρ^{224}_{tok} outperforms ρ^{224}_{CLS} by 0.020 ± 0.014 mean Pearson, and crop-and-resize is worse than every read-out at 224 px. The relative gain (11%) is smaller than for classification (17–27% for UNI2). This can be due to the transcript spillover problem in Xenium where transcripts can be assigned to adjacent cells (Bilous et al., 2026), so expression targets partly reflect the neighborhood, which favors read-outs that summarize the tile. Unlike in classification, central tokens are the weakest read-out at 100 px: after upsampling, they cover only about 12 px of the

Table 3: Single-cell gene-expression prediction with DeepCell on lung (300 genes, UNI2, $K = 30$ neighbors). Mean Pearson correlation across genes; values are mean $\pm$ standard deviation over $n = 15$ LOSO splits. The difference to ρ^{224}_{CLS} is computed per split and then averaged. Best read-out per field of view in bold.

Read-out	**Mean Pearson**	**Δ vs. ρ^{224}_{CLS}**
ρ^{224}_{CLS}	0.180 ± 0.056	ref.
ρ^{224}_{tok}	0.200 ± 0.055	$+0.020 \pm 0.014$
ρ^{224}_{mix}	$\mathbf{0.205 \pm 0.056}$	$\mathbf{+0.025 \pm 0.011}$
ρ^{100}_{CLS}	0.160 ± 0.051	-0.020 ± 0.009
ρ^{100}_{tok}	0.142 ± 0.047	-0.038 ± 0.021
ρ^{100}_{mix}	$\mathbf{0.161 \pm 0.050}$	$\mathbf{-0.019 \pm 0.012}$

original image, less than a typical nucleus (Appendix C). Adding the CLS token compensates: ρ^{100}_{mix} recovers the loss (0.161 vs. 0.142), and ρ^{224}_{mix} is slightly above ρ^{224}_{tok} (0.205 vs. 0.200). For classification, adding the CLS token brings no gain (Appendix C), so $\rho^{\text{P}}_{\text{mix}}$ is mainly useful for expression prediction.

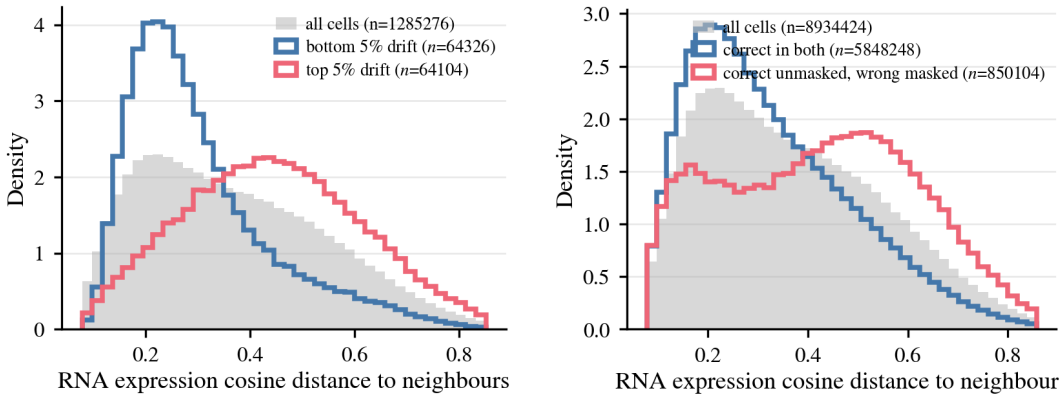
4.2 What drives the improvement brought by ρ^{224}_{TOK}?

A cell's type is reflected both in cell-intrinsic features, such as nucleus shape, size and staining, and in its context, such as neighboring cells and tissue architecture. We hypothesize that ρ^{224}_{tok} outperforms the other read-outs because it integrates both, whereas each of the others sacrifices one: crop-and-resize loses the context, and the CLS token dilutes the cell among its neighbors. We first verify that individual tokens encode both kinds of information, then show that both contribute to classification. All analyses use colon, with UNI2 for probing and Phikon-v2, the only model whose iBOT mask token is available, for masking.

4.2.1 Individual tokens encode context and cell information

Context. The cell-type composition of the 224 px square around a cell is recovered at least as well from ρ^{224}_{tok} as from ρ^{224}_{CLS} for UNI2 ($R^2 = 0.42$ vs. 0.39), and far better than from 100 px crops (0.10 and 0.14), which do not see the whole square (Table 7). For the other three models, central tokens and CLS are comparable at 224,px, and both far exceed the 100,px crops (Appendix D). Consistently, adding the CLS token to the central tokens does not improve classification ($+0.005 \pm 0.010$ macro-F1 at 224 px; Appendix C). Central tokens thus encode the surrounding context as well as the CLS token does, despite being indexed to a single location.

Cell. To isolate cell-intrinsic information, we compare masked and unmasked embeddings of Phikon-v2 at 448 px (Table 10). Nucleus area is predictable from the unmasked embedding ($R^2 = 0.266$) but barely from the masked one ($R^2 = 0.057$): masking removes cell-intrinsic information, and the unmasked tokens encode it. We define the *drift* of a cell as the cosine distance between its masked and unmasked embeddings. Since iBOT trains the encoder to predict masked tokens from their surroundings, drift measures how much the visible cell departs from what its context predicts (Appendix G). We compare it with the cell's *expression divergence*: the cosine distance between its expression profile and the mean profile of the cells in a 448 px square centered on it, over the genes shared by all colon panels. We measure the difference between groups with Cliff's δ (Cliff, 1993), the probability that a value from one group exceeds a value from the other minus the reverse, which ranges from -1 to 1. Cells in the top 5% of drift have higher expression divergence than cells in the bottom 5% (Figure 3a; $\delta = 0.62 \pm 0.17$ over $n = 5$ slides). Conversely, cells in the top 5% of expression divergence have higher drift than those in the bottom 5% ($\delta = 0.43 \pm 0.23$). When a cell differs molecularly from its surroundings, the encoder thus registers a difference in its appearance. Across all cells the relation is moderate (per-slide Pearson $r = 0.26 \pm 0.10$; Appendix G).



(a) High- vs. low-drift cells.

(b) Cells recovered by the unmasked setting.

Figure 3: Drift analysis (Phikon-v2, colon). Drift is the cosine distance between masked and unmasked embeddings; expression divergence is the cosine distance between a cell's expression and the mean expression of its neighbors. (a) High-drift cells (top 5%) have higher expression divergence than low-drift cells (bottom 5%; Cliff's $\delta = 0.62 \pm 0.17$, $n = 5$ slides). (b) Cells misclassified when masked but correctly classified when visible have higher expression divergence than other cells (Cliff's $\delta = 0.46 \pm 0.24$, $n = 4$ slides; testing is not performed on the healthy slide).

4.2.2 Both context and cell information contribute to classification

Increasing the field of view from 100 to 224 px improves UNI2 by 0.078 ± 0.028 macro-F1 (Table 9). Enlarging it further brings no gain, and performance declines beyond 448 px (0.687 at 448 px, 0.670 at 896 px, 0.648 at 1344 px). Field of view and scale are confounded in both directions: the 100 px crop is upsampled $2.24\times$ and lies further from the pretraining distribution, while crops of 448 px and above are downsampled, so each token covers more tissue and the cell occupies a smaller part of it. The sweep alone therefore cannot separate the contribution of context from that of scale.

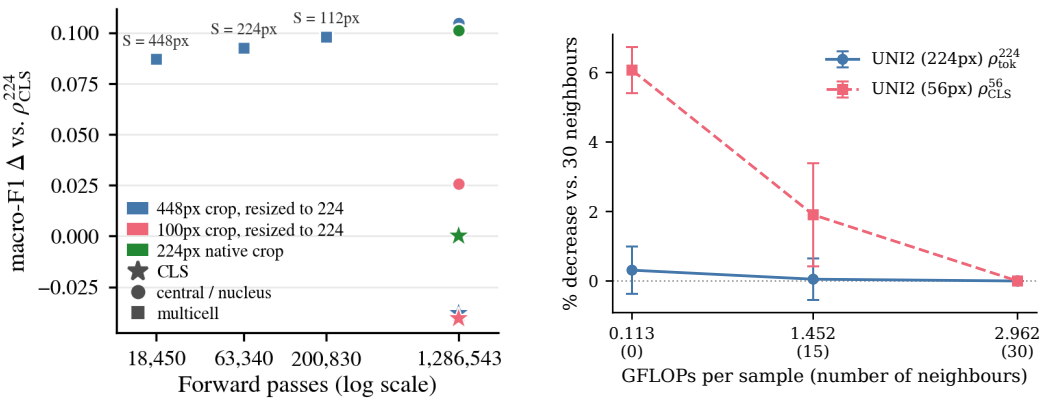
The masking experiment isolates the context. With the cell replaced by the iBOT mask token, Phikon-v2 retains 0.597 macro-F1, 93% of its unmasked performance (0.639; Table 10) despite the nucleus-area probe dropping to $R^2 = 0.057$. Context alone thus carries most of the signal.

The cell itself still matters: unmasking it adds 0.042 ± 0.008 macro-F1. Cells that are misclassified when masked but correctly classified when visible have higher expression divergence than other cells (Figure 3b; $\delta = 0.46 \pm 0.24$ over $n = 4$ slides; see Appendix A). Cell-intrinsic information therefore matters most for the cells that differ from their surroundings, precisely those for which context alone is misleading. Classification relies on both the cell and its context, which spatially indexed tokens combine and the other read-outs trade against each other.

4.3 Efficiency gains from spatial indexing

Multi-cell extraction. Since a tile contains many cells, spatial indexing can read out all cells in its central square from a single forward pass, at the price of cells no longer being centered (Section 3). On colon, with $P = 448$ and $S = 448$, multi-cell extraction needs 1.43% of the forward passes of one-cell extraction ($70\times$ fewer) and loses only 0.018 macro-F1 relative to ρ^{224}_{tok} (Figure 4a). It still ρ^{224}_{CLS} and ρ^{100}_{tok}.

Neighbor aggregation. Because common read-outs trade context for cell specificity, single-cell models add explicit modules that aggregate neighboring cells, e.g. DeepSpot (Nonchev et al., 2025a), whose DeepCell variant pools the embeddings of the $K = 30$ nearest cells at a cost that grows linearly with K (Appendix A). If the read-out already integrates context, this module should be unnecessary. We test this on colon (Section 3) by varying K for two read-outs: ρ^{56}_{CLS}, a crop-and-resize read-out that sees almost no context, and ρ^{224}_{tok} (Figure 4b). Removing the neighbors lowers the mean Pearson of ρ^{56}_{CLS} by 6% (0.162 vs. 0.152 Pearson correlation) from $K = 30$ to $K = 0$ but



(a) Multi-cell extraction.

(b) Neighbor aggregation in DeepCell.

Figure 4: Efficiency of spatial indexing (UNI2, colon). (a) Cell-type classification against the number of forward passes for one-cell and multi-cell extraction; S is the side of the central extraction square. (b) Gene-expression prediction (300 most variable genes) against the per-sample training FLOPs of DeepCell when varying the number of neighbors K: removing neighbors degrades ρ^{56}_{CLS} but not ρ^{224}_{tok}.

leaves ρ^{224}_{tok} unchanged (0.226 vs. 0.225 Pearson correlation), while reducing the per-sample training cost of DeepCell 26×. Without any neighbors, ρ^{224}_{tok} thus matches the with 30 neighbors pipeline.

5 Discussion

Read-out matters as much as the encoder. Across four PFMs and two classification benchmarks, replacing the CLS token of a cell-centered tile by its central tokens improves macro-F1 by about 0.10, more than the spread between the four encoders under the same read-out (0.05 on colon, 0.06 on pan-cancer). **Practical implications.** For single-cell tasks we recommend the central tokens of a cell-centered tile at the pretraining scale: this needs no segmentation or encoder change, supports multi-cell extraction (70× fewer passes), and makes neighbor-aggregation modules unnecessary; for expression prediction, adding the CLS token helps slightly. **A resource beyond the present benchmarks.** Our experiments use three controlled benchmarks drawn from the resource. Because every cell combines registered H&E, gene expression and a molecularly derived cell type, the full resource also supports larger pan-tissue and disease-specific evaluations, finer-grained cell types, gene-expression prediction and multimodal representation learning.

Limitations. *Evaluation scope.* Our benchmarks hold out 7 (pan-cancer), 4 (colon) and 15 (lung) slides, use five coarse cell types, and evaluate frozen embeddings with linear classifiers and a single downstream expression model (DeepCell). Broader evaluations across organs, finer cell types and state-of-the-art cell classification pipelines would strengthen our conclusions. *Labels.* Cell types and expression targets come from Xenium and can be affected by transcript spillover (Bilous et al., 2026), the benchmarks also mix 40× and 20× scans. *Models.* Follow-up analyses use UNI2 only, and masking uses Phikon-v2, the only model whose mask token we could obtain. Because Phikon-v2 is the weakest of the four encoders, the relative contributions of cell and context may differ for stronger models. *Mechanistic analysis.* Changing the field of view also changes the input scale, through upsampling below 224 px and downsampling above it, so the effects of context and scale cannot be fully separated. More fundamentally, cell and context are coupled: a cell's neighborhood is informative about its type, and its appearance is informative about its surroundings. Our analysis should therefore be read as evidence that both sources contribute, not as a causal attribution. The 5% drift threshold is arbitrary, and the full correlations, while pointing in the same direction, are moderate (Appendix G).

Spatial indexing is a simple, segmentation-free change to how cell embeddings are extracted. It turns PFMs trained on tiles into context-aware cell encoders, and we hope it, together with our resource, will serve as a default for cell-level pathology.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, to support qualitative and thematic data analysis (help in plotting and the model had access to results tables when reviewing the paper so it could make suggestion based on these results), provide feedback on methodology or experiments (We asked an LLM for reviews on the paper), and to assist with the FLOP counts in Appendix A. We have not used generative AI tools to propose or refine hypotheses, develop theoretical models or conceptual frameworks, design methodology or experiments, clean or reformat datasets, interpret results, or translate. Generating synthetic data and writing proofs are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, create and modify figures, summarize existing literature, rephrase parts of the paper, and edit and format the paper for readability. We have reviewed all AI-assisted work. All AI-generated code was manually reviewed by the authors. The FLOP counts in Appendix A were independently re-derived by hand. All AI-suggested literature was read and checked for accuracy and relevance by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We make all code used for this project available at: `https://anonymous.4open.science/r/Cell-level-representations-in-pathology-foundation-models-are-a-question-of-read-out-code-6B59`
The curated single-cell H&E–transcriptomics resource is available on huggingface (hidden for anonymity)

REFERENCES

Paul H. Acosta, Pingjun Chen, Simon P. Castillo, Maria Esther Salvatierra, Yinyin Yuan, and Xiaoxi Pan. CellSymphony: Deciphering the molecular and phenotypic orchestration of cells with single-cell pathomics. *arXiv preprint arXiv:2508.10232*, 2025.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature Methods*, 18(11):1352–1362, 2021. doi: 10.1038/s41592-021-01264-7.

Mariia Bilous, Daria Buszta, Jonathan Bac, Senbai Kang, Yixing Dong, Stephanie Tissot, Sylvie Andre, Marina Alexandre Gaveta, Christel Voize, Solange Peters, Krisztian Homicsko, and Raphael Gottardo. Resolving sensitivity, specificity and signal contamination in Xenium spatial transcriptomics. *Nature Methods*, 23:1152–1162, 2026. doi: 10.1038/s41592-026-03089-8.

Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Ivan Shujski, Noora Neittaanmäki, Kuan-lin Huang, Ricky Kwan, Jane Houldsworth, Adam J. Schoenfeld, and Chad Vanderbilt. A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16:3640, 2025. doi: 10.1038/s41467-025-58796-1.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993. doi: 10.1037/0033-2909.114.3.494.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel

Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction, 2024. URL https://arxiv.org/abs/2409.09173.

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

Marc Glettig, Tim Ehrensperger, Josephine Yates, and Valentina Boeva. H&Enium, applying foundation models to computational pathology and spatial transcriptomics to learn an aligned latent space. *bioRxiv*, 2025. doi: 10.1101/2025.07.22.665986.

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 multitissue histology images. *Medical Image Analysis*, 58:101563, 2019. doi: 10.1016/j.media.2019.101563.

Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, Noorul Wahab, Fayyaz Minhas, Shan E Ahmed Raza, Hesham El Daly, Kishore Gopalakrishnan, David Snead, and Nasir Rajpoot. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 684–693, 2021.

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

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

Amanda Janesick et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023. doi: 10.1038/s41467-023-43458-x.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Perez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis. *arXiv*, June 2024. URL https://arxiv.org/abs/2406.16192v1.

Kelli F. Johnson, Xiangning Dong, Yu-Hwai Tsai, Angeline Wu, Sydney G. Clark, Abigail Vallie, Sha Huang, Charlie J. Childs, Rachel K. Zwick, Ian Glass, Katherine D. Walton, Ophir D. Klein, and Jason R. Spence. Mapping mesenchymal diversity in the human small intestine and organoids. *Nature Cell Biology*, 28(8):1758–1770, 2026. doi: 10.1038/s41556-026-02027-2.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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*, 30:863–874, 2024.

Jérôme Lüscher, Nora Koreuber, Jannik Franzen, Fabian H. Reith, Claudia Winklmayr, Elias Baumann, Christian M. Schürch, Dagmar Kainmüller, and Josef Lorenz Rumberger. PathoCellBench: A Comprehensive Benchmark for Cell Phenotyping . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15966. Springer Nature Switzerland, September 2025.

Simon Mages, Noa Moriel, Inbal Avraham-Davidi, Evan Murray, Jan Watter, Fei Chen, Orit Rozenblatt-Rosen, Johanna Klughammer, Aviv Regev, and Mor Nitzan. TACCO unifies annotation transfer and decomposition of cell identities for single-cell and spatial omics. *Nature Biotechnology*, 41(10):1465–1473, 2023. doi: 10.1038/s41587-023-01657-3.

Kelli A. McCord, Emerald Kan, Sean Hyslop, Amanda Y. Xia, Colby J. Hofferek, James S. Jr. Lewis, Andreas Wieland, David J. Hernandez, Vlad C. Sandulache, and William H. Hudson. Single-cell TCR mapping reveals spatially coordinated T cell states in head and neck cancer. *Science Immunology*, 11(118):eaec3133, 2026. doi: 10.1126/sciimmunol.aec3133.

Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature Biomedical Engineering*, 10(6):1113–1123, 2026. doi: 10.1038/s41551-025-01516-3.

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*, 2025a. doi: 10.1101/2025.02.09.25321567. Version posted July 9, 2025.

Kalin Nonchev, Sebastian Dawo, Karina Silina, Holger Moch, Sonali Andani, Tumor Profiler Consortium, Viktor H. Koelzer, and Gunnar Rätsch. he2st: Code for "DeepSpot: Leveraging spatial context for enhanced spatial transcriptomics prediction from H&E images". https://github.com/ratschlab/he2st, 2025b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Marin Scalbert, Charlie Saillard, Thomas Peeters, Liam Gonzalez, Dasha Valter, Felipe Llinares-López, Zelda E. Mariet, and Rodolphe Jenatton. H-optimus-1: A foundation model for computational histopathology. In *Proceedings of the American Association for Cancer Research Annual Meeting 2026; Part 2 (Late-Breaking, Clinical Trial, and Invited Abstracts)*, volume 86, pp. LB174. AACR, 2026. doi: 10.1158/1538-7445.AM2026-LB174.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016. doi: 10.1126/science.aaf2403.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019. doi: 10.1016/j.cell.2019.05.031.

Michael Tyler, Avishai Gavish, Chaya Barbolin, Roi Tschernichovsky, Rouven Hoefflin, Michael Mints, Sidharth V. Puram, and Itay Tirosh. The curated cancer cell atlas provides a comprehensive characterization of tumors at single-cell resolution. *Nature Cancer*, 6:1088–1101, 2025. doi: 10.1038/s43018-025-00957-8.

Annika Vannan, Ruqian Lyu, Arianna L. Williams, Nicholas M. Negretti, Evan D. Mee, Joseph Hirsh, Samuel Hirsh, Niran Hadad, David S. Nichols, Carla L. Calvi, Chase J. Taylor, Vasiliy V. Polosukhin, Ana P. M. Serezani, A. Scott McCall, Jason J. Gokey, Heejung Shim, Lorraine B. Ware, Matthew J. Bacchetta, Ciara M. Shaver, Timothy S. Blackwell, Rajat Walia, Jennifer M. S. Sucre, Jonathan A. Kropski, Davis J. McCarthy, and Nicholas E. Banovich. Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nature Genetics*, 57(3):647–658, 2025. doi: 10.1038/s41588-025-02080-x.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

Huasheng Yu, Saad S. Nagi, Dmitry Usoskin, Yizhou Hu, Jussi Kupari, Otmane Bouchatta, Hanying Yan, Suna Li Cranfill, Mayank Gautam, Yijing Su, You Lu, James Wymer, Max Glanz, Phillip Albrecht, Hongjun Song, Guo-Li Ming, Stephen Prouty, John Seykora, Hao Wu, Minghong Ma, Andrew Marshall, Frank L. Rice, Mingyao Li, Håkan Olausson, Patrik Ernfors, and Wenqin Luo. Leveraging deep single-soma RNA sequencing to explore the neural basis of human somatosensation. *Nature Neuroscience*, 27(12):2326–2340, 2024. doi: 10.1038/s41593-024-01794-1.

Bokai Zhao, Yiyang Zhang, Hanqing Chao, Yawei Ma, Long Bai, Tai Ma, Minfeng Xu, Ming Song, and Tianzi Jiang. CellPath-Bench: A multidimensional benchmark for whole-slide cellular representations in pathology foundation models. *arXiv preprint arXiv:2608.21060*, 2026.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Implementation details

Resource assembly. We assembled 111 publicly available Xenium samples. Samples were generated with the 10x Genomics Xenium platform and had matched H&E histology available that could be spatially registered to the Xenium image. The resulting resource comprises 111 samples and 20.75 million cells across multiple tissues, diseases and magnification levels. For every retained cell, the resource links the Xenium spatial coordinates with its gene-expression profile and the location in the registered H&E image.

Importantly, the complete resource is substantially larger and more heterogeneous than the benchmark subsets used in the main experiments. For controlled evaluation, we selected three datasets: a pan-cancer benchmark, a colon cancer benchmark, and a lung fibrosis benchmark. These subsets were chosen to isolate specific evaluation settings rather than to exhaust the available data. The full

resource allows users to construct considerably larger datasets spanning additional tissues, disease contexts, and combinations thereof.

We constructed our dataset from 62 Xenium samples available in HEST-1K (Jaume et al., 2024), 29 samples available in SPLIT (Bilous et al., 2026), 11 samples available in McCord (McCord et al., 2026), 5 samples available in Kelli (Johnson et al., 2026) and 4 samples available in Yu (Yu et al., 2024).

Reference curation. We used manually curated appropriate single-cell RNA-sequencing reference datasets for each tissue and disease context. Xenium panels, tissue composition, and relevant cell populations differ substantially across tissues and disease contexts. We therefore did not use a single universal reference for cell-type annotation. Each context represented in our dataset was matched by a corresponding reference. Reference cell-type identities were harmonized to a common set of cell-types and annotated with the genes available in the corresponding Xenium panel. Reference datasets were obtained from the 3CA (Tyler et al., 2025) and CellXGene (CZI Cell Science Program et al., 2025) atlas.

High-confidence majority-voting annotation. Each cell was annotated individually using three complementary strategies. First, Tangram (Biancalani et al., 2021) was used for reference-based mapping between the curated scRNA-seq reference and the cell's expression profile. Second, reference labels were transferred independently using TACCO (Mages et al., 2023). Third, cells were clustered within each Xenium sample and clusters were annotated using canonical, reference-derived gene signatures. The results of the three independent annotation procedures were then combined by majority voting to obtain the final cell-type assignment. Cells for which the annotation evidence was ambiguous or inconsistent were excluded. We additionally performed image-quality control by embedding the corresponding H&E regions with pathology foundation models and identifying out-of-distribution image embeddings. This enabled the removal of cells associated with low-quality H&E regions, including blurry tissue, absent or black image regions, and empty crops resulting from misalignment. For the benchmarks used in this study, fine-grained labels were subsequently mapped to five broad classes: *Lymphoid, Stromal, Epithelial, Malignant,* and *Myeloid*, providing a consistent taxonomy across tissues and disease contexts.

Data splits. The cell type classification task aims to predict the cell type from frozen embeddings. We have 5 classes (*Lymphoid, Stromal, Epithelial, Malignant, Myeloid*); a residual *Unknown* bin is dropped from training and evaluation entirely, so all reported metrics concern this five-class problem. We evaluate on two independent leave-one-WSI-out (LOSO) regimes, splitting strictly at the whole-slide level (no cell appears on both sides of a fold):

- **Pan-cancer** (7 Xenium WSIs spanning lung, colon cancer, ovarian, liver and skin cancer): each WSI is held out as the test slide exactly once; for each test slide, two of the remaining six slides are used in turn as the validation slide (train = other five), giving $7 \times 2 = 14$ folds.
- **Colon cancer** (in-distribution, single cancer type; 5 colon cancer WSIs — 4 cancer sections plus 1 non-diseased section that is always kept in train): each of the 4 cancer sections is held out as the test slide once, with two choices of validation slide among the remaining three cancer sections (train = other two cancer sections + the non-diseased section), giving $4 \times 2 = 8$ folds.
- **Lung fibrosis** (15 lung fibrosis Xenium WSIs from Vannan et al. (2025)) This cohort contains 15 WSIs and is used for the gene expression prediction task. As we fix the hyperparameters for this task, we train on the 14 slides and test on the remaining one: 15 folds

The pan-cancer and colon datasets share one slide in total (detail of the individual slides used for each dataset and split is available in the configs/ folder of the code). Every other source of randomness is driven by a single fixed seed (`seed=42`).

Cell-type classifier. The classifier head is a simple linear head applied to the cell embedding ($d = 1280$ for Virchow2, $d = 1536$ for UNI2/H-Optimus-1, $d = 1024$ Phikon-v2), trained with soft-

max cross-entropy for the full 30-epoch budget and the final-epoch weights are evaluated. Optimization uses AdamW (Loshchilov & Hutter, 2019). Class imbalance is handled by inverse-frequency reweighting of the per-sample cross-entropy loss, and *which* reweighting scheme to use is itself part of the hyperparameter search rather than fixed a priori: we grid-search 2 learning rates (5×10^{-4}, 1×10^{-4}) $\times$ 3 loss-reweighting settings (none, inverse WSI-frequency, inverse cell-type-frequency) $= 6$ configurations per fold. For each fold, the configuration with the highest *validation*-slide macro-F1 is selected, and we report its macro-F1 on the held-out *test* slide; hyperparameter selection never has access to the test slide.

Gene-expression regressor (DeepSpot). For the embedding-to-expression evaluation, we train one DeepSpot regressor per fold on a 15-fold leave-one-sample-out split over 15 Xenium samples from Vannan et al. (2025) (14 train, 1 test) and colon samples (4 train, 1 test), predicting the top 300 highly-variable genes (selected with scanpy (Wolf et al., 2018), Seurat v3 flavor (Stuart et al., 2019), batched by sample) from single-cell embeddings aggregated over each cell's $K = 30$ nearest spatial neighbors (max-pooled). DeepSpot's architecture is fixed across folds: we reuse the base parameters and implementation from (Nonchev et al., 2025b) an ensemble of $n_{\text{ensemble}} = 10$ phi/rho sub-networks with hidden size 1024 and embedding size 1024, trained with MSE loss, AdamW, weight decay 10^{-6}, dropout 0.2, batch size 1024, for at most 6 epochs with early stopping on validation loss (patience 3 epochs, minimum delta 0.01). We fixed the learning rate to 1×10^{-3} and the gene normalization the *standard* as we observed that tuning these hyperparameter did not bring different results (so all hyperparameter are fixed, no validation). The only difference we have with the available implementation is that we set the resolution to 0 for efficiency reasons. As with the classifier, a single fixed seed (2024) is used throughout.

DeepCell FLOPs calculation We report the FLOPs per sample in Figure 4b rather than wall-clock training time, as the latter is confounded by hardware, data-loading throughput and scheduler contention. DeepCell applies an ensemble of E linear ϕ modules, $\phi_e : \mathbb{R}^D \to \mathbb{R}^H$, to the embedding of the target cell and, with shared weights, to each of the K neighbor embeddings in its bag. The per-element ensemble median and the permutation-invariant reduction over the bag are parameter-free, after which an ensemble of ρ modules $\rho_e : \mathbb{R}^{M \cdot H} \to \mathbb{R}^G$ predicts gene expression, where M is the number of concatenated context branches.

Counting one multiply accumulate (MAC) as two floating-point operations, a linear layer $\mathbb{R}^a \to \mathbb{R}^b$ costs $a \cdot b$ MACs per input vector. The forward cost per sample is therefore

$$\mathrm{MACs}(K) = \underbrace{E_\phi \cdot D \cdot H \cdot (1 + K)}_{\phi \text{ on cell and neighbors}} + \underbrace{E_\rho \cdot (M \cdot H) \cdot G}_{\rho}, \tag{1}$$

with $M = 2$ when neighbors are used and $M = 1$ for the neighbor-free baseline, which omits the concatenation entirely and is thus not the $K \to 0$ limit of the neighbor model. The backward pass computes gradients with respect to both the layer inputs and the layer weights, each at approximately the cost of the forward pass (Kaplan et al., 2020), giving

$$\mathrm{FLOPs}_{\text{train}}(K) = 6 \cdot \mathrm{MACs}(K). \tag{2}$$

We use $D = 1536$ (UNI2), $H = 1024$, $G = 300$ and $E_\phi = E_\rho = 10$, as reported in the previous paragraph

Probing. To complement the end-to-end classifier, we also fit closed-form linear probes on top of the frozen embeddings, following the standard probing paradigm: how much of a given cell-level or neighborhood-level property is *linearly* recoverable from a representation, independent of any trained head. Probes are fit per read-out condition under the same 8-fold Leave-One-WSI-Out colon-cancer splits used for the classifier. All targets for a probe are computed identically across train/val/test from that split's own population; features are standardized (zero mean, unit variance, fit on train only) before fitting.

Solver. Probes are fit by closed-form ridge regression on the centered normal equations, $w = (X_c^\top X_c + \lambda I)^{-1} X_c^\top y_c$, $b = \bar{y} - \bar{x}^\top w$, so that the bias term is never itself penalized. For a probe with multiple output columns (e.g. per-class composition fractions) every column shares the same λ and feature matrix and is solved in one batched linear system. λ is swept over

$\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$; for each λ a model is fit once on train and scored on val, the λ with the best validation *selection metric* (defined per probe below) is kept, and that same train-only fit is evaluated once on the held-out test WSI (never refit on train∪val). All folds use a fixed seed (42).

Probes. Three probes are fit independently per embedding condition:

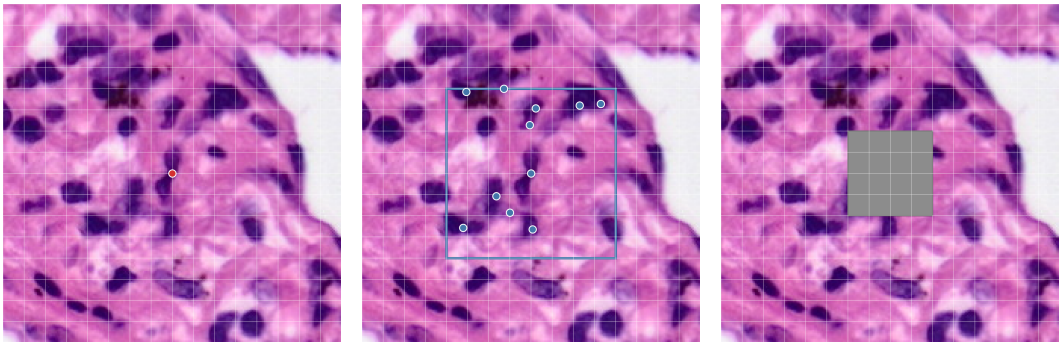
- **Area** (regression): $\log(1 + \text{area})$ of the cell's own Xenium nucleus boundary polygon.
- **Cell-type composition** (regression): per-class cell-type fractions among every detected cell (regardless of whether that cell itself has an embedding) within a $c \times c$ px window centered on the probed cell, for $c = 224$.
- **Cell count** (regression): $\log(1 + n)$, n the raw number of detected cells in the same $c \times c$ px windows.

Area is derived analytically from the continuous Xenium boundary polygon vertices via the shoelace formula.

Metrics. Every probe reports per-output-column R^2 (coefficient of determination) plus their mean, and the mean R^2 is the λ-selection metric.

B Extraction settings

Figure 5 illustrates the three extraction settings described in Section 3.



(a) One cell (b) Multi-cell (c) Masked

Figure 5: **Extraction settings. (a)** One cell: one forward pass per cell, with the crop centered on the nucleus centroid so that every cell gets a symmetric, complete context window. This is the default setting. **(b)** Multi-cell: every cell whose centroid falls inside a central extraction square of side $S \leq P$ is read out from the same forward pass, trading context symmetry for fewer forward passes (Section 4.3). **(c)** Masked: as in (a), but the central 4×4 tokens are replaced by the learned iBOT (Zhou et al., 2022) mask token, hiding the cell while keeping its surroundings (Phikon-v2 only; Appendix F).

C Additional read-out comparisons

We compare $\rho^{\text{P}}_{\text{mix}}$ and $\rho^{\text{P}}_{\text{nuc}}$ with $\rho^{\text{P}}_{\text{tok}}$ to test whether their additional complexity improves performance.

Adding the CLS token to the central tokens ($\rho^{\text{P}}_{\text{mix}}$). For UNI2 on colon, adding the CLS token to the central tokens does not improve classification at either field of view (Table 4), unlike for gene-expression prediction (Table 3). This is consistent with the probes of Section 4.2: the central tokens already encode the context captured by the CLS token.

Table 4: Effect of adding the CLS token to the central read-out, for UNI2 on colon. Values are mean $\pm$ standard deviation of macro-F1 over $n = 8$ LOSO splits; the difference is computed per split and then averaged. Best read-out per field of view in bold.

Field of view	Read-out	Macro-F1	Δ vs. $\rho^{\mathrm{P}}_{\mathrm{tok}}$
100 px (upsampled)	$\rho^{100}_{\mathrm{tok}}$	$\mathbf{0.608 \pm 0.090}$	ref.
	$\rho^{100}_{\mathrm{mix}}$	0.606 ± 0.094	-0.002 ± 0.012
224 px (native)	$\rho^{224}_{\mathrm{tok}}$	0.684 ± 0.099	ref.
	$\rho^{224}_{\mathrm{mix}}$	$\mathbf{0.688 \pm 0.101}$	$\mathbf{+0.005 \pm 0.010}$

Table 5: Central versus nucleus read-out for UNI2 on colon. Values are mean $\pm$ standard deviation of macro-F1 over $n = 8$ LOSO splits; the difference is computed per split and then averaged. Best read-out per field of view in bold; all differences are within one standard deviation.

Field of view	Read-out	Macro-F1	Δ vs. $\rho^{\mathrm{P}}_{\mathrm{tok}}$
100 px (upsampled)	$\rho^{100}_{\mathrm{tok}}$	$\mathbf{0.608 \pm 0.090}$	ref.
	$\rho^{100}_{\mathrm{nuc}}$	$\mathbf{0.608 \pm 0.094}$	0.000 ± 0.007
224 px (native)	$\rho^{224}_{\mathrm{tok}}$	0.684 ± 0.099	ref.
	$\rho^{224}_{\mathrm{nuc}}$	$\mathbf{0.686 \pm 0.098}$	$\mathbf{+0.003 \pm 0.011}$

Central versus nucleus tokens. Table 6 reports how many tokens each nucleus overlaps and how often it lies entirely within the four central tokens. At 448 px this holds for 95–98% of cells, so $\rho^{448}_{\mathrm{tok}}$ and $\rho^{448}_{\mathrm{nuc}}$ select nearly the same tokens. At 224 px it holds for 50–62% of cells. At 100 px, the nucleus overlaps a median of 15–18 tokens, and the central tokens cover only a small part of it, so $\rho^{100}_{\mathrm{tok}}$ and $\rho^{100}_{\mathrm{nuc}}$ select very different tokens. Classification performance is nevertheless the same at both fields of view (Table 5). A likely explanation is that self-attention spreads information about the nucleus to the tokens around it, as it does for the surrounding context (Section 4.2). Since $\rho^{\mathrm{P}}_{\mathrm{tok}}$ performs on par with $\rho^{\mathrm{P}}_{\mathrm{nuc}}$ and requires no segmentation, we use it for the main evaluations (Section 4.1).

D Context probes

We report in table 7 the results for the two context probing tasks on our 4 models across read-outs. We discuss the performance of UNI2 in the main text and don't further analyze the results for the cell count probes as their R^2 is unstable across models.

E Field of view

In order to assess the importance of context and obtain an upper bound on the performance from neighboring cell types only, we perform classification by doing majority voting on the ground truth cell types present in a 112 and 224 px square around the cell. This is not a perfect context upper bound as it disregards spatial organization inside of the square and majority voting also fails to capture more complex structure containing different cell types. However, it remains a strong baseline as it assumes knowledge of all the surrounding cell types. We see in table 8 that at 112 px context, the macro F1 is 0.671 performing similarly to our best model and read-outs and showing that patch context is a main driver of cell type classification performance.

Table 9 reports the full field-of-view sweep discussed in Section 4.2. Above 224 pixels, a single token covers more area than a typical nucleus (Table 6), so we shift the crop so that the nucleus falls inside one token rather than straddling the four central tokens, and read out with $\rho^{\mathrm{P}}_{\mathrm{nuc}}$.

Table 6: Token overlap by field of view, for the UNI2 token grid ($p = 14$). For each cell, the number of tokens its nucleus overlaps is reported as a mean and median; the last column gives the fraction of cells whose nucleus lies entirely within the four central tokens. Based on 1,286,543 cells (colon) and 1,106,713 cells (pan-cancer).

		Tokens per nucleus		
Dataset	**Field of view**	Mean	Median	**All in central**
Colon	448 px	4.14	4	0.950
	224 px	6.67	4	0.504
	100 px	21.27	18	0.021
Pan-cancer	448 px	4.04	4	0.981
	224 px	5.53	4	0.615
	100 px	17.17	15	0.027

Table 7: Linear-probe performance (R^2) on colon for cell count and cell type composition at 224 px context. Values are mean $\pm$ standard deviation over $n = 8$ LOSO splits.

Model	**Token**	**Cell count**	**Cell type composition**
UNI2	ρ^{224}_{tok}	0.43 ± 0.33	0.42 ± 0.12
	ρ^{224}_{CLS}	0.04 ± 0.92	0.39 ± 0.17
	ρ^{100}_{tok}	-0.45 ± 1.41	0.10 ± 0.25
	ρ^{100}_{CLS}	-1.22 ± 3.03	0.14 ± 0.22
H-Optimus-1	ρ^{224}_{tok}	0.41 ± 0.35	0.37 ± 0.15
	ρ^{224}_{CLS}	-0.10 ± 1.33	0.38 ± 0.18
	ρ^{100}_{tok}	-0.38 ± 1.40	0.09 ± 0.24
	ρ^{100}_{CLS}	-1.06 ± 2.79	0.04 ± 0.34
Phikon-v2	ρ^{224}_{tok}	0.09 ± 0.72	0.23 ± 0.26
	ρ^{224}_{CLS}	0.08 ± 0.89	0.21 ± 0.25
	ρ^{100}_{tok}	-1.00 ± 2.39	0.02 ± 0.22
	ρ^{100}_{CLS}	-1.37 ± 3.30	0.02 ± 0.23
Virchow2	ρ^{224}_{tok}	0.21 ± 0.66	0.25 ± 0.33
	ρ^{224}_{CLS}	0.08 ± 1.04	0.29 ± 0.25
	ρ^{100}_{tok}	-1.32 ± 3.15	0.02 ± 0.22
	ρ^{100}_{CLS}	-2.11 ± 4.66	-0.20 ± 0.50

F MASKING RESULTS

We mask the central 4×4 tokens of Phikon-v2 with its learned iBOT mask token (Section 3) at 448 px, where 95% of nuclei lie within the four central tokens (Table 6), so that the mask reliably covers the whole nucleus. The masked region spans about 128×128 slide pixels and therefore also hides the cell's immediate neighbors; the masked performance is thus a conservative estimate of what the context alone provides. Both conditions use the same read-out (ρ^{448}_{nuc}), so they differ only in whether the cell is visible.

G DRIFT ANALYSIS

Interpretation. During iBOT pretraining, the encoder learns to predict the representation of masked tokens from their surroundings. The masked embedding of a cell can therefore be read as what the encoder expects at that location given the context alone, and the drift as how much the visible cell departs from that expectation. In H&E, most cells look as their surroundings predict and have low drift. We test whether the cells with high drift also differ molecularly from their neighbors, using the expression divergence defined in Section 4.2.

Table 8: Cell type classification on colon by neighborhood window side.

Window side	Macro-F1	Balanced acc.	Accuracy
112 px	0.671	0.653	0.786
224 px	0.609	0.595	0.747

Table 9: UNI2 on colon across field of view (then resized to 224px). Macro-F1 is reported as mean $\pm$ standard deviation over $n = 8$ LOSO splits; improvement is measured against the 100 px setting, paired within split.

Patch size (px)	Token	Macro-F1	Δ vs. 100 px
100	ρ_{tok}^{100}	0.608 ± 0.090	ref.
224	ρ_{nuc}^{224}	0.686 ± 0.098	0.078 ± 0.028
448	ρ_{nuc}^{448}	0.687 ± 0.096	0.079 ± 0.035
896	ρ_{nuc}^{896}	0.670 ± 0.093	0.062 ± 0.035
1344	ρ_{nuc}^{1344}	0.648 ± 0.091	0.040 ± 0.038

Correlation over all cells. Drift and expression divergence are moderately correlated within slides (Figure 6; per-slide Pearson $r = 0.26 \pm 0.10$ and Spearman $\rho = 0.29 \pm 0.10$, $n = 5$ slides). Because most cells have low drift, this correlation is dominated by cells for which the image carries little information beyond the context.

Tails of the drift distribution. We therefore compare the cells with the highest and lowest 5% of drift (Figure 3a). High-drift cells have clearly higher expression divergence (Cliff's $\delta = 0.62 \pm 0.17$), showing that the cells whose appearance departs most from the context are also those that differ most in expression. The 5% threshold is arbitrary; the full correlations above point in the same direction.

H EMBEDDING EXTRACTION FOR OTHER MODELS

PCA baseline. We flatten the raw RGB values of the cell-centered 224 px crop and project them onto their first 50 principal components, fit on all the slides (train+test) of each split. The linear classifier is trained as for the PFMs.

CONCH. We extract the CLS token of CONCH (Lu et al., 2024) at its native 448 px input. We do not apply spatial indexing, because CONCH's contrastive objective does not supervise individual tokens.

CellViT. We follow Hörst et al. (2025) and use the fine-tuned SAM (Kirillov et al., 2023) encoder of CellViT as a feature extractor, discarding the segmentation decoder. CellViT++ pools the tokens covered by its own predicted nucleus masks; since Xenium provides nucleus masks, we read out the encoder tokens directly as for the PFMs.

Table 10: Effect of masking for Phikon-v2 at 448 px on colon: cell-type classification (macro-F1) and nucleus-area probe (R^2). Values are mean ± standard deviation over $n = 8$ LOSO splits; the difference is computed per split and then averaged. Best condition in bold.

Condition	**Macro-F1**	**Nucleus area R^2**
Unmasked (ρ^{448}_{nuc})	$\mathbf{0.639 \pm 0.104}$	$\mathbf{0.266 \pm 0.047}$
Masked (ρ^{448}_{nuc})	0.597 ± 0.103	0.057 ± 0.039
Unmasked − masked	0.042 ± 0.008	0.209 ± 0.041

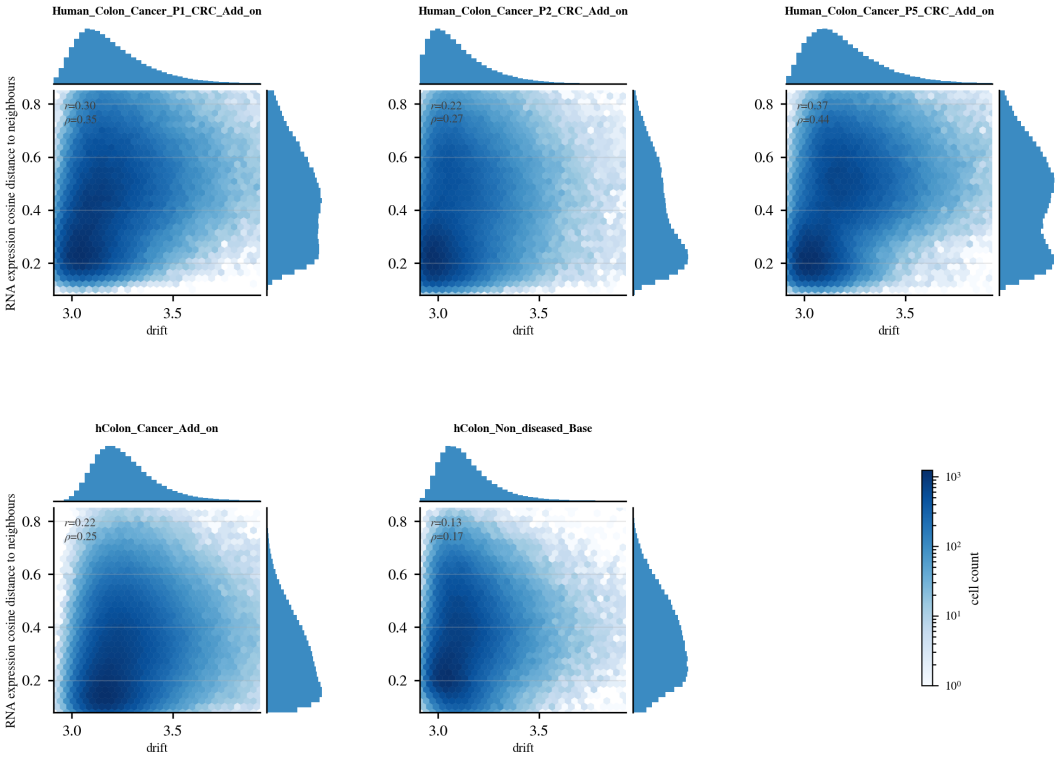


Figure 6: Drift versus expression divergence (cosine distance between a cell's expression and the mean expression of its neighbors) for the five colon slides (Phikon-v2). Pearson's r and Spearman's ρ are reported in the top left corner of each panel.

Table 11: Macro-F1 gain over the PCA baseline on pan-cancer. Values are mean ± standard deviation over the 14 LOSO splits. Best model in bold. †CONCH is run at its native 448 px input without resizing.

Model	**Read-out**	**Δ macro-F1 vs. PCA**
UNI2	ρ^{224}_{tok}	$\mathbf{0.381 \pm 0.118}$
H-Optimus-1	ρ^{224}_{tok}	0.343 ± 0.109
Phikon-v2	ρ^{224}_{tok}	0.318 ± 0.106
Virchow2	ρ^{224}_{tok}	0.341 ± 0.109
CONCH (Lu et al., 2024)	ρ^{448}_{CLS} †	0.135 ± 0.077
CellViT (SAM) (Hörst et al., 2024; Kirillov et al., 2023)	ρ^{224}_{tok}	0.217 ± 0.082