# DIAGNOSING SPATIAL PREDICTION IN HISTOLOGY REPRESENTATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Predicting the cellular and molecular state of surrounding tissue offers a test of what histology representations encode beyond a local image. A spatial prediction score, however, can rise for reasons other than prediction of unseen tissue: the comparison can hide the baseline, the target can favor one encoder, and the input can see part of the target. We develop a diagnostic protocol that isolates each effect with matched evaluations across thirteen TCGA cancers and colorectal (Xenium) and renal (Visium) spatial-omics cohorts. Retrained TRIPLEX and MERGE show 74–92% smaller context gains on raw expression than on targets averaged over neighboring tissue already visible to the model, and neither method beats the training mean on raw renal targets. With six frozen encoders, an advantage in predicting nearby over distant tissue coexists with both predictions below the training mean, and the UNI2-H versus ResNet50 gap shrinks by 77–78% on shared measured targets while pathology encoders keep their ranking. Notably, images of the target region predict it where local crops cannot, and two slide encoders show no gain that clears a prespecified criterion once the target is withheld. Local UNI2-H features still improve nearby cell-composition prediction beyond a reference from the crop's measured gene profile, and a case study of a spatially supervised encoder shows that its neighborhood gain depends on the target, readout, and baseline. We distill these diagnostics into three reporting requirements: absolute scores for both compared targets, shared measured targets for encoder comparisons, and a declaration of the tissue each input and target can see.

## 1 INTRODUCTION

A local H&E image captures a limited field of tissue, yet its morphology may carry information about the cellular and molecular state of its surroundings. Predicting that state from the image offers a way to examine the spatial information encoded by a histology representation, and paired histology and spatial transcriptomics now make such tests possible at scale (He et al., 2020; Jaume et al., 2024). Recent methods use spatial transcriptomic measurements to supervise image encoders (Han et al., 2026; Hemker et al., 2026), and incorporate neighborhood or whole-slide views into expression predictors (Chung et al., 2024; Ganguly et al., 2025). These approaches affect different aspects of the task: supervision shapes what a representation learns, while additional views change what tissue the model observes at prediction time. Whether the resulting gains provide evidence about tissue beyond the input depends on how the target is constructed and which tissue the model can see.

Spatially averaged expression targets can include measurements from neighboring tissue already visible to the model. TRIPLEX (Chung et al., 2024) predicts expression at a Visium spot from images of the spot, its neighborhood, and the whole slide. Its training target averages expression over the spot and its immediate neighbors, so the neighborhood view contains tissue that contributes directly to the quantity being predicted. A gain from this view can therefore reflect better prediction of a neighborhood average, without establishing how much context helps predict the spot's own expression. We examine this dependence by retraining TRIPLEX and MERGE (Ganguly et al., 2025) on smoothed and raw expression targets. Across our renal and colorectal cohorts, TRIPLEX's neighborhood-view gain and MERGE's spatial-edge component are 74–92% smaller on raw than on smoothed targets, measured by mean gene-wise Pearson correlation, and on raw renal targets neither model predicts better than the training mean (Section 4.1).



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

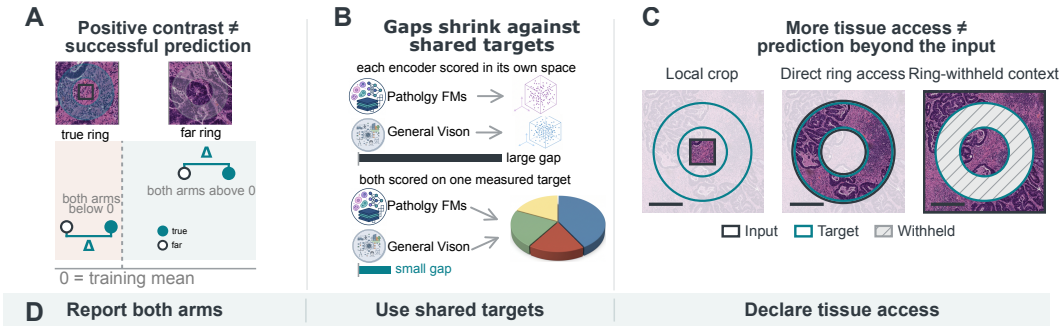



Figure 1: Three ways a spatial prediction score can mislead, and the reporting requirement for each (**D**). **A:** The same positive true-minus-far contrast ($\Delta$) can occur with both scores above or below the training-mean baseline. **B:** Gaps between pathology and general-vision encoders shrink when encoder-specific embedding targets are replaced by a shared measured target. **C:** Three inputs for one target ring: a local crop, direct ring access, and ring-withheld section context (H&E: Xenium CA560, chosen for geometry without inspecting predictions).

The dependence on target smoothing is one instance of a broader difficulty. A spatial prediction score reflects both the representation and the evaluation used to probe it, and it can rise for at least three reasons besides prediction of unseen tissue. (i) *The comparison hides the baseline*: a true-minus-far difference can be positive when both targets are predicted worse than a constant. (ii) *The target favors one model*: an encoder that predicts its own embeddings of the neighborhood holds an advantage that need not transfer to measured biology. (iii) *The input sees the target*: wider crops, slide context, and smoothed targets can each place part of the target inside the input. Each reason produces the same kind of number, so a single score cannot separate them. We therefore develop a diagnostic protocol that isolates each effect through matched comparisons of existing encoders without retraining. Figure 1 summarizes the three diagnostics and reporting requirements.

**Contributions.** Our contributions are threefold. **(i)** We re-examine TRIPLEX and MERGE and show that 74–92% of their context gain disappears once the target no longer averages over tissue the model can see, and that on raw renal targets neither beats the training mean (Section 4.1). **(ii)** We develop one diagnostic for each failure mode above and test it in matched experiments on six frozen tile encoders and two slide encoders across thirteen TCGA cancers and two spatial-omics cohorts (Sections 4.2–4.5): positive neighborhood contrasts coexist with below-baseline prediction; the UNI2-H versus ResNet50 gap shrinks by 77–78% on shared measured targets while the pathology-encoder ranking is preserved; and images of the target tissue support above-baseline prediction where local crops do not. **(iii)** With endpoints fixed before embedding extraction, we apply the diagnostics to SEAL-UNIv2, a released spatially supervised encoder, and show that its neighborhood gain over its base encoder depends on the target, readout, and baseline (Section 4.6). The diagnostics are consolidated into a reporting checklist (Appendix B.5, Table S4).

## 2 Related Work

**Histology-to-expression prediction and its evaluation.** Histology-to-expression methods relate morphology to measured tissue state (He et al., 2020; Chung et al., 2024; Ganguly et al., 2025; Zhu et al., 2025; Weng et al., 2026), increasingly on top of pathology foundation models (Chen et al., 2024; Filiot et al., 2024; Xu et al., 2024), and expression measurements are now also used to supervise the encoders themselves (Han et al., 2026; Hemker et al., 2026; Xu et al., 2025). HEST-1k pairs histology with spatial transcriptomics and benchmarks expression prediction by Pearson correlation over highly variable genes (Jaume et al., 2024); SpaPath-Bench addresses spatial-domain identification (Zhao et al., 2026). Prior work has noted the risk of mean-like predictions (Zhu et al., 2025) and shown that metric choice can reverse model comparisons (Chung et al., 2024), yet none of the expression-prediction studies in our literature sample reports a training-mean or constant-predictor comparison (Appendix G, Table S25). Spatial-component ablations can change supervision, geometry, and image context at once, and therefore need not isolate the contribution of neighboring images; in TRIPLEX, MERGE, and SEPAL, the context input includes tissue whose measurements

enter the smoothed training target (Chung et al., 2024; Ganguly et al., 2025; Mejia et al., 2023), and larger context windows can also reduce performance (Weng et al., 2026). Unlike these works, we do not propose a predictor; we ask which controls are needed before a spatial prediction gain can be read as prediction of measured targets outside the visible input.

**Leakage and validation of spatially structured data.** Leakage, in which information about the evaluation target reaches the model through its features or the train–test split, is a recurring source of overoptimistic results in ML-based science (Kaufman et al., 2012; Kapoor & Narayanan, 2023). For spatially, temporally, or hierarchically structured data, random splits overstate performance, and blocked or grouped validation is recommended (Roberts et al., 2017; Ploton et al., 2020); our patient-grouped splits follow this practice. Tissue access within an example is a distinct problem: the input may contain tissue that contributes directly to its own target, and a correctly grouped split does not remove this dependence. Target smoothing does not by itself invalidate an evaluation, but it changes what a spatial gain measures. We therefore treat tissue access as a question of attribution that must be declared alongside the split. Appendix I relates these diagnostics to context prediction, matched controls, and probe reliability.

## 3 Evaluation setup

Our evaluation examines spatial prediction through three comparisons, one for each diagnostic in Section 1: performance against a training-mean baseline, encoder differences on shared targets, and prediction under varying tissue access. We first define the matched prediction task and its score (3.1), then specify the cohorts and encoders (3.2), shared targets and crop references (3.3), and tissue-access comparisons (3.4), and close with the reporting requirements (3.5).

### 3.1 Matched spatial prediction

The reference task evaluates prediction from a local image to its surrounding tissue. A *query* is a location on a tissue section, and the *input* is an H&E crop centered at that location and encoded by a frozen encoder. The *true target* summarizes a quantity in a surrounding annulus that does not overlap the crop, such as cell-type composition 80 to 160 µm from the query center. The *far target* summarizes the same quantity in an annulus of the same size centered at a distant location on the same slide. Separate probes predict each target from the same input. Section 3.4 varies tissue access.

The matched far target retains shared slide-level context but samples tissue outside the query's neighborhood. Information specific to that neighborhood should therefore favor the true target. We quantify this advantage as the *neighborhood contrast*, the difference between the true- and far-target scores. Matching slide identity and target eligibility controls for slide-level information; evaluating each arm against its own baseline additionally establishes absolute performance (Section 4.2).

Representation-space targets average patch embeddings within each annulus, and measured targets replace embeddings with cell-composition or gene-expression measurements (Section 3.3). For annulus $b = (r_1, r_2]$, far centers lie more than $2r_2$ from the query, and both rings satisfy identical eligibility criteria and are subsampled to the smaller patch count (Appendix B.2).

**Absolute performance and neighborhood contrast.** We score each arm by $R^2$ relative to its training-mean predictor. For a held-out patient group $s$ and arm $a \in \{\mathrm{true}, \mathrm{far}\}$,

$$R^2_{s,a} = 1 - \frac{\sum_{i \in s} \|y_{i,a} - \hat{y}_{i,a}\|^2}{\sum_{i \in s} \|y_{i,a} - \bar{y}_{\mathrm{train},a}\|^2}, \qquad \delta_b = R^2_{\mathrm{true},b} - R^2_{\mathrm{far},b}.$$

Here, $\hat{y}_{i,a}$ is the probe prediction and $\bar{y}_{\mathrm{train},a}$ is the training mean for that arm. Zero denotes performance equal to the corresponding training-mean predictor; negative values denote greater error than that predictor. The contrast $\delta_b$ compares the two arms at ring $b$.

**Probe fitting.** Separate ridge probes for the true and far arms use patient-grouped splits and training-only preprocessing. Appendices B and E.1 specify sampling, donor controls, transformations, and model selection.

**Scoring conventions.** Reported $R^2$ values average patient scores, giving each patient equal weight regardless of query count. Unless stated otherwise, we report the *global* score relative to the training mean, which captures differences between patients and sections as well as variation within them;

the *within-section* score centers targets and predictions by their test-section means and isolates variation within a section. Because the two scores can differ in their conclusions, each result identifies the score used. Intervals resample patients conditional on the fitted predictions; spatial-cohort comparisons use paired patient resampling and exact paired sign-flip tests, and TCGA encoder-family comparisons pair encoders within cancer (Appendices B.1 and B.3).

### 3.2 Cohorts and encoder comparisons

**Encoders.** The main panel comprises three pathology encoders, UNI2-H (Chen et al., 2024), Phikon-v2 (Filiot et al., 2024), and the Prov-GigaPath tile encoder (Xu et al., 2024), and three natural-image encoders, supervised ResNet50 (He et al., 2016), supervised ViT-B/16 (Dosovitskiy et al., 2021), and DINOv3 ViT-B/16 (Siméoni et al., 2025), each receiving an $80 \times 80$ µm H&E field. TITAN (Ding et al., 2025) and the Prov-GigaPath slide encoder (Xu et al., 2024) evaluate section context (Section 3.4). A case study compares SEAL-UNIv2, a released spatial-expression-aligned fine-tune of UNI2-H (Hemker et al., 2026), with its base encoder using identical preprocessing and colorectal sections outside SEAL-UNIv2's training data. Appendix B.1 specifies checkpoints, sampling, and capacity controls; Appendix F.8 describes the SEAL implementation.

**Cohorts.** Three cohorts support complementary comparisons. TCGA provides H&E-only data from thirteen cancers for the distance sweep, with targets defined as ring averages of encoder embeddings (Appendix C). The main biological cohort has sixteen Xenium colorectal cancer patients, each with a tumor and a matched-normal section. Paired H&E and single-cell Xenium measurements provide cell-composition and gene-expression targets around each query. The Visium clear-cell renal cell carcinoma subset of HEST-Benchmark (Jaume et al., 2024) provides one section for each of twenty-four patients, extending evaluation to another organ and assay. Patients are the independent units: all splits retain each patient's sections together, and all intervals resample patients.

**Matched query sets.** Eligibility criteria are defined for each comparison. Within a comparison, all arms share identical queries, far partners, and patient splits. Table S1 in Appendix B lists the study comparisons with complete query-set definitions, counts, and eligibility criteria.

### 3.3 Shared measured targets and crop-conditioned prediction

**Cell composition (Xenium).** The main measured target is the centered-log-ratio composition of four cell classes (epithelial cells, fibroblasts, myeloid cells, and lymphocytes) in the (80,160] µm ring. Each encoder predicts both its own ring embeddings and this shared composition using identical queries, folds, and far partners. The primary endpoint is the patient-paired change in the UNI2-H versus ResNet50 contrast gap from the embedding targets to this shared composition. A cross-encoder matrix further evaluates all source encoders within each common embedding target space. Appendices B.3 and F.3 specify target construction and matrix fitting.

**Gene expression (Visium renal).** The renal cohort repeats the primary endpoint using mean log-normalized expression of the benchmark's fifty genes over spots in the (80,160] µm ring. Spot measurement footprints lie outside the crop, and far rings are count-matched. Ridge penalty candidates include the training-mean predictor. Appendix B.3 details assay geometry and target construction.

**Prediction conditional on crop measurements.** Spatial autocorrelation makes a crop's measured tissue state a reference for predicting its surroundings. We fit reference models using the crop's measured composition or its 372-gene profile, each augmented with log cell count, and quantify the increment from adding image features. For reference features $X$, the *conditional image gain* subtracts the far-ring increment from the true-ring increment,

$$G(X) = \left(R^2_{X+\text{image}} - R^2_X\right)_{\text{true}} - \left(R^2_{X+\text{image}} - R^2_X\right)_{\text{far}},$$

using the same penalty and block-weight search for all models. The primary endpoint is $G$ for the gene-profile reference. Practical equivalence to zero requires its 90% interval to lie within $\pm 0.02$.

### 3.4 Controlled variation in tissue access

**Crop expansion.** UNI2-H, Phikon-v2, and ResNet50 receive re-extracted 80, 160, or 320 µm crops, evaluated with ridge and MLP readouts under equal tuning budgets. Fixed-resolution tile means control for changes in image resolution. The nearby (80,160] µm ring lies entirely within the 320 µm crop. Appendix F.5 specifies input construction and fitting.

**Direct target access.** The farther (320,640] μm ring lies outside every crop. Its *direct-access* input averages stored tile embeddings centered in the ring, exposing the target tissue. This input is a reference for prediction from the target's own H&E under the tested readout, not an upper bound.

**Section context with target withholding.** TITAN (Ding et al., 2025) and the Prov-GigaPath slide encoder (Xu et al., 2024) process sections at native tile scales; the readout uses the token of the tile containing the query. We compare four inputs: the tile feature; the slide encoder on that tile alone; the whole-section token, which observes both rings; and a ring-withheld token, built by removing every tile overlapping the farther true ring before encoding. Far-ring tissue remains available. Appendix F.6 specifies tile geometry, masking, and the prespecified context-gain criterion.

### 3.5 Diagnostic reporting requirements

The protocol translates the three diagnostics into corresponding reporting requirements. First (*report both arms*), report absolute scores for both compared arms against their respective training-mean predictors, together with the between-arm difference. Second (*use shared targets*), compare encoders on shared measured targets and report target reliability; for embedding targets, evaluate all encoders within each common target space. Third (*declare tissue access*), specify each input's tissue footprint and the target construction, including whether spatial averaging incorporates measurements from visible tissue, and include a target-withheld variant for contextual models. Table S4 in Appendix B.5 summarizes these requirements and the alternative explanations they address.

## 4 Results

We first examine target construction in published spatial-context methods (Section 4.1), then use frozen encoders to distinguish relative gains from absolute prediction and assess the effects of target space and tissue access (Sections 4.2–4.5). A final case study applies these diagnostics to spatially supervised representations (Section 4.6).

### 4.1 Published context gains depend on target smoothing

**Replication as a prerequisite for attribution.** We evaluate the tissue-access requirement in TRIPLEX (Chung et al., 2024) and MERGE (Ganguly et al., 2025), retraining both on the renal cohort and on Xenium cells aggregated into pseudo-spots with Visium geometry. Because these cohorts differ from the original data, we decompose a published gain only after it replicates with a patient-bootstrap interval above zero; the TRIPLEX whole-slide branch does not pass this gate (Appendix F.9). Controls, replication gates, endpoints, and decision rules were fixed before training. We report mean Pearson correlation over fifty genes, the metric used by both methods, alongside global $R^2$ relative to the training mean. Figure 2 summarizes context effects and absolute scores; Appendix F.9 and Table S15 provide full results and complementary checks.

**TRIPLEX: dependence on visible neighbors in the target.** TRIPLEX's neighborhood-view gain replicates in both cohorts when the target averages each spot with the same immediate neighbors visible in that view, but decreases substantially on raw spot targets. Masking adjacent capture areas removes 77–80% of the smoothed-target gain. Smoothing-footprint and reliability-normalized analyses support this dependence (Appendix F.9), whereas the residual raw-target gain does not separate from section-level information.

**MERGE: dependence on smoothing under fixed image features.** MERGE provides a complementary test through its spatial edges between spots. We compare SPCS-smoothed targets (Liu et al., 2022) with raw spot targets, permuting coordinates within each section while retaining every image feature. The spatial-edge gain is substantially larger on smoothed targets; on raw targets, its correlation intervals span zero.

**Context gains attenuate on raw targets.** Across both methods and cohorts, replacing smoothed with raw targets reduces context gains by 74–92% in correlation and 79–94% in global $R^2$ (Tables S16 and S18). Full results, cross-metric interactions, and patient-level checks are in Appendix F.9.

**Absolute performance qualifies the context gains.** On raw renal targets, both TRIPLEX models and all three MERGE variants score below the training mean. Only two renal configurations exceed

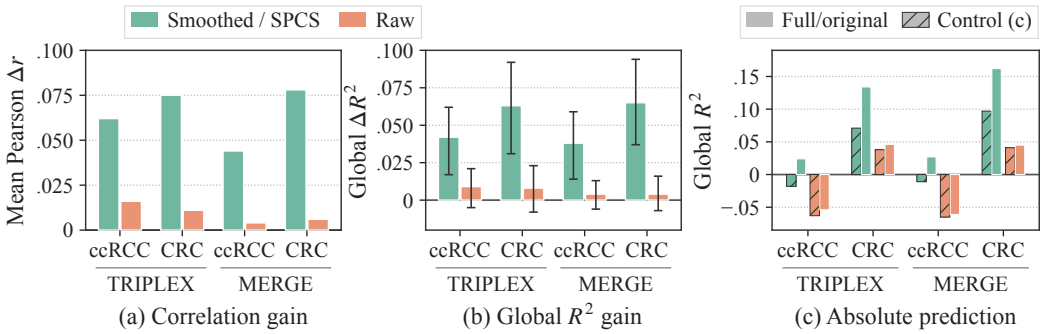



Figure 2: Context gains depend on target smoothing. (a) Changes in mean gene-wise Pearson correlation. (b) Changes in global $R^2$, with 95% paired patient-bootstrap intervals. (c) Absolute global $R^2$; zero denotes the training mean. TRIPLEX contrasts full with target-view-only (TEM), and MERGE contrasts original with permuted coordinates. CRC denotes Xenium colorectal pseudo-spots; MERGE smoothing uses SPCS. Panels (a) and (c) show point estimates; full results and interactions are in Appendix F.9.

this baseline: TRIPLEX with its neighborhood view and MERGE with intact spatial edges, both using smoothed targets. Above-baseline prediction in this cohort therefore occurs only when the input observes tissue contributing to the target average.

**Target construction defines the prediction task.** Together, these results show that much of the gain attributed to spatial context reflects prediction of an average that includes neighbors already visible to the model (Appendix F.9). The following comparisons isolate these distinctions with frozen encoders, varying the input, target, and readout one at a time.

### 4.2 Absolute scores distinguish prediction from relative advantage

**Above-baseline prediction with a positive contrast.** For nearby cell composition, a positive neighborhood contrast accompanies above-baseline prediction. On the shared-target set, all six encoders beat the training mean on average, with true-target $R^2$ of 0.07–0.13 and far-target scores near zero (Figure S4A). Patient variation nevertheless yields global contrast intervals that cross zero (Appendix F.3).

**Below-baseline prediction despite a positive contrast.** For the farther ring, a positive contrast frequently coexists with below-baseline prediction. Most configurations predict the true target better than the far target while remaining worse than the training mean. For UNI2-H with a 320 µm crop and an MLP readout, true- and far-target scores are $-0.03$ and $-0.22$, giving a contrast of $+0.19$ (Table S23). This relative advantage alone could be interpreted as information about the farther ring, although neither arm outperforms a constant predictor.

**Above-baseline prediction despite a small contrast.** A small contrast can conversely obscure above-baseline prediction in both arms. With the wider penalty search for renal gene prediction, UNI2-H achieves true- and far-target scores of 0.12 and 0.09. Both exceed the training mean, despite their small difference. Reporting both absolute scores therefore distinguishes relative neighborhood advantage from predictive performance in either direction. Passing the HEST correlation criterion (Jaume et al., 2024) likewise does not ensure above-baseline gene prediction (Appendix F.9).

### 4.3 Shared measured targets narrow encoder gaps while preserving their ordering

**Encoder-specific targets yield large pathology-encoder advantages.** When each encoder predicts the mean of its own embeddings in the surrounding ring, every pathology encoder exceeds every natural-image encoder at short range in all thirteen TCGA cancers (Appendix B.1). This ordering also replicates in Xenium (Table S3), but these comparisons assign a different target to each encoder.

**Shared measurements reduce the gap by 77–78%.** On the shared cell-composition target, the UNI2-H versus ResNet50 contrast gap falls from 0.23 to 0.05: a patient-paired reduction of 0.180

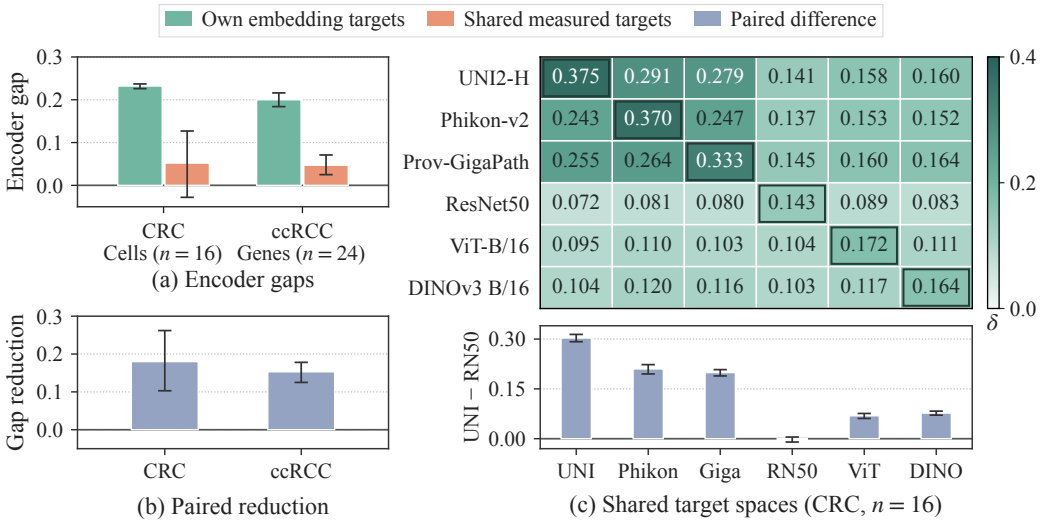



Figure 3: Target-space dependence of encoder differences. (a–b) UNI2-H minus ResNet50 gaps in contrast $\delta$ and their paired reduction from own embeddings to shared measured targets (CRC: Xenium CRC). (c) Cross-encoder contrasts on Xenium CRC ($n = 16$): rows denote input encoders; columns follow the same order of target spaces, with boxes marking self-space. Aligned bars compare UNI2-H with ResNet50 within each space. All results use the (80,160] µm ring; whiskers are 95% paired patient-bootstrap intervals.

[0.103, 0.262], or 78% (Figure 3(a–b); Table S7). The renal cohort repeats this pattern in another organ and assay: the gap falls from 0.20 on ring embeddings to 0.05 on measured gene expression, a 77% reduction. The reduction persists after target-reliability normalization in both cohorts (Table S24; Appendix F.7); in the renal cohort, unequal embedding-target reliability explains part of it.

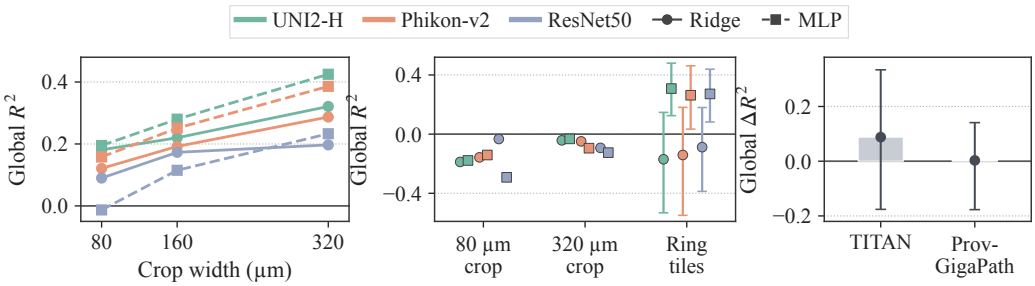
**Gap magnitude and encoder ordering are distinct.** The smaller measured-target gap retains an overall pathology-encoder advantage: patient-averaged cross-family win rates are 65% on cell composition and 79% on gene expression (Table S14). The cross-encoder matrix further separates this ordering from target-space effects (Figure 3(c); Table S8). Each input encoder attains its largest contrast in its own target space (outlined diagonal). Excluding self-space entries, every pathology encoder exceeds every natural-image encoder in all six spaces. The pathology-encoder advantage thus persists under shared targets, while encoder-specific targets amplify its magnitude.

**Image features contribute beyond the crop's measured state.** Shared measurements also permit a complementary test of the information carried by image features. Spatial autocorrelation makes the crop's measured composition predictive of its surroundings (true-target $R^2 = 0.31$; Table 1); the relevant comparison is the additional information provided by the image. Adding UNI2-H features to the crop's measured 372-gene profile improves true-target prediction more than far-target prediction, yielding conditional image gain $G = 0.041$ [0.011, 0.074] (Table 1). The representation therefore contributes neighborhood information beyond the crop's measured molecular state.

Table 1: Image gain beyond the crop's measured state. UNI2-H features are added to each reference to predict (80,160] µm four-class cell composition on X-CropRef. $G$ is the true-arm increment minus the far-arm increment; intervals are 95% paired patient bootstraps.

| Crop reference | True $R^2$: ref. → +image | Far $R^2$: ref. → +image | $G$ [95% CI] |
|---|---|---|---|
| Composition | 0.3057 → 0.3299 | 0.0162 → 0.0118 | 0.0285 [0.0032, 0.0491] |
| Gene profile (primary) | 0.2440 → 0.2741 | 0.0119 → 0.0010 | 0.0411 [0.0108, 0.0741] |

Paired sign-flip $p = 0.033$ (composition) and 0.028 (gene profile). The primary 90% interval [0.0147, 0.0681] is not contained within the $\pm 0.02$ equivalence margin.




(a) Nearby target (b) Direct ring access (c) Target withheld

Figure 4: Prediction depends on access to the target tissue. (a) Nearby-ring composition across crop widths on X-Near; the 320 µm crop contains the ring. (b) Farther-ring composition from local crops or ring tiles on X-Ring, with ridge and MLP readouts. (c) Ring-withheld minus center-only global MLP scores on X-Context. Zero denotes the training mean in (a–b) and no increment in (c). These are distinct matched query sets from sixteen patients; displayed whiskers are 95% intervals.

### 4.4 Access to target tissue explains gains from expanded inputs

**Crop expansion improves prediction when it includes the target.** The 320 µm crop contains the entire nearby ring. Increasing crop width from 80 to 320 µm raises UNI2-H ridge $R^2$ for nearby composition from 0.18 to 0.32 (Figure 4(a)), with similar improvements for Phikon-v2 and ResNet50. Increasing readout flexibility on the smaller crop does not recover this gain: ridge on the 320 µm crop outperforms an MLP on the 80 µm crop. This comparison attributes the improvement to the additional visible tissue rather than readout capacity.

**Direct target access reveals predictability beyond local crops.** The farther ring lies outside every crop. On the direct-access set, both crop widths give negative true-target scores for all encoders, whereas an MLP on tiles within the ring reaches $R^2$ of 0.26–0.31 (Figure 4(b)). For UNI2-H, direct access improves on the matched 320 µm crop by 0.34, with a patient interval above zero (Table S23). The target is thus predictable from its own H&E even when the tested local crops remain below baseline. Ridge readouts on ring tiles score lower (penalty selection: Appendix F.5).

### 4.5 Evaluating slide context with the target withheld

**Separating context from direct target access.** Slide encoders extend tissue access to the whole section, including the target itself. To distinguish contextual prediction from direct target access, we remove every tile overlapping the farther ring before encoding and compare the resulting token with the same model applied to the center tile alone.

**Primary evaluation of withheld context.** Neither TITAN nor the Prov-GigaPath slide encoder meets the prespecified context-gain criterion. Under the primary global MLP evaluation, ring-withheld minus single-tile differences are 0.088 [−0.176, 0.334] for TITAN and 0.003 [−0.177, 0.141] for Prov-GigaPath, with Holm-adjusted $p = 1.0$ for both. Absolute global scores remain negative for every ring-withheld input under both readouts (Figure 4(c); Table S13). The wide intervals bound the effect size but do not exclude a context benefit. Secondary within-section scores improve with ring-withheld context but remain negative for both MLP readouts, and the ring-withheld input retains the inner 320 µm neighborhood that the single-tile input lacks (Table S13; Appendix F.6).

### 4.6 Case study: spatial supervision under matched evaluation

**Spatial supervision includes the predicted neighborhood.** SEAL provides a direct test of whether spatial supervision improves neighborhood prediction beyond a general predictive gain. It fine-tunes UNI2-H with low-rank adapters to predict Visium expression averaged over each spot and its immediate neighbors (Hemker et al., 2026). Those neighbors lie approximately 100 µm away, inside our nearby ring, so the supervision explicitly includes neighborhood information.

**Matched comparison against the base encoder.** We compare released SEAL-UNIv2 with its base encoder, UNI2-H, on the crop-reference set, with the adapters as the only difference between arms.

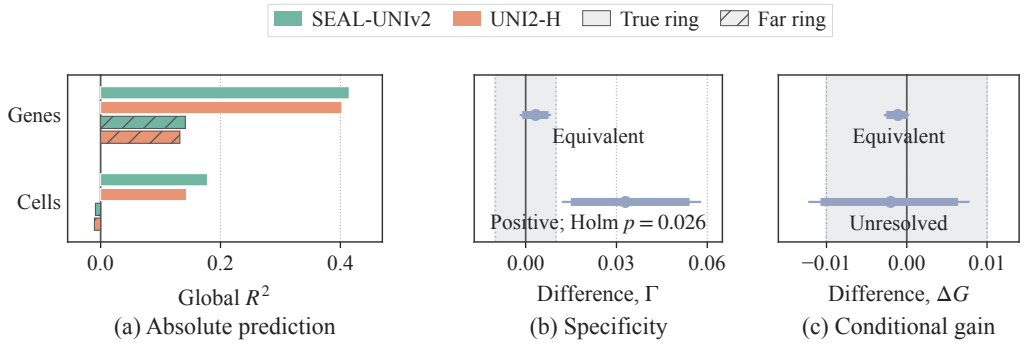



Figure 5: SEAL-UNIv2 versus UNI2-H on nearby targets with global ridge readouts (X-CropRef). (a) Absolute true/far $R^2$ scores. (b) Neighborhood-specific difference $\Gamma$, the SEAL advantage on the true ring minus its advantage on the far ring. (c) Difference in conditional image gain, $\Delta G = G_S - G_U$, over the crop's measured gene profile. In (b–c), thin/thick intervals are 95%/90% patient bootstraps; gray bands mark $\pm 0.01$. Equivalence requires the whole 90% interval inside the band; (b–c) use different scales.

The comparison, endpoints, $\pm 0.01$ equivalence margin, and decision rules were fixed before SEAL embedding extraction (Appendix F.8). The *neighborhood-specific difference* subtracts SEAL's advantage on the far ring from its advantage on the true ring (Figure 5(b)).

**Transfer does not imply uniform neighborhood specificity.** SEAL transfers to this unseen cohort and assay, predicting the crop's own gene profile slightly better than UNI2-H. With global ridge readouts, its advantage for cell composition is neighborhood-specific: +0.033 [+0.012, +0.058] (Holm-adjusted $p = 0.026$). For gene expression, SEAL improves true- and far-ring prediction almost equally; the neighborhood-specific difference of +0.003 is equivalent to zero under the fixed margin. This equivalence is conditional on the prespecified $\pm 0.01$ tolerance, which was not calibrated to a biologically meaningful effect, and does not establish absence of spatial information.

**Conditioning on crop measurements distinguishes incremental image gain.** Under the same primary evaluation, the difference in conditional image gain beyond the crop's gene profile is equivalent to zero for genes under the margin, while the cell result remains unresolved (Figure 5(c)).

**The comparison depends on the evaluation.** The gene neighborhood-specific difference becomes positive with an MLP readout or within-section scoring, and the encoder ordering reverses between SEAL's and UNI2-H's embedding target spaces. Appendix F.8 reports these secondary comparisons and the sensitivity to the equivalence margin. For this released SEAL checkpoint, evaluating the target, readout, baseline, and embedding space together distinguishes transfer, neighborhood specificity, and conditional image gain, exposing dependencies rarely reported jointly.

## 5 Conclusion

We present a diagnostic protocol for spatial prediction in histology representations that separates prediction of unseen tissue from three ways an evaluation can inflate a score: a contrast that hides the baseline, a target that favors one encoder, and an input that sees part of the target. Replacing targets averaged over visible tissue with raw expression removes 74–92% of the context gain of TRIPLEX and MERGE and leaves neither above the training mean in the renal cohort. Across six frozen tile encoders and two slide encoders, it shows that a positive neighborhood contrast can coexist with below-baseline prediction, that the UNI2-H versus ResNet50 gap shrinks by 77–78% on a shared measured target while the pathology-encoder ordering survives, and that direct access to the target tissue supports prediction that local crops cannot. These findings reduce to three reporting requirements that need neither a new model nor new data: score both arms against their training means, compare encoders on the same measured target, and state which tissue each input can see, withholding the target from contextual inputs. Smoothed targets remain valid endpoints for neighborhood-averaged expression; read with these checks, a spatial prediction score becomes evidence about the tissue a representation has predicted, not about the evaluation that produced it.

## Reproducibility Statement

All analysis code is provided in the supplementary materials: patch sampling, feature extraction, neighborhood and measured-target construction, matched far-target controls, probe fitting, scoring, statistics, and figures, with a mapping from each script to the section it produces. Input data are obtained separately: the whole-slide images are public TCGA diagnostic slides and the renal cohort is the public HEST-Benchmark ccRCC subset, while the paired colorectal H&E and Xenium sections are restricted-access and are not redistributed. Analysis sets, queries, patient folds, far partners and eligibility rules are written to hashed manifests and frozen before scoring, so all model arms within a comparison are fitted on identical rows; patients are the unit of splitting and inference. Given the same inputs and manifests, the reported results are reproduced under fixed seeds. Data access and ethical considerations are summarized in Appendix J.

## AI Use Statement

AI tools were used to assist with language editing and formatting. All study design, analyses, statistical decisions, result interpretation, and final claims were developed and verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the manuscript.

## References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197, 2020.

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.

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

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 31(11):3749–3761, 2025.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

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.

Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15611–15620. IEEE, 2025.

Minghao Han, Dingkang Yang, Linhao Qu, Zizhi Chen, Gang Li, Han Wang, Jiacong Wang, and Lihua Zhang. Fusing pixels and genes: Spatially-aware learning in computational pathology. In *International Conference on Learning Representations*, volume 2026, pp. 130377–130410, 2026.

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.

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

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

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

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.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

Agustin Lage-Castellanos, Giancarlo Valente, Elia Formisano, and Federico De Martino. Methods for computing the maximum performance of computational models of fmri responses. *PLoS computational biology*, 15(3):e1006397, 2019.

Yusong Liu, Tongxin Wang, Ben Duggan, Michael Sharpnack, Kun Huang, Jie Zhang, Xiufen Ye, and Travis S Johnson. Spcs: a spatial and pattern combined smoothing method for spatial transcriptomic expression. *Briefings in Bioinformatics*, 23(3):bbac116, 2022.

Gabriel Mejia, Paula Cárdenas, Daniela Ruiz, Angela Castillo, and Pablo Arbeláez. Sepal: spatial gene expression prediction from local graphs. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2286–2295. IEEE, 2023.

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

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11(1):4540, 2020.

David R Roberts, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346, 2018.

Ziqiao Weng, Yaoyu Fang, Jiahe Qian, Xinkun Wang, Lee AD Cooper, Weidong Cai, and Bo Zhou. Hifusion: hierarchical intra-spot alignment and regional context fusion for spatial gene expression prediction from histopathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10630–10637, 2026.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gl3D-xY7wLq.

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.

Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Cheng Jin, Shu Yang, Jinbang Li, Zhengyu Zhang, Chenglong Zhao, Huajun Zhou, et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature communications*, 16(1):11406, 2025.

Bokai Zhao, Yiyang Zhang, Yuanchi Zhu, Hanqing Chao, Long Bai, Tai Ma, Minfeng Xu, Ming Song, and Tianzi Jiang. Benchmarking pathology foundation models for spatial domain understanding. *arXiv preprint arXiv:2605.25764*, 2026.

Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *International Conference on Learning Representations*, volume 2025, pp. 19666–19690, 2025.

# A SCOPE AND LIMITATIONS

The study evaluates coarse cell classes and selected gene panels, with two slide encoders assessed in one spatial cohort. Published methods were retrained on our cohorts rather than on their original data, so their outcomes remain conditional on passing the replication gate. Sixteen and twenty-four patients limit statistical power, and several intervals are wide.

TCGA embedding targets alone do not validate biological prediction. Renal scores depend on the penalty range, and split-half reliability estimates depend on sampling assumptions. Additional encoders, finer cell states, and hierarchical models remain settings in which to apply the protocol.

# B SAMPLING AND SCORING

Table S1 defines the analysis sets, their eligibility criteria, and the comparisons they support. The sets overlap and must not be pooled by summing their query counts. Within each comparison, all arms share queries, far partners, and patient folds. Crop-width comparisons require the widest crop to fit within tissue. The crop-reference set is denoted X-CropRef in the main figures and table and X-Crop in the original analysis registry; both names identify the same matched set.

Table S1: Analysis sets and matched comparisons. Each row identifies the input–target comparison, cohort, and query eligibility. Query counts refer to overlapping subsets of the same sections and must not be summed. Parenthetical codes identify the sets throughout the appendix.

| Analysis set | Prediction task | Cohort and eligibility |
|---|---|---|
| **Distance sweep** (T-Band, T-Common) | Distance dependence: six crop encoders → ring embeddings | TCGA, 13 cancers; cross-band controls keep 13 cancers through 4480 µm and 10 at 5120 µm |
| **Shared-target set** (X-Shared) | Encoder gaps on shared measured targets: six crop encoders → own embeddings or cell composition, (80,160] µm | Xenium CRC, 16 patients, 91,861 queries; valid targets for every encoder |
| **Renal gene set** (H-Gene) | Shared-target comparison across organ and assay: six crop encoders → own embeddings or 50-gene expression, (80,160] µm | Visium ccRCC, 24 patients, 71,684 queries |
| **Crop-reference set** (X-Crop; X-CropRef) | Crop measurements with or without image features; SEAL versus UNI2-H → (80,160] µm composition or gene profile | Xenium CRC, 16 patients, 82,053 queries; $\geq$10 crop cells, $\geq$50 ring cells |
| **Crop-width set** (X-Near) | Target-visible crop expansion: 80/160/320 µm crops or fixed-resolution tile means → (80,160] µm composition | Xenium CRC, 16 patients, 80,857 queries; full 2×2 and $\geq$75% 4×4 tile coverage |
| **Farther-ring set** (X-Wide) | Full crop-width and tiling scan → (320,640] µm composition | Xenium CRC, 16 patients, 98,555 queries; farther-ring validity and tiling coverage |
| **Direct-access set** (X-Ring) | Prediction from the target's own H&E: 80/320 µm crops or ring tiles → (320,640] µm composition | Xenium CRC, 16 patients, 91,651 queries; subset of X-Wide with $\geq$50% ring coverage |
| **Slide-context set** (X-Context) | Section context with the target withheld: TITAN and Prov-GigaPath slide tokens → (320,640] µm composition | Xenium CRC, 16 patients, 87,387 queries; subset of X-Wide with TITAN center tiles inside 320 µm |

**Protocol variants.** Distance analyses vary the number of far targets (1/2/4/8) and ring patches (12/24/48), and examine donor nulls, feature dimension, and target reliability. Shared cellular-target analyses evaluate the full source-by-target encoder matrix, target scaling, ridge penalty range, split-half reliability normalization, and per-patient ordering. Shared expression-target analyses add the (160,320] µm ring and assess penalty range, reliability normalization, and per-patient ordering.

Crop-reference variants use six-label composition, six labels with genes, and class-pair interactions; they also examine the original tuning scheme, reference footprint, and target definition. Nearby tissue-access analyses compare fixed-resolution tile means and ridge/MLP readouts. Farther tissue-access analyses use both readouts and the complete crop-width scan in Appendix F.5. Section-

context comparisons use tile, single-tile, whole-section, and ring-withheld tokens, UNI2-H crop references, and both readouts. An additional Prov-GigaPath-only analysis uses all 98,555 farther-ring queries in X-Wide.

**Distance support and donor construction.** Donor comparisons require eight eligible far rings within each band; distance comparisons additionally retain common rows across the supported bands. T-Common support is selected from training/validation geometry: thirteen cancers through 4480 µm and ten at 5120 µm. LIHC, COAD, and GBM contribute through the sixth fixed-width band. This restriction applies to the cross-band controls; the original T-Band sweep retains thirteen cancers in all seven bands (Table S5).

Multiple-donor comparisons distinguish score averaging, target averaging, and pooled fixed-count sampling. Score averaging fits each donor target separately; target averaging predicts the mean of the donor targets; pooled sampling draws the same total number of distinct patches from the pooled donors as from the true ring. Target averaging and pooled sampling therefore change the spatially averaged target.

Fixed-count distance comparisons use twelve patches per ring, angularly stratified sampling, five repeats, and ten random-half assignments to assess reliability. Sensitivities use 24 or 48 patches, spatial-block halves, pooled or within-slide centering, and exclusion of pixel-overlapping patches. Split-half reliability is a diagnostic and does not rescale these scores. Geometric nulls use 199 physically separated same-slide donor draws with replacement, refitting the target basis and penalties for each draw. Block-donor, frozen-basis/shared-penalty, and fixed-count variants use twenty draws.

### B.1 Encoder panel and sampling

Table S2 lists the encoder panel, feature dimensions, and roles. Checkpoint sources are the UNI2-H model card, the Phikon-v2 and Prov-GigaPath publications, and the original ResNet, ViT, and DINOv3 papers. Representation probes use only local tile features, including those from the Prov-GigaPath tile encoder.

Table S2: Encoder panel for representation probes. The primary panel contains six crop encoders; DINOv3 ViT-L/16 provides a supplementary architecture and capacity control. Dimensions refer to the extracted local features. Natural denotes natural-image pretraining.

| Encoder | Pretraining | Dim. | Role |
|---|---|---|---|
| UNI2-H | Histology | 1536 | Primary pathology |
| Phikon-v2 | Histology | 1024 | Primary pathology |
| Prov-GigaPath (tile) | Histology | 1536 | Primary pathology |
| ResNet50 | Natural, supervised | 2048 | Convolutional reference |
| ViT-B/16 | Natural, supervised | 768 | Transformer reference |
| DINOv3 ViT-B/16 | Natural, self-supervised | 768 | Self-supervised reference |
| DINOv3 ViT-L/16 | Natural, self-supervised | 1024 | Architecture/capacity control |

Crop fields are sampled on a 40 µm grid. In the renal Visium cohort, crops are centered on spots with at least 200 UMIs.

The fixed-width TCGA sweep covers ACC, BLCA, BRCA, CHOL, COAD, DLBC, ESCA, GBM, LGG, LIHC, LUAD, MESO, and PCPG, with patient-grouped 60/20/20 training/validation/test splits. Each split includes up to thirty sampled slides, with up to 400 query centers per slide. Slide–band pairs without valid far targets are excluded. A 128-component PCA basis fitted to the training true targets transforms both arms; scaling is also fitted on training data. Each arm selects its ridge penalty separately by validation error from $\{10, 100, 10^3, 10^4, 10^5\}$. TCGA intervals use 4,000 patient bootstraps. Encoder-family gap comparisons pair encoders within cancer and use the prespecified $\pm 0.02$ stability margin.

All 546 original T-Band encoder–cancer–band cells (6 encoders $\times$ 13 cancers $\times$ 7 bands) satisfy equal true/far patch counts, disjoint patch identities, and the required donor separation. Reliability assessments use random and spatial-block halves under both pooled and within-slide centering.

The dimension-matched control standardizes feature coordinates before orthogonal PCA to 768 dimensions and retains the projected component variances. Encoder-family separation holds in all thirteen cancers for the (80,160] and (320,640] µm rings. DINOv3 and supervised ViT-B/16 have similar fixed-width contrasts (all $p \geq 0.216$). Supplementary DINOv3 ViT-L/16 results provide a capacity control and do not enter the primary six-encoder distance fits.

### B.2 Distinct annulus protocols

**Matched ring targets.** For a crop $x_i$ centered at $u_i$, let $z_i = f(x_i)$ be its frozen representation and $\mathcal{N}_b(u_i)$ the set of patches in annulus $b = (r_1, r_2]$. The true representation-space target is the mean embedding of those patches:

$$y_{i,b}^{\text{true}} = \frac{1}{|\mathcal{N}_b(u_i)|} \sum_{j \in \mathcal{N}_b(u_i)} z_j,$$

where the far target applies the same operation around a center more than $2r_2$ from the query. Both rings satisfy identical eligibility criteria and are subsampled to the smaller patch count. Measured targets replace embeddings with cell-composition or gene-expression measurements (Section 3.3).

**Short-range replication.** Table S3 compares the original short-range self-space contrasts in TCGA COAD and Xenium CRC. These comparisons are distinct from the matched target-space analysis in Section 4.3.

Table S3: Short-range self-space contrasts in TCGA COAD and Xenium CRC. All contrasts use the (80,160] µm ring. Xenium entries give the mean and minimum–maximum range over five patient splits; these ranges are not confidence intervals.

| Encoder | TCGA COAD | Xenium CRC | |
|---|---|---|---|
| | $\delta$ | Mean $\delta$ | Five-split range |
| UNI2-H | 0.473 | 0.401 | [0.382, 0.425] |
| Phikon-v2 | 0.548 | 0.397 | [0.384, 0.422] |
| Prov-GigaPath | 0.453 | 0.352 | [0.334, 0.373] |
| ResNet50 | 0.252 | 0.144 | [0.135, 0.157] |
| ViT-B/16 | 0.232 | 0.174 | [0.160, 0.190] |
| DINOv3 B/16 | 0.282 | 0.173 | [0.166, 0.183] |

The short-ring replication uses a far (80,160] µm ring with its center at least 2560 µm from the query. Five 10/3/3 patient splits include nine distinct test patients. Logarithmic annuli are analyzed separately from the fixed-width sweep. The longest replication ring does not satisfy the far-separation criterion and is excluded from spatial interpretation.

**Cross-cancer adaptation.** Leave-one-cancer-out adaptation estimates the target basis on the held-out cancer's training tissue. At short range, encoder-specific mean differences between adapted and within-cancer contrasts range from $-0.024$ to +0.013; the encoder-family difference is $-0.003$ ($p = 0.32$). For the (1280,2560] µm ring, the family difference is +0.269 across eleven cancers ($p = 0.00098$).

### B.3 Biological scoring

**Patient-level aggregation.** Reported $R^2$ values give equal weight to patients. TCGA scores pool rows within each patient; Xenium scores first average the two sections within each patient; renal scores use each patient's single section. A negative patient-average score means that the mean ratio of prediction error to baseline error exceeds one. The fraction of patients with negative scores is a separate quantity.

Four-fold evaluation assigns two folds to training, one to validation, and one to testing, yielding held-out predictions for every patient. Xenium uses 8/4/4 patients per training/validation/test split and keeps paired tumor and normal sections together; renal splits use 12/6/6 patients. Intervals use 20,000 paired patient bootstraps conditional on the fitted predictions. Paired differences use exact sign-flip tests over the sixteen Xenium or twenty-four renal patients.

**Measured target construction.** The primary Xenium composition target contains at least fifty cells from the four classes in the (80,160] µm ring. Far centers lie at least 320 µm from the query. For class counts $n_c$, the centered-log-ratio transformation adds a pseudocount of 0.5:

$$\mathrm{clr}(n)_c = \log(n_c + 0.5) - \frac{1}{4}\sum_{c'=1}^{4} \log(n_{c'} + 0.5).$$

For the renal target, 100 µm spot spacing places the six nearest spots in this ring, with their 55 µm capture areas outside the crop. Expression is averaged after log normalization over the benchmark's fifty genes; far rings are count-matched.

**Within-section scoring.** For test section $s$, let $\bar{y}_s$ and $\bar{\hat{y}}_s$ denote the section means of the targets and predictions. The within-section score is

$$R^2_{s,\mathrm{within}} = 1 - \frac{\sum_{i\in s} \|(y_i - \bar{y}_s) - (\hat{y}_i - \bar{\hat{y}}_s)\|^2}{\sum_{i\in s} \|y_i - \bar{y}_s\|^2}.$$

Centering targets and predictions separately removes section offsets from both the prediction error and target variation. The score therefore evaluates within-section variation and equals zero for a section-constant prediction.

### B.4 Predictive distance curves and spatial covariance

SpatialDE characterizes distance-dependent covariance in measured expression. The distance curves here summarize held-out prediction relative to a matched same-slide target. Their shape depends jointly on prediction error, target variance, feature statistics, and donor construction. Accordingly, the predictive summaries in Appendix C do not estimate a covariance length scale.

### B.5 Reporting checklist

The three diagnostics require complementary evidence: absolute prediction scores, shared-target comparisons, and explicit input–target geometry. Table S4 links the reporting requirements in Section 3.5 to the alternative explanations each comparison addresses.

Table S4: Reporting requirements for spatial prediction claims. Each row links an alternative explanation to the comparison needed to assess it.

| Alternative explanation | Required reporting | Interpretation assessed |
|---|---|---|
| The comparison hides the baseline | Absolute scores for both arms against their respective training means, together with the arm difference | A positive true-minus-far contrast despite below-baseline prediction in both arms |
| The target favors one model | Shared measured targets and target reliability; embedding-target comparisons within each common space | Advantages specific to predicting an encoder's own representation |
| The input sees the target | Each input's tissue footprint and any target averaging over visible tissue; target-withheld variants for contextual models | Gains attributable to direct access to tissue contributing to the target |

## C Distance curves and exploratory summaries

Distance analyses distinguish the original bandwise cohort (T-Band) from the common-query cohort used for cross-band controls (T-Common). Fixed-count comparisons standardize ring patch counts and assess target reliability. Distance summaries, including fitted half-lives, are exploratory.

Neighborhood contrast declines with distance even after patch counts are standardized on matched queries (Figure S1). Target reliability also changes: with twelve patches per ring, within-slide split-half reliability falls from approximately 0.70 to 0.35 for UNI2-H and from 0.45 to 0.21 for ResNet50. Equal patch counts therefore preserve the decline in contrast without equalizing reliability. These curves describe predictive performance rather than a physical interaction range.

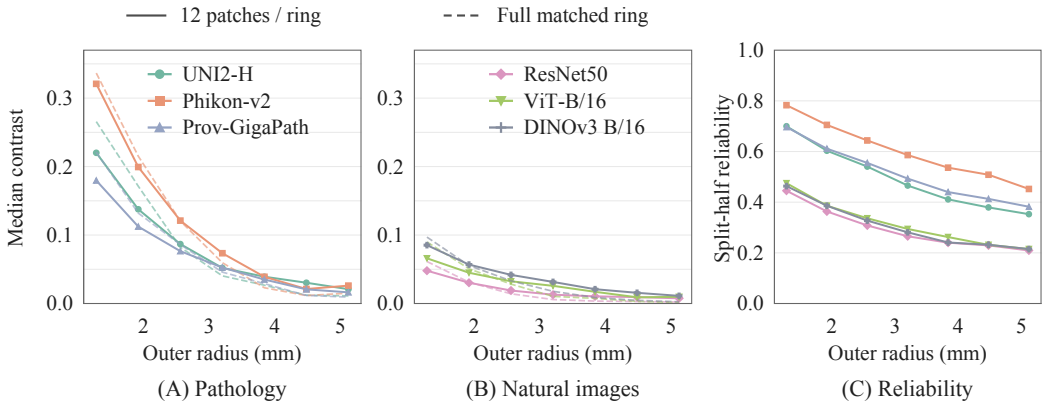



Figure S1: Distance-dependent contrast and reliability under fixed patch counts. **A–B:** Median contrasts for pathology and natural-image encoders on identical T-Common cross-band rows: solid lines use twelve patches per ring, and dashed lines use all matched patches. **C:** Median within-slide random-half reliability across cancers for twelve-patch true targets. Lines connect seven fixed-width bands; thirteen cancers contribute through 4480 µm and ten at 5120 µm. Table S1 defines the analysis sets; Table S6 reports geometric-null summaries.

The original T-Band analysis matches true/far patch counts within each band, while allowing queries, patch counts, and target reliability to vary across distance (Table S5; Figure S2). It does not impose T-Common cross-band support. The median outer radius of the last band with contrast above 0.05 is 3.20 mm for UNI2-H and Phikon-v2, 2.56 mm for Prov-GigaPath, and 1.92 mm for both natural-image ViT-B/16 models. ResNet50 has no band above this threshold in seven of thirteen cancers.

Table S5: Bandwise distance summaries for the original T-Band cohort. The seven bands comprise 546 encoder–cancer–band cells (6 encoders × 13 cancers × 7 bands), including all 78 cells from thirteen cancers at 5120 µm. The ten-cancer outermost restriction applies only to T-Common (Figure S1). Complete family separation means that, within a cancer, the lowest pathology-encoder contrast exceeds the highest natural-image-encoder contrast. The fourth column counts cells whose patient-level confidence interval lies above zero; these are intervals within cancer–encoder cells, not uncertainty across 78 cells.

| Outer radius (µm) | Mean contrast | Family separation | Patient CI above zero | Contrast above 0.05 |
|---|---|---|---|---|
| 1280 | 0.1620 | 13/13 | 78/78 | 67/78 |
| 1920 | 0.0954 | 13/13 | 76/78 | 50/78 |
| 2560 | 0.0590 | 11/13 | 69/78 | 30/78 |
| 3200 | 0.0366 | 10/13 | 59/78 | 20/78 |
| 3840 | 0.0223 | 11/13 | 60/78 | 14/78 |
| 4480 | 0.0127 | 6/13 | 47/78 | 5/78 |
| 5120 | 0.0076 | 7/13 | 47/78 | 0/78 |

A log-linear fit to the mean contrast curve yields a decay factor of 0.603 per 640 µm band, a half-life of 0.88 mm, and $R^2 = 0.999$. Individual common-support fits give encoder-wise median half-lives of 0.76–0.90 mm. The fitted curve level is higher for the pathology family in all eleven cancers with common support ($p = 0.00098$).

For common-support Theil–Sen slopes of log contrast, the pathology-minus-natural-image difference is +0.059 per mm (95% CI −0.062 to 0.162; $p = 0.21$). Restricting the fit to the first four

bands gives +0.056 per mm (−0.272 to 0.242), with Wilcoxon $p = 0.0342$. The bootstrap evaluates the mean difference, whereas Wilcoxon uses paired ranks. The slope comparison consequently depends on both the fitting window and the statistic.

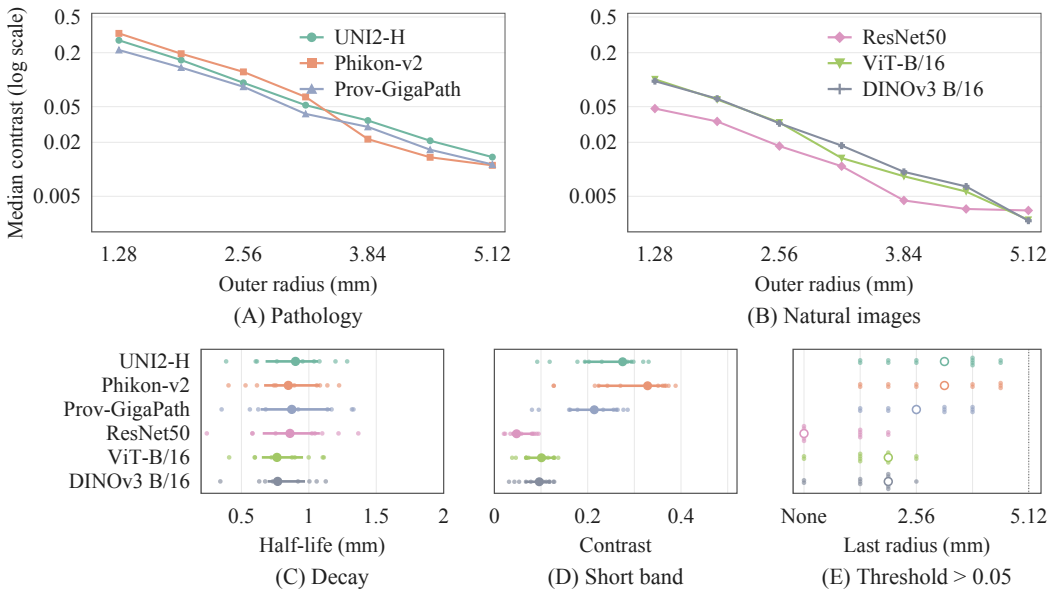



Figure S2: Exploratory distance summaries from the original T-Band analysis. **A–B:** Median contrasts across thirteen cancers for pathology and natural-image encoders, with a logarithmic vertical scale and true/far counts matched within each band. **C:** Theil–Sen contrast half-lives on common positive support (eleven cancers per encoder). **D:** Contrasts for the (640,1280] µm ring. In C–D, small points denote cancers, large filled circles denote medians, and bars denote interquartile ranges. **E:** Last band with contrast above 0.05; small points denote cancers and hollow circles denote medians. "None" indicates no band above threshold; the dotted line marks the 5.12 mm measurement limit. C–E use the same encoder order. Figure S1 compares full and fixed-count targets on the common-query cohort.

## D Prediction uncertainty and ring heterogeneity

**Prediction uncertainty.** A two-stage model first fits a mean predictor and generates residuals by cross-fitting across slides. A variance predictor is then trained on those residuals with the mean predictor held fixed. Predictive negative log-likelihood improves in 378 of the 390 cells in the five-encoder sweep, while interval coverage decreases with distance. These intervals quantify prediction-residual variation.

**Within-ring heterogeneity.** A separate probe predicts the log of within-ring embedding variance. At the first short ring, supervised ViT-B/16 attains within-section $R^2 = 0.386$, compared with 0.228–0.287 for pathology encoders. Complete encoder-family separation occurs in zero of thirteen cancers. Thus, predicting encoder-specific variance targets yields a different ordering from predicting ring means.

## E Implementation details and sensitivity analyses

### E.1 Target centering and model selection

**Training-only transformations.** Embedding targets in both arms use a common basis fitted to the training true targets. Scaling is fitted on training data, and each arm selects its penalty separately by validation error.

Ridge fitting centers targets on training rows and restores the training mean to predictions, leaving the intercept unpenalized. Validation includes the training-mean predictor as the infinite-penalty candidate. Model selection minimizes query-pooled validation MSE; reporting instead averages section-normalized scores within each patient and then across patients. Including the mean candidate therefore does not guarantee a nonnegative patient-average score on either validation or held-out data. Appendix F.5 examines this selection–scoring distinction for direct-access ridge.

Initial Xenium CLR fits used a penalized intercept. After identifying this implementation issue, we recomputed the cellular shared-target, crop-reference ridge, and input-expansion ridge fits with target centering and the mean candidate, retaining the same queries, folds, targets, and endpoints. Self-space PCA targets and MLP targets were already centered or standardized, and TCGA used a separate implementation. Renal and direct-access analyses used the corrected implementation throughout.

### E.2 Penalty and reference sensitivities

The primary shared-target ridge search spans 10 to $10^5$; the extended search spans $10^{-2}$–$10^7$. Both include the mean candidate. Extending the grid raises UNI2-H's cellular true-target score from 0.127 to 0.179, close to 0.181 on the nearby-access rows, while preserving its advantage over ResNet50. This comparison evaluates penalty range with the mean candidate included; a corrected finite-only ablation was not run.

Crop-reference models search nine finite penalties from 0.01 to $10^6$, five block weights from 1/16 to 16, and the mean candidate. Each feature block is standardized on training data and divided by the square root of its dimension. Reference-only and image-augmented models use the same search. The original-tuning sensitivity uses the narrower penalty grid without searching block weights (Table S10).

### E.3 Endpoint specification and implementation sensitivity

Primary endpoints were specified internally before scoring. For SEAL, the protocol was fixed and hashed before embedding extraction, and the hash was verified at scoring. These specifications were internal; no public preregistration was performed. Family averages, within-section scores, alternate rings, and penalty ranges are secondary analyses. Reliability normalization and ordering analyses were added after the primary results.

The intercept correction and the separate addition of the mean candidate followed initial scoring while retaining the specified endpoints. Together, they changed the Xenium target-gap difference from 0.103 ($p = 0.067$) to 0.180 ($p < 0.001$) and the gene-reference gain from 0.0296 ($p = 0.144$) to 0.0411 ($p = 0.028$).

## F Robustness and practical-input experiments

This appendix provides the reproducibility checks, complete results, and sensitivity analyses underlying the main diagnostics. Sections F.1–F.2 examine sampling and representation-space calibration; Sections F.3–F.4 report matched-target and crop-reference comparisons; Sections F.5–F.6 examine tissue access. Sections F.7–F.9 provide reliability and ordering analyses, the SEAL case study, and the published-method comparisons.

### F.1 Sampling and numerical reproducibility

TCGA reproducibility checks preserve the original patient splits, query locations, and first far donor; additional donors are sampled through independent digest-seeded streams. MESO retains the slide universe defined at feature extraction. Geometry checks reproduce all 546 T-Band cells, and numerical checks reproduce fourteen band scores from two encoder–cancer configurations. All matched-donor analyses enforce physical separation and retain the fixed patient splits.

### F.2 Calibration and representation diagnostics

**Feature statistics and encoder-family differences.** Pathology features exhibit stronger short-range autocorrelation. At 40–80 μm, median feature trace correlations are 0.58–0.66 for pathology

encoders and 0.25–0.50 for natural-image encoders. Adjusting for 80–160 µm autocorrelation reduces the encoder-family coefficient from 0.164 to 0.108, whereas adjustment for participation ratio leaves it at 0.157 (Figure S3). These descriptive associations span six encoders that differ in training data, architecture, and capacity (Appendix B.1).

The adjustments in Figure S3 use standardized feature trace correlation at 80–160 µm or standardized center participation ratio, both measured on training tissue. Within-cancer ordinary least squares fits include six encoders and an intercept; pooled fits additionally include cancer fixed effects.

Complete family separation persists at target ranks of 32, 64, and 128 principal components. At 256 and 16 components, separation holds in twelve and ten cancers, respectively.

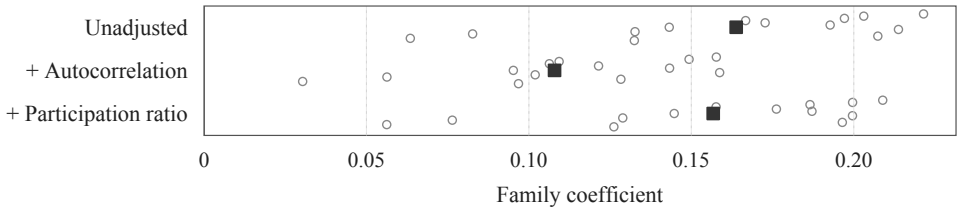



Figure S3: Encoder-family coefficients and feature statistics for true-minus-far contrast at (640,1280] µm. Squares denote pooled estimates across 78 cancer–encoder cells with cancer fixed effects; open points denote separate six-encoder fits within each cancer. Adjustments use feature autocorrelation at 80–160 µm or center participation ratio. Coefficients describe associations.

**Fixed-count rings and geometric-null calibration.** Using twelve patches per ring preserves complete family separation in all thirteen cancers at the first fixed-width band and in seven of ten supported cancers at the outermost band (Table S6). The geometric null evaluates each donor construction through pooled offsets and departures within individual encoder–cancer cells.

Table S6: Geometric-null offsets for a refitted single far target. We denote the offset by $o$. Each band includes 78 encoder–cancer cells. Offset summaries describe between-cell variation; the final column gives the number of supported cancers with complete family separation using twelve patches per ring.

| Ring (µm) | Median offset | Offset min–max | Cells $\lvert o\rvert > 0.01$ | 95% CI excludes 0 | Largest $\lvert o\rvert$ Encoder / cancer | Family separation |
|---|---|---|---|---|---|---|
| (80,160] | $-0.0004$ | $[-0.0051, +0.0095]$ | 0/78 | 11/78 | Phikon-v2 COAD | 13/13 |
| (640,1280] | $-0.0004$ | $[-0.0073, +0.0024]$ | 0/78 | 20/78 | Phikon-v2 ESCA | 13/13 |
| (1280,1920] | $+0.0000$ | $[-0.0078, +0.0037]$ | 0/78 | 16/78 | DINOv3-B PCPG | 12/13 |
| (1920,2560] | $+0.0000$ | $[-0.0019, +0.0204]$ | 1/78 | 19/78 | DINOv3-B PCPG | 12/13 |
| (2560,3200] | $+0.0002$ | $[-0.0023, +0.0148]$ | 1/78 | 25/78 | ResNet50 PCPG | 10/13 |
| (3200,3840] | $-0.0002$ | $[-0.0038, +0.0021]$ | 0/78 | 15/78 | DINOv3-B PCPG | 10/13 |
| (3840,4480] | $+0.0001$ | $[-0.0073, +0.0015]$ | 0/78 | 15/78 | DINOv3-B PCPG | 7/13 |
| (4480,5120] | $+0.0000$ | $[-0.0035, +0.0027]$ | 0/78 | 20/78 | ViT-B/16 PCPG | 7/10 |

DINOv3-B denotes DINOv3 ViT-B/16.

For each band, the geometric null comprises 78 encoder–cancer cells and 199 donor draws, with target bases refitted and the two arms tuned separately. Table S6 reports the offset range across cells

and the number of cells whose 95% offset interval excludes zero. Null comparisons use matched rows within each band. Fixed-count separation instead uses rows common across bands, with ten cancers supported at the outermost band.

Varying the donor count to two, four, or eight under the three constructions in Appendix B changes family gaps by at most $\pm 0.02$; cross-band intervals span $-0.0105$ to 0.0041. Null offsets nevertheless depend on how donors are combined. At the short ring, median offsets for averaged targets reach $-0.016$ and $-0.020$ with four and eight donors, respectively, whereas averaging separately fitted scores keeps them near $-0.001$. Offsets are wider with a frozen basis and shared penalty than with full refitting.

Across cells, the median fraction of individual null draws whose patient interval excludes zero is 12–18%, above the nominal 5%. Null-centered sensitivity analyses resample both patients and draws; their interpretation remains conditional on the null construction. Fixed-count null medians range from approximately $-0.0009$ to zero. Two cell-specific departures are larger: +0.0708 (95% CI 0.0529–0.0897) for DINOv3/PCPG at band 4 and +0.0122 for Prov-GigaPath/GBM at band 2. In the full-ring analysis, PCPG has the largest absolute null offset at six of eight bands and contributes two test patients per cell.

**Injection response and reliability.** Injection responses are monotonic in all eighteen encoder–cancer curves through (2560,3200] µm and in seventeen curves at each of the two farthest tested bands. At the short ring, an injection amplitude of 0.25 yields a median contrast of 0.058, compared with 0.339 at amplitude 1. Contrast therefore responds nonlinearly to injected variance.

Pooled reliability from random halves approaches one at large radii, whereas spatial-block and overlap-excluded halves yield lower estimates. For twelve-patch rings, within-slide reliability declines with distance. Using 24 or 48 patches produces similar family gaps where those counts are supported.

### F.3 Matched self-space and measured targets

The matched-target comparison asks whether an encoder advantage persists when all encoders predict the same measured quantity. Table S7 reports the primary UNI2-H–ResNet50 comparisons summarized in Figure 3 and adds the secondary family means. Table S8 fixes each embedding target space in turn and compares all six source encoders within that space; Figure S4 complements these contrasts with absolute true- and far-target scores.

Table S7: Encoder gaps on self-space and shared measured targets. Gaps are differences in true-minus-far contrast at (80,160] µm; the paired difference subtracts the measured-target gap from the self-space gap. Intervals are 95% paired patient bootstraps, and $p$ values use exact paired sign flips. Family means compare pathology and natural-image encoders and are secondary comparisons.

| Comparison | Self-space gap | Measured-target gap | Paired difference | $p$ |
|---|---|---|---|---|
| *Xenium CRC, cell composition* | | | | |
| UNI2-H − ResNet50 (primary) | 0.232 [0.226, 0.237] | 0.052 [−0.028, 0.127] | 0.180 [0.103, 0.262] | $< 0.001$ |
| Pathology − natural-image family mean (secondary) | 0.199 [0.193, 0.205] | 0.008 [−0.068, 0.076] | 0.191 [0.121, 0.270] | $< 0.001$ |
| *HEST ccRCC, gene expression* | | | | |
| UNI2-H − ResNet50 (primary) | 0.200 [0.184, 0.216] | 0.047 [0.025, 0.071] | 0.153 [0.125, 0.178] | $< 0.001$ |
| Pathology − natural-image family mean (secondary) | 0.208 [0.195, 0.221] | 0.039 [0.021, 0.060] | 0.169 [0.146, 0.192] | $< 0.001$ |

All six encoders use the common rows and folds in Table S1. Table S9 gives their individual self-space and measured-target contrasts, with intervals paired over sixteen Xenium patients or twenty-four renal patients. For the cross-encoder matrix (Table S8), each source encoder predicts every

Table S8: Cross-encoder prediction of ring embeddings. Rows denote input encoders and columns denote target embedding spaces. Entries are true-minus-far contrasts at (80,160] µm on the shared-target set; bold entries mark self-space prediction. GigaPath and DINOv3 columns denote Prov-GigaPath and DINOv3 B/16, respectively. The final three rows report patient-paired UNI2-H minus ResNet50 differences and the lower and upper bounds of their 95% intervals across sixteen patients. Excluding the diagonal, every target space retains the pathology-over-natural-image ordering.

| **Input encoder** | **Target embedding space** | | | | | |
|---|---|---|---|---|---|---|
| | **UNI2-H** | **Phikon-v2** | **GigaPath** | **ResNet50** | **ViT-B/16** | **DINOv3** |
| UNI2-H | **0.375** | 0.291 | 0.279 | 0.141 | 0.158 | 0.160 |
| Phikon-v2 | 0.243 | **0.370** | 0.247 | 0.137 | 0.153 | 0.152 |
| Prov-GigaPath | 0.255 | 0.264 | **0.333** | 0.145 | 0.160 | 0.164 |
| ResNet50 | 0.072 | 0.081 | 0.080 | **0.143** | 0.089 | 0.083 |
| ViT-B/16 | 0.095 | 0.110 | 0.103 | 0.104 | **0.172** | 0.111 |
| DINOv3 B/16 | 0.104 | 0.120 | 0.116 | 0.103 | 0.117 | **0.164** |
| Paired difference | 0.303 | 0.210 | 0.199 | −0.002 | 0.069 | 0.077 |
| 95% CI: lower | 0.292 | 0.195 | 0.189 | −0.010 | 0.061 | 0.071 |
| 95% CI: upper | 0.314 | 0.223 | 0.208 | 0.005 | 0.076 | 0.083 |

encoder's ring embeddings. A separate 128-component basis is fitted to the training true targets in each target space and then applied to both arms.

Table S9: Global true-minus-far contrasts at (80,160] µm under the matched self-space and measured-target protocols. Entries give estimates with 95% patient-bootstrap intervals. The specified primary test compares the UNI2-H–ResNet50 gap between target spaces; family averages are secondary.

| **Encoder** | **Self-space $\delta$** | | **Measured-target $\delta$** | |
|---|---|---|---|---|
| | **Estimate** | **95% CI** | **Estimate** | **95% CI** |
| *Xenium CRC, cell composition* | | | | |
| UNI2-H | 0.375 | [0.364, 0.387] | 0.133 | [−0.052, 0.292] |
| Phikon-v2 | 0.370 | [0.355, 0.385] | 0.079 | [−0.119, 0.247] |
| Prov-GigaPath | 0.333 | [0.320, 0.345] | 0.099 | [−0.084, 0.255] |
| ResNet50 | 0.143 | [0.134, 0.153] | 0.081 | [−0.050, 0.187] |
| ViT-B/16 | 0.172 | [0.159, 0.185] | 0.106 | [−0.016, 0.208] |
| DINOv3 B/16 | 0.164 | [0.156, 0.174] | 0.099 | [−0.030, 0.208] |
| *HEST ccRCC, gene expression* | | | | |
| UNI2-H | 0.375 | [0.358, 0.393] | 0.070 | [0.036, 0.109] |
| Phikon-v2 | 0.434 | [0.420, 0.448] | 0.069 | [0.039, 0.103] |
| Prov-GigaPath | 0.385 | [0.370, 0.399] | 0.047 | [0.024, 0.071] |
| ResNet50 | 0.176 | [0.162, 0.190] | 0.023 | [−0.000, 0.051] |
| ViT-B/16 | 0.195 | [0.177, 0.212] | 0.019 | [−0.002, 0.043] |
| DINOv3 B/16 | 0.200 | [0.187, 0.213] | 0.027 | [0.005, 0.054] |

For Xenium, the primary self-space-minus-measured gap difference is 0.180 [0.103, 0.262] ($p = 0.0005$). Its within-section counterpart is 0.066 [0.011, 0.124], and standardizing cellular coordinates yields 0.178 [0.117, 0.246]. Extending the penalty grid raises UNI2-H's true-target score to 0.179 (far-target score −0.005), while the cell-target encoder gap is 0.054 [−0.025, 0.126].

Renal rings contain medians of six spots at (80,160] µm and thirty spots at (160,320] µm. The corresponding sets retain 71,684 queries in H-Gene and 72,590 queries in the secondary wider-ring analysis. On the first ring, extending the penalty grid raises UNI2-H's true/far scores to 0.122/0.089 and ResNet50's to 0.062/0.055. The measured-target encoder gap is then 0.026 [0.015, 0.039], and

the self-space-minus-measured gap difference is 0.172 [0.155, 0.189]. Within-section measured-target true scores remain near or below zero under both grids.

At the second ring, the global gap difference is 0.101 [0.070, 0.131]. Its within-section counterpart changes from 0.008 [−0.070, 0.063] to 0.089 [0.053, 0.124] when the penalty grid is extended.

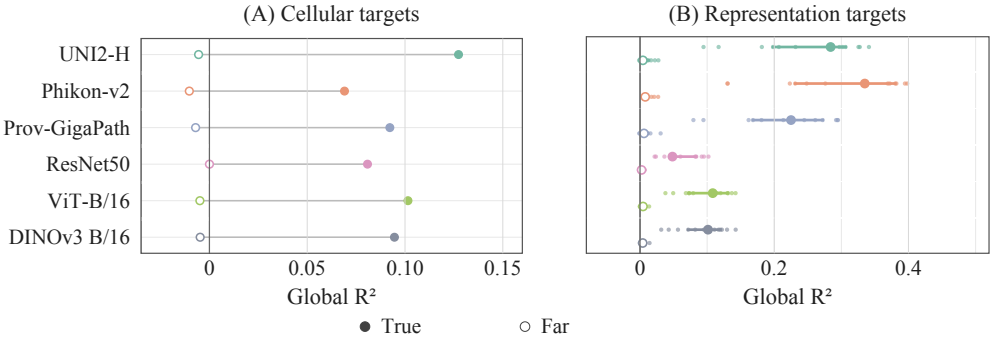



Figure S4: Absolute prediction of cellular and representation targets. **A:** Global true- and far-target scores for (80,160] µm cell composition on X-Shared (sixteen patients), shown by filled and hollow circles, respectively. **B:** Representation-space scores at (640,1280] µm on T-Band. Small points represent thirteen cancers; large filled and hollow circles mark true- and far-target medians, and bars show interquartile ranges. Colors identify encoders. Zero denotes the corresponding training-mean reference.

### F.4 Crop-reference and target sensitivities

These analyses vary the crop reference while retaining disjoint crop and target cell sets. Beyond coarse composition, the references include six-label composition, genes, their combination, and composition augmented with class-pair interactions. All references include log cell count. Table S10 distinguishes the original tuning protocol from the common tuning protocol.

Under common tuning, the gene-reference 90% interval, [0.0147, 0.0681], is not contained within the ±0.02 equivalence margin. Relative to the composition reference, the global conditional image gain changes by 0.0126 [−0.0191, 0.0455], and the within-section gain changes by −0.0091 [−0.0355, 0.0184]. The gene reference has lower accuracy at every tested training fraction; each fit uses eight training patients.

The final column of Table S10 identifies contrasts driven primarily by reduced far-arm accuracy. Only the composition reference with class-pair interactions meets this criterion. Six-label and six-label-plus-gene references yield Holm-adjusted $p = 0.018$ and $0.013$, respectively. With an MLP readout, the gene-reference result is $G = 0.048$ [−0.025, 0.160].

Changing the reference footprint to a radius of 80 µm yields first-ring gains of 0.005 [−0.009, 0.019] for composition and 0.034 [0.011, 0.058] for genes. For the second ring, the exact-crop gene-reference gain is 0.019 [−0.043, 0.098].

For a gene-profile ring target, the reference-only and image-augmented true scores are 0.5549 and 0.5636, respectively; the corresponding far scores are 0.1676 and 0.1659. The conditional image gain is 0.0104 [0.0064, 0.0146] ($p = 0.0004$).

### F.5 Input expansion, resolution, and cost

**Crop expansion and fixed-resolution tiling.** Re-extracted crops of width 80, 160, and 320 µm are resized to 224 pixels, corresponding to 0.357, 0.714, and 1.429 µm per pixel. In contrast, means of $2 \times 2$ or $4 \times 4$ stored 80 µm tile embeddings retain the original spatial sampling resolution while increasing token count. Ridge and two-layer MLP readouts receive equal tuning budgets; finite ridge penalties range from $10^{-2}$ to $10^{6}$. Inputs within each scan share the queries satisfying the coverage rules in Table S1. Tables S11 and S22 report the global and within-section results, respectively.

Figure S5 shows absolute true-target within-section $R^2$: targets and predictions are centered separately by their respective test-section means, and scores average paired sections within each patient

Table S10: Exact-crop references for first-ring four-class composition. Image increments compare each reference with the same reference augmented by image features. $G$ is the true-arm increment minus the far-arm increment. Scores are global unless marked within-section; all intervals resample patients. The final column identifies whether far-arm degradation dominates $G$.

| Reference | True increment | Far increment | Global $G$ 95% CI | Within-section $G$ 95% CI | Far-arm dominated |
|---|---|---|---|---|---|
| *Original tuning* | | | | | |
| Composition | $+0.0784$ | $+0.0063$ | $0.0722$ $[0.0325, 0.1144]$ | $0.1483$ $[0.1094, 0.1934]$ | No |
| Genes | $+0.0180$ | $-0.0058$ | $0.0237$ $[-0.0266, 0.0657]$ | $0.0521$ $[0.0311, 0.0709]$ | No |
| *Common tuning* | | | | | |
| Composition | $+0.0242$ | $-0.0043$ | $0.0285$ $[0.0032, 0.0491]$ | $0.0520$ $[0.0378, 0.0653]$ | No |
| Six labels | $+0.0378$ | $+0.0016$ | $0.0362$ $[0.0098, 0.0622]$ | $0.0660$ $[0.0373, 0.0977]$ | No |
| Genes | $+0.0302$ | $-0.0109$ | $0.0411$ $[0.0108, 0.0741]$ | $0.0429$ $[0.0254, 0.0615]$ | No |
| Six labels + genes | $+0.0314$ | $-0.0022$ | $0.0336$ $[0.0138, 0.0547]$ | $0.0643$ $[0.0370, 0.0934]$ | No |
| Composition + interactions | $+0.0013$ | $-0.0123$ | $0.0136$ $[-0.0091, 0.0334]$ | $0.0306$ $[0.0141, 0.0475]$ | Yes |

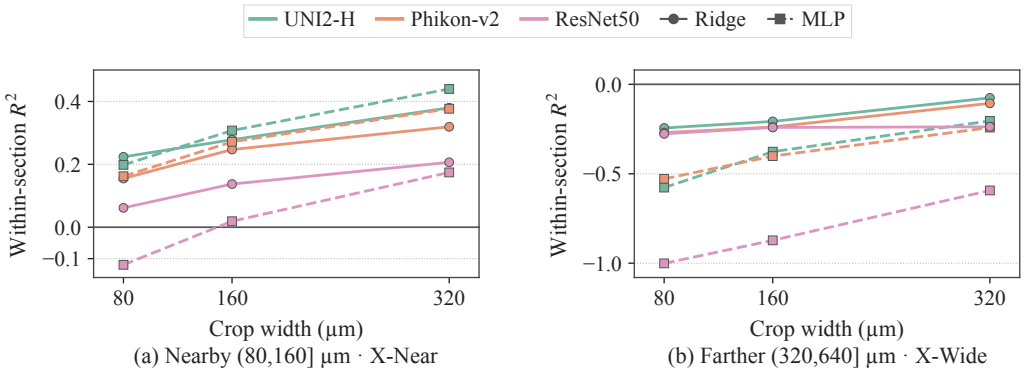



Figure S5: Within-section prediction across crop widths. True-target $R^2$ for cell composition is shown for nearby $(80, 160]$ µm rings on X-Near and farther $(320, 640]$ µm rings on X-Wide. Targets and predictions are centered separately within each test section; scores average sections within patients and then sixteen patients. Colors identify encoders; circles and solid lines denote ridge, and squares and dashed lines denote MLP. Horizontal positions are measured crop widths; the panels use separate vertical scales. Exact values are retained in Table S22.

before averaging over sixteen patients. The nearby scan uses X-Near (80,857 queries), and the farther scan uses X-Wide (98,555 queries). Figure 4(a) shows the nearby global scores; Table S11 retains both complete crop-width scans and the tiling controls. Exact within-section values are retained in Table S22.

The lowest within-section score is $-1.0009$ for ResNet50. Such large negative values can arise when low section-level target variation amplifies normalized prediction error.

For the nearby target, ridge on the 320 µm crop improves over ridge on the 80 µm crop by 0.141 [0.087, 0.196], 0.166 [0.101, 0.223], and 0.107 [0.013, 0.220] for UNI2-H, Phikon-v2, and

Table S11: Global true/far $R^2$ across crop-width and tiling scans. True and far scores occupy separate rows; column headings give input widths in μm. Nearby targets are (80,160] μm rings on X-Near (80,857 queries); farther targets are (320,640] μm rings on X-Wide (98,555 queries). Each scan matches all inputs on its own query set and retains sixteen patients. Scores average sections within patients and then patients. The farther scan requires crop-tiling eligibility; the direct-access comparison in Figure 4(b) and Table S23 instead requires ring-grid coverage.

| Encoder | Readout | Arm | Crop width (μm) | | | Tile width (μm) | |
|---|---|---|---|---|---|---|---|
| | | | 80 | 160 | 320 | 160 | 320 |
| **Nearby: (80,160] μm, X-Near** | | | | | | | |
| UNI2-H | Ridge | True | 0.181 | 0.220 | 0.321 | 0.228 | 0.244 |
| | | Far | −0.030 | −0.028 | 0.009 | −0.048 | −0.058 |
| | MLP | True | 0.195 | 0.280 | 0.425 | 0.303 | 0.411 |
| | | Far | −0.240 | −0.206 | −0.105 | −0.189 | −0.154 |
| Phikon-v2 | Ridge | True | 0.121 | 0.192 | 0.287 | 0.179 | 0.236 |
| | | Far | −0.022 | −0.031 | 0.003 | −0.027 | −0.030 |
| | MLP | True | 0.158 | 0.251 | 0.386 | 0.268 | 0.381 |
| | | Far | −0.244 | −0.220 | −0.145 | −0.186 | −0.193 |
| ResNet50 | Ridge | True | 0.090 | 0.173 | 0.197 | 0.156 | 0.199 |
| | | Far | −0.002 | −0.003 | 0.002 | −0.014 | −0.021 |
| | MLP | True | −0.013 | 0.115 | 0.233 | 0.114 | 0.297 |
| | | Far | −0.145 | −0.185 | −0.140 | −0.159 | −0.225 |
| **Farther: (320,640] μm, X-Wide** | | | | | | | |
| UNI2-H | Ridge | True | −0.165 | −0.157 | −0.038 | −0.205 | −0.221 |
| | | Far | −0.133 | −0.212 | −0.208 | −0.255 | −0.204 |
| | MLP | True | −0.208 | −0.134 | −0.006 | −0.120 | −0.049 |
| | | Far | −0.441 | −0.224 | −0.184 | −0.397 | −0.335 |
| Phikon-v2 | Ridge | True | −0.154 | −0.173 | −0.046 | −0.167 | −0.154 |
| | | Far | −0.147 | −0.164 | −0.227 | −0.241 | −0.258 |
| | MLP | True | −0.133 | −0.118 | −0.059 | −0.053 | −0.045 |
| | | Far | −0.373 | −0.284 | −0.215 | −0.258 | −0.250 |
| ResNet50 | Ridge | True | −0.028 | −0.020 | −0.062 | −0.035 | −0.053 |
| | | Far | −0.047 | −0.102 | −0.119 | −0.085 | −0.083 |
| | MLP | True | −0.264 | −0.278 | −0.114 | −0.177 | −0.109 |
| | | Far | −0.403 | −0.475 | −0.348 | −0.374 | −0.226 |

ResNet50, respectively. Relative to an MLP on the 80 μm crop, the corresponding ridge gains are 0.126 [0.020, 0.252], 0.130 [0.054, 0.225], and 0.210 [0.078, 0.410]. Ridge on the 320 μm tile mean improves over ridge on the 80 μm crop by 0.063 [−0.031, 0.142], 0.115 [0.072, 0.160], and 0.110 [0.046, 0.178], respectively.

For the farther target, ridge on the 320 μm crop improves over ridge on the 80 μm crop by 0.127 [0.070, 0.192] for UNI2-H and 0.108 [0.047, 0.173] for Phikon-v2; the ResNet50 difference is −0.034 [−0.134, 0.050]. All global true-target means in this scan remain negative. Among the 33 configurations, six global and fifteen within-section contrasts are negative.

**Direct access to the target ring.** Figure S6 compares local crops with tiles sampled directly from the farther ring on a separate matched query set.

The direct-access comparison uses X-Ring: 91,651 queries from sixteen patients, each with at least half of the ring grid present. Gains pair inputs within patient, and intervals use 20,000 patient bootstraps. Figure 4(b) displays the global true-target scores in Table S23. For UNI2-H with the 320 μm crop and an MLP readout, far-target $R^2$ is −0.222 and the true-minus-far contrast is 0.190.

Direct-access rings contain a median of 89% of grid positions. Tile centers lie in the ring, although tile footprints may extend up to 56.6 μm beyond its boundaries. Ring-tile MLP readouts achieve

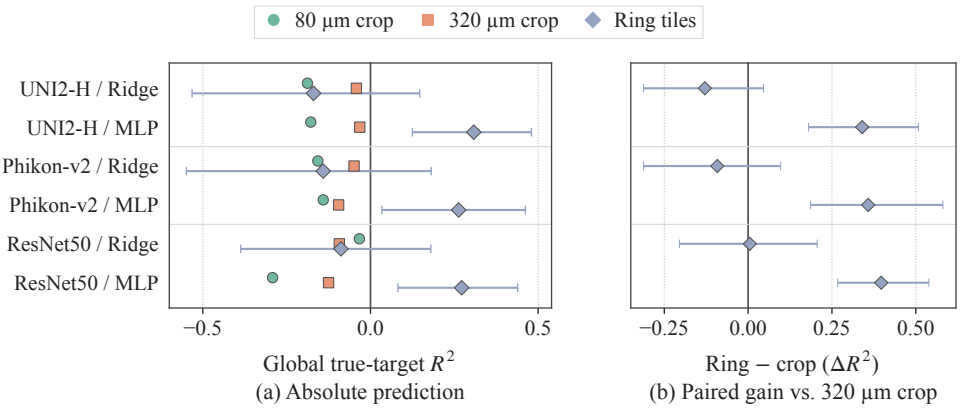



Figure S6: Prediction with direct access to the target ring. On X-Ring (91,651 queries, sixteen patients), (a) global true-target $R^2$ is compared for 80 and 320 µm crops and direct access to the $(320, 640]$ µm ring through ring tiles; (b) paired gains compare ring tiles with the matched 320 µm crop. Whiskers show the reported 95% intervals for ring-tile scores and gains, obtained from 20,000 paired patient bootstraps. Exact values remain in Table S23.

within-section true scores of 0.286, 0.279, and 0.241 for UNI2-H, Phikon-v2, and ResNet50, respectively, exceeding the matched 320 µm crop MLP scores by 0.515–0.880.

**Ridge selection and score aggregation.** Ridge readouts on ring tiles score lower in part because the validation objective and the reported patient-level score favor different penalties. For UNI2-H on X-Ring, rebuilding the ring means and refitting the four frozen folds reproduces every selected penalty and reproduces patient scores to within $2.3 \times 10^{-5}$. The selected penalties are $10^5$, $10^4$, $10^6$, and $10^4$; none selects the mean predictor. Each selected fit reduces query-pooled validation MSE by 15–36% relative to the mean candidate, yet three folds already have negative validation patient-macro $R^2$ (Table S12). The mean predictor was therefore included, but selection and reporting assigned different weights to prediction errors.

On held-out rows, the same predictions score +0.176 when residual sums and training-reference sums are pooled, compared with $-0.170$ when section scores are averaged within patients and then across patients. Eight of sixteen patients have negative ridge scores; the lowest three are $-1.864$, $-1.329$, and $-0.837$. Section-mean prediction errors account for 76% of pooled squared error, identifying a substantial offset component. These results demonstrate heterogeneous errors and sensitivity to aggregation. The pooled score uses a different weighting and does not replace the primary patient-macro score.

An exploratory re-selection on the identical candidate grid instead maximizes validation patient-macro $R^2$. It retains $10^5$ in fold 0 and selects the mean predictor in folds 1–3. The resulting held-out patient-macro score is +0.063, reducing the ridge–MLP gap from 0.478 to 0.245. Table S23 retains the primary scores from the original selection rule. This post-hoc sensitivity accounts for part of the discrepancy; it does not identify nonlinearity as the sole source of the remaining gap.

In Table S12, validation pooled $R^2$ equals one minus the selected model's MSE divided by the training-mean predictor's MSE on the same validation rows. The final column reports exploratory selection using validation patient-macro $R^2$, without access to test scores; all preceding columns describe the original selection rule. Each test fold contains four patients. This analysis covers UNI2-H alone, rather than all three direct-access encoders.

**Image access and computational cost.** Crop expansion increases nominal pixel reads sixteenfold and unique image area from 499 to 838 mm$^2$ across 123,912 centers. Resized crops still require one encoder pass per input, whereas tiling reuses shared embeddings. UNI2-H GPU times range from 445 to 516 seconds across crop widths; these timings do not constitute a controlled I/O comparison.

Table S12: UNI2-H ridge selection for direct-access inputs on X-Ring. The original rule selects penalties by query-pooled validation MSE. The final column gives an exploratory alternative selected by validation patient-macro $R^2$ on the same candidate grid. Each test fold contains four patients.

| Test fold | Original selection rule | | | | Exploratory |
|---|---|---|---|---|---|
| | Penalty | Validation pooled $R^2$ | Validation patient-macro $R^2$ | Test patient-macro $R^2$ | Penalty |
| 0 | $10^5$ | 0.303 | 0.472 | 0.251 | $10^5$ |
| 1 | $10^4$ | 0.150 | −0.197 | 0.069 | Mean predictor |
| 2 | $10^6$ | 0.359 | −0.597 | −0.481 | Mean predictor |
| 3 | $10^4$ | 0.215 | −0.104 | −0.521 | Mean predictor |

### F.6 SECTION-CONTEXT SLIDE ENCODERS

**Tile inputs and contextual tokens.** TITAN uses CONCH v1.5 features from 256 µm tiles at 448 pixels. Prov-GigaPath reads 128 µm tiles, with a center crop spanning 112 µm. Tiles must have more than half tissue coverage, except that every query-center tile is retained. Median section token counts are 342 for TITAN and 1,212 for Prov-GigaPath; the readout uses the query-center token. TITAN applies a mask to absent grid cells under its distance bias, whereas Prov-GigaPath retains fixed positional embeddings. Prov-GigaPath attention is evaluated with explicit float32 softmax and log-normalizers; equivalence to the fused CUDA kernel was not tested.

Ring withholding removes every tile overlapping the (320,640] µm annulus, retaining only tiles fully inside 320 µm or beyond 640 µm. Removing 11,168 queries whose TITAN center tile overlaps that annulus yields the common X-Context set of 87,387 queries for both models. The mask follows the true ring, so far-ring tissue remains available. Nearby targets overlap the center tile under every input condition. Consequently, the farther-ring comparison combines access to the inner neighborhood and more distant context. Table S13 reports the resulting global and within-section scores.

The first four score columns of Table S13 give global true/far $R^2$; the final two give absolute true-target within-section $R^2$. Nearby targets are (80,160] µm compositions on X-Near (80,857 queries), and farther targets are (320,640] µm compositions on X-Context (87,387 queries), as defined in Table S1. Within each target comparison, all inputs share queries, folds, far partners, and tuning budgets. Whole-section tokens observe both rings. UNI2-H crops provide practical references with different tile encoders, footprints, and resolution.

**Primary context-gain criterion.** The primary criterion requires a positive ring-withheld-minus-single-tile global MLP gain whose 95% interval excludes zero, Holm-adjusted $p < 0.05$, and a positive change in true-minus-far contrast. Neither model satisfies this joint criterion (Section 4.5). Contrast changes are −0.080 [−0.246, 0.069] for TITAN and −0.008 [−0.124, 0.098] for Prov-GigaPath. On X-Wide (98,555 queries), Prov-GigaPath's global MLP gain is 0.045 [−0.141, 0.216].

**Secondary scores and comparisons.** All four within-section ring-withheld-minus-single-tile gains are positive: TITAN gains 0.660 [0.372, 1.030] with an MLP and 0.320 [0.194, 0.461] with ridge; Prov-GigaPath gains 0.208 [0.108, 0.334] and 0.201 [0.105, 0.310], respectively. Prov-GigaPath ridge also improves globally by 0.093 [0.056, 0.135]. These gains have different implications for absolute performance. MLP scores improve from −0.730 to −0.070 for TITAN and from −0.509 to −0.301 for Prov-GigaPath, remaining below zero. TITAN ridge improves from −0.298 to 0.022, whereas Prov-GigaPath ridge improves from −0.241 to −0.040.

Relative to UNI2-H with the 320 µm crop, global MLP differences are −0.261 [−0.781, 0.108] for TITAN and −0.164 [−0.320, −0.029] for Prov-GigaPath. This comparison changes the tile encoder and sampling as well as contextual access.

For the nearby target, whole-section MLP scores are below the 320 µm UNI2-H crop score by 0.298 [0.110, 0.535] for TITAN and 0.234 [0.173, 0.297] for Prov-GigaPath. For the farther target, restoring the withheld ring tiles improves TITAN ridge by 0.051 [0.029, 0.074] and Prov-GigaPath MLP by 0.029 [0.001, 0.055].

Table S13: Section-context prediction of cell composition. Global $R^2$ is shown in separate true- and far-target rows; within-section columns report absolute true-target $R^2$ only. Dashes mark inapplicable nearby comparisons and unreported within-section far scores. Nearby (80,160] μm targets use X-Near, and farther (320,640] μm targets use X-Context. Inputs within each target comparison share queries, folds, far partners, and tuning budgets.

| Input | Arm | Nearby, global | | Farther, global | | Farther, within-section | |
|---|---|---|---|---|---|---|---|
| | | Ridge | MLP | Ridge | MLP | Ridge | MLP |
| **TITAN** | | | | | | | |
| Tile feature | True | 0.141 | 0.206 | −0.238 | −0.302 | −0.364 | −0.662 |
| | Far | 0.002 | −0.135 | −0.223 | −0.349 | — | — |
| Single tile | True | 0.185 | 0.184 | −0.151 | −0.380 | −0.298 | −0.730 |
| | Far | 0.006 | −0.170 | −0.167 | −0.568 | — | — |
| Ring withheld | True | — | — | −0.118 | −0.292 | 0.022 | −0.070 |
| | Far | — | — | −0.108 | −0.400 | — | — |
| Whole section | True | 0.151 | 0.126 | −0.067 | −0.184 | 0.084 | 0.081 |
| | Far | −0.016 | −0.171 | −0.116 | −0.371 | — | — |
| **Prov-GigaPath** | | | | | | | |
| Tile feature | True | 0.152 | 0.213 | −0.213 | −0.130 | −0.239 | −0.425 |
| | Far | 0.001 | −0.233 | −0.225 | −0.376 | — | — |
| Single tile | True | 0.167 | 0.182 | −0.110 | −0.199 | −0.241 | −0.509 |
| | Far | −0.010 | −0.181 | −0.175 | −0.250 | — | — |
| Ring withheld | True | — | — | −0.017 | −0.196 | −0.040 | −0.301 |
| | Far | — | — | −0.138 | −0.239 | — | — |
| Whole section | True | 0.110 | 0.191 | −0.141 | −0.166 | −0.071 | −0.204 |
| | Far | −0.009 | −0.163 | −0.141 | −0.263 | — | — |
| **UNI2-H** | | | | | | | |
| 80 μm crop | True | 0.181 | 0.195 | −0.152 | −0.188 | −0.230 | −0.541 |
| | Far | −0.030 | −0.240 | −0.169 | −0.534 | — | — |
| 320 μm crop | True | 0.321 | 0.425 | −0.029 | −0.032 | −0.066 | −0.261 |
| | Far | 0.009 | −0.105 | −0.203 | −0.267 | — | — |

Within TITAN, nearby within-section MLP scores are 0.179 for tile features, 0.156 for single-tile tokens, and 0.293 for whole-section tokens. The single-tile transformation improves TITAN's nearby within-section ridge score, but reduces Prov-GigaPath's nearby MLP score under both scoring conventions; the corresponding patient intervals exclude zero.

UNI2-H comparisons also change the underlying tile features: TITAN uses CONCH v1.5 over 256 μm tiles, while Prov-GigaPath uses its own 112 μm tile features. Within TITAN, nearby global MLP scores decrease from 0.206 for tile features to 0.184 for single-tile tokens and 0.126 for whole-section tokens, even as whole-section context improves within-section prediction. This pattern suggests that the usefulness of slide-level tokens depends on the local readout and scoring convention.

### F.7 Target reliability and scale-free ordering

These analyses examine whether differences between target-space gaps reflect target reliability or score scale. Reliability normalization and per-patient ordering retain the queries, folds, far partners, penalties, and target bases used in Table S7.

**Split-half reliability.** Each target is a mean over ring members: patches, cells, or spots. We partition those members into halves $A$ and $B$, containing $n_A$ and $n_B$ members. Under the exchangeable-member approximation,

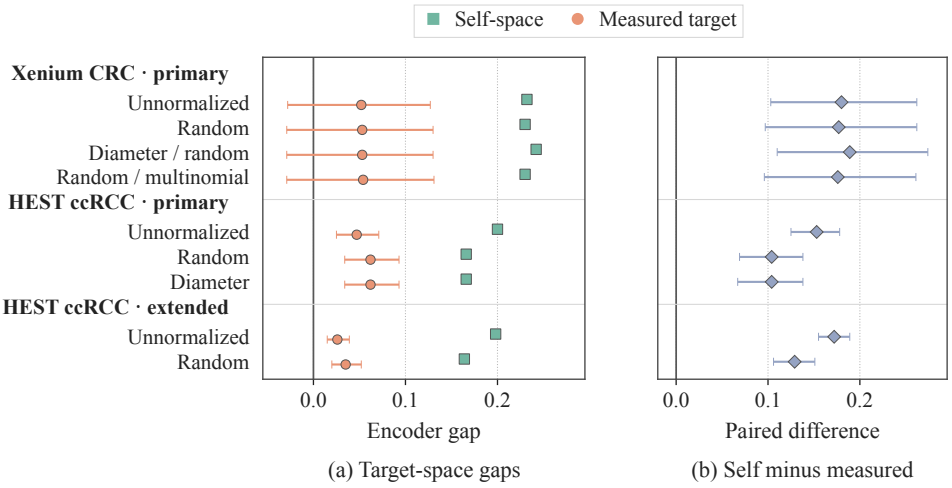
$$e_i = \left(\bar{y}_{i,A} - \bar{y}_{i,B}\right) \frac{\sqrt{n_A n_B}}{n_A + n_B}$$

estimates the sampling error of the full-ring mean. For test section $s$, we compute a reliability ceiling and normalize its score in the coordinates used for evaluation: the fold-specific 128-PC basis for self-space targets, CLR composition for cellular targets, and log-normalized expression for gene targets:

$$c_s = 1 - \frac{\sum_{i \in s} \|e_i\|^2}{\sum_{i \in s} \|y_i - \bar{y}_{\text{train}}\|^2}, \qquad \tilde{R}^2_s = \frac{R^2_s}{c_s}.$$

Normalization is applied separately to each arm and section before aggregation as in Section 3.1. For within-section ceilings, both targets and split-half error vectors are centered by section; multinomial noise sums receive the factor $1 - 1/n_s$, where $n_s$ is the section query count.

Random halves provide the reference estimate. Overlapping Xenium patches can inflate this estimate, whereas diameter halves also treat spatial heterogeneity as sampling error. Empty diameter halves in 571 rings contribute zero error. Cell splits use hypergeometric sampling of class counts; a 40-draw multinomial sensitivity uses pseudocount-smoothed proportions. Cell positions are unavailable, so the cell-sampling model does not represent spatial clustering. Renal image crops are pixel-disjoint.




(a) Target-space gaps (b) Self minus measured

Figure S7: Encoder gaps under split-half reliability normalization. (a) UNI2-H minus ResNet50 gaps on self-space and shared measured targets; (b) their paired difference, self-space minus measured target. Measured targets are cell composition for Xenium CRC and gene expression for HEST ccRCC. Primary and extended denote penalty searches. Xenium split labels separated by a slash specify self-space / cellular splits; Random applies to both. Whiskers show reported 95% paired-patient bootstrap intervals for measured-target gaps and paired differences. All comparisons use global scoring at (80,160] µm. Full scores, ceilings, and inference are in Table S24.

Table S24 reports true-arm patient-mean ceilings. Ranges span the three encoders in each family; measured-target ceilings are shared by all encoders. Far-arm ceilings differ from true-arm ceilings by less than 0.002, and neither depends on the penalty search. Encoder gaps compare UNI2-H with ResNet50; intervals use paired patient bootstraps, and $p$ values use exact paired sign flips.

Global measured-target ceilings are 0.95–0.96 for cells and 0.76–0.77 for genes. Reliability normalization preserves the positive direction of every global paired gap difference (Figure S7; Table S24). It reduces the renal difference by about one third, principally because self-space reliability differs between encoder families. With random halves, the normalized family-average difference is 0.188 [0.115, 0.271] in Xenium and 0.104 [0.074, 0.134] and 0.134 [0.111, 0.159] in renal tissue under the primary and extended grids, respectively.

Within-section normalization yields a Xenium gap difference of 0.066 [0.005, 0.130] ($p = 0.061$) with random halves and 0.115 [0.052, 0.179] with diameter halves. Renal differences with random halves are 0.027 [−0.079, 0.100] on the primary grid and 0.122 [0.086, 0.163] on the extended grid. The renal gene ceiling averages 0.395 after within-section centering, compared with 0.767 under the global training-mean reference in Table S24 (random halves). Removing section means changes the target-variance denominator and lowers the estimated ceiling; normalization consequently amplifies negative absolute within-section scores. Neither ceiling depends on the penalty search. Bootstrap intervals and sign-flip $p$ values follow different procedures and may fall on different sides of their nominal thresholds.

**Per-patient encoder ordering.** Table S14 complements score-based gaps with encoder orderings within each patient.

The win rate averages the nine pathology–natural-image encoder-pair comparisons within each patient, assigning half a win to a tie. Complete family separation requires every pathology contrast to exceed every natural-image contrast. On the primary grid, six renal patients have all-zero gene contrasts because every arm selects the mean predictor. Table S14 therefore reports both the full cohort and the subset with ordered contrasts. Reliability normalization preserves self-space ordering and changes measured-target win rates by at most 0.009.

Table S14: Per-patient encoder ordering for global (80,160] µm self-space and measured-target contrasts. Win rates average the nine cross-family encoder pairs within each patient, with ties counted as one half. Complete separation requires every pathology encoder to exceed every natural-image encoder. Count columns report patients meeting each condition.

| Target | Win rate | | Complete family separation | UNI2-H > ResNet50 |
|---|---|---|---|---|
| | Mean | 95% CI | | |
| **Xenium CRC, primary** | | | | |
| Self-space | 1.000 | — | 16/16 | 16/16 |
| Cell composition | 0.653 | [0.438, 0.854] | 9/16 | 12/16 |
| **HEST ccRCC, primary** | | | | |
| Self-space | 1.000 | — | 24/24 | 24/24 |
| Gene expression, all patients | 0.792 | [0.704, 0.875] | 11/24 | 17/24 |
| Gene expression, patients with ordered contrasts | 0.889 | [0.815, 0.957] | 11/18 | 17/18 |
| **HEST ccRCC, extended** | | | | |
| Self-space | 1.000 | — | 24/24 | 24/24 |
| Gene expression | 0.806 | [0.699, 0.898] | 12/24 | 20/24 |

Self-space-minus-measured win-rate differences are 0.347 [0.146, 0.562] in Xenium and 0.208 [0.125, 0.296] and 0.194 [0.102, 0.301] in renal tissue under the primary and extended grids, respectively. Within the common embedding target spaces of Table S8, off-diagonal win rates are 1.000 in the pathology and DINOv3 spaces, 0.958 [0.885, 1.000] in ResNet50 space, and 0.969 [0.917, 1.000] in ViT-B/16 space. Complete separation holds for 14–16 patients. These rankings remove dependence on score units while retaining uncertainty in individual patient contrasts.

### F.8 Spatially supervised case study

**Matched encoder implementation.** SEAL-UNIv2 provides low-rank adapters for UNI2-H blocks 21–23 (rank 8, scale 1). Merging the adapters into the base weights makes those weights the sole difference between the case-study arms. Both arms share the tissue mask, query grid, crops, resize, normalization, and mixed-precision settings used throughout the encoder evaluation. The merged model reproduces an unmerged-adapter forward pass to $3.2 \times 10^{-6}$, and re-extraction with the base encoder reproduces its stored embeddings exactly. These checks attribute the reported between-arm differences to the adapters.

Features are taken before the projection head, following the reference implementation. Inputs are the standard 80 μm crops resized to 224 pixels, within SEAL's training resolutions. The crop-reference comparison uses X-CropRef (also denoted X-Crop in the analysis registry), comprising 82,053 queries from sixteen patients.

**Fixed endpoints and decision rules.** The endpoints, $\pm 0.01$ equivalence margin, Holm adjustment across four primary tests, manipulation check, and outcome interpretations were specified in an internal protocol and hashed before embedding extraction. Scoring verifies that hash; the protocol was not publicly preregistered. Primary endpoints use X-CropRef with the folds, far partners, and penalty search of Appendix F.4; embedding targets use X-Shared.

Under these fixed rules, genes meet the non-specific-gain category at $\pm 0.01$. Cells show a positive neighborhood-specific gain, while the conditional-increment endpoint remains unresolved. The 90% intervals for gene and cell neighborhood-specific differences are [−0.00113, +0.00754] and [+0.01492, +0.05416], respectively; the corresponding conditional-increment intervals are [−0.00255, +0.00015] and [−0.01075, +0.00638].

The $\pm 0.01$ tolerance was fixed in the archived protocol without empirical or biological calibration. Comparisons with Table 1 and the manipulation check in Section 4.6 contextualize its scale after scoring; they do not describe how the margin was originally chosen. Applying a post-hoc $\pm 0.005$ margin to the same intervals preserves gene conditional-increment equivalence but removes gene neighborhood-specific equivalence. Neither cell endpoint meets equivalence under either margin. This sensitivity changes neither the estimates nor the primary endpoint rules.

**Manipulation check and secondary analyses.** Figure 5 reports all four primary endpoints. Predicting the crop's own target provides a manipulation check: the gene-profile difference is +0.0039 [+0.0019, +0.0058], and the cell-composition difference is −0.0014 [−0.0066, +0.0034]. These targets lie inside the input, so this check has no far arm.

Secondary analyses examine how the neighborhood-specific difference varies with the readout, score, query set, and distance. An MLP with the same tuning budget yields a positive gene interval and a negative cell point estimate whose interval spans zero. Within-section ridge differences are +0.0131 [+0.0084, +0.0178] for genes and +0.0244 [+0.0140, +0.0354] for cells. X-Shared reproduces the cell result (+0.0288 [+0.0122, +0.0484]). On the farther (320,640] μm ring, neither difference separates from zero: +0.0126 [−0.0113, +0.0347] for genes and +0.0035 [−0.0250, +0.0267] for cells.

**Embedding-space dependence.** For ring-embedding prediction, SEAL exceeds UNI2-H in SEAL's target space for all sixteen patients and falls below it in UNI2-H's space for all sixteen. This reversal is consistent with the target-space dependence in Section 4.3: the apparent advantage of the spatially supervised encoder changes with which encoder defines the prediction target.

### F.9 Published-method re-examination

**Evaluation protocol and summary.** TRIPLEX and MERGE were retrained with four patient-grouped folds. Controls, replication gates, endpoints, and decision rules were fixed and hashed before training. Xenium pseudo-spots aggregate cells into 55 μm spots on a 100 μm grid. Each model is evaluated by mean Pearson correlation over fifty genes, the metric used by both methods, and by global $R^2$ relative to the training mean, averaged over patients as in Section 3.1. Intervals use 20,000 paired patient bootstraps, and $p$ values use exact paired sign flips.

Table S15 provides the numerical summary shown in Figure 2. The subsequent tables report the full controls, interactions, and patient-level results.

**TRIPLEX: target construction and neighborhood access.** Table S16 compares the target-view-only model with full and adjacent-area-masked neighborhood inputs under smoothed and raw targets. The smoothed target averages each spot with the same immediate neighbors visible in the neighborhood view. The neighborhood-view gain replicates on these targets in both cohorts, with correlation gains of +0.062 in ccRCC and +0.075 in Xenium, compared with +0.016 and +0.011, respectively, on raw targets.

Table S15: Context effects under smoothed and raw expression targets. $\Delta r$ is the change in mean Pearson correlation over fifty genes; $\Delta R^2$ is the change in global $R^2$, with 95% patient-bootstrap intervals. Contrasts compare full with target-view-only TRIPLEX and original with permuted-coordinate MERGE. The final column reports absolute global $R^2$ for full TRIPLEX or original-graph MERGE relative to the training mean. Subsequent tables report additional variants, interactions, and patient-level counts.

| Cohort | Target | $\Delta r$ | $\Delta R^2$ | | Full $R^2$ |
|---|---|---|---|---|---|
| | | | Estimate | 95% CI | |
| **TRIPLEX** | | | | | |
| ccRCC | Smoothed | 0.062 | 0.042 | [0.017, 0.062] | 0.024 |
| | Raw | 0.016 | 0.009 | [−0.005, 0.021] | −0.054 |
| Xenium | Smoothed | 0.075 | 0.063 | [0.031, 0.092] | 0.134 |
| | Raw | 0.011 | 0.008 | [−0.008, 0.023] | 0.046 |
| **MERGE** | | | | | |
| ccRCC | SPCS | 0.044 | 0.038 | [0.014, 0.059] | 0.027 |
| | Raw | 0.004 | 0.004 | [−0.006, 0.013] | −0.061 |
| Xenium | SPCS | 0.078 | 0.065 | [0.037, 0.094] | 0.162 |
| | Raw | 0.006 | 0.004 | [−0.007, 0.016] | 0.045 |

Table S16: TRIPLEX neighborhood-view gains under smoothed and raw targets. TEM denotes the target-view-only model; "adj. masked" removes adjacent capture areas from the neighborhood view. Full and masked gains are relative to TEM. Table S19 reports intervals for the global $R^2$ gains.

| Cohort | Target | Model score | | | Gain over TEM | |
|---|---|---|---|---|---|---|
| | | TEM | Full | Adj. masked | Full | Masked |
| **Mean Pearson correlation** | | | | | | |
| ccRCC | Smoothed | 0.200 | 0.262 | 0.214 | 0.062 | 0.014 |
| | Raw | 0.180 | 0.196 | 0.184 | 0.016 | 0.004 |
| Xenium | Smoothed | 0.310 | 0.385 | 0.325 | 0.075 | 0.015 |
| | Raw | 0.284 | 0.295 | 0.287 | 0.011 | 0.003 |
| **Global $R^2$** | | | | | | |
| ccRCC | Smoothed | −0.018 | 0.024 | −0.009 | 0.042 | 0.009 |
| | Raw | −0.063 | −0.054 | −0.061 | 0.009 | 0.002 |
| Xenium | Smoothed | 0.071 | 0.134 | 0.084 | 0.063 | 0.013 |
| | Raw | 0.038 | 0.046 | 0.040 | 0.008 | 0.002 |

The target-by-view interaction is the smoothed-target gain minus the raw-target gain. In correlation, it is 0.046 [0.029, 0.064] in ccRCC and 0.064 [0.038, 0.089] in Xenium; after reliability normalization, the corresponding interactions are 0.037 [0.019, 0.056] and 0.053 [0.028, 0.078]. Masking adjacent capture areas removes 79% of the smoothed-target gain in global $R^2$ in both cohorts, consistent with the 77% reduction in correlation for ccRCC and 80% for Xenium.

The residual raw-target gain does not separate from section-level information: the remaining proximity component is +0.002 [−0.004, 0.009] (Table S21).

In Xenium, increasing the smoothing footprint increases the neighborhood-view gain in mean Pearson correlation: +0.011, +0.075, and +0.094 for targets averaging 1, 7, and 19 spots, respectively (Table S17).

**MERGE: spatial edges and target smoothing.** Table S18 compares the original graph with within-section coordinate permutation and a no-graph reference, retaining the distinction between SPCS-smoothed and raw targets. Coordinate permutation preserves every image feature. The spatial-edge component in correlation is 0.044 in ccRCC and 0.078 in Xenium on smoothed targets, compared with 0.004 and 0.006, respectively, on raw targets, whose intervals span zero.

Table S17: TRIPLEX neighborhood-view gain across target-smoothing footprints in Xenium pseudo-spots. Gains compare the full model with the target-view-only model (TEM). The first-ring target matches the smoothed condition in Table S16.

| Footprint | Spots | Reliability | Mean Pearson | | | Global $R^2$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | TEM | Full | Gain | TEM | Full | Gain |
| None (raw) | 1 | 0.52 | 0.284 | 0.295 | 0.011 | 0.038 | 0.046 | 0.008 |
| First ring | 7 | 0.71 | 0.310 | 0.385 | 0.075 | 0.071 | 0.134 | 0.063 |
| Second ring | 19 | 0.80 | 0.318 | 0.412 | 0.094 | 0.075 | 0.158 | 0.083 |

Table S18: MERGE spatial-edge effects under within-section coordinate permutation. The component is the original-graph score minus the permuted-coordinate score; entries include 95% patient-bootstrap intervals. The no-graph model provides a separate reference.

| Cohort | Target | Model score | | | Spatial-edge component | |
|---|---|---|---|---|---|---|
| | | No graph | Permuted | Original | Estimate | 95% CI |
| **Mean Pearson correlation** | | | | | | |
| ccRCC | SPCS | 0.210 | 0.224 | 0.268 | 0.044 | $[0.027, 0.062]$ |
| | Raw | 0.174 | 0.183 | 0.187 | 0.004 | $[-0.004, 0.012]$ |
| Xenium | SPCS | 0.330 | 0.352 | 0.430 | 0.078 | $[0.055, 0.103]$ |
| | Raw | 0.277 | 0.286 | 0.292 | 0.006 | $[-0.003, 0.015]$ |
| **Global $R^2$** | | | | | | |
| ccRCC | SPCS | $-0.012$ | $-0.011$ | 0.027 | 0.038 | $[0.014, 0.059]$ |
| | Raw | $-0.071$ | $-0.065$ | $-0.061$ | 0.004 | $[-0.006, 0.013]$ |
| Xenium | SPCS | 0.081 | 0.097 | 0.162 | 0.065 | $[0.037, 0.094]$ |
| | Raw | 0.031 | 0.041 | 0.045 | 0.004 | $[-0.007, 0.016]$ |

In correlation, the target-by-permutation interaction is 0.040 [0.022, 0.059] in ccRCC and 0.072 [0.045, 0.098] in Xenium.

**Global contrasts and absolute prediction.** Rescoring every model with global $R^2$ relative to the training mean preserves the dependence on smoothing (Tables S16–S20). In both cohorts, TRIPLEX's neighborhood gain and MERGE's spatial-edge component have patient intervals above zero on smoothed targets and spanning zero on raw targets; the smoothed-minus-raw interaction is positive for both methods (Table S19).

Absolute scores additionally establish whether these gains yield prediction beyond the training mean. On raw ccRCC targets, the TRIPLEX target-view-only and full models score $-0.063$ $[-0.115, -0.021]$ and $-0.054$ $[-0.106, -0.012]$, respectively, and all three MERGE variants also score below the baseline (Table S18). The full TRIPLEX model falls below the baseline in 19 of 24 patients; Table S20 distinguishes these individual-patient counts from the patient-average scores. Only full TRIPLEX and MERGE with intact spatial edges exceed the baseline in ccRCC, both on smoothed targets (Tables S16 and S18). Smoothed targets remain legitimate endpoints, but their gains characterize neighborhood-averaged expression rather than the contribution of surrounding tissue to predicting an individual spot; target construction is therefore integral to tissue access, because averaging over the input footprint incorporates measurements from tissue the model already observes.

Table S19: Global $R^2$ contrasts for TRIPLEX and MERGE. Gains and spatial-edge components are paired within patient. Interactions subtract the raw-target contrast from the smoothed-target contrast. Intervals are 95% paired patient bootstraps, and $p$ values use exact paired sign flips.

| Contrast | Cohort, target | Estimate | 95% CI | $p$ |
|---|---|---|---|---|
| **TRIPLEX** | | | | |
| Full − TEM | ccRCC, smoothed | 0.042 | $[0.017, 0.062]$ | 0.003 |
| Full − TEM | ccRCC, raw | 0.009 | $[-0.005, 0.021]$ | 0.215 |
| Full − TEM | Xenium, smoothed | 0.063 | $[0.031, 0.092]$ | $< 0.001$ |
| Full − TEM | Xenium, raw | 0.008 | $[-0.008, 0.023]$ | 0.280 |
| TRIPLEX interaction | ccRCC | 0.033 | $[0.012, 0.051]$ | 0.006 |
| TRIPLEX interaction | Xenium | 0.055 | $[0.024, 0.081]$ | $< 0.001$ |
| **MERGE** | | | | |
| Original − permuted | ccRCC, SPCS | 0.038 | $[0.014, 0.059]$ | $< 0.001$ |
| Original − permuted | ccRCC, raw | 0.004 | $[-0.006, 0.013]$ | 0.390 |
| Original − permuted | Xenium, SPCS | 0.065 | $[0.037, 0.094]$ | $< 0.001$ |
| Original − permuted | Xenium, raw | 0.004 | $[-0.007, 0.016]$ | 0.365 |
| MERGE interaction | ccRCC | 0.034 | $[0.015, 0.056]$ | 0.002 |
| MERGE interaction | Xenium | 0.061 | $[0.029, 0.091]$ | $< 0.001$ |

Table S20: TRIPLEX patients below the training-mean baseline (global $R^2 < 0$). Counts report individual patients and complement the patient-macro averages, which summarize normalized prediction error.

| Cohort | Target | TEM | Full |
|---|---|---|---|
| ccRCC | Raw | 20/24 | 19/24 |
| | Smoothed | 16/24 | 9/24 |
| Xenium | Raw | 5/16 | 4/16 |
| | Smoothed | 2/16 | 1/16 |

**Replication scope and complementary checks.** The TRIPLEX whole-slide branch does not meet the replication gate: its ccRCC gain is +0.004 $[-0.006, 0.013]$ in mean Pearson correlation. Accordingly, this branch was not decomposed (Table S21).

Under the HEST protocol (Jaume et al., 2024), UNI2-H meets the correlation criterion for 42 of 50 genes in the renal cohort. Nevertheless, 14 of 50 genes (nine of them among those 42) and 12 of 24 patients are predicted less accurately than the training mean. Encoder rankings remain largely preserved across metrics ($\tau = 0.87$).

Changing Xenium splits from section level to patient level yields a TRIPLEX mean Pearson correlation difference of $-0.007$ $[-0.018, 0.004]$, with no resolved performance change. Because paired tumor and normal sections share less tissue than serial sections, this result need not extend to serial-section cohorts. Table S21 brings these checks together with the published-method and SEAL-UNIv2 comparisons.

Table S21: Published evaluation claims under the study diagnostics. TRIPLEX and MERGE were retrained with released code and hyperparameters. Decomposition required replication of the original gain with a patient interval above zero. TRIPLEX and MERGE values are mean Pearson correlations over fifty genes unless marked as global $R^2$; intervals are patient bootstraps. Tables S16–S20 provide the global $R^2$ results.

| Original claim | Finding under the diagnostics |
|---|---|
| **HEST ccRCC ranking** *Status:* Narrowed | |
| Pearson ranking reflects representation quality<br>*Replicated:* n/a | Rankings largely agree with global $R^2$ ($\tau = 0.87$), but 14/50 genes (9 of the 42 passing HEST's correlation criterion) and 12/24 patients fall below the training mean |
| **TRIPLEX neighbor branch** *Status:* Narrowed | |
| Neighborhood view gain reflects spatial context<br>*Replicated:* Yes, both | Gain on raw targets is 74–85% smaller than on smoothed targets; remaining proximity component $+0.002$ [$-0.004$, 0.009]; on raw renal targets both views score below the training mean (global $R^2$) |
| **TRIPLEX global branch** *Status:* Not replicated | |
| Whole-slide view adds information<br>*Replicated:* No | $+0.004$ [$-0.006$, 0.013] on ccRCC; not decomposed |
| **MERGE spatial edges** *Status:* Narrowed | |
| Spatial edges improve prediction<br>*Replicated:* Yes, both | Coordinate permutation removes 0.044–0.078 on SPCS targets and 0.004–0.006 (intervals span zero) on raw targets; same pattern in global $R^2$ (0.038–0.065 vs. 0.004) |
| **SEAL-UNIv2** *Status:* Narrowed | |
| Smoothed supervision improves neighborhood prediction<br>*Replicated:* Transfers ($+0.0039$ on crop profile) | Cells $+0.033$ [0.012, 0.058]; genes equivalent at $\pm 0.01$; reverses across embedding spaces |
| **Section-level split** *Status:* No resolved change | |
| Estimates generalization to new patients<br>*Replicated:* n/a | Patient resplit: $-0.007$ [$-0.018$, 0.004] |

### F.10 Numerical records for the supplementary figures

Tables S22, S23, and S24 retain the complete reported values for Figures S5, S6, and S7, respectively. The crop-width records include all three encoders and both readouts; the direct-access records retain absolute scores and paired gains; the reliability records additionally give split-half ceilings and inference.

Table S22: Within-section true-target $R^2$ for cell composition across crop widths. Nearby and farther rings use X-Near and X-Wide, respectively. Targets and predictions are centered separately within each test section; scores average sections within patients and then sixteen patients.

| Encoder | Readout | Nearby (X-Near) | | | Farther (X-Wide) | | |
|---|---|---|---|---|---|---|---|
| | | **80** | **160** | **320** | **80** | **160** | **320** |
| UNI2-H | Ridge | 0.2238 | 0.2783 | 0.3799 | −0.2440 | −0.2074 | −0.0757 |
| | MLP | 0.1983 | 0.3075 | 0.4398 | −0.5771 | −0.3765 | −0.2044 |
| Phikon-v2 | Ridge | 0.1550 | 0.2471 | 0.3194 | −0.2704 | −0.2385 | −0.1062 |
| | MLP | 0.1621 | 0.2716 | 0.3768 | −0.5285 | −0.4011 | −0.2414 |
| ResNet50 | Ridge | 0.0616 | 0.1373 | 0.2065 | −0.2767 | −0.2411 | −0.2374 |
| | MLP | −0.1196 | 0.0190 | 0.1743 | −1.0009 | −0.8722 | −0.5932 |

Column headings specify crop width in µm. Nearby and farther rings are (80,160] and (320,640] µm, respectively. Figure S5 plots these scores.

Table S23: Direct access to the (320,640] µm target ring on X-Ring (91,651 queries, sixteen patients). Scores are global true-target $R^2$; gains compare ring tiles with the matched 320 µm crop. Intervals use 20,000 paired patient bootstraps.

| Encoder | Readout | Crop width (µm) | | Ring tiles | Ring tiles minus 320 µm crop |
|---|---|---|---|---|---|
| | | 80 | 320 | Estimate / 95% CI | Estimate / 95% CI |
| UNI2-H | Ridge | −0.188 | −0.042 | −0.170 [−0.532, 0.147] | −0.129 [−0.312, 0.046] |
| | MLP | −0.178 | −0.032 | 0.308 [0.125, 0.480] | 0.340 [0.180, 0.508] |
| Phikon-v2 | Ridge | −0.157 | −0.049 | −0.141 [−0.549, 0.181] | −0.092 [−0.312, 0.097] |
| | MLP | −0.141 | −0.095 | 0.263 [0.034, 0.462] | 0.358 [0.186, 0.581] |
| ResNet50 | Ridge | −0.033 | −0.093 | −0.088 [−0.387, 0.180] | 0.005 [−0.205, 0.206] |
| | MLP | −0.292 | −0.125 | 0.272 [0.082, 0.439] | 0.397 [0.267, 0.539] |

Figure S6 plots the absolute scores and paired gains.

Table S24: Split-half ceilings and reliability-normalized encoder gaps for global (80,160] µm prediction. Ceilings are true-arm patient means; family ranges span three encoders. Gaps compare UNI2-H with ResNet50, and the paired difference is the self-space gap minus the measured-target gap. Intervals are 95% paired patient bootstraps; $p$ values use exact paired sign flips. Figure S7 plots the encoder gaps.

| Halves | Self-space ceiling | | Measured ceiling |
|---|---|---|---|
| | Pathology | Natural-image | |
| **Xenium CRC, primary** | | | |
| None | — | — | — |
| Random | 0.942–0.949 | 0.863–0.866 | 0.964 |
| Diameter / random | 0.777–0.800 | 0.629–0.635 | 0.964 |
| Random / multinomial | 0.942–0.949 | 0.863–0.866 | 0.954 |
| **HEST ccRCC, primary** | | | |
| None | — | — | — |
| Random | 0.832–0.862 | 0.589–0.613 | 0.767 |
| Diameter | 0.814–0.843 | 0.565–0.591 | 0.763 |
| **HEST ccRCC, extended** | | | |
| None | — | — | — |
| Random | 0.832–0.862 | 0.589–0.613 | 0.767 |

Table S24 (continued)

| Halves | Self gap | Measured gap Estimate [95% CI] | Paired difference Estimate [95% CI] |
|---|---|---|---|
| **Xenium CRC, primary** | | | |
| None | 0.232 | 0.052 [−0.028, 0.127] | 0.180 [0.103, 0.262] |
| Random | 0.230 | 0.053 [−0.029, 0.130] | 0.177 [0.097, 0.262] |
| Diameter / random | 0.242 | 0.053 [−0.029, 0.130] | 0.189 [0.110, 0.274] |
| Random / multinomial | 0.230 | 0.054 [−0.029, 0.131] | 0.176 [0.096, 0.261] |
| **HEST ccRCC, primary** | | | |
| None | 0.200 | 0.047 [0.025, 0.071] | 0.153 [0.125, 0.178] |
| Random | 0.166 | 0.062 [0.034, 0.093] | 0.104 [0.069, 0.138] |
| Diameter | 0.166 | 0.062 [0.034, 0.093] | 0.104 [0.067, 0.138] |
| **HEST ccRCC, extended** | | | |
| None | 0.198 | 0.026 [0.015, 0.039] | 0.172 [0.155, 0.189] |
| Random | 0.164 | 0.035 [0.020, 0.052] | 0.129 [0.106, 0.151] |

None denotes unnormalized scores; primary unnormalized rows reproduce Table S7. For Xenium CRC, paired scheme names specify self-space / cellular splits; Random alone applies random splitting to both. Every paired difference has $p < 0.001$.

## G Comparison of H&E-to-ST papers

**Scope and selection.** We manually identified 28 related papers published or posted between 2020 and 2026. Of these, 27 concern prediction of spatial transcriptomic measurements from H&E or benchmarks of these predictors. SpaPath-Bench addresses spatial-domain identification and is listed separately, outside the expression-prediction counts. This manually assembled sample is not an exhaustive systematic review; its counts describe the reviewed papers, not field-wide prevalence. Previously unresolved entries were checked against available articles, supplements, and public code on 21 September 2026. This date records verification, not a systematic search cutoff.

Two pre-2022 foundational methods, ST-Net and HisToGene, are retained and flagged. TRIPLEX, MERGE, and SEAL, which we re-examine quantitatively, serve as calibration cases: their codes agree with Section 4.1 and Section 4.6.

**Coding.** We coded six items from each paper's main text and, where available, its supplement and released code (Table S25). *Smoothed* indicates that the training or evaluation target averages the queried spot with spatial neighbors. It excludes feature aggregation, retrieval-weighted predictions, and aggregation within a single spot's measurement support; evaluation-only smoothing is marked explicitly. *Covers* indicates that context inputs contain neighboring locations whose measurements enter the smoothed target. It is coded only for smoothed or possibly smoothed targets.

*Ctx* indicates that an ablation attributes a gain to a spatial component, including spatial supervision or geometry. A positive code need not isolate the addition of neighboring images. *Mean* records an explicit comparison with a training-mean or constant predictor; negative codes refer to the inspected reports.

*Split* records the unit or setting of expression-prediction evaluation. Multiple settings are retained, and splits from unrelated downstream tasks are not substituted. *Code* records public implementation coverage of the method or benchmark pipeline. Full coverage denotes availability of the core pipeline on inspection, not successful execution or independent reproduction. Partial coverage denotes a limited release; a negative code indicates that no public implementation was identified at verification. A question mark denotes unresolved availability or completeness. Denominators are item-specific, excluding unclear or inapplicable entries. Counts were recomputed from the table.

**Results.** Counts exclude unclear or inapplicable cells. No reviewed expression-prediction study with determinate baseline reporting ($n = 27$) reported a training-mean or constant-predictor comparison. Among studies attributing gains to spatial components through ablation ($n = 18$), three trained on smoothed targets incorporating measurements from neighboring tissue visible in their context inputs. Smoothed training or evaluation targets occur in 5 of 27 studies: TRIPLEX, MERGE, SEAL,

SEPAL, and ST-Net. ST-Net's positive code refers to its additional evaluation against neighbor-averaged measurements (He et al., 2020), not to smoothed training labels. SEAL supervises a spot-only encoder with smoothed targets. In TRIPLEX, MERGE, and SEPAL, context inputs contain neighbors contributing to the target; SEAL and ST-Net lack this neighboring input access in the coded settings. The split codes comprise 18 slide-level, 3 patient-level, 1 within-slide, 2 mixed-setting, and 1 external-evaluation entries; 2 remain unclear.

**Interpretation.** The absence of reported training-mean comparisons among papers with a determinate code supports the first checklist item for this sample. Overlap between smoothed targets and context inputs occurs in 3 of the 18 reviewed papers attributing gains to spatial components, including SEPAL, which we did not re-examine. Identifying this overlap requires target-construction details that are rarely stated alongside the ablation. Slide-level splits predominate in the sample. In our Xenium cohort, replacing section-level with patient-level splits did not resolve a change in TRIPLEX performance (Table S21); this finding may not extend to cohorts with serial sections.

Table S25: Comparison of H&E-to-spatial-transcriptomics papers (28 related papers; 27 concerning expression prediction).

| Paper | Smoothed | Covers | Ctx | Mean | Split | Code |
|---|---|---|---|---|---|---|
| *Methods* | | | | | | |
| TRIPLEX (Chung et al., 2024) | ✓ | ✓ | ✓ | ✗ | P | ✓ |
| MERGE (Ganguly et al., 2025) | ✓ | ✓ | ✓ | ✗ | S | ✓ |
| SEAL[‡] (Hemker et al., 2026) | ✓ | ✗ | ✓ | ✗ | S | ✓ |
| SEPAL (Mejia et al., 2023) | ✓ | ✓ | ✓ | ✗ | S | ✓ |
| ST-Net[†] (He et al., 2020) | ✓[e] | ✗ | ✗ | ✗ | P | ✓ |
| HisToGene[†] (Pang et al., 2021) | ✗ | – | ✓ | ✗ | S | ✓ |
| Hist2ST (Zeng et al., 2022) | ✗ | – | ✓ | ✗ | S | ✓ |
| BLEEP (Xie et al., 2023) | ✗ | – | ✗ | ✗ | S | ✓ |
| THItoGene (Jia et al., 2024) | ✗ | – | ✗ | ✗ | S | ✓ |
| mclSTExp (Min et al., 2024) | ✗ | – | ✓ | ✗ | S | ✓ |
| EGN/EGGN (Yang et al., 2023, 2024) | ✗ | – | ✓ | ✗ | S | ✓ |
| iStar (Zhang et al., 2024) | ✗ | – | ✗ | ✗ | S* | ✓ |
| HisToSGE (Shi et al., 2024) | ✗ | – | ✓ | ✗ | S | ✓ |
| GeneQuery (Xiong et al., 2025) | ✗ | – | ✓ | ✗ | S | ✓ |
| DeepSGE (Yuan et al., 2026) | ✗ | – | ✓ | ✗ | S | ✓ |
| FEAST (Jeong et al., 2026) | ✗ | – | ✓ | ✗ | S | ✓ |
| MSGR[‡] (Xu et al., 2026) | ✗ | – | ✗ | ✗ | S | ✓ |
| HiFusion (Weng et al., 2026) | ✗ | – | ✓ | ✗ | P/S | ○ |
| mSTAR (Xu et al., 2025) | ✗ | – | ✓ | ✗ | S | ✓ |
| STAMP (Han et al., 2026) | ✗ | – | ✓ | ✗ | U | ○ |
| DeepSpot[‡] (Nonchev et al., 2025) | ✗ | – | ✓ | ✗ | P | ✓ |
| GHIST (Fu et al., 2025) | ✗ | – | ✓ | ✗ | P/S* | ✓ |
| sCellST (Chadoutaud et al., 2026) | ✗ | – | ✗ | ✗ | S | ✓ |
| STPath (Huang et al., 2025) | ✗ | – | ✓[g] | ✗ | E | ✓ |
| OmiCLIP/Loki (Chen et al., 2025) | ✗ | – | ✗ | ✗ | U | ○ |
| *Expression-prediction benchmarks* | | | | | | |
| HEST-1k (Jaume et al., 2024) | ✗ | – | ✗ | ✗ | S | ✓ |
| HEtoSGEBench (Wang et al., 2025) | ✗ | – | ✗ | ✗ | S | ✓ |
| *Related spatial-domain benchmark (excluded from expression-prediction counts)* | | | | | | |
| SpaPath-Bench[‡] (Zhao et al., 2026) | – | – | – | – | – | ✗ |

✓ = yes, ✗ = no, ? = unclear, – = not applicable, ○ = partial. Smoothed: the training or evaluation target averages the queried spot with spatial neighbors. Covers: the context input contains neighboring locations whose measurements enter the target (coded only for smoothed or possibly smoothed targets). Ctx: a gain is attributed to a spatial component via ablation. Mean: comparison with a training-mean or constant predictor. Split: P = patient, S = slide, S* = within-slide fit or spatial-block cross-validation, E = external held-out dataset evaluation, U = unclear; combined codes retain multiple settings. Code: ✓ core pipeline available, ○ partial, ✗ no public implementation identified, ? unresolved. Implementations were not rerun for this comparison.

[†] Pre-2022 foundational method, retained as a flagged subgroup. [‡] Preprint. [e] ST-Net reports additional evaluation against smoothed measurements; this does not indicate smoothed training labels. [g] STPath

compares spatial architectures whose controls retain context. HiFusion: patient-grouped 2D evaluation and within-patient cross-section 3D evaluation. GHIST: patient-grouped HER2ST and within-section Xenium blocks. STAMP and Loki: expression-prediction fold units remain unclear. STPath's E code establishes sample exclusion from pretraining, not patient exclusion. HEtoSGEBench's Code entry refers to its benchmark analysis pipeline.

## H Analysis reconstruction and extension

**Unit of data.** Each analysis example identifies a source patient and section, query location, input footprint, true ring, and matched far partner. Reconstruction requires physical coordinates and units, ring bounds, member indices or a deterministic membership rule, target transformation, eligibility mask, and patient fold assignment. Input coverage defines comparison-specific subsets (Table S1); these subsets must be retained to reproduce paired comparisons.

**Unit of evaluation.** Model records specify the encoder checkpoint, image preprocessing and physical scale, readout, target transformation fitted only on training data, and hyperparameters selected on validation data. Score records identify the cohort, analysis set, target space, input access, fold, patient, and true/far arm. Global and within-section scores use distinct references and remain separate. The reported paired contrasts and uncertainty calculations use patient scores.

**Reconstruction.** Starting from source images and spatial measurements, construct coordinates and targets, apply eligibility rules, and assign the recorded patient splits. Fit transformations and readouts using the designated training and validation partitions, then score held-out patients. The supplementary code covers this sequence and figure/table generation.

**Extension.** Comparisons with additional encoders should retain the same queries, target values, far partners, and patient partitions. For a different input footprint, report whether it contains the target; changes in coverage require a newly declared matched subset and rescoring of the reference models. Retain all three diagnostic outputs in Section 3.5 to evaluate new representations with an explicit biological target and sampling design.

## I Additional related work

**Spatial dependence and context prediction.** SpatialDE characterizes distance-dependent covariance in gene expression (Svensson et al., 2018), while CPC, I-JEPA, and relative-position prediction learn representations through context-based objectives over spatial neighbors (Oord et al., 2018; Assran et al., 2023; Doersch et al., 2015). Our study evaluates frozen histology representations through held-out prediction from images to spatially disjoint targets, including measured tissue states, using matched same-slide far controls. This predictive evaluation is distinct from spatial covariance estimation (Appendix B.4).

**Testing context correspondence.** Background replacement tests how surrounding content affects image classification (Xiao et al., 2021). Conditional permutation methods formalize controls that preserve dependence on conditioning variables (Berrett et al., 2020). These matched comparisons assess context dependence beyond a single predictive score and inform our matched far targets and same-slide donor reassignments for testing spatial correspondence.

**Probe interpretation and measurement reliability.** Control tasks show that probe performance can reflect what the probe learns as well as information encoded in the representation (Hewitt & Liang, 2019). In neural encoding, noise-ceiling analyses relate achievable prediction to response reliability (Lage-Castellanos et al., 2019). These perspectives inform our matched controls and split-half reliability checks. Because estimated ceilings depend on sampling assumptions, we report unnormalized scores and reserve normalization for sensitivity analysis (Appendix F.7).

## J Data use and ethical considerations

This study evaluates computational predictions using existing histology images and spatial-omics measurements. The TCGA diagnostic slides and the HEST-Benchmark ccRCC subset are publicly available. The paired colorectal H&E and Xenium sections are restricted-access and are not redistributed. The analyses address methodological questions about histology representations and spatial prediction.