EVERY GENE COUNT HAS A PLATFORM AND A SLIDE: MEASUREMENT-AWARE OBSERVATION MODEL FOR H&E-TO-SPATIAL-TRANSCRIPTOMICS PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting spatial transcriptomics (ST) from H&E images is moving toward foundation models trained by pooling data across cohorts and platforms, on the assumption that a larger pool makes the model more generalizable in a new cohort. We tested this assumption with five H&E-to-ST models and found that the larger pool gave no consistent gain on the HEST benchmark or across downstream tasks. The cause, we argue, lies in the target, not the model. We first clarify that gene counts, the prediction target, are measurements rather than expression itself, and show how far this shifts the target: expression is read at a scale that differs by platform and gene, at a level that varies from slide to slide, and as counts whose noise varies by gene and platform. The standard objective, squared error on log counts, has a term for none of these, so the output must account for everything the measurement contributed. We then propose an observation model that attaches to any regressor without changing its architecture, with a term for each of the three: the per-gene scale of each platform, estimated once and held fixed; the variation of each slide, absorbed in training by an intercept of its own; and the counts themselves, modeled with a negative binomial likelihood. Under our observation model, the same five models all improve with pooling on the benchmark. The improvement carries over to three further tasks: detection of tertiary lymphoid structures, spatial clustering, and evaluation on a new platform, with a gain over the standard objective in nearly every comparison. Our code is available at https://anonymous.4open.science/r/ST-Obs/.

1 INTRODUCTION

Spatial transcriptomics (ST) (Ståhl et al., 2016) measures gene expression while preserving its spatial context, allowing tissue architecture and molecular state to be examined together. This has made it possible to study what bulk or single-cell assays cannot resolve, such as the spatial organization of the tumor microenvironment (Moncada et al., 2020). Wider adoption, however, is limited by practical constraints: ST assays are costly and low in throughput, which places them out of reach for many studies (Jain & Eadon, 2024). By contrast, slides stained with hematoxylin and eosin (H&E) are abundant. They are produced routinely in clinical practice and already exist in large archives. Predicting ST directly from an H&E image has therefore attracted growing interest. Early studies trained a predictor for one cohort, one ST platform, and one gene panel at a time (He et al., 2020; Chung et al., 2024). Recent work has moved beyond this, pooling slides across cohorts and platforms at increasing scale and building toward pan-organ and pan-cancer H&E-to-ST foundation models (Tran et al., 2026; Xiang et al., 2026).

Pooling slides at this scale rests on a simple assumption: the more, and more diverse, data a model sees, the better it should generalize to a cohort it has never seen (Huang et al., 2025b). We tested this assumption directly, training five different H&E-to-ST models on HEST-1k (Jaume et al., 2024) under two data regimes and evaluating them on the HEST benchmark and several downstream tasks. Each model was trained twice: once on slides from the same ST platform as the evaluation, and once on human slides from every ST platform in HEST-1k, so the second run should be at least as accurate if the assumption holds. However, this is not what we observed: the larger pool gave no consistent improvement on the benchmark, nor on downstream tasks such as spatial clustering and

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

the detection of tertiary lymphoid structures (TLS). Conditioning the model on the platform identity (i.e., providing it as an input) did not change this outcome. Why does a model that has seen more slides not improve as expected? We hypothesize that the cause lies not in the model but in the target.

What these models predict is not expression itself but a measurement of it. The gene counts used as targets are the output of a measurement process: tissue expression read out by a particular platform on a particular slide, and recorded as counts. Each of the three affects the target. One platform reads several times more counts than another from the same tissue, and the difference varies from gene to gene (§2.1). Within a platform, the count level of a slide varies with the conditions under which it was measured, such as sequencing depth and tissue preservation (Leek et al., 2010) (§2.2). The counts themselves are integer-valued, mostly zero, and increasingly variable as they grow. The standard objective, squared error on log counts, takes none of the three into account: it compares the model output directly with the measured value, so the output must account for everything the measurement has contributed (§2.3).

What the output must account for differs across the three. For the platform, it is a scale that differs by gene. A model trained on several platforms must therefore hold a separate scale for every gene on every platform, and giving the platform as an input does not remove this need: the input says which platform a spot came from, but the scale of each gene on that platform still has to be learned from the counts. For the slide, it is the level at which that slide was read. This level is set by the conditions of the measurement. Those conditions are not observed for a new slide, and the image reflects them only in part (§2.2). For the counts, squared error on their logarithm treats a Visium spot whose counts are mostly zero and a Xenium spot whose counts are in the tens as equally precise, although the noise level differs by platform and by gene. Much of this does not come from the tissue, yet the standard objective leaves all of it to the output.

The three differ in kind, and so should their treatment. We therefore propose an observation model that gives each a term of its own, modifying the objective rather than the input or the architecture (§3). The backbone is left as it is, and the measurement process is modeled on top of the backbone's prediction. The platform scale is a property of the platform rather than of any one slide, so it can be estimated from the training data: we fit one value per gene and platform before training and hold it fixed. The slide level is not observed for a new slide, so each training slide receives an intercept that absorbs its deviation, and the intercept is integrated out at inference. The counts are modeled with a negative binomial likelihood in place of squared error on their logarithm. Because the backbone is left unchanged, this observation model can be attached to any regressor in the same manner.

Our contributions are threefold. First, we clarify that the target of H&E-to-ST prediction is itself a measurement, and show that both the platform and the slide shift it substantially, such that the standard objective cannot disentangle this measurement variation from the underlying tissue signal (§2). Second, we propose an observation model that attaches to any regressor without changing its architecture, with a dedicated term for each property of the measurement: a per-gene scale for each platform, estimated once and held fixed; an intercept for each slide's level, absorbed during training and integrated out at inference; and a negative binomial likelihood for the counts (§3). Third, we show that under the proposed observation model, pooling raises the benchmark mean of all five H&E-to-ST models, by more than it does under MSE, and that the observation model improves their performance on three downstream evaluations (§5).

2 The Target Is a Measurement

We first measure how much two properties of the measurement, the platform and the slide, shift the target. All statistics below are computed on HEST-1k for the two platforms most commonly used to train these models, Visium and Xenium. Further analyses are given in Appendix A.

2.1 Same gene, different platform

The ST platforms in common use fall into two families. Sequencing-based platforms such as Visium capture the whole transcriptome on an array of spatially barcoded spots and read it out by sequencing. Imaging-based platforms such as Xenium hybridize probes to a predefined panel of a few hundred to a few thousand genes and image them, counting individual molecules in place. This difference carries straight through to the detection rate and the sensitivity of each gene. A study

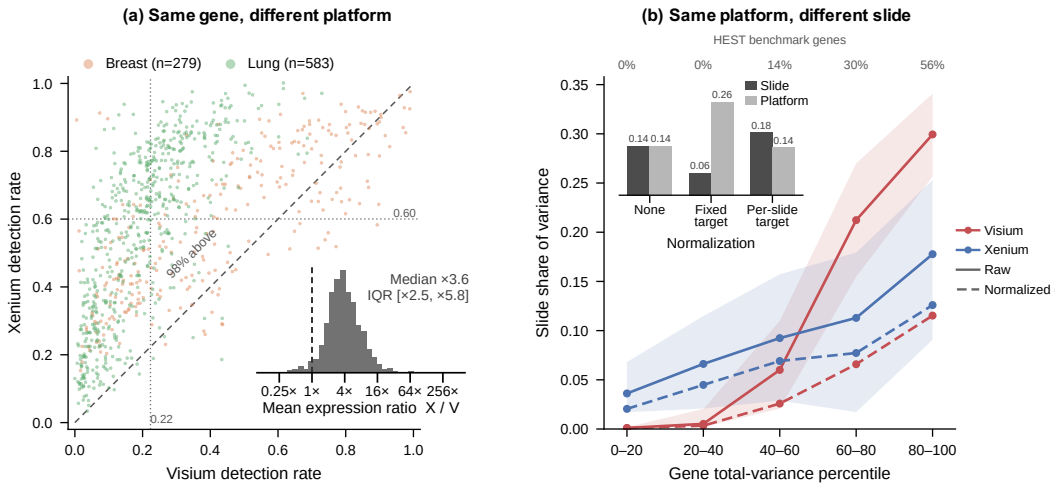

Figure 1: (a) Detection rate of the shared genes by two platforms, one point per gene and organ; the dotted lines, the 98% and the inset use the 787 genes of both organs pooled. Inset: per-gene ratio of mean $\log(1+y)$, Xenium over Visium. (b) Slide share by quintile of a gene's total variance, on raw counts (band: interquartile range over genes) and after library-size normalization to a fixed target; top row: fraction of HEST benchmark genes per Visium quintile. Inset: slide and platform share of organ-matched data, unnormalized and with fixed- or per-slide-target normalization (median total count per slide).

that read adjacent sections of the same tissue block on both platforms reported a median per-gene sensitivity $8.4\times$ higher on Xenium at matched area (Janesick et al., 2023).

To quantify this difference in our data, we take the slides from organs profiled on both platforms and aggregate the cell-level Xenium counts into $48\,\mu\mathrm{m}$ squares, so that the two platforms are compared at matched area. Over the 787 genes shared by the two platforms, the median detection rate, the fraction of spots in which a gene is read as non-zero, is 0.22 on Visium and 0.60 on Xenium, and Xenium is the higher of the two for 98% of genes (Figure 1a). This gap, moreover, is not a single factor shared by all genes. The per-gene ratio of mean $\log(1+y)$, Xenium over Visium, has a median of $3.6\times$, and its interquartile range runs from $2.5\times$ to $5.8\times$ (Figure 1a, inset). In short, a platform difference calls for a term that varies by gene.

2.2 Same platform, different slide

Within one platform and one organ, the level of the gene counts y still varies by slide. We decompose the sum of squares of the spot-level $\log(1+y)$ into platform, organ, slide, and spot. For each gene, we call the variance fraction explained by slide (or platform) identity its *slide (or platform) share*. A larger share means more of the gene's variation lies between slides (or platforms).

On both platforms, the slide share grows with the gene's variability. Dividing genes into quintiles by total variance, the top quintile on the Visium slides has a share of 0.30 (Figure 1b). Of the 430 genes the HEST benchmark selects as prediction targets, 86% fall in the top two Visium quintiles, so the genes usually chosen lie where the slide share is largest (Figure 1b, top row). On the organ-matched data, the two are comparable, with gene medians of 0.140 for the slide and 0.139 for the platform.

Nor is the share removed by library-size normalization. Dividing each spot by its total count and rescaling to a fixed target lowers the share of the top quintile on the Visium slides from 0.30 to 0.12 (Figure 1b, dashed), but on the organ-matched data it lowers the slide share from 0.140 to 0.063 and raises the platform share from 0.139 to 0.264 (inset). Rescaling to a per-slide target instead leaves the platform share almost unchanged and raises the slide share, and neither lowers both.

One quantity that changes by slide is library size, the total count of a spot, summarized per slide by its median. A slide with a larger library size reads more counts of nearly every gene, so its targets sit at a higher level throughout. Whether a model could read this condition from the image, and so

account for the slide share on its own, we test with a ridge probe on the frozen image embedding that predicts each spot's log library size, read per slide. It can, in part, and mostly within a cohort: the probe recovers the level well when other slides of the same study were seen in training (R^2 of 0.63 on Visium and 0.82 on Xenium), but less when the whole study is held out (0.27 and 0.53; Appendix A.3). For a slide from a new cohort, the level is thus not observed, and the probe recovers it only in part.

2.3 THE STANDARD OBJECTIVE HAS NO MEASUREMENT TERM

The target y is therefore a count set by the tissue and by a measurement process, and that process depends on the platform p and the slide s. The standard objective fits $f(x)$ to $\log(1+y)$ by mean squared error, and has no term for p or s. Whether or not the platform or the slide is given to the backbone as an input, the objective compares the output directly with the measured value. The output is thus left to account for what the measurement contributed, a scale that belongs to the platform and a level that belongs to the slide, neither of which the image determines; and as the preceding subsections showed, normalization removes neither. In addition, squared error on $\log(1+y)$ corresponds to a homoscedastic Gaussian likelihood: every spot, gene, and platform is fitted with the same variance, although the noise level differs by platform and by gene.

3 METHOD

3.1 THREE TERMS ON TOP OF THE BACKBONE

Consider a spot i on slide s, profiled on platform p, with image patch x_i and count y_{ig} for gene g. The backbone can be any regression model that produces one value per gene from the patch. We take that value to be the log of the tissue's expression intensity, and the observation model writes down how this intensity is observed as the count y_{ig} on platform p and slide s:

$$\log \mu_{ig} = \overbrace{\underbrace{\eta_g(x_i)}_{\substack{\text{tissue signal,}\\ \text{predicted from the patch}}}}^{\text{backbone (unchanged)}} + \overbrace{\underbrace{A_s}_{\substack{\text{area: known,}\\ \text{not learned (§3.2)}}} + \underbrace{c_{gp}}_{\substack{\text{platform scale: estimated}\\ \text{once, then fixed (§3.2)}}} + \underbrace{b_{sg}}_{\substack{\text{slide: absorbed in training,}\\ \text{integrated out at inference (§3.3)}}}}^{\text{observation model (this paper)}}$$

$$\underbrace{y_{ig} \sim \mathrm{NB}(\mu_{ig}, \theta_{gp})}_{\text{count likelihood (§3.4)}}, \qquad \underbrace{\theta_{gp} = \theta_g \, e^{\delta_p}}_{\text{dispersion: a gene factor and a platform factor (§3.4)}} \tag{1}$$

Here, $\eta_g(x_i)$ is the output of the backbone for gene g at patch x_i, the log expression intensity per reference area, and the remaining three terms are the logarithms of factors that scale μ. We address the platform and slide effects of §2 with a *platform term*, a *slide term*, and a *count likelihood*. The platform term has two parts, each the log of a factor defined relative to the reference platform, Visium[1]: for A_s, the area over which the transcripts of a target are aggregated, relative to the area of a reference spot; for c_{gp}, the amount of gene g that platform p reads per unit area. Both are set before training and held fixed. The slide term, b_{sg}, carries the slide difference: the level at which slide s was read. It is absorbed during training and integrated out at inference. The count likelihood is negative binomial, $y_{ig} \sim \mathrm{NB}(\mu_{ig}, \theta_{gp})$, with dispersion $\theta_{gp} = \theta_g e^{\delta_p}$: a gene factor θ_g and a platform factor e^{δ_p}, both estimated jointly with the backbone.

At inference, the observation model adds a single choice: the platform scale on which the prediction is read, which is the evaluation platform, or the most similar training platform if it is unseen.

3.2 THE PLATFORM TERM: ESTIMATED ONCE, THEN FIXED

Platforms differ in resolution. A Visium spot is a disc of $55\ \mu\mathrm{m}$ diameter fixed by the assay, whereas high-resolution platforms such as Xenium resolve far finer units, so the area over which their counts are aggregated into one target is a choice we make. Because counts scale in proportion to area, this area enters as the known term A_s. η is then defined as a log intensity per reference area.

[1] Visium is the reference because it has the most slides in HEST-1k and covers the whole transcriptome.

Matching the area does not remove the difference in scale. We estimate it as the platform offset c_{gp}, the log ratio of the mean count per unit area of gene g on platform p to that on the reference. The ratio is computed within strata defined by organ and by clusters of the image embedding, and averaged across strata. The estimation procedure and the resulting values of c are given in Appendix B.2.

We estimate c before training and hold it fixed, rather than learning it jointly with the backbone, because a fixed offset puts η on a known scale. If the offset were learned, the data would constrain only the sum of η and the offset, and how the per-gene scale of a platform is split between the two would be settled by the optimizer rather than by the data. Holding the offset fixed settles this in advance: η is learned on the scale of the reference platform, and the per-gene scale of each platform is supplied by c rather than learned.

3.3 The slide term: absorbed in training, integrated out at inference

The level of a slide shifts the target about as much as the platform does (§2.2), but it is not observed for a new slide. We therefore absorb it rather than predict it. Each training slide receives an intercept of its own, shrunk toward zero, that absorbs the deviation of that slide. For a new slide, the intercept is integrated out at inference (§3.5). Any slide-level effect enters through this same term.

Under the standard objective, the level of each training slide has to be reproduced by η; under Equation 1, it can be carried by b_s instead.

Giving each slide as many free parameters as there are genes would let the model memorize slides, so the intercept is kept low-rank and shrunk by a quadratic penalty. With G genes in total, the intercept vector $b_s \in \mathbb{R}^G$ of slide s is

$$b_s = U v_s, \qquad \sum_{s \in p} v_s = 0, \tag{2}$$

where $U \in \mathbb{R}^{G \times k}$ is a projection shared across slides, only $v_s \in \mathbb{R}^k$ is slide-specific, and $k \ll G$. The last condition keeps this term separate from the platform term. Every slide belongs to one platform, so the slides of a platform could shift their b together. We therefore center v_s over the slides of the same platform before use. b can then express only how slides differ within a platform, and carries no component shared by a whole platform.

3.4 A count likelihood

The target y is a count, and how widely it scatters differs from gene to gene and platform to platform. We therefore model it with a negative binomial likelihood (Zhao et al., 2022), whose dispersion $\theta_{gp} = \theta_g e^{\delta_p}$ is the product of a gene factor θ_g and a platform factor e^{δ_p} (Appendix B.3).

The likelihood also settles the detection gap between platforms without a separate mechanism for zeros. Platforms differ both in the magnitude of the values they read and in how often a gene is detected at all. Under the negative binomial, the probability of a zero, $P(y = 0) = (\theta/(\theta + \mu))^\theta$, is a function of μ, so when c lowers μ, the values become smaller and zeros become more frequent at the same time. A single scale factor thus accounts for both faces of the gap.

The dispersion does not move the mean; it sets only the weight of each observation. It decides how far the counts of gene g on platform p are trusted to move the shared output, so that spots with mostly zero counts and spots with counts in the tens are no longer treated with the same variance.

3.5 Training and inference

Training minimizes the negative log-likelihood of the negative binomial in Equation 1. At inference, the slide of an evaluation spot is new, so its intercept b_s is not observed. The natural value to put in its place is zero, the mean of the intercepts within a platform (Equation 2). In the model, however, the spread of the target across slides is carried by b and the spread within a slide by θ, so a prediction that fixed b at zero would understate the spread of counts on a new slide. We therefore integrate the intercept out: the intercept of a new slide is assumed to be distributed as the fitted intercepts of the training slides, and the prediction is averaged over that spread:

$$\hat{y}_{ig} = \mathbb{E}_{b \sim \mathcal{N}(0,\, \sigma^2_{gp})} \Big[\mathbb{E}\big[\log(1 + y) \,\big|\, \mu_{ig}\, e^{b},\, \theta_{gp} \big] \Big], \tag{3}$$

where μ_{ig} is that of Equation 1 without b_{sg}, and σ_{gp} is the standard deviation of b_{sg} over the training slides of the same platform. Appendix B.3 describes how Equation 3 is evaluated numerically.

4 Experimental Setup

This section briefly summarizes the experimental setup; further details, such as the data splits and model hyperparameters, are in Appendix C and Appendix D, respectively.

4.1 Data

We train on HEST-1k (Jaume et al., 2024), a large public collection of paired H&E and ST slides. We remove every slide used in any evaluation and hold out 32 slides for validation. This leaves 546 training slides, spanning six platforms: Spatial Transcriptomics, Visium, Visium HD, Visium HD 3′, Xenium, and Xenium Prime 5K.

The genes to predict are the union of the training slide panels after removing control probes, harmonizing gene names, and, as in STPath (Huang et al., 2025b), requiring at least 3 total counts on some slide: 38,605 genes remain. For Xenium and Xenium Prime 5K, we use the HEST-1k release, in which transcripts are aggregated into $100\,\mu\mathrm{m}$ squares; for Visium HD and Visium HD 3′, we aggregate into $112\,\mu\mathrm{m}$ squares.

4.2 Backbones

We use five backbones. The patch encoder is UNI2-h (Chen et al., 2024b), frozen except in Probe+LoRA, where LoRA adapts it; otherwise, the backbones differ only in what sits on top of it.[2] **Probe** is an MLP that sees the embedding of a single spot. **Probe+LoRA** is the same probe, with the encoder adapted by LoRA (Hu et al., 2022) during training. **Simple-TF** adds an organ and platform embedding to each spot embedding, then passes all spots through self-attention. **STPath**† (Huang et al., 2025b)[3] is an H&E-to-ST foundation model that is pretrained by masked modeling of spot expression and, once pretrained, predicts the whole transcriptome directly, without a separate fine-tuning stage. **STORM**† (Xiang et al., 2026) is a multimodal H&E-to-ST foundation model, pretrained by masked reconstruction of a spot and its neighborhood in both modalities, after which a prediction head is trained for H&E-only prediction.

4.3 Evaluation

Each backbone is evaluated in two ways, on the HEST benchmark and on downstream validation. The HEST benchmark consists of ten tasks, five on Visium and five on Xenium, each corresponding to one cancer type. For each task, the model predicts the expression of 50 designated genes. Downstream validation covers three evaluations. **TLS detection** asks whether the predicted expression locates TLS, the organized immune aggregates within a tumor, on 17 clear cell renal cell carcinoma (ccRCC) Visium slides (Meylan et al., 2022). **Spatial clustering** clusters the predicted expression of each of 27 slides drawn from three cohorts, human dorsolateral prefrontal cortex (DLPFC) (Maynard et al., 2021), breast cancer (Andersson et al., 2021), and prostate cancer (Erickson et al., 2022). **External validation** uses slides from Atera (10x Genomics), a platform absent from HEST-1k, and asks whether a model can be applied to a new platform.

4.4 Training conditions

We test two training pools with two objectives. **Expert** is trained only on slides from the same platform as the evaluation slides (313 for Visium and 48 for Xenium). **Pooled** is trained on all 546 slides across the six platforms. The two objectives are **MSE** on $\log(1 + y)$, which is the standard baseline, and the **observation model** (Obs.) of §3. Each condition is trained with five random seeds, and we report the mean and standard deviation. Training stops when the validation score stops improving; the schedule of each backbone is in Appendix D.4.

[2] Simple-TF and STPath† take a platform embedding as input; Probe, Probe+LoRA, and STORM† do not.

[3] † indicates our reimplementation, trained from scratch under our protocol.

Table 1: Two training pools and two objectives on the HEST benchmark. PCC, mean$\pm$sd over five seeds. Δ: Pooled $-$ Expert over the ten tasks, computed before rounding. Bold: best per backbone and platform.

		Visium (5 tasks)		Xenium (5 tasks)		All (10 tasks)		
Backbone	Obj.	Expert	Pooled	Expert	Pooled	Expert	Pooled	Δ
Probe	MSE	0.170±0.007	0.164±0.009	0.499±0.004	0.524±0.025	0.335±0.005	0.344±0.009	+0.009
	Obs.	0.186±0.008	**0.199**±0.005	0.496±0.004	**0.556**±0.009	0.341±0.004	**0.378**±0.007	+0.037
Probe+LoRA	MSE	0.192±0.006	0.194±0.015	0.515±0.005	**0.579**±0.008	0.353±0.003	0.386±0.010	+0.033
	Obs.	0.195±0.006	**0.215**±0.012	0.497±0.006	0.577±0.005	0.346±0.004	**0.396**±0.004	+0.050
Simple-TF	MSE	0.204±0.006	0.205±0.020	0.500±0.012	0.526±0.069	0.352±0.003	0.366±0.043	+0.014
	Obs.	**0.209**±0.006	0.208±0.011	0.501±0.018	**0.584**±0.009	0.355±0.008	**0.396**±0.008	+0.041
STPath†	MSE	0.158±0.008	0.132±0.019	0.496±0.016	0.480±0.045	0.327±0.009	0.306±0.025	−0.022
	Obs.	**0.197**±0.011	0.186±0.009	0.497±0.011	**0.549**±0.032	0.347±0.007	**0.367**±0.019	+0.021
STORM†	MSE	0.174±0.014	0.163±0.011	0.449±0.003	0.491±0.020	0.311±0.007	0.327±0.012	+0.016
	Obs.	0.182±0.010	**0.190**±0.008	0.431±0.006	**0.507**±0.003	0.306±0.007	**0.349**±0.005	+0.042

5 RESULTS

5.1 HEST BENCHMARK

Table 1 reports the Pearson correlation coefficient (PCC)[4] of the five backbones on the HEST benchmark. Under MSE, the larger pool brings no consistent gain: it raises the correlation in only 16 of the 50 task-backbone cells (Appendix E.1). On the Visium tasks, Pooled has seen 1.7 times as many slides as the Visium expert and shows no meaningful gain on any of the five backbones: it loses on three, and on the other two it moves by at most 0.002. On the Xenium tasks it has seen 11.4 times as many, and four backbones do gain, but on two of them the gain is within one standard deviation. STPath†, which already takes a platform embedding as input, loses on both platforms. More data, then, does not guarantee a better prediction under the standard objective.

The proposed observation model changes this on the same two pools. On the Visium tasks, Pooled now gains on three backbones and does not gain on the other two. On the Xenium tasks every backbone gains, each by more than under MSE. Over the ten tasks, the pooled model is ahead of the expert on all five backbones, by 6 to 14%, and the gain from pooling is larger than under MSE on all five. Part of the Xenium gain comes from supervision rather than generalization, most of it on SKCM. On SKCM, COAD, and IDC, some of the 50 genes are measured by no Xenium training slide, so only the pooled model is trained on them. Rescored on the genes the Xenium expert does supervise, pooling still gains more under the observation model than under MSE in 13 of the 15 task-backbone cells (Appendix E.1). Per-task scores are in Appendix E, results with other patch encoders in Appendix F.

5.2 DOWNSTREAM VALIDATION

Table 2 reports the three downstream validations for the pooled model under MSE and under the observation model, and for the Visium expert under MSE; further results are in Appendix G.

TLS detection. A spot is scored by the ranks of a signature gene set within its predicted expression, with two signatures, the 29-gene Meylan (Meylan et al., 2022) and the 12-gene chemokine (Xu et al., 2022). Under MSE, pooling lowers the AUROC of Probe with both signatures, leaves Probe+LoRA where the expert is, and raises the other three backbones by at most 0.055. Under the observation model, the pooled model scores above the pooled MSE model on all five backbones and with both signatures, by the most on Probe, 9% and 11%. On the chemokine signature it exceeds GT on all five backbones, plausibly because measured counts tie many signature genes at zero within a spot, whereas a prediction ranks them continuously.

[4]Every metric we report is computed within one slide and then averaged over slides with equal weight.

Table 2: Downstream validation, mean$\pm$sd over five seeds; *: the Visium expert. TLS: AUROC of the signature score, M: Meylan, C: chemokine, GT: the score on the measured expression. Clustering: AMI and homogeneity (Homo.). Atera: PCC over all genes. Bold: best per backbone and column.

		TLS AUROC (17 slides)		Clustering (27 slides)		Atera PCC (2 slides)	
Backbone	Obj.	M, GT 0.821	C, GT 0.726	AMI	Homo.	Breast	Cervix
Probe	MSE*	0.740±0.018	0.709±0.024	0.195±0.002	0.287±0.002	0.494±0.023	0.341±0.015
	MSE	0.725±0.020	0.664±0.015	0.207±0.003	0.301±0.004	0.495±0.056	0.404±0.025
	Obs.	**0.789**±0.015	**0.741**±0.009	**0.222**±0.002	**0.329**±0.004	**0.635**±0.005	**0.507**±0.004
Probe+LoRA	MSE*	0.766±0.014	0.727±0.039	0.197±0.011	0.292±0.012	0.538±0.013	0.389±0.008
	MSE	0.766±0.007	0.724±0.017	0.206±0.005	0.305±0.006	0.607±0.023	0.477±0.014
	Obs.	**0.793**±0.007	**0.748**±0.014	**0.224**±0.005	**0.335**±0.006	**0.669**±0.003	**0.529**±0.005
Simple-TF	MSE*	0.768±0.019	0.684±0.036	0.184±0.006	0.284±0.008	0.519±0.038	0.363±0.024
	MSE	0.779±0.016	0.719±0.029	0.197±0.037	0.319±0.036	0.612±0.014	0.474±0.028
	Obs.	**0.798**±0.010	**0.757**±0.014	**0.224**±0.020	**0.347**±0.030	**0.693**±0.005	**0.526**±0.008
STPath†	MSE*	0.715±0.118	0.696±0.148	**0.261**±0.023	0.385±0.024	0.434±0.062	0.274±0.045
	MSE	0.770±0.023	0.732±0.030	0.260±0.023	**0.411**±0.022	0.594±0.018	0.426±0.026
	Obs.	**0.807**±0.017	**0.736**±0.027	**0.261**±0.011	0.402±0.015	**0.667**±0.013	**0.487**±0.011
STORM†	MSE*	0.752±0.038	0.694±0.031	0.277±0.003	0.438±0.018	0.438±0.039	0.277±0.027
	MSE	0.756±0.018	0.707±0.031	**0.283**±0.006	0.471±0.020	0.548±0.017	0.399±0.047
	Obs.	**0.799**±0.016	**0.753**±0.016	**0.283**±0.003	**0.474**±0.005	**0.626**±0.009	**0.452**±0.002

Spatial clustering. The predicted expression of a slide is reduced by PCA, clustered by Leiden, and compared by adjusted mutual information (AMI) and homogeneity with its annotation: cortical layers and white matter for the 12 DLPFC sections, pathological regions for the 8 breast and 7 prostate cancer sections. Against pooled MSE, the observation model raises AMI, and homogeneity with it, on three of the five backbones, by 7% or more and on 23 of the 27 slides for each; on STPath† and STORM† both metrics stay within one standard deviation of MSE. Under MSE, pooling raises most of the 27-slide means, but its sign depends on the cohort (Appendix G.2).

External validation. Two slides of breast and cervical cancer (10x Genomics, 2026a;b) were measured on Atera, which images transcripts in situ as Xenium does but over the whole transcriptome; the metric is the PCC over all 18,023 genes. The expert is the Visium one, since a Xenium expert predicts only its own panel; a Visium expert covers the genes but not the imaging readout, and only pooling has seen both. This is accordingly the validation on which pooling helps most under MSE: on the breast sample, for example, the pooled model is above the Visium expert by 13 to 37% on four of the five backbones and equal to it on Probe. The observation model improves on the pooled MSE model further, by 10 to 28% on all five backbones and both slides (Appendix G.3).

5.3 Ablation study

Table 3 removes from the pooled models each group of terms of Equation 1: the platform group, A_s and c_{gp}, together with δ_p, and the slide term, b_{sg}; NB has neither and differs from MSE only in the likelihood, and MSE (CPM) normalizes the target by library size first, the usual preprocessing.

Normalizing the target does not stand in for the terms. Over the ten tasks, MSE (CPM) is below MSE on four of the five backbones, and on the Xenium tasks it is below on every backbone, by 0.019 to 0.098. This is the outcome §2 anticipated: normalization removes neither the platform nor the slide from the target. The likelihood alone is not the source either: task by task, NB is above MSE on only 32 of the 50 task-backbone cells, against 44 for the full model (Appendix H). The two groups of terms are what raise the mean: added to NB, the platform terms raise the ten-task mean on all five backbones, and so does the slide term. Once one group is on, the other still adds on four of the five backbones, the two exceptions within one standard deviation. The two groups thus overlap in part, and the full model is the best row on three backbones and within one standard deviation of the best on the other two. Further results are in Appendix H.

Table 3: Ablation study in pooled training. PCC over the ten benchmark tasks, mean$\pm$sd over five seeds. Platform group: A_s, c_{gp}, δ_p; slide term: b_{sg}. CPM: the target is normalized by library size to 10^6. Bold: best per column. ✓: included, –: removed, ·: not included.

	Group		All (10 tasks)				
Model	Platform	Slide	Probe	Probe+LoRA	Simple-TF	STPath†	STORM†
MSE	·	·	0.344±0.009	0.386±0.010	0.366±0.043	0.306±0.025	0.327±0.012
MSE (CPM)	·	·	0.325±0.010	0.372±0.003	0.309±0.034	0.317±0.028	0.286±0.004
NB	–	–	0.358±0.006	0.391±0.005	0.338±0.038	0.335±0.027	0.328±0.006
w/o platform	–	✓	0.372±0.004	**0.398**±0.002	0.385±0.016	0.344±0.025	0.339±0.006
w/o slide	✓	–	0.372±0.006	0.395±0.006	0.378±0.019	**0.375**±0.009	0.338±0.006
Obs.	✓	✓	**0.378**±0.007	0.396±0.004	**0.396**±0.008	0.367±0.019	**0.349**±0.005

6 Related Work

H&E-to-ST foundation models. Early predictors of ST from H&E, from ST-Net (He et al., 2020) on, were each trained on one cohort measured on one platform, with a fixed panel of a few hundred genes (Appendix I.1). Recent work aims instead at one model that serves many cohorts at once. STPath (Huang et al., 2025b) is pretrained on about a thousand paired slides, with the platform given as an input token, and predicts a new dataset's expression without fine-tuning. STORM (Xiang et al., 2026) learns joint histology-expression representations before specializing to H&E-only prediction. Phoenix (Tran et al., 2026) sets Visium aside, citing its lower sensitivity and sparser transcripts, and trains on 22.2 million Xenium cell pairs, predicting at single-cell resolution across cancers. In these models the remedy sits on the input side or in the choice of data, and the objective is left as it is: the target is fitted as it was measured, with no term for the platform it was read on and none for the slide it came from.

Count likelihoods and batch correction. Count likelihoods have been used in this task before: Hist2ST (Zeng et al., 2022) fits a zero-inflated negative binomial and STimage (Tan et al., 2026) a negative binomial per spot, but both model how the counts scatter, not the platform or the slide they were read on. XFuse (Bergenstråhle et al., 2022) comes closest, with covariates per sample that shift the rate of its negative binomial gene by gene, but it is fitted to the sections of one study. Outside histology, batch effects are usually removed from the expression itself, as in ComBat (Johnson et al., 2007) and scVI (Lopez et al., 2018). A pooled collection resists this for two reasons: panels differ between platforms, and on Xenium between organs, so few genes are measured in every batch; and each platform was applied to different organs, so correcting for the platform removes biology along with it (Appendix I.3). We instead keep the target as it was measured and model the measurement in the objective: a count likelihood, as in Hist2ST and STimage, together with a per-gene platform scale fixed across the pooled platforms and a slide intercept integrated out for a new slide. Further discussion is in Appendix I.

7 Conclusion

We placed the reason pooling does not pay off in H&E-to-ST prediction in the target: gene counts are measurements, read at a per-gene scale that differs by platform, at a level that differs from slide to slide, and as counts whose spread grows with their mean. We proposed an observation model that gives each of the three a place outside the backbone and attaches to any regressor. With it, all five backbones improve with pooling on the HEST benchmark, and the gain over MSE carries over to the three downstream tasks, on every backbone for TLS detection and the new platform and on three for clustering, providing a basis for training H&E-to-ST foundation models on pooled data.

Limitations. The observation model replaces a regression loss, so generative models (Zhu et al., 2025) need a different formulation. The platform terms are estimated from HEST-1k, so they may not hold for the same platform measured elsewhere. The slide intercept absorbs whatever differs between slides, tissue composition included, so part of what it takes from the backbone may be biology. Sequencing-based platforms make up 90% of the pool, so how the terms behave with a different composition is unclear. Our external validation covers one new platform with two slides.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the observation model as a conceptual framework, to discuss the design of the experiments and give feedback on it, to implement the method and the experiment code, to write the code that preprocesses the datasets and converts them between formats, to interpret results, and to assist with translation. We have not used generative AI tools to propose or refine hypotheses. Generating synthetic data, formulating and proving mathematical claims, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft parts of this paper and edit it for readability, to write and edit other code, such as the scripts that generate the tables, to create and modify figures, and to identify and summarize related literature.

We have reviewed all AI-assisted work. The code, including the implementation of the method and the data processing, was reviewed by the authors. The authors revised every passage drafted with AI assistance, checked each cited work against its original publication, and checked each translated passage against its source. The observation model and the interpretations of the results were adopted only as far as the experiments reported in this paper support them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly released data, and no human tissue was collected for it. The paired H&E and spatial transcriptomics slides come from HEST-1k (Jaume et al., 2024), the annotations used for TLS detection and spatial clustering from STimage-1K4M (Chen et al., 2024a), and the two external samples from public releases by 10x Genomics (10x Genomics, 2026a;b). The predicted expression is a research output that has not been validated for clinical use, and it should not stand in for measured expression or for a pathologist's assessment in decisions about patients. The metadata we use also record no patient demographics, so we could not test whether the error of the model differs across groups of patients.

REPRODUCIBILITY STATEMENT

The anonymous repository linked at the end of the abstract holds the code that prepares the public data, trains the models and evaluates them, along with the configuration files that fix every split. Appendix A describes how the measurement statistics of §2 were computed. The observation model is specified in §3, and Appendix B gives the details of each of its terms. Appendix C identifies the data: the version of HEST-1k, the training pool and how it is separated from the evaluation slides, the benchmark tasks, and the downstream cohorts, all of them publicly available. Appendix D gives the preprocessing of the targets, the gene sets (Table 12), the hyperparameters of each backbone (Table 13), and the training protocol, including the seeds and the hardware. Every trained condition is run with five seeds, and Appendices E to H report the results per task, per backbone and per patch encoder.

REFERENCES

10x Genomics. Preview Data: Atera In Situ Gene Expression, FFPE Human Breast Cancer. https://www.10xgenomics.com/datasets/atera-wta-ffpe-human-breast-cancer, 2026a. Public dataset.

10x Genomics. Preview Data: Atera In Situ Gene Expression, FFPE Human Cervical Cancer. https://www.10xgenomics.com/datasets/atera-wta-ffpe-human-cervical-cancer, 2026b. Public dataset.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):6012, 2021.

Ludvig Bergenstråhle, Bryan He, Joseph Bergenstråhle, Xesús Abalo, Reza Mirzazadeh, Kim Thrane, Andrew L Ji, Alma Andersson, Ludvig Larsson, Nathalie Stakenborg, Guy Boeckxs-

taens, Paul Khavari, James Zou, Joakim Lundeberg, and Jonas Maaskola. Super-resolved spatial transcriptomics by deep data fusion. *Nature Biotechnology*, 40(4):476–479, 2022.

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

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. STimage-1K4M: A histopathology image-gene expression dataset for spatial transcriptomics. *Advances in Neural Information Processing Systems*, 37:35796–35823, 2024a.

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, 2024b.

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.

Andrew Erickson, Mengxiao He, Emelie Berglund, Maja Marklund, Reza Mirzazadeh, Niklas Schultz, Linda Kvastad, Alma Andersson, Ludvig Bergenstråhle, Joseph Bergenstråhle, et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*, 608 (7922):360–367, 2022.

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

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.

Gene H Golub and John H Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969.

Tiantian Guo, Zhiyuan Yuan, Yan Pan, Jiakang Wang, Fengling Chen, Michael Q Zhang, and Xiangyu Li. SPIRAL: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. *Genome Biology*, 24(1):241, 2023.

Ellie Haber, Ajinkya Deshpande, Jian Ma, and Spencer Krieger. Unified integration of spatial transcriptomics across platforms with LLOKI. *Genome Research*, 35(12):2722–2733, 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 *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=uVXO6gzVzj`.

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.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Susu Hu and Stefanie Speidel. MoLF: Mixture-of-latent-flow for pan-cancer spatial gene expression prediction from histology. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SNRYGh6n69.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=Ossg1IbHDT.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Rex Ying, and Wengong Jin. STPath: a generative foundation model for integrating spatial transcriptomics and whole-slide images. *NPJ Digital Medicine*, 8(1):659, 2025b.

Sanjay Jain and Michael T Eadon. Spatial transcriptomics in health and disease. *Nature Reviews Nephrology*, 20(10):659–671, 2024.

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

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.

Taejin Jeong, Joohyeok Kim, Jinyeong Kim, Chanyoung Kim, and Seong Jae Hwang. FEAST: Fully connected expressive attention for spatial transcriptomics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26793–26802, June 2026.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Joohyeok Kim, Taejin Jeong, Jinyeong Kim, and Seong Jae Hwang. Contrastive and adaptive multimodal masked autoencoder for spatial transcriptomics. *arXiv preprint arXiv:2606.21156*, 2026.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pp. 963–974, 1982.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10): 733–739, 2010.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021.

Maxime Meylan, Florent Petitprez, Etienne Becht, Antoine Bougoüin, Guilhem Pupier, Anne Calvez, Ilenia Giglioli, Virginie Verkarre, Guillaume Lacroix, Johanna Verneau, et al. Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. *Immunity*, 55(3):527–541, 2022.

Wenwen Min, Zhiceng Shi, Jun Zhang, Jun Wan, and Changmiao Wang. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551, 2024.

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

Yi Niu, Jiashuai Liu, Yingkang Zhan, Jiangbo Shi, Di Zhang, Marika Reinius, Ines Machado, Mireia Crispin-Ortuzar, Jialun Wu, Chen Li, et al. PH2ST: Prompt-guided hypergraph learning for spatial transcriptomics prediction in whole slide images. *Medical Image Analysis*, pp. 104008, 2026.

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

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

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

Hang Shi, Changxi Chi, Peng Wan, Daoqiang Zhang, and Wei Shao. Multi-modal topology-embedded graph learning for spatially resolved genes prediction from pathology images with prior gene similarity information. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20810–20819. IEEE, 2025.

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

Xiao Tan, Onkar Mulay, Jacky Xie, Samual MacDonald, Taehyun Kim, Chenhao Zhou, Zherui Xiong, Samuel X. Tan, Nan Ye, Amy McCart Reed, Kiarash Khosrotehrani, Fred Roosta, Maciej Trzaskowski, Peter T. Simpson, and Quan Nguyen. Robust and interpretable prediction of gene markers and cell types from spatial transcriptomics data. *Nature Communications*, 17(1):1781, 2026.

Manuel Tran, Rushin H Gindra, Philipp Putze, Kang Senbai, Giovanni Palla, Tina Kos, Chiara Falcomatà, Chen Wang, Ruifeng Guo, Melanie Boxberg, et al. Pan-cancer virtual spatial transcriptomics from routine histology with Phoenix. *bioRxiv*, pp. 2026–04, 2026.

Hongyi Wang, Xiuju Du, Jing Liu, Shuyi Ouyang, Yen-Wei Chen, and Lanfen Lin. M2OST: Many-to-one regression for predicting spatial transcriptomics from digital pathology images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7709–7717, 2025.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Shaocheng Wu, Mohita Mahajan, Maarten van der Linde, ChunFang Zhu, Robert B West, David van IJzendoorn, and Magdalena Matusiak. Resolving the immune response across clonal cancer evolution in situ with Atera whole transcriptome profiling. *bioRxiv*, pp. 2026–08, 2026.

Jinxi Xiang, Siyu Hou, Yuchen Li, Ryan Quinton, Xiaoming Zhang, Feyisope Eweje, Xiangde Luo, Yijiang Chen, Zhe Li, Colin Bergstrom, et al. A multimodal foundation model of spatial transcriptomics and histology for biological discovery and clinical prediction. *arXiv preprint arXiv:2604.03630*, 2026.

Xinyu Xiao, Yunfei Li, Zheni Zeng, Yukun Yan, Zhiyuan Liu, Zhenghao Liu, Yujia Xiang, Zhenbang Ye, Jianming Ying, Yang Li, et al. SciCore-Omics: a tri-modal foundation model unifying histology, spatial transcriptomics and language for spatial biology. *bioRxiv*, pp. 2026–05, 2026.

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

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.

Wenhao Xu, Chunguang Ma, Wangrui Liu, Aihetaimujiang Anwaier, Xi Tian, Guohai Shi, Yuanyuan Qu, Shiyin Wei, Hailiang Zhang, and Dingwei Ye. Prognostic value, DNA variation and immunologic features of a tertiary lymphoid structure-related chemokine signature in clear cell renal cell carcinoma. *Cancer Immunology, Immunotherapy*, 71(8):1923–1935, 2022.

Fang Yan, Jianfeng Wu, Jiawen Li, Wei Wang, Yirong Chen, Linda Wei, Jiaxuan Lu, Wen Chen, Zizhao Gao, Jianan Li, et al. PathOrchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. *NPJ Digital Medicine*, 8(1):695, 2025.

Changchun Yang, Haoyang Li, Yushuai Wu, Yilan Zhang, Yifeng Jiao, Yu Zhang, Rihan Huang, Yuan Cheng, Yuan Qi, Xin Guo, et al. PAST: A multimodal single-cell foundation model for histopathology and spatial transcriptomics in cancer. *arXiv preprint arXiv:2507.06418*, 2025.

Yuhang Yang, Yonggan Bu, Shengyuan Zhou, Yiming Luo, and Kai Zhang. DriftST: One-step generative inference of spatial transcriptomics from H&E histology. *arXiv preprint arXiv:2607.04740*, 2026.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.

Andrew Zhang, Lijia Yu, Beilei Bian, Yue Cao, Shuchang Ye, Ethan Han, Harry Robertson, Yuhao Dong, Yixiao Mao, Boxiang Liu, et al. Generalisable tissue-wide molecular reconstruction from histology. *bioRxiv*, pp. 2026–06, 2026.

Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078, 2020.

Peiyao Zhao, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biology*, 23(1):118, 2022.

Xiang Zhou, Kangning Dong, and Shihua Zhang. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nature Computational Science*, 3 (10):894–906, 2023.

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.

Appendix: Table of Contents

A Measurement Statistics **15**
A.1 Variance decomposition . . . 15
A.2 Library-size normalization . . . 17
A.3 Predicting library size from the image . . . 18

B Observation Model Details **18**
B.1 Aggregation aperture . . . 19
B.2 Estimating the platform offset . . . 20
B.3 Dispersion, the slide intercept, and the readout . . . 20

C Datasets and Splits **21**
C.1 Training pool and split . . . 22
C.2 HEST benchmark tasks . . . 22
C.3 Downstream cohorts . . . 22

D Training Details **23**
D.1 Targets and preprocessing . . . 23
D.2 Gene sets . . . 24
D.3 Backbones and hyperparameters . . . 24
D.4 Training protocol . . . 27

E HEST Benchmark, Full Results **27**
E.1 Per-task correlation . . . 27
E.2 Other metrics . . . 29

F Other Pathology Foundation Models **33**
F.1 Encoders . . . 33
F.2 Benchmark correlation by encoder . . . 33

G Downstream Validation, Full Results **34**
G.1 TLS detection . . . 34
G.2 Spatial clustering . . . 35
G.3 A new platform . . . 36

H Ablation, Full Results **38**
H.1 The count-likelihood arms, task by task . . . 38
H.2 The same terms under a squared error . . . 39

I Extended Related Work **40**
I.1 Predicting expression from histology . . . 40
I.2 Foundation models . . . 42
I.3 Batch correction and data harmonization . . . 43

A Measurement Statistics

This section reports how the statistics of §2 were computed, all of them on HEST-1k and on two platforms, Visium and Xenium. They are the two on which models of this kind are most commonly trained, and both hold enough slides per organ in HEST-1k to tell a slide effect from an organ effect.

A.1 Variance decomposition

For each gene, we decompose the sum of squares of the spot-level $\log(1+y)$ into four components, removed in this order: platform, organ within platform, slide within organ, and spot within slide. The four sum to the gene's total, so we report each as a fraction of that total and call it the *share* of that component: the *slide share* is the fraction explained by slide identity within one platform

Table 4: Median per-gene share of the total sum of squares, by component, on raw counts. Only the organ-matched set carries both platforms and so a platform component. *slide (bench50)* restricts the slide share to those of the benchmark's 430 genes present in the set, and *slide (rest)* holds every other gene.

			Share of the total sum of squares				Slide share, split by gene set	
Set	Genes	Slides	Platform	Organ	Slide	Spot	Slide (bench50)	Slide (rest)
Organ-matched	787	74	0.139	0.007	**0.140**	0.632	0.133	0.140
Visium	35513	310	–	0.017	**0.058**	0.922	0.262	0.056
Xenium	1082	36	–	0.000	**0.080**	0.904	0.098	0.076

Table 5: The decomposition with the platform removed before the organ and the other way round, on the organ-matched set. *Target* is the normalization applied before the decomposition: raw counts, a fixed total, or a per-slide total (Table 6).

Target	Order	Genes	Platform	Organ	Slide	Spot
Raw	Platform first	787	0.139	0.007	0.140	0.632
	Organ first	787	0.096	0.026	0.140	0.632
Fixed total	Platform first	787	0.264	0.003	0.063	0.605
	Organ first	787	0.180	0.039	0.063	0.605
Per-slide total	Platform first	787	0.136	0.007	0.177	0.577
	Organ first	787	0.094	0.026	0.177	0.577

and one organ. Each share is computed over only those slides that measured that gene, and we summarize a set of genes by the median of its shares; because each median is taken at a different gene, the four medians of a set need not sum to one.

We run the decomposition on three sets of training slides: the organs profiled on both platforms, which are breast and lung (41 Visium and 33 Xenium slides), Visium alone, and Xenium alone. Only the first has a platform component; the other two decompose into three. An organ enters a set only if that platform has at least three slides of it, and a gene only if at least two slides of that platform measured it. Because Xenium's panel differs by organ, a gene measured in a single organ has a structurally zero organ component; only 277 of the 1,082 Xenium genes were measured in more than one organ, and their median organ share is 0.029, so the 0.000 that Table 4 shows for the full set reflects the panels rather than an absence of organ effect. To compare the two platforms at matched area, Visium is used at its $55\,\mu$m-diameter spot ($2{,}376\,\mu\mathrm{m}^2$) and the cell-level Xenium counts are aggregated into a $48\,\mu$m square ($2{,}304\,\mu\mathrm{m}^2$).

Table 4 gives the four shares for the three sets of slides. On the organ-matched set, the platform share and the slide share are of almost the same size, 0.139 and 0.140, which is the pair §2.2 quotes. On the single-platform sets the median gene has a smaller slide share, 0.058 on Visium and 0.080 on Xenium; the genes the benchmark selects sit above that median on both platforms, far above on Visium (0.262) and modestly on Xenium (0.098).

The benchmark takes the most variable genes of each task, and the slide share grows with a gene's variability. In Figure 1b (§2.2), the genes of each single-platform set are divided into quintiles of their total variance on raw counts, and each point is the median slide share within one quintile. The quintiles are cut once, on raw counts, and kept for the normalized line, so that the raw and the normalized lines compare the same genes. The median rises across the quintiles on both platforms, from 0.001 to 0.299 on Visium and from 0.036 to 0.178 on Xenium, and 370 of the 430 benchmark genes fall in the top two Visium quintiles; these are the 0.30 and the 86% of §2.2.

The decomposition is sequential, so we check the two choices it rests on. The first is order: platform and organ are crossed rather than nested, since both matched organs appear on both platforms in different proportions, so the part they share is assigned to whichever is removed first. The second is weighting: the identity counts one spot as one observation, as the training loss does, so the numbers we report are spot-weighted, and we also report a weighting under which every slide counts equally.

Table 6: Where the slide share and the platform share sit under each normalization. A fixed total scales every spot to one constant for the whole corpus, a per-slide total to the median total of that slide. Δ is the change in the gene's own slide share, raw $-$ variant, paired by gene, as a median and an interquartile range; the column beside it is the fraction of genes on which that change is positive.

		Slide share			Paired by gene	
Set	Variant	All genes	bench50	Platform share	Δ slide	$\Delta > 0$
Organ-matched	Raw	0.140	0.133	0.139	–	–
	Fixed total	0.063	0.099	0.264	0.044 [0.013, 0.112]	0.87
	Per-slide total	0.177	0.157	0.136	-0.025 [-0.066, -0.004]	0.18
Visium	Raw	0.058	0.262	–	–	–
	Fixed total	0.025	0.147	–	0.025 [0.001, 0.135]	0.91
	Per-slide total	0.048	0.303	–	0.000 [-0.028, 0.001]	0.49
Xenium	Raw	0.080	0.098	–	–	–
	Fixed total	0.057	0.084	–	0.009 [-0.002, 0.048]	0.65
	Per-slide total	0.096	0.117	–	-0.015 [-0.058, -0.001]	0.18

Removing the organ first lowers the platform share from 0.139 to 0.096 and leaves the slide share at 0.140 (Table 5), since the slide and the spot components sit below both factors and the order cannot reach them. Weighting every slide equally moves the two shares as little: the platform share reads 0.179 and the slide share 0.145 on the same set. What matters in each case is not that the shares are unchanged but that the slide share remains comparable to the platform share.

A.2 Library-size normalization

The *library size* of a spot is the sum of the counts of every gene measured at that spot. It varies by orders of magnitude, both between the spots of one slide and between slides. Rescaling each spot to remove that variation is standard preprocessing: each spot's counts are divided by that spot's library size and multiplied by a common constant, so that every spot sums to the same total, and the result is passed through $\log(1 + \cdot)$ (Wolf et al., 2018; Butler et al., 2018); recent work on paired histology and expression prepares its expression data the same way (Hemker et al., 2026; Han et al., 2026). The constant is a free choice, and the two conventions in common use behave differently here. A *fixed target* uses one constant for the whole corpus, commonly 10^4 or 10^6 and 10^4 here, whereas a *per-slide target* reads it off the slide, taking the median library size of that slide's spots, so different slides are rescaled to different totals. We compare three variants of the target: raw counts, a fixed target, and a per-slide target, each followed by $\log(1 + y)$.

On the sets with a single platform there is no platform axis, and normalization does lower the slide share (Table 6): on Visium a fixed target lowers the median slide share from 0.058 to 0.025, and on the genes the benchmark selects from 0.262 to 0.147, with the Xenium set moving the same way, from 0.080 to 0.057 and from 0.098 to 0.084. In Figure 1b, it lowers the median of the highest Visium quintile from 0.299 to 0.115, the 0.30 and 0.12 of §2.2. A per-slide target does not consistently lower it: reading the constant off the slide itself leaves the difference in depth between slides in place and normalizes only the spots within one slide, which shrinks the spot component and so can raise the slide share. On Visium the median moves from 0.058 to 0.048 with no consistent direction across genes (the change is positive for 49% of them) and rises on the benchmark genes from 0.262 to 0.303; on Xenium it rises from 0.080 to 0.096 and from 0.098 to 0.117.

Once two platforms are present the two axes no longer move together. On the organ-matched set the raw shares are almost equal, 0.140 for the slide and 0.139 for the platform. A fixed target lowers the slide share to 0.063 but raises the platform share to 0.264: what it removes on the slide axis reappears on the platform axis. A per-slide target leaves the platform share at 0.136 and raises the slide share to 0.177. Neither movement is a property of the median alone: pairing genes between variants, a fixed target lowers the slide share of 87% of genes, while a per-slide target raises it for 82% of them (the change is positive for only 18%), which are the last two columns of Table 6. Choosing the constant therefore does not remove the variation: a fixed target trades the slide axis for the platform axis, a per-slide target raises the slide axis and leaves the platform axis where it was,

and neither choice lowers both. Part of the rise in the platform share under a fixed target is built into the preprocessing. A spot's library size is summed over the genes its platform measured, the whole transcriptome on Visium and a targeted panel on Xenium, so the same gene is divided by totals of very different sizes on the two platforms; that is how the preprocessing is applied when panels differ, and it is one reason no single constant can put both platforms on one scale.

Even with the platform and the organ both held fixed, the median library size per slide can span two orders of magnitude. Across the 55 Visium bowel slides it runs from 252 to $59{,}061$ counts, a factor of 234, and the 34 lung slides span a factor of 409; over the fourteen Visium organs the factor ranges from 1.9 to 409, with a median of about 11. A difference of this size within one organ on one platform is unlikely to be a property of the tissue alone, yet it is the level that a model trained on the raw target would have to reproduce.

A.3 Predicting library size from the image

One quantity that changes from slide to slide is the library size (§2.2), which reflects the depth at which the slide was read: a slide read more deeply carries more counts of nearly every gene, so its targets sit at a higher level throughout. That level is not given to the model as an input. If it were legible in the image, however, the backbone could read it off the patches and would already be absorbing the difference between slides during training, which would leave a separate slide term with little to do. We therefore test whether it is legible, by fitting a ridge regression from the frozen UNI2-h embedding of a spot's image patch to the logarithm of that spot's library size.

We read the fit in two ways, because a spot's log library size is the sum of two quantities. The *between-slide* reading takes each slide's mean as one observation and asks how much of the variation among those means the prediction recovers; that mean reflects the depth at which the slide as a whole was read. The *within-slide* reading subtracts each slide's own mean from both the target and the prediction, leaving the spot-to-spot pattern inside one section. It is the between-slide reading that §2.2 quotes, because what a model trained on the raw target would have to reproduce is the level of the slide, not the pattern inside it.

The probe runs on every Visium and Xenium slide of our data, in the training, validation, and test splits alike (Appendix C), which is 398 Visium slides and 67 Xenium slides. The two platforms are not pooled, since library size is a sum over that platform's own panel and is not the same quantity across panels. We build five folds along each of two axes. Along the *slide* axis the folds group slides, so a held-out slide is never itself seen in training but other slides of its own study may be: the slide is new while the cohort is not. Along the *study* axis whole studies are held out, which is the situation of meeting a new cohort. The ridge penalty is selected by generalized cross-validation on the training fold, and the centering and feature scaling are computed from the training fold only; Table 7 reports how many slides each fold held out.

The level of a slide is recovered well when a sibling slide of its study is in training, with a between-slide R^2 of 0.63 on Visium and 0.82 on Xenium, and less well when the whole study is held out, 0.27 and 0.53 (Table 7); these are the four values §2.2 quotes. On Visium the within-slide reading is the smaller of the two (0.22 along the slide axis against 0.63 between), so there the image speaks to the level at which a slide was read more than to the spot-to-spot pattern inside it; on Xenium the order of the two readings differs between the axes, so we do not read the comparison beyond Visium. Because the probe is linear and the embedding is frozen, the between-slide values bound from below what the image carries about the level; a backbone trained end to end may recover more.

B Observation Model Details

In Equation 1 the area term A_s, the platform offset c_{gp}, and the slide intercept b_{sg} are all logarithms of factors that multiply the mean, which is why they enter additively. This section defines A_s (§B.1), describes how c_{gp} is estimated (§B.2), and shows how b_{sg} and the dispersion enter the readout (§B.3).

Table 7: How much of a spot's library size a ridge probe recovers from the frozen H&E embedding alone, as R^2 on held-out slides. *within*: each side centered by its own slide mean. *between*: the per-slide means, one slide one observation. Under `study` the held-out slides come from studies absent from training; under `slide` a sibling slide of the same study is in training.

Platform	Fold axis	Folds (slides)	Within	Between
Visium	`slide`	5 (79/80/80/80/79)	$+0.215$	$+0.629$
	`study`	5 (57/89/67/83/102)	$+0.072$	$+0.265$
Xenium	`slide`	5 (13/13/13/14/14)	$+0.756$	$+0.815$
	`study`	5 (12/7/10/27/11)	$+0.640$	$+0.531$

B.1 Aggregation aperture

A Visium spot is fixed by the assay, whereas high-resolution platforms (e.g., Xenium, Visium HD) resolve finer units, so the area over which their transcripts are aggregated into one target is a choice we make; we call that area the *aperture*. A wider aperture collects more counts, and Equation 1 receives this through A_s, the log of the slide's aperture area divided by the area of the reference Visium spot: a slide aggregated over the area of a Visium spot adds nothing, one aggregated over four times that area adds $\log 4$. It enters $\log \mu$ with no learned weight in front of it, which assumes the captured count grows in proportion to the area.

The assumption holds closely enough in the data. We test it by re-aggregating the same transcripts of a slide at 16, 48 and $80\,\mu\mathrm{m}$ and at the aperture the targets are built at, $100\,\mu\mathrm{m}$ squares on the Xenium platforms and $112\,\mu\mathrm{m}$ squares on the Visium HD platforms (§4.1), which we call the *native* aperture below. We then ask whether the counts a narrower aperture collects fall by as much as its area does. Under proportionality the two fractions of what the native aperture collects, the fraction of the area and the fraction of the counts, are equal. They agree to within one percent on all four platforms whose aperture we choose: at the $48\,\mu\mathrm{m}$ square, whose area is about 3% smaller than a Visium spot's, Visium HD keeps 0.1846 of its counts against an area fraction of 0.1837, and Xenium 0.2311 against 0.2304. The log gap between the two fractions stays within 0.006 of zero at the $48\,\mu\mathrm{m}$ square and within 0.007 at every aperture, and the slope of log counts on log area across the four apertures stays within 0.002 of one, both medians over slides. The validation and test slides of Xenium, the one platform whose slides are split, stay within the same bounds. The offset is applied per gene, so we also refit per gene, over the genes with at least 100 counts at the native aperture: the medians stay where they are and the interquartile range of the slope stays inside $[0.977, 1.028]$ on every platform.

The same term also shapes how the training targets are built. On the platforms whose aperture can be chosen, we do not fix one aperture. Targets are built at the 48 and $80\,\mu\mathrm{m}$ apertures and at the native one, all at once, so that a single η is not fitted to a single area. What η should stand for is a log intensity per reference area, with everything about the area left to A_s. A η that only ever saw one aperture is free to absorb the constant belonging to that aperture instead. During training the two places are indistinguishable, but on a slide aggregated over a different area the absorbed constant and A_s both apply and the prediction is off by their overlap. Showing the same tissue at several apertures removes that freedom, since no one constant fits all of them.

What is at stake is therefore the level of the predicted mean, not the order of the spots: a constant misplaced between η and A_s shifts $\log \mu$ as a whole. A per-gene correlation is largely blind to such a shift, whereas a count likelihood charges for it directly, so we check the apertures on held-out negative log-likelihood. Five backbones are trained twice over, once with targets at the three apertures and once at the native aperture alone, five seeds each and everything else held fixed, and scored on the held-out slides of the ten benchmark tasks, 57 on Visium and 15 on Xenium. On Xenium, where the apertures are applied, the negative log-likelihood is lower with them on all five backbones, by 0.17 to 0.59 nats. On Visium, where they are not applied, it is lower on one of five and the differences stay between -0.11 and $+0.07$, with no consistent direction. This is the pattern expected if the gain comes from the apertures: present where they are applied and absent where they are not. The two readings of the zero probability, the Brier score and the gap between predicted and observed zero rates, show no consistent difference between the arms on either platform. Nor do the apertures move the benchmark correlation, as expected of a shift in level: over the ten tasks the two

Table 8: The platform offset as a multiplier per target: the mean count per target at the platform's native aperture over that of a Visium spot, at the median covered gene. *Both*, the average of the two strata, is what we use; Visium is one by definition. *Covered* is the fraction of (gene, platform) pairs measured on both platforms that received an estimate.

Stratum	ST	HD	HD 3′	Xenium	Xen. 5K	Covered
Organ	0.48	3.50	1.68	28.44	11.22	76.7%
Cluster	0.40	5.19	2.66	15.21	10.41	80.7%
Both	**0.45**	**4.26**	**2.25**	**20.70**	**10.90**	**83.8%**

arms differ by -0.007 to $+0.004$, the arm at the native aperture alone being the higher on three of the five backbones, and that arm stays above the pooled MSE model on all five, by 0.017 to 0.069. The apertures therefore do not account for the gain of the observation model over squared error.

B.2 Estimating the platform offset

The platform offset is the log ratio of the amount of gene g that platform p reads per unit area to the amount the reference platform reads, the reference being Visium. Taking that ratio over all the data at once would confound the platform with the tissue, since a platform and the reference were not applied to the same mixture of organs. We therefore form the ratio within a stratum and average across strata, a stratum being either an organ or a cluster of the patch embedding.

The area and the offset are kept as two terms because not every unit is ours to choose. A Visium spot is fixed by the assay, and so are the spots of the original Spatial Transcriptomics platform, $100\,\mu\mathrm{m}$ in diameter on the slides we use and over three times the area of a Visium spot. Their counts arrive already pooled, with no way to subdivide them after the fact. The area term carries the differences in area that the assay has already settled. The ratio itself is measured on the targets each platform is evaluated on, at its native aperture, as a mean count per target. That target does not have the area of a Visium spot, so c_{gp} in Equation 1 is the log of that ratio less the log of the two areas' ratio, which keeps the area from being counted twice. Table 8 reports the per-target multiplier as measured; the area ratios it contains are 4.21 on the Xenium platforms, 5.28 on the Visium HD platforms and 3.31 on Spatial Transcriptomics.

The estimate is built in two stages, both using the training slides only. The first fits the morphology strata: we L_2-normalize the frozen patch embeddings, draw 800 spots from each of the 546 training slides, and fit a mini-batch k-means with $K = 50$. A single fit leaves the cluster stratum too unstable to use, so we fit 20 seeds and average their estimates in log space. The organ stratum reads its labels from the metadata and does not depend on a seed.

The second stage estimates the ratio itself. A *cell* is one (platform, stratum, gene) combination. Dividing the counts of that gene a cell holds by the spots they were collected over gives its mean count per spot. The estimate a cell contributes is the log ratio of that mean to the mean of the reference platform's cell at the same stratum and gene. Averaging over the cells of one stratification gives one estimate per (gene, platform) pair under it. A cell is counted only if both it and the reference cell it is paired with hold at least 200 spots and at least 5 counts of the gene. The organ and the cluster estimate are averaged in log space where both exist, and whichever exists is taken where only one does. What remains is a single (gene, platform) table with no stratum index. Pairs left with no estimate fall into two cases. When the platform's training slides measured the gene, the pair's per-target multiplier is set to one, so that neither c_{gp} nor the area ratio enters, in training and in every evaluation on HEST-1k; the external platform is read differently (§G.3). When no training slide of the platform measured the gene, the pair is met only at evaluation, and its multiplier is the area ratio alone, with $c_{gp} = 0$. Table 8 gives the resulting multipliers.

B.3 Dispersion, the slide intercept, and the readout

The dispersion is factored into a gene term and a platform term, $\theta_{gp} = \theta_g e^{\delta_p}$, both learned jointly with the backbone, the platform term starting at zero. We clamp $\log \mu$ and $\log \theta$ to ± 12 so that $\exp$ cannot overflow, $e^{12} \approx 1.6 \times 10^5$ sitting far above the largest count in the data. Only $\log \theta_g + \delta_p$ is

identified, the split between them carrying one free constant, so θ_g is not the marginal overdispersion of that gene.

The slide term is close in spirit to a random effect (Laird & Ware, 1982): the intercepts are fitted under a penalty rather than integrated over during training, and only a new slide's intercept is integrated out, at inference. The penalty on the slide term is quadratic and is measured as the mean square of b_{sg} over genes rather than as the norm of v_s. The intercept is $b_s = Uv_s$ and the scale of U is free, so a penalty on $\|v_s\|$ could be driven down by enlarging U and shrinking v_s while the intercepts it is meant to restrain stay where they are. We take $k = 16$ and a penalty coefficient of 0.01. Where a platform contributes a single training slide, the centering in Equation 2 sets that slide's intercept to zero: the slide effect and the platform effect are then perfectly confounded and the platform offset alone carries that slide's level.

The readout is where the dispersion enters. The mean of a negative binomial is μ whatever the dispersion, so the $\log(1 + \mathbb{E}[y])$ a mean model reads off carries no θ at all, and a platform axis on θ could then reach a metric only through the fitted μ, never through the readout itself. We therefore predict $\mathbb{E}[\log(1 + y)]$, which is a function of θ and which is also the optimal point estimate under squared error in $\log(1 + y)$ space. Since $\log(1 + \cdot)$ is concave, Jensen's inequality orders the two,

$$\log\big(1 + \mathbb{E}[y]\big) \;\geq\; \mathbb{E}\big[\log(1 + y)\big], \tag{4}$$

the first being the larger, so moving to the second lowers every prediction by a systematic amount. Whether that helps squared error depends on where the predictions already sit: it is the right estimate under the fitted model, but predictions that are already low against the observed counts are only made worse by lowering them further. Pearson correlation is little affected either way: both readouts increase with μ, and their gap, near zero for $\mu \ll 1$, tends to a constant that depends only on θ as μ grows, so a per-gene correlation, which is invariant to a constant shift, moves little.

The expectation has no closed form, and the sum $\sum_k \log(1+k)\, P(k)$ needs more terms as μ grows, which reaches the thousands. Frullani's identity $\log a = \int_0^\infty (e^{-t} - e^{-at})\, t^{-1}\, \mathrm{d}t$ with $a = 1 + y$ turns it into a single integral over the Laplace transform of the count distribution, which the negative binomial has in closed form:

$$\mathbb{E}\big[\log(1 + y)\big] = \int_0^\infty \frac{e^{-t}\big(1 - G(t)\big)}{t}\, \mathrm{d}t, \qquad G(t) = \mathbb{E}\big[e^{-ty}\big] = \left(\frac{\theta}{\theta + \mu\,(1 - e^{-t})}\right)^{\theta}. \tag{5}$$

Substituting $t = e^u$ cancels the $1/t$ factor and leaves an integrand on the whole line whose derivatives vanish at both ends, so a trapezoidal rule converges exponentially and a fixed grid suffices whatever the mean; we take 128 nodes on $u \in [-30, 5]$, the lower limit chosen so that the mass it discards, about $\mu\, e^{-30}$, stays orders of magnitude below the precision of the metric. Against cancellation, $1 - e^{-t}$ and $1 - G$ are written with $-\mathrm{expm1}$.

The slide intercept is integrated out on top of this. Training sees each slide's own intercept while it fits θ, so leaving it at zero for a new slide would hand the readout a θ fitted with the spread already absorbed. For a new slide, the intercept is taken as normal with mean zero and a standard deviation σ computed, for each gene, over the fitted intercepts of the training slides of that platform. The mean is zero because Equation 2 centers v_s within a platform. The standard deviation is a plug-in from intercepts that the penalty has shrunk, so it may understate the spread of a new slide's level, and the normal is likewise an assumption that keeps the integral a quadrature. Substituting $b = \sqrt{2}\,\sigma z$ turns the outer integral into a Gauss–Hermite quadrature over seven nodes (Golub & Welsch, 1969),

$$\mathbb{E}_b\big[\log(1 + y)\big] = \int \mathbb{E}\big[\log(1 + y) \,\big|\, b\big]\, \phi_\sigma(b)\, \mathrm{d}b \;\approx\; \sum_{j=1}^{7} w_j\, \mathbb{E}\big[\log(1 + y) \,\big|\, \mu\, e^{\sqrt{2}\sigma z_j},\, \theta\big], \tag{6}$$

whose inner expectation is Equation 5 at a shifted mean. The weights are normalized to sum to one, so a platform contributing a single training slide, where $\sigma = 0$, reduces to the conditional readout exactly.

C Datasets and Splits

This section reports where the data come from and how they were divided. §C.1 gives the training pool and the rules that keep every evaluation slide out of it, §C.2 identifies the benchmark tasks, and §C.3 the three cohorts of the downstream evaluation.

Table 9: Composition of the training pool by organ and platform. A dash is zero.

	Visium family				Xenium family		
Organ	ST	Visium	Visium HD	HD 3′	Xenium	Prime 5K	Total
Bladder	–	5	–	–	–	–	5
Bone	–	–	–	–	1	–	1
Bowel	–	55	1	–	2	–	58
Brain	–	27	–	–	1	–	28
Breast	100	7	3	–	12	–	122
Cervix	–	4	–	–	–	1	5
Eye	–	5	–	–	–	–	5
Heart	19	39	–	–	1	–	59
Kidney	–	35	1	–	2	–	38
Liver	–	12	–	–	2	–	14
Lung	–	34	5	–	21	3	63
Lymph node	–	1	–	–	–	–	1
Lymphoid	–	–	3	–	3	1	7
Ovary	–	2	2	3	1	2	10
Pancreas	–	3	1	1	–	–	5
Prostate	24	10	1	–	–	1	36
Skin	12	60	–	–	2	1	75
Uterus	–	14	–	–	–	–	14
Total	155	313	17	4	48	9	546

C.1 Training pool and split

We use the human slides of HEST-1k (Jaume et al., 2024) at version 1.3.0, 689 slides in all. We first set aside the benchmark's test slides, 72 of them, 57 on Visium and 15 on Xenium. From what remains we then remove every slide that overlaps with one of those test slides, under three rules: any slide sharing a patient with one of them, 6 slides; any slide sharing an organ, a patient and a tissue sample with one of them, 1 slide; and any slide from the same study as one of them, 12 slides, even where no patient identifier is recorded. We make one exception to the last rule: TENX114, a healthy Xenium bowel slide from the study of one COAD test slide and with no patient identifier, which we keep because removing it would leave Xenium bowel with a single training slide. Last, we remove the annotated slides of the downstream evaluation that are not already test slides: the 12 DLPFC and 8 breast-cancer slides of §C.3, 20 in all. The breast-cancer slides are one section from each of eight patients, and the other sections of these patients remain in the training pool. The TLS and prostate cohorts need no removal, as their slides belong to the CCRCC and PRAD tasks. The remaining 578 slides are divided into 546 for training and 32 for validation. The validation slides lie on Visium and Xenium alone, 28 and 4 of them, the same two platforms as the test set. Table 9 gives which platform and which organ the training slides came from.

C.2 HEST benchmark tasks

We take the benchmark as published: which slides belong to each task and which genes it asks for are fixed in advance, and we use them as given. There are ten tasks, five on Visium and five on Xenium, and each names one cancer type in one organ; Table 10 identifies them. Fifty genes are scored per task. What differs on our side is training rather than evaluation: the benchmark envisages a model fitted to each task's own training slides, whereas the models here are fitted once on a pool spanning all ten tasks and evaluated per task.

C.3 Downstream cohorts

The downstream evaluation uses three cohorts, set side by side in Table 11. The annotations the first two rely on come from STimage-1K4M (Chen et al., 2024a), joined to the HEST slides by spot barcode on Visium and by grid coordinate on legacy Spatial Transcriptomics.

Table 10: The ten HEST benchmark tasks. *Genes measured* is the panel the task's slides were read with, which the fifty scored genes are drawn from. One COAD slide carries no patient identifier.

Task	Cancer type	Organ	Slides	Patients	Genes measured
Visium, transcriptome-wide panel					
CCRCC	Clear cell renal cell carcinoma	Kidney	24	24	17,943–36,601
HCC	Hepatocellular carcinoma	Liver	2	2	36,601
LYMPH_IDC	Breast IDC, lymph-node metastasis	Lymph node	4	4	33,931
PRAD	Prostate adenocarcinoma	Prostate	23	2	33,538
READ	Rectal adenocarcinoma	Bowel	4	2	36,601
Xenium, targeted panel					
COAD	Colon adenocarcinoma	Bowel	4	3	422–425
IDC	Breast invasive ductal carcinoma	Breast	4	4	280–313
LUNG	Lung adenocarcinoma	Lung	2	2	377–480
PAAD	Pancreatic ductal adenocarcinoma	Pancreas	3	3	377–474
SKCM	Cutaneous melanoma	Skin	2	2	282–382

Table 11: The three downstream cohorts. *Spot* is the capture area each is read at.

Cohort	Platform	Organ	Slides	Spot	Annotation
TLS detection	Visium	Kidney	17 of 23	55 μm	TLS vs. no TLS, per spot
Clustering, DLPFC	Visium	Brain	12	55 μm	cortical layers (L1–L6, WM)
Clustering, breast	ST	Breast	8	100 μm	pathology regions
Clustering, prostate	Visium	Prostate	7	55 μm	benign, Gleason groups, stroma
External validation	Atera	Breast, Cervix	2	100 μm sq.	none

The first is a clear cell renal cell carcinoma (ccRCC) Visium cohort of 24 slides (Meylan et al., 2022), the same slides as the CCRCC task, carrying tertiary lymphoid structure annotations at the level of individual spots; 23 of the 24 have an annotation file. A slide is scored only where an AUROC is defined, which needs both classes to occur and at least ten annotated TLS spots; 17 slides meet this.

The second is the set of annotated cohorts used as ground truth for spatial clustering, 27 slides in three cohorts: 12 Visium slides of dorsolateral prefrontal cortex (Maynard et al., 2021), 8 legacy Spatial Transcriptomics slides of breast cancer (Andersson et al., 2021) and 7 Visium slides of prostate (Erickson et al., 2022).

The third is an external cohort, one breast-cancer (10x Genomics, 2026a) and one cervical-cancer (10x Genomics, 2026b) sample measured on Atera (10x Genomics), a platform absent from HEST-1k. The cervical sample was cut from the same tissue block as TENX188, a Xenium Prime 5K slide in the training pool (10x Genomics, 2026b). Because the measurement is imaging-based, its transcripts are aggregated into $100\,\mu\mathrm{m}$ square pseudo-spots, the same area used for Xenium. Its panel is transcriptome-wide, 18,028 genes alongside 9,076 control probes, and the control probes are discarded. Five further genes lie outside the gene space the models predict (§D.1) and are dropped as well, leaving the 18,023 genes each sample is scored over.

D Training Details

This section gives the training procedure of §4. §D.1 states what the targets are and what preprocessing produced them, §D.2 says which genes the evaluations are scored on, §D.3 describes what the five backbones are made of and the values they were trained with, and §D.4 reports the schedule, the stopping rule and the compute.

D.1 Targets and preprocessing

The targets are the raw counts, with no library-size normalization. A model with the count likelihood is fitted to those counts as they are, and a model with the squared-error objective to $\log(1+y)$ of them. Every arm is scored against $\log(1+y)$ of the raw counts: a count-likelihood arm through

the readout of §3.5, and a squared-error arm on its output, which for the arms fitted to a rescaled target (§H.2) is first mapped back to the raw scale; MSE (CPM) is scored on its output as it is. Image patches are extracted at $0.5\,\mu\mathrm{m}$ per pixel and are 224 pixels wide, so one patch sees $112\,\mu\mathrm{m}$ of tissue, and a patch whose tissue fraction falls below 0.15 is dropped. The area over which transcripts are aggregated into one target follows the platform: $100\,\mu\mathrm{m}$ squares on Xenium and Xenium Prime 5K, $112\,\mu\mathrm{m}$ squares on Visium HD and Visium HD 3′, while Visium and the legacy Spatial Transcriptomics slides use the spot the assay defines. On the platforms whose aperture we choose, targets are built at the 48 and $80\,\mu\mathrm{m}$ apertures as well, for the reason §B.1 gives. Only the arms that carry the platform terms use them, under either loss; the others, plain MSE among them, see the native aperture alone, since without A_s the same patch would carry targets that differ by area with nothing in the model to account for it. Each draw of a slide or a window takes one aperture, so the apertures change what a step sees but not how many steps there are, and every evaluation is at the native aperture.

The gene space is assembled from one gene list per training slide, each built in two steps. First, control probes are removed and gene names are harmonized: reference prefixes are stripped, names are uppercased, and an Ensembl identifier is replaced by its gene symbol where exactly one symbol in our gene list matches it. When two genes of a slide end up under the same name, their counts are added. Second, as in STPath (Huang et al., 2025b), each slide keeps only the genes with at least three counts in total over its spots. The gene space is the union of the genes the training slides keep, $38{,}605$ in all. The legacy Spatial Transcriptomics slides do not add to this union: their clone identifiers do not harmonize onto current symbols, and letting them in would add $9{,}737$ entries. Those slides are still used for training, supervised on every gene they share with the union. Because panels differ from slide to slide, the training loss is computed only over the genes a slide measured.

D.2 Gene sets

The model predicts the whole gene space: it has one output for each of the $38{,}605$ genes §D.1 defines. Which of those outputs an evaluation is scored on depends on the evaluation. The HEST benchmark scores each task on the 50 genes predefined for it, 430 in the union of the ten tasks. Tertiary lymphoid structure (TLS) detection scores two published signatures, a 29-gene signature from Meylan et al. (2022) and a 12-gene chemokine signature from Xu et al. (2022). Spatial clustering runs over every gene the model's training slides of the Visium family supervised, which is narrower than the gene space (§G.2). The external cohort is scored on the genes its own panel shares with the gene space. Table 12 lists the benchmark lists and the two signatures.

D.3 Backbones and hyperparameters

Four of the five backbones sit on the same frozen patch encoder and differ only in what is placed on top of it; the rest adapts that same encoder with LoRA. We use UNI2-h (Chen et al., 2024b) for the patch encoder, and its patch embeddings are $1{,}536$-dimensional. Table 13 gives the architecture, the optimization and the schedule each backbone was trained with, and the paragraphs below add further detail.

Probe. An MLP over the embedding of a single spot. Weight decay is applied to the first linear layer's weight alone, and the gene axis, meaning the output layer and the dispersion, is excluded from it. AdamW decays a parameter every step whether or not it received a gradient, and genes are measured at very different rates, so decay on the gene axis would act as shrinkage in proportion to how often a gene is measured rather than as regularization. Simple-TF excludes its gene head for the same reason.

Probe+LoRA. The same probe, with the patch encoder adapted by LoRA (Hu et al., 2022) and trained jointly with the head. Training the adapters against a freshly initialized head is unstable early on, so no run starts from scratch: each takes the weights of the frozen probe trained under the same condition, training pool and seed, and trains the adapters from there. What LoRA adapts is the last four blocks of the encoder: in each of them a weight is left as it is and a trainable low-rank update is added beside it, of rank 8 here, scaled by $\alpha = 16$, with dropout 0.1 on the input to that branch.

Table 12: The gene sets the evaluations are scored on: the ten lists HEST-1k registers for its tasks, 50 genes each and 430 in their union, and the two tertiary lymphoid structure signatures.

Task	Organ	Genes
HEST benchmark, Visium		
READ	Bowel	AC010883.3, ADAMDEC1, ANGPTL2, APOC1, BTNL3, CA1, CA2, CCL14, CCL21, CD177, CEMIP, CHI3L1, CLCA1, CLU, CXCL10, CXCL8, CXCL9, EFEMP1, EGLN3, FRZB, GUCA2A, HMOX1, IGHG2, IGHG4, ITGB2, ITIH5, LAMA3, LOX, MMP28, MMP7, MMP9, MT1F, MT1M, MXI1, P3H3, PADI2, PDE9A, PLAAT3, PRRG3, PTGDS, SECTM1, SGK1, SLC26A2, SLC26A3, SMPDL3A, SPHK1, SPP1, TPSB2, TSPAN9, UBD
CCRCC	Kidney	AKAP8, BLVRB, BTF3, CD109, CD14, CHST15, CIB1, CKB, CPD, CPE, DERL1, DNAJB12, EGFL7, FLNB, GADD45B, GPNMB, H3F3A, HMOX1, HNRNPK, HSPA1A, HTRA1, IFI27, IFI6, IRF7, ISG15, KRTCAP2, LAPTM4B, LRRC41, LY6E, NAMPT, NOL3, NOTCH3, NT5DC2, PARP10, PTGS1, PXN, RXRA, S100A9, SCARB1, SEMA4B, SLC15A4, SLC39A14, STK39, TFRC, TM2D3, TMEM91, TP53INP1, TRBC1, VMP1, YWHAE
HCC	Liver	AC092069.1, ACTA2, AEBP1, ASPM, BEST1, C11orf96, CAVIN3, CCL21, CCN1, CXCL10, CXCL9, CYP2E1, ELN, ETNPPL, FBLN2, FBLN5, GJA4, HAMP, HBB, IGFBP5, IGHA1, IGHG1, IGHM, IGKC, IGLC1, IGLC2, JAG1, LRRC75A, LTBP2, MCAM, MFAP4, MGP, MUSTN1, MYL9, NDUFA4L2, PDE9A, PDZK1IP1, PHKG1, PPP1R14A, PRELP, PTGDS, RBP7, RND1, S100A4, SPARCL1, TAGLN, TCEAL9, TPM2, TRBC2, TSPAN13
LYMPH_IDC	Lymph node	AAMDC, AKR1E2, ANKS1B, ANO8, AREG, CCL28, CCNO, CD22, CD72, CLNS1A, CROT, CSTA, CXCL10, CXCL11, CYP4Z1, DAPK1, EPHB3, FABP4, GTPBP3, HAAO, HAUS8, IFI44L, IFIT2, IGHV3-11, IGHV6-1, IGLV9-49, MAPK8IP2, MX1, MX2, MYH11, NARS2, NUDT1, OASL, PARD6B, PATZ1, PDZK1IP1, POLD3, PTRH2, QSOX1, RAB38, RSAD2, RSF1, SCRN1, SERPINI1, SHANK2, SNCG, SPP1, TCF7, TMEM64, ZMAT5
PRAD	Prostate	AMD1, C1orf43, C6orf48, C8orf59, CAT, CCDC47, COX17, COX5A, COX7B, DNAJC3, EIF3E, EIF3H, GLO1, GNG5, HLA-DRA, HSPD1, IMPDH2, MAOA, MIF, MPC2, MRPL33, MT1E, MT1X, NUCB2, NUDT4, PCBD1, PDIA4, PERP, REXO2, RPL37A, RPN1, RPS29, RSL24D1, SEC11C, SELENOT, SGK1, SLIRP, SNHG8, SNHG9, SNRPE, SOD2, SPCS3, SPON2, SRRM1, SVIP, TMCO1, TOB1, TXNL4A, XBP1, ZFAS1
HEST benchmark, Xenium		
COAD	Bowel	AQP8, AREG, BANK1, CA1, CA4, CALB2, CCL20, CCR7, CD177, CEACAM6, CEACAM7, CHGA, CHGB, CHI3L2, CLCA1, CLCA4, COL11A1, CTSE, CXCL10, CXCL13, DMBT1, DUOX2, FCRL1, GATA2, GCG, GUCA2A, GUCA2B, HEPACAM2, IL1B, INSL5, ITLN1, LCN2, MMP1, MMP3, MS4A1, MS4A12, NXPE4, OLFM4, OTOP2, PI3, PRPH, RETNLB, SCG2, SCGN, SLC26A3, SPIB, TCL1A, TMIGD1, TYMS, UGT2B17
IDC	Breast	ABCC11, ADH1B, ADIPOQ, ANKRD30A, AQP1, AQP3, CCR7, CD3E, CEACAM6, CEACAM8, CLIC6, CYTIP, DST, ERBB2, ESR1, FASN, GATA3, IL2RG, IL7R, KIT, KLF5, KRT14, KRT5, KRT6B, MMP1, MMP12, MS4A1, MUC6, MYBPC1, MYH11, MYLK, OPRPN, OXTR, PIGR, PTGDS, PTN, PTPRC, SCD, SCGB2A1, SERHL2, SERPINA3, SFRP1, SLAMF7, TACSTD2, TCL1A, TENT5C, TOP2A, TPSAB1, TRAC, VWF
LUNG	Lung	ACE2, ALDH1A3, BBOX1, CAPN8, CCL19, CCR7, CD163, CD19, CD1A, CD2, CD274, CD79A, CDK1, CNN1, COL5A2, CTSG, CTSK, CXCL10, CXCL9, CXCR4, CYP2A7, DMBT1, DNAAF1, FCER1A, FGFBP1, GPRC5A, INMT, LIF, MALL, MARCO, MCEMP1, MEDAG, MET, MKI67, MS4A1, MZB1, OGN, PCSK2, PECAM1, PLA2G7, PTPRC, SERPINB2, SNCG, SNTN, SOX2, SPIB, SRPX, TCL1A, TOP2A, UBE2C
PAAD	Pancreas	ACE2, ACTA2, AIF1, ANPEP, BANK1, CCL19, CCL5, CCR7, CD14, CD19, CD1A, CD1C, CD2, CD27, CD3E, CD79A, CD8A, CDK1, CPA3, CXCL10, CXCL2, CXCL6, CXCL9, CXCR4, DERL3, DMBT1, EPCAM, ESR1, FCGR3A, FOXP3, GPR183, GZMA, GZMK, HPGDS, IL7R, IRF8, KLRC1, LIF, LILRA4, MARCO, MKI67, MS4A1, PDCD1, PLA2G7, PTPRC, SELL, TCL1A, TNFRSF17, UBE2C, VSIG4
SKCM	Skin	ACER1, AHNAK2, AQP3, BIRC3, CAPNS2, CCL19, CCL22, CD1B, CD52, COL17A1, CRABP2, DEFB1, DMKN, DSC1, DSP, FGFBP1, GZMK, HAPLN1, HBD, IDO1, IL1B, ITM2A, KRT16, KRT17, KRT2, KRT5, KRTDAP, LAMB3, LCK, LY6D, LYZ, MAL2, MGP, MIR205HG, MMRN1, MYH11, POSTN, RHOV, S100A14, SBSN, SELE, SERPINB2, SERPINB5, SLPI, TFF3, TP63, TYRP1, VIM, VWF, WDFY4
Tertiary lymphoid structure signatures		
Meylan	–	IGHA1, IGHG1, IGHG2, IGHG3, IGHG4, IGHGP, IGHM, IGKC, IGLC1, IGLC2, IGLC3, JCHAIN, CD79A, FCRL5, MZB1, SSR4, XBP1, TRBC2, IL7R, CXCL12, LUM, C1QA, C7, CD52, APOE, PLTP, PTGDS, PIM2, DERL3
Chemokine	–	CCL2, CCL3, CCL4, CCL5, CCL8, CCL18, CCL19, CCL21, CXCL9, CXCL10, CXCL11, CXCL13

Table 13: Architecture, optimization and schedule of the five backbones. *Depth* counts hidden layers of the head or transformer blocks, with a two-layer decoder of width 512 above that on STORM†; *Scope* is the set of spots one forward pass attends over; *Patience* counts consecutive validations without improvement. An arrow separates STORM†'s two stages, and Probe+LoRA's two learning rates are the head's and the adapters'.

	Probe	Probe+LoRA	Simple-TF	STPath†	STORM†
Architecture					
Patch encoder	frozen	LoRA, last 4	frozen	frozen	frozen
Width d	512	512	512	512	512
Depth	1	1	4	4	4
Attention heads	–	–	8	4	8
MLP ratio	–	–	2.0	2.0	2.0
Dropout	0.1	0.1	0.1	0.1	0.1
Scope	spot	spot	slide	region	window
Platform token	no	no	yes	yes	no
Optimization					
Optimizer	AdamW	AdamW	AdamW	Adam	AdamW
Learning rate	2×10^{-4}	$2/1\times10^{-4}$	3×10^{-4}	5×10^{-4}	$1\rightarrow5\times10^{-4}$
Weight decay	0.05	0.05	0.05	0	0.05
Batch	1,024 spots	128 spots	1 slide	2 slides	256 windows
Schedule					
Maximum epochs	500	10	500	500	$500\rightarrow100$
Validate every	5	1	5	5	5
Patience (evals)	5	3	5	5	5
Seeds	5	5	5	5	5

Simple-Transformer (Simple-TF). An organ embedding and a platform embedding are added to each spot embedding, and all spots of a slide pass together through transformer blocks of self-attention and a feed-forward network. There are no positional embeddings: position enters as a static, non-learned bias on the attention logits, as proposed in ALiBi (Press et al., 2022; Ding et al., 2025) and applied here to the distance between a pair of spots, subtracted with a slope that differs from head to head. That distance is Euclidean, measured in units of the slide's own median nearest-neighbor spacing rather than in micrometers, so that one slope reaches the same number of spots on every platform; measured in micrometers it would reach across different amounts of tissue depending on how closely a platform samples. The platform token is dropped per spot with probability one half during training and is always supplied at evaluation.

STPath†. Our reimplementation, following the published implementation as closely as our setting allows. Masked expression tokens and the image embedding are encoded together by a spatial transformer, which reconstructs the masked positions. The masked fraction is drawn from a $\mathrm{Beta}(10, 1)$ and the region's spot count uniformly between 64 and 1,024, the region itself gathered by nearest neighbors. Three things depart from the upstream recipe. The patch encoder is the one every backbone here shares rather than the Prov-GigaPath the original uses (Xu et al., 2024). The platform token is dropped with probability one half rather than the upstream four fifths, which is the rate Simple-TF uses and puts the only two models that receive the token on the same terms. And at inference the slide is cut into tiles of 1,024 spots, the size training uses, rather than passed through in one piece: attention here grows with the square of the spot count, and the largest slide in this pool does not fit at once.

STORM†. Our reimplementation, written from the paper's description alone since no code accompanies it. Where the description fixes a value we follow it; where it leaves one open, such as the width and depth of the encoder or the width of the decoder, we set it to match Simple-TF and STPath†, so that the comparison is not a comparison of size. Two encoders feed it: the frozen patch encoder the other backbones use, and, on the gene expression side and as in the original, frozen 64-dimensional Novae (Blampey et al., 2025) spot embeddings. Spots are grouped into windows of the 25 nearest spots. Pretraining is the first of two stages: five spots per modality are left visible and the remaining twenty are masked, the two modalities pass together through transformer blocks that share their self-attention and give each modality its own feed-forward expert, and lightweight

decoders reconstruct the masked positions of both modalities before being discarded. The second stage attaches an MLP head for prediction from H&E alone and trains that head: the projections, the modality and mask embeddings and every spatial block stop receiving gradient, while the parameters of the observation model are fitted alongside the head. As in the original, the model is given no platform token.

What the dagger marks is common to both: the pretraining data and the gene space are ours rather than theirs, the training budget is ours, and the patch encoder is the frozen one the five backbones share rather than the one the original names. The original STORM trains its H&E encoder jointly with the rest; here it is frozen, as for every other backbone, so that the patch encoder is the same across the comparison.

Single-platform experts. An expert trains on one platform, so a platform input or a platform term has nothing to tell apart. The two backbones that take a platform token receive the padding token in its place, which keeps the implementation simple, in training and in every evaluation, the HEST benchmark included; no platform's token is therefore ever trained in an expert. The terms of the observation model are kept. On the Visium expert they sit at their reference values by construction: Visium is the reference platform, so $c_{gp} = 0$; its spot is the reference area, so $A_s = 0$; and only the Visium row of the dispersion is fitted. On the Xenium expert the Xenium offset and area term apply as in pooled training. The slide term stays in both experts, centered over the expert's own slides.

D.4 Training protocol

Training runs for at most 500 epochs with validation every five, and stops when five consecutive validations, 25 epochs, bring no improvement. The validation slides are grouped by organ and platform family. Within a group we take the genes its slides all measured, rank them by variance on the training slides alone, and score the group on the Pearson correlation over the top 50. Group scores are then averaged over the organs of a platform family, and those averages over the families. The checkpoint we keep is the one that scores highest on that. Probe+LoRA departs from this schedule: continuing from an already-fitted probe, it validates after every epoch and stops sooner. STORM† departs from it in both stages: the first is validated on held-out masked reconstruction and keeps the weights that reconstruct best, and the second starts from those weights and runs for at most 100 epochs rather than 500 (Table 13). Every condition is run with five seeds, 0 through 4. Every run was trained on a single NVIDIA A100 40 GB GPU.

E HEST Benchmark, Full Results

E.1 Per-task correlation

This section reports the score of each of the ten benchmark tasks. Beyond the correlation over the predefined 50 genes, we report five further metrics. Every metric we report is computed within one slide and then averaged over slides with equal weight, as the footnote in §5.1 states. The alternative is to concatenate the test slides of a task and score them once, but each slide carries an offset of its own, and a task holds anywhere from two to twenty-four test slides. Concatenating therefore moves the correlation by an amount that differs from task to task in direction as well as in size. A clustering score that compares the predicted with the measured clusters is distorted more overtly: the slide boundaries become cluster boundaries on both sides, so the score begins to measure which slide a spot came from rather than how well a prediction reproduces the structure. We therefore score each slide on its own and average those scores over slides.

Table 14 gives the per-task correlation over the 50 genes. With both arms trained on the pooled data, the observation model has the higher correlation on 44 of the 50 task and backbone cells: on all ten tasks for Probe and for STPath†, on nine for STORM†, on eight for Simple-TF and on seven for Probe+LoRA. The six that go the other way are LYMPH_IDC and PRAD on Simple-TF, PAAD on STORM†, and LUNG, PAAD and SKCM on Probe+LoRA. The pool contrast depends on which objective is held fixed: pooling raises the correlation on 37 of the 50 cells when the observation model is fitted but on only 16 when squared error is.

Notably, one Xenium task gains a great deal when the training pool changes from the expert one to the pooled one, and that single task is the main reason the Xenium aggregate moves. On SKCM, the

Table 14: Per-task Pearson correlation over the 50 benchmark genes on the HEST benchmark. Rows are the four conditions of each backbone, and bold marks the best of the four (higher is better). Means over five seeds; the standard deviations are omitted for space (median 0.015 over the cells here). Tasks are named as the benchmark names them, except that LYMPH_IDC is written LYMPH.

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE Exp.	0.206	0.142	0.081	0.198	0.224	0.472	0.579	**0.609**	**0.494**	0.343	0.335
MSE Pool.	0.182	0.139	0.062	0.195	0.245	0.433	0.572	0.575	0.443	0.596	0.344
Obs. Exp.	0.217	0.158	**0.087**	0.201	0.268	0.445	0.599	0.594	0.490	0.349	0.341
Obs. Pool.	**0.227**	**0.181**	0.086	**0.213**	**0.287**	**0.490**	**0.602**	0.605	0.463	**0.622**	**0.378**
Probe+LoRA											
MSE Exp.	0.208	0.166	0.083	0.230	0.270	0.487	0.599	**0.625**	**0.512**	0.354	0.353
MSE Pool.	0.229	0.176	0.080	0.224	0.260	0.515	0.611	0.622	0.497	**0.650**	0.386
Obs. Exp.	0.221	0.165	0.077	0.222	0.290	0.457	0.599	0.597	0.488	0.341	0.346
Obs. Pool.	**0.244**	**0.190**	**0.093**	**0.238**	**0.309**	**0.525**	**0.613**	0.620	0.487	0.641	**0.396**
Simple-TF											
MSE Exp.	0.246	0.170	0.092	0.200	**0.310**	0.479	0.592	0.606	**0.494**	0.328	0.352
MSE Pool.	0.214	0.184	0.088	**0.236**	0.305	0.490	0.570	0.585	0.427	0.557	0.366
Obs. Exp.	0.245	0.188	0.099	0.218	0.295	0.470	0.588	0.605	**0.494**	0.347	0.355
Obs. Pool.	**0.258**	**0.189**	**0.117**	0.233	0.243	**0.532**	**0.625**	**0.629**	0.486	**0.647**	**0.396**
STPath†											
MSE Exp.	0.171	0.126	0.055	0.140	0.299	0.453	0.591	**0.601**	0.484	0.353	0.327
MSE Pool.	0.158	0.125	0.053	0.152	0.170	0.428	0.536	0.521	0.365	0.551	0.306
Obs. Exp.	0.223	**0.171**	**0.067**	0.201	**0.321**	0.442	0.589	0.596	**0.496**	0.361	0.347
Obs. Pool.	**0.244**	0.147	0.064	**0.215**	0.259	**0.505**	**0.595**	0.581	0.439	**0.625**	**0.367**
STORM†											
MSE Exp.	0.195	0.163	0.060	0.161	**0.290**	0.439	0.525	**0.558**	**0.454**	0.267	0.311
MSE Pool.	0.174	0.148	0.060	0.176	0.255	0.459	0.518	0.530	0.422	0.525	0.327
Obs. Exp.	0.215	**0.164**	0.067	0.178	0.286	0.422	0.514	0.525	0.437	0.257	0.306
Obs. Pool.	**0.224**	0.160	**0.069**	**0.213**	0.285	**0.488**	**0.556**	**0.558**	0.385	**0.548**	**0.349**

Xenium skin task, the gain is $+0.273$ under the observation model and $+0.252$ under squared error, while no other task in that block moves by more than 0.051 in either direction (the Probe block of Table 14); in each of the four other backbones below it, SKCM again moves more than any other task. On closer inspection, the gain reflects which genes are supervised rather than how well the model predicts them. Of that task's 50 target genes, 17 are measured by no slide the Xenium expert trains on, and since the loss is computed only over the genes a slide measured (§D.1), no gradient ever reaches them while training is restricted to Xenium. They are supervised only once the whole-transcriptome Visium slides enter the pool. The score therefore rises not because the model has generalized from a wider variety of slides, but because genes it could not predict at all have become predictable. Three of the five Xenium tasks hold genes in this position: SKCM with 17, COAD with 3 and IDC with 2. Rescoring both pools on the genes the Xenium expert does supervise separates the two effects.

On SKCM those 33 genes give the expert a correlation of 0.530 rather than 0.349 (Probe under the observation model), and the expert-to-pooled gain falls from $+0.273$ to $+0.103$: across the five backbones and both objectives, between 50 and 84% of the gain measured over the 50 genes is supervision being filled in rather than anything the model learned. On COAD, where 47 of the 50 are supervised, that share is 36 to 57% under the observation model, and on IDC, where 48 are, the gain clears the seed spread in only two of the ten cells, both under the observation model, with shares of 46 and 64%. These three tasks therefore have to be read differently from the rest. The contrast between the objectives, however, survives the rescoring. On the supervised genes of these three tasks, pooling gains more under the observation model than under squared error in 13 of the 15 task and backbone cells, the two exceptions being SKCM on Probe and on Probe+LoRA; on COAD and IDC, pooling lowers the correlation on those genes in 9 of 10 cells under squared error but in only 3 under the observation model. The gap in supervision is, at the same time, what makes pooled

Table 15: Per-task squared error on $\log(1+y)$ over the 50 benchmark genes on the HEST benchmark (lower is better; median standard deviation 0.039).

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE Exp.	**0.321**	0.649	**0.438**	0.570	0.543	**1.565**	2.067	1.219	**0.971**	2.968	1.131
MSE Pool.	0.333	0.709	0.459	0.577	0.557	1.694	2.328	1.527	1.121	2.340	1.165
Obs. Exp.	0.322	0.660	0.496	0.569	0.551	2.609	**1.803**	**1.167**	1.020	2.200	1.140
Obs. Pool.	0.345	**0.639**	0.572	**0.551**	**0.542**	1.679	1.858	1.301	0.982	**1.924**	**1.039**
Probe+LoRA											
MSE Exp.	0.332	0.650	**0.465**	0.567	**0.518**	**1.459**	1.996	1.223	0.954	2.953	1.112
MSE Pool.	**0.330**	0.766	0.530	0.640	0.576	1.545	2.147	1.526	1.041	2.129	1.123
Obs. Exp.	0.342	0.668	0.569	0.559	0.534	2.726	**1.835**	1.134	1.057	2.243	1.167
Obs. Pool.	0.343	**0.639**	0.635	**0.544**	0.522	1.762	1.841	**1.132**	**0.949**	**1.833**	**1.020**
Simple-TF											
MSE Exp.	**0.312**	0.684	0.523	0.609	0.624	2.043	2.024	1.294	1.105	3.236	1.245
MSE Pool.	0.314	0.709	0.522	0.633	**0.567**	**1.518**	1.736	1.240	1.065	3.009	1.131
Obs. Exp.	0.323	**0.650**	0.486	0.594	0.660	1.975	1.857	1.317	1.142	2.237	1.124
Obs. Pool.	0.353	0.652	**0.448**	**0.569**	0.614	1.535	**1.668**	**1.065**	**0.886**	**1.977**	**0.977**
STPath†											
MSE Exp.	0.360	0.866	0.452	0.657	0.771	2.286	1.987	**1.294**	0.996	2.994	1.266
MSE Pool.	0.357	0.847	0.454	0.607	0.969	**1.638**	2.071	1.881	1.366	2.458	1.265
Obs. Exp.	**0.330**	0.881	0.439	0.668	**0.706**	2.694	1.875	1.331	**0.994**	2.144	1.206
Obs. Pool.	0.333	**0.816**	**0.438**	**0.594**	0.813	1.671	**1.777**	1.345	1.130	**1.960**	**1.088**
STORM†											
MSE Exp.	0.330	**0.661**	**0.471**	0.620	**0.528**	2.041	2.209	1.354	1.076	3.199	1.249
MSE Pool.	**0.327**	0.707	0.557	0.614	0.557	**1.468**	2.319	1.533	1.164	2.278	1.153
Obs. Exp.	0.352	0.676	0.508	0.608	0.565	2.249	2.058	1.432	**1.034**	2.383	1.187
Obs. Pool.	0.358	0.693	0.570	**0.581**	0.553	1.635	**1.919**	**1.322**	1.092	**2.206**	**1.093**

training necessary: a platform with a narrow panel offers no path to a gene outside that panel, and no change to the architecture will supply one.

E.2 Other metrics

Five further metrics are reported per task: the squared error on $\log(1 + y)$ over the same 50 genes (Table 15); the correlation and the squared error over every gene shared by the task's test slides, which we call the shared-panel metrics (Tables 16 and 17); and the same two over every gene each slide measured, which we call the slide-panel metrics (Tables 18 and 19). Both leave out the genes outside the gene space of §D.1. A Visium task is then scored on 17,943 to 31,463 genes either way, whereas a Xenium task is scored on 98 to 351 shared genes but on 280 to 480 per slide, because the panels of its slides differ. With both trained on the pooled data, the observation model has the lower squared error over the 50 genes on all five backbones, although it is the other arm that is fitted to reduce error on exactly that scale. The same holds beyond the 50 genes the benchmark selects: on both panel metrics, in correlation and in squared error, the pooled observation model is ahead of the pooled squared-error arm on all five backbones and on 43 to 47 of the 50 task and backbone cells. The two panel metrics differ only where the slides of a task were read with different panels, and of the Visium tasks only CCRCC changes.

The pool contrast depends on the genes scored. Over the 50 genes, pooling is ahead of the expert pool in all ten cells of the two metrics and five backbones under the observation model, and in 7 under squared error. On the panel metrics it is ahead under both objectives on all five backbones, and on the Xenium tasks its gain in correlation is concentrated on two of them: IDC, by $+0.153$ to $+0.185$ under the observation model and $+0.091$ to $+0.168$ under squared error across the two panels, and SKCM, by $+0.242$ to $+0.294$ and $+0.200$ to $+0.259$, against -0.116 to $+0.092$ on the other three. On these two tasks many of the panel genes were never supervised by the slides the Xenium expert trains on, so their gains carry the supervision effect described for SKCM in §E.1: they measure the genes pooling brings within reach rather than how much better it predicts the genes

Table 16: Per-task Pearson correlation over every gene shared by the task's test slides on the HEST benchmark (higher is better; median standard deviation 0.009).

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE Exp.	0.090	0.068	0.080	0.164	0.059	0.517	0.439	**0.603**	0.485	0.399	0.290
MSE Pool.	0.086	0.062	0.069	0.154	0.061	0.452	0.576	0.568	0.431	0.605	0.306
Obs. Exp.	**0.108**	0.080	0.092	0.163	0.083	0.518	0.446	0.592	**0.487**	0.411	0.298
Obs. Pool.	0.101	**0.094**	**0.096**	**0.169**	**0.086**	**0.545**	**0.602**	0.598	0.451	**0.661**	**0.340**
Probe+LoRA											
MSE Exp.	0.107	0.080	0.089	0.177	0.076	0.534	0.454	**0.619**	**0.507**	0.411	0.305
MSE Pool.	0.104	0.086	0.085	0.178	0.074	0.557	**0.623**	0.618	0.491	0.669	0.349
Obs. Exp.	0.110	0.082	0.086	0.171	0.092	0.535	0.450	0.599	0.484	0.416	0.303
Obs. Pool.	**0.114**	**0.097**	**0.100**	**0.183**	**0.096**	**0.573**	0.614	0.616	0.479	**0.687**	**0.356**
Simple-TF											
MSE Exp.	0.110	0.084	0.092	0.167	0.085	0.539	0.446	0.600	**0.487**	0.396	0.301
MSE Pool.	0.099	0.093	0.095	**0.182**	0.084	0.552	0.587	0.577	0.422	0.641	0.333
Obs. Exp.	0.117	0.097	**0.102**	0.169	**0.091**	0.537	0.448	0.599	0.485	0.400	0.304
Obs. Pool.	**0.122**	**0.100**	**0.102**	0.169	0.075	**0.588**	**0.633**	**0.619**	0.474	**0.694**	**0.358**
STPath†											
MSE Exp.	0.081	0.058	0.052	0.125	0.072	0.513	0.452	**0.594**	0.478	0.381	0.280
MSE Pool.	0.059	0.064	**0.070**	0.126	0.048	0.484	0.544	0.514	0.362	0.593	0.287
Obs. Exp.	**0.106**	**0.088**	0.066	0.159	**0.096**	0.506	0.442	0.589	**0.488**	0.394	0.294
Obs. Pool.	0.084	0.081	0.067	**0.160**	0.082	**0.540**	**0.598**	0.571	0.433	**0.656**	**0.327**
STORM†											
MSE Exp.	0.088	0.074	0.077	0.145	0.072	0.476	0.390	**0.550**	**0.454**	0.344	0.267
MSE Pool.	0.070	0.070	0.065	0.148	0.065	0.486	0.509	0.518	0.416	0.588	0.294
Obs. Exp.	**0.105**	**0.083**	**0.082**	0.151	**0.083**	0.461	0.372	0.519	0.429	0.320	0.261
Obs. Pool.	0.102	**0.083**	0.079	**0.169**	0.082	**0.519**	**0.552**	0.538	0.376	**0.601**	**0.310**

the expert already had, and we have not rescored the panel metrics on the supervised genes alone. The objective contrast above is not affected, since both of its arms were trained on the same pooled data.

Table 17: Per-task squared error on $\log(1+y)$ over every gene shared by the task's test slides on the HEST benchmark (lower is better; median standard deviation 0.020).

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE Exp.	0.091	0.157	0.197	**0.170**	**0.059**	1.412	2.782	**0.852**	**0.845**	3.887	1.045
MSE Pool.	0.097	0.169	0.204	0.180	0.062	2.063	2.346	1.113	0.979	2.937	1.015
Obs. Exp.	0.093	0.160	0.190	0.179	0.065	1.655	**1.620**	0.882	0.915	2.456	0.821
Obs. Pool.	**0.086**	**0.152**	**0.182**	0.174	0.061	**1.255**	1.637	0.998	0.850	**2.259**	**0.765**
Probe+LoRA											
MSE Exp.	0.090	0.156	0.190	**0.174**	**0.057**	1.470	2.705	0.860	**0.822**	3.819	1.034
MSE Pool.	0.094	0.189	0.190	0.228	0.066	1.782	1.971	1.153	0.903	2.510	0.909
Obs. Exp.	0.093	0.159	0.186	0.186	0.064	1.658	**1.587**	**0.844**	0.950	2.430	0.816
Obs. Pool.	**0.085**	**0.152**	**0.179**	0.175	0.061	**1.260**	1.616	0.865	**0.822**	**2.123**	**0.734**
Simple-TF											
MSE Exp.	0.087	0.161	0.143	0.207	0.065	1.437	2.752	0.921	0.941	4.095	1.081
MSE Pool.	0.103	0.169	**0.139**	0.239	0.062	1.157	**1.325**	0.890	0.923	2.333	0.734
Obs. Exp.	0.089	0.155	0.162	0.195	0.067	1.342	1.645	0.916	0.977	2.605	0.815
Obs. Pool.	**0.085**	**0.150**	0.148	**0.179**	**0.059**	**1.103**	1.403	**0.837**	**0.768**	**2.109**	**0.684**
STPath†											
MSE Exp.	0.158	0.216	0.293	0.279	0.082	1.572	2.687	**0.908**	0.876	3.975	1.105
MSE Pool.	0.153	0.202	0.254	0.226	0.089	1.938	**1.597**	1.430	1.195	3.131	1.021
Obs. Exp.	0.150	0.215	0.277	0.281	**0.072**	1.603	1.628	0.934	**0.860**	2.564	0.858
Obs. Pool.	**0.137**	**0.194**	**0.250**	**0.213**	0.075	**1.350**	1.619	1.046	0.984	**2.389**	**0.826**
STORM†											
MSE Exp.	0.094	0.165	0.195	0.194	**0.060**	1.570	2.923	**0.992**	0.942	4.120	1.125
MSE Pool.	0.103	0.171	0.199	0.200	0.062	1.903	2.247	1.173	1.010	2.580	0.965
Obs. Exp.	0.091	0.162	0.201	0.195	0.064	1.572	1.814	1.058	**0.905**	2.767	0.883
Obs. Pool.	**0.086**	**0.158**	**0.188**	**0.184**	**0.060**	**1.249**	**1.694**	1.053	0.950	**2.566**	**0.819**

Table 18: Per-task Pearson correlation over every gene the slide measured on the HEST benchmark (higher is better; median standard deviation 0.009).

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE Exp.	0.090	0.059	0.080	0.164	0.059	0.481	0.435	0.583	0.459	0.416	0.283
MSE Pool.	0.086	0.056	0.069	0.154	0.061	0.454	0.570	0.553	0.413	0.616	0.303
Obs. Exp.	**0.108**	0.071	0.092	0.163	0.083	0.482	0.443	0.580	**0.483**	0.428	0.293
Obs. Pool.	0.101	**0.082**	**0.096**	**0.169**	**0.086**	**0.548**	**0.597**	**0.592**	0.462	**0.670**	**0.340**
Probe+LoRA											
MSE Exp.	0.107	0.070	0.089	0.177	0.076	0.499	0.451	0.602	0.485	0.427	0.298
MSE Pool.	0.104	0.077	0.085	0.178	0.074	0.559	**0.618**	0.610	0.483	0.680	0.347
Obs. Exp.	0.110	0.072	0.086	0.171	0.092	0.499	0.447	0.586	0.481	0.432	0.298
Obs. Pool.	**0.114**	**0.087**	**0.100**	**0.183**	**0.096**	**0.576**	0.609	**0.611**	**0.487**	**0.696**	**0.356**
Simple-TF											
MSE Exp.	0.110	0.074	0.092	0.167	0.085	0.500	0.442	0.579	0.462	0.410	0.292
MSE Pool.	0.099	0.082	0.095	**0.182**	0.084	0.553	0.582	0.570	0.431	0.651	0.333
Obs. Exp.	0.117	0.086	**0.102**	0.169	**0.091**	0.501	0.444	0.581	0.471	0.416	0.298
Obs. Pool.	**0.122**	**0.089**	**0.102**	0.169	0.075	**0.593**	**0.627**	**0.613**	**0.481**	**0.702**	**0.357**
STPath†											
MSE Exp.	0.081	0.050	0.052	0.125	0.072	0.479	0.448	**0.569**	0.452	0.400	0.273
MSE Pool.	0.059	0.055	**0.070**	0.126	0.048	0.481	0.539	0.508	0.370	0.602	0.286
Obs. Exp.	**0.106**	**0.079**	0.066	0.159	**0.096**	0.474	0.439	0.568	**0.468**	0.413	0.287
Obs. Pool.	0.084	0.073	0.067	**0.160**	0.082	**0.543**	**0.592**	0.565	0.437	**0.666**	**0.327**
STORM†											
MSE Exp.	0.088	0.065	0.077	0.145	0.072	0.441	0.387	0.533	**0.424**	0.367	0.260
MSE Pool.	0.070	0.061	0.065	0.148	0.065	0.486	0.504	0.505	0.398	0.599	0.290
Obs. Exp.	**0.105**	**0.074**	**0.082**	0.151	**0.083**	0.429	0.370	0.501	0.408	0.338	0.254
Obs. Pool.	0.102	**0.074**	0.079	**0.169**	0.082	**0.520**	**0.547**	**0.537**	0.401	**0.610**	**0.312**

Table 19: Per-task squared error on $\log(1 + y)$ over every gene the slide measured on the HEST benchmark (lower is better; median standard deviation 0.019).

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE Exp.	0.091	0.130	0.197	**0.170**	**0.059**	1.984	2.730	0.988	1.032	3.871	1.125
MSE Pool.	0.097	0.141	0.204	0.180	0.062	2.353	2.299	1.177	1.310	2.925	1.075
Obs. Exp.	0.093	0.135	0.190	0.179	0.065	1.832	**1.602**	**0.975**	1.130	2.443	0.865
Obs. Pool.	**0.086**	**0.126**	**0.182**	0.174	0.061	**1.404**	1.610	1.006	**1.008**	**2.254**	**0.791**
Probe+LoRA											
MSE Exp.	0.090	0.128	0.190	**0.174**	**0.057**	2.016	2.654	0.966	0.991	3.797	1.106
MSE Pool.	0.094	0.161	0.190	0.228	0.066	1.993	1.934	1.179	1.174	2.500	0.952
Obs. Exp.	0.093	0.135	0.186	0.186	0.064	1.820	**1.569**	0.941	1.158	2.409	0.856
Obs. Pool.	**0.085**	**0.126**	**0.179**	0.175	0.061	**1.388**	1.588	**0.878**	**0.967**	**2.101**	**0.755**
Simple-TF											
MSE Exp.	0.087	0.134	0.143	0.207	0.065	2.004	2.705	1.022	1.133	4.107	1.161
MSE Pool.	0.103	0.141	**0.139**	0.239	0.062	1.303	**1.309**	0.930	1.024	2.304	0.755
Obs. Exp.	0.089	0.130	0.162	0.195	0.067	1.572	1.625	0.985	1.117	2.624	0.857
Obs. Pool.	**0.085**	**0.123**	0.148	**0.179**	**0.059**	**1.226**	1.382	**0.857**	**0.942**	**2.096**	**0.710**
STPath†											
MSE Exp.	0.158	0.186	0.293	0.279	0.082	2.117	2.641	**1.005**	**1.069**	3.984	1.181
MSE Pool.	0.153	0.173	0.254	0.226	0.089	2.190	**1.572**	1.508	1.466	3.178	1.081
Obs. Exp.	0.150	0.185	0.277	0.281	**0.072**	1.802	1.608	1.029	1.071	2.571	0.904
Obs. Pool.	**0.137**	**0.163**	**0.250**	**0.213**	0.075	**1.542**	1.592	1.123	1.132	**2.383**	**0.861**
STORM†											
MSE Exp.	0.094	0.138	0.195	0.194	**0.060**	2.143	2.873	1.103	1.146	4.074	1.202
MSE Pool.	0.103	0.143	0.199	0.200	0.062	2.132	2.212	1.299	1.402	**2.559**	1.031
Obs. Exp.	0.091	0.137	0.201	0.195	0.064	1.797	1.788	1.179	1.109	2.784	0.934
Obs. Pool.	**0.086**	**0.130**	**0.188**	**0.184**	**0.060**	**1.396**	**1.665**	**1.075**	**1.058**	2.570	**0.841**

Table 20: The four patch encoders, as their published sources describe them. The last column is the tile each model was pretrained on, not what we feed it: our patches are the same for all four (§F.1). A dash marks a value the source does not state.

Encoder	Architecture	Dim.	Slides	Tiles	Pretraining tile
UNI2-h (Chen et al., 2024b)	ViT-H/14, 681M	1536	>300k	>200M	–
H-optimus-1 (Scalbert et al., 2026)	ViT, 1.1B	1536	>1M	–	224 px, 0.5 μm/px
Prov-GigaPath (Xu et al., 2024)	ViT-g/14	1536	171,189	1.3B	256 px, 0.5 μm/px
PathOrchestra (Yan et al., 2025)	ViT-L/16	1024	287,424	141M	256 px, 20×

Table 21: The benchmark with UNI2-h as the patch encoder. PCC over the 50 benchmark genes, mean$\pm$sd over the five seeds of Table 1. Bold: best per backbone and platform. These are the numbers of Table 1.

		Visium (5 tasks)		Xenium (5 tasks)		All (10 tasks)	
Backbone	Obj.	Expert	Pooled	Expert	Pooled	Expert	Pooled
Probe	MSE	0.170±0.007	0.164±0.009	0.499±0.004	0.524±0.025	0.335±0.005	0.344±0.009
	Obs.	0.186±0.008	**0.199**±0.005	0.496±0.004	**0.556**±0.009	0.341±0.004	**0.378**±0.007
Probe+LoRA	MSE	0.192±0.006	0.194±0.015	0.515±0.005	**0.579**±0.008	0.353±0.003	0.386±0.010
	Obs.	0.195±0.006	**0.215**±0.012	0.497±0.006	0.577±0.005	0.346±0.004	**0.396**±0.004
Simple-TF	MSE	0.204±0.006	0.205±0.020	0.500±0.012	0.526±0.069	0.352±0.003	0.366±0.043
	Obs.	**0.209**±0.006	0.208±0.011	0.501±0.018	**0.584**±0.009	0.355±0.008	**0.396**±0.008
STPath†	MSE	0.158±0.008	0.132±0.019	0.496±0.016	0.480±0.045	0.327±0.009	0.306±0.025
	Obs.	**0.197**±0.011	0.186±0.009	0.497±0.011	**0.549**±0.032	0.347±0.007	**0.367**±0.019
STORM†	MSE	0.174±0.014	0.163±0.011	0.449±0.003	0.491±0.020	0.311±0.007	0.327±0.012
	Obs.	0.182±0.010	**0.190**±0.008	0.431±0.006	**0.507**±0.003	0.306±0.007	**0.349**±0.005

F Other Pathology Foundation Models

We rerun the same experiments with other pathology foundation models in place of UNI2-h. Nothing but the encoder changes, every other condition following §D, and we report the PCC of the HEST benchmark for each encoder. It is not a comparison of encoders, and we make no claim about which is better.

F.1 Encoders

The four encoders are UNI2-h (Chen et al., 2024b), H-optimus-1 (Scalbert et al., 2026), Prov-GigaPath (Xu et al., 2024) and PathOrchestra (Yan et al., 2025). All of them are patch encoders pretrained on whole-slide pathology images, and each one replaces UNI2-h at the same point in the pipeline. The backbones built on frozen embeddings take their embeddings from the new encoder, and Probe+LoRA attaches its adapters to the new encoder instead. They differ in the width of the embedding and in the data they were pretrained on, which Table 20 lists. Patch extraction is identical across the four, which read the same cached patches of 224 pixels at $0.5\,\mu\mathrm{m}$ per pixel, a field of view of $112\,\mu\mathrm{m}$. Only the normalization differs, and it uses each model's published constants: H-optimus-1 has its own and the other three use ImageNet statistics. No patch is resized or cropped, since 224 pixels is already the input size of all four. The published transform of Prov-GigaPath resizes to 256 pixels and crops back to 224, which on our patches would narrow the field of view to $98\,\mu\mathrm{m}$. The platform offset is re-estimated for each encoder (§B.2).

F.2 Benchmark correlation by encoder

Tables 21 to 24 give the benchmark PCC under each encoder. Under squared error the larger pool again brings little, though less sharply than with UNI2-h, where no backbone gains more than 0.002 on the Visium tasks. On the Visium tasks, 16 of the 20 backbone and encoder cells gain less than 0.010 or lose, and STPath† loses on all four encoders. On the Xenium tasks, where the pool holds

Table 22: The benchmark with H-optimus-1 as the patch encoder. PCC over the 50 benchmark genes, mean±sd over the five seeds of Table 1. Bold: best per backbone and platform.

		Visium (5 tasks)		Xenium (5 tasks)		All (10 tasks)	
Backbone	Obj.	Expert	Pooled	Expert	Pooled	Expert	Pooled
Probe	MSE	0.175±0.012	0.176±0.009	0.483±0.008	0.505±0.011	0.329±0.007	0.341±0.009
	Obs.	0.200±0.011	**0.214**±0.002	0.492±0.007	**0.528**±0.012	0.346±0.007	**0.371**±0.006
Probe+LoRA	MSE	0.211±0.008	0.211±0.009	0.512±0.005	**0.569**±0.004	0.361±0.004	0.390±0.006
	Obs.	0.214±0.014	**0.238**±0.008	0.494±0.007	0.563±0.011	0.354±0.010	**0.401**±0.008
Simple-TF	MSE	0.208±0.016	0.227±0.010	0.499±0.018	0.582±0.011	0.353±0.011	0.405±0.002
	Obs.	0.214±0.014	**0.232**±0.004	0.515±0.005	**0.601**±0.004	0.364±0.009	**0.417**±0.004
STPath†	MSE	0.170±0.017	0.144±0.021	0.503±0.013	0.508±0.033	0.336±0.009	0.326±0.018
	Obs.	**0.204**±0.014	0.197±0.012	0.510±0.007	**0.573**±0.015	0.357±0.008	**0.385**±0.013
STORM†	MSE	0.180±0.019	0.194±0.006	0.456±0.006	0.510±0.012	0.318±0.010	0.352±0.007
	Obs.	0.203±0.007	**0.209**±0.006	0.458±0.014	**0.534**±0.007	0.331±0.009	**0.372**±0.005

Table 23: The benchmark with Prov-GigaPath as the patch encoder. PCC over the 50 benchmark genes, mean±sd over the five seeds of Table 1. Bold: best per backbone and platform.

		Visium (5 tasks)		Xenium (5 tasks)		All (10 tasks)	
Backbone	Obj.	Expert	Pooled	Expert	Pooled	Expert	Pooled
Probe	MSE	0.147±0.009	0.147±0.007	0.474±0.004	0.489±0.010	0.310±0.004	0.318±0.008
	Obs.	0.185±0.002	**0.191**±0.010	0.476±0.008	**0.525**±0.005	0.330±0.005	**0.358**±0.005
Probe+LoRA	MSE	0.183±0.008	0.185±0.015	0.495±0.005	**0.544**±0.005	0.339±0.005	0.364±0.009
	Obs.	0.191±0.005	**0.208**±0.010	0.481±0.010	0.542±0.008	0.336±0.007	**0.375**±0.007
Simple-TF	MSE	0.186±0.009	**0.213**±0.003	0.493±0.011	0.563±0.003	0.340±0.010	0.388±0.003
	Obs.	0.197±0.005	0.212±0.014	0.496±0.007	**0.565**±0.009	0.346±0.006	**0.389**±0.011
STPath†	MSE	0.156±0.008	0.122±0.021	0.493±0.014	0.499±0.015	0.324±0.009	0.311±0.013
	Obs.	**0.175**±0.025	0.170±0.012	0.489±0.009	**0.556**±0.016	0.332±0.013	**0.363**±0.011
STORM†	MSE	0.151±0.008	0.158±0.003	0.462±0.005	0.507±0.012	0.307±0.006	0.333±0.006
	Obs.	0.162±0.009	**0.184**±0.006	0.455±0.006	**0.531**±0.005	0.308±0.004	**0.358**±0.005

11.4 times as many slides, 18 of the 20 cells gain. With both trained on the pooled data, the observation model is ahead of squared error on the ten-task mean in 19 of the 20 cells. The exception is Simple-TF on PathOrchestra (-0.003), and Simple-TF on Prov-GigaPath is ahead by only 0.001. The other foundation models therefore show the same pattern as UNI2-h.

G Downstream Validation, Full Results

Table 2 reduces each of the three downstream evaluations, TLS detection, spatial clustering and prediction on a new platform, to one or two numbers. This section gives the full results behind them and states how each score is computed.

G.1 TLS detection

The TLS score of a spot is computed from where a signature's genes rank within the predicted expression of that spot. We rank the genes of the spot by predicted expression, sum the ranks of the signature genes, and normalize by the size of the signature and of the background. The score therefore depends only on the ordering of genes within a spot and not on the absolute level of expression. Spots annotated as TLS are positive and spots annotated as not TLS are negative; spots carrying neither annotation are excluded from scoring. We use two signatures over the gene space of §D.2, the 29 genes of Meylan (Meylan et al., 2022) and a 12-chemokine signature (Xu et al.,

Table 24: The benchmark with PathOrchestra as the patch encoder. PCC over the 50 benchmark genes, mean$\pm$sd over the five seeds of Table 1. Bold: best per backbone and platform.

		Visium (5 tasks)		Xenium (5 tasks)		All (10 tasks)	
Backbone	Obj.	Expert	Pooled	Expert	Pooled	Expert	Pooled
Probe	MSE	0.152±0.015	0.147±0.006	0.457±0.003	0.485±0.009	0.304±0.008	0.316±0.006
	Obs.	0.159±0.012	**0.176**±0.004	0.457±0.006	**0.510**±0.004	0.308±0.008	**0.343**±0.003
Probe+LoRA	MSE	0.177±0.014	0.181±0.012	0.482±0.005	**0.539**±0.004	0.330±0.007	0.360±0.008
	Obs.	0.174±0.009	**0.205**±0.008	0.469±0.005	0.536±0.008	0.322±0.005	**0.370**±0.005
Simple-TF	MSE	0.176±0.006	**0.201**±0.009	0.469±0.002	0.538±0.017	0.323±0.003	**0.370**±0.009
	Obs.	0.187±0.012	0.180±0.012	0.471±0.006	**0.554**±0.006	0.329±0.005	0.367±0.009
STPath†	MSE	0.139±0.006	0.136±0.013	0.455±0.009	0.444±0.015	0.297±0.006	0.290±0.007
	Obs.	0.167±0.004	**0.175**±0.008	0.464±0.010	**0.495**±0.024	0.315±0.007	**0.335**±0.013
STORM†	MSE	0.149±0.009	0.154±0.003	0.429±0.008	0.488±0.008	0.289±0.003	0.321±0.005
	Obs.	0.168±0.004	**0.176**±0.007	0.430±0.007	**0.508**±0.007	0.299±0.003	**0.342**±0.006

2022), on the 17 ccRCC slides of §C.3. The background they are ranked against is every gene the slide measured among those the model's training slides of the Visium family supervised, which is slightly narrower for the Visium expert than for the pooled model (§G.2). The ground truth of the tables, GT, is the same procedure run on the measured expression of those slides over the same background: the score a model would be given for reproducing the measurement exactly. It moves with the background by less than 0.001, and the tables give the pooled model's. It is a scale rather than an upper bound, since it says how well the signature separates TLS in the measurement itself, and on the chemokine signature the observation model exceeds it. We also report a control in which the score is computed from a random gene set of the same size. A control far from 0.5 would mean the score came from the procedure rather than from the content of the signature, and that row could not be trusted.

Table 25 gives the four regimes on the two signatures. Every control sits between 0.489 and 0.514, so no row of the table is carrying a number produced by the procedure itself. Under squared error, pooling has no consistent sign: it moves the AUROC by -0.015 to $+0.055$ on the Meylan signature and by -0.045 to $+0.036$ on the chemokine one. Under the observation model it does. Pooling helps on all five backbones with the Meylan signature, by $+0.013$ to $+0.058$, and on four of the five with the chemokine one, STPath† being the exception at -0.005. The pooled observation model is accordingly the best of the four regimes on every backbone with the Meylan signature and on four of the five with the chemokine one. The two signatures differ in where they sit against GT. No regime reaches the Meylan ground truth of 0.821, whereas on the chemokine signature every pooled observation model is above its ground truth of 0.726, as are four of the five expert ones.

G.2 Spatial clustering

The predicted expression of a slide is reduced by principal components, clustered by Leiden, and compared with the annotation by adjusted mutual information (AMI) and homogeneity. Every hyperparameter of that pipeline is held at one setting for all slides and all models: 50 principal components, a 30-nearest-neighbor graph on the first 30 of them, and Leiden at resolution 1.0. The clustering reads every gene supervised by the model's training slides of the Visium family, the four platforms Table 9 groups under that name, so that the pooled model is compared with the Visium expert within the platform family the cohorts are read on. That is $38{,}584$ genes for the pooled model, all of its $38{,}605$ but 21 targets found only on Xenium panels, and $34{,}968$ for the Visium expert, which trained on Visium slides alone and so never fitted the genes measured only on the other three. The expert reads every slide as Visium, the eight legacy Spatial Transcriptomics slides of the breast cohort included, since Visium is the one platform whose terms it has fitted; the pooled model reads each slide as its own platform.

The three cohorts are not annotated in the same sense. The prostate slides are labeled benign, Gleason group and stroma, and the breast slides invasive cancer, cancer in situ and immune infiltrate,

Table 25: TLS detection on the 17 ccRCC slides; AUROC of the signature score, mean$\pm$sd over five seeds. *: the Visium expert. *GT* is the same procedure run on the measured expression over the pooled model's genes; over the expert's it moves by less than 0.001. *Control* is the score from a random gene set of the same size. Bold: the best AUROC per backbone and signature.

		Meylan (GT 0.821)		Chemokine (GT 0.726)	
Backbone	Obj.	AUROC	Control	AUROC	Control
Probe	MSE*	0.7402$\pm$0.0177	0.5035	0.7094$\pm$0.0238	0.5013
	MSE	0.7252$\pm$0.0202	0.4977	0.6644$\pm$0.0152	0.4990
	Obs.*	0.7624$\pm$0.0104	0.5015	0.7318$\pm$0.0118	0.5120
	Obs.	**0.7893**$\pm$0.0148	0.5067	**0.7405**$\pm$0.0090	0.4913
Probe+LoRA	MSE*	0.7655$\pm$0.0141	0.4983	0.7265$\pm$0.0392	0.5074
	MSE	0.7657$\pm$0.0070	0.5014	0.7240$\pm$0.0173	0.4958
	Obs.*	0.7800$\pm$0.0077	0.5046	0.7334$\pm$0.0139	0.5134
	Obs.	**0.7933**$\pm$0.0067	0.5075	**0.7484**$\pm$0.0138	0.4963
Simple-TF	MSE*	0.7680$\pm$0.0189	0.5018	0.6839$\pm$0.0360	0.4996
	MSE	0.7787$\pm$0.0164	0.5025	0.7189$\pm$0.0291	0.4952
	Obs.*	0.7396$\pm$0.0249	0.5046	0.7215$\pm$0.0295	0.4888
	Obs.	**0.7978**$\pm$0.0097	0.5037	**0.7567**$\pm$0.0141	0.4928
STPath†	MSE*	0.7147$\pm$0.1184	0.5043	0.6961$\pm$0.1475	0.4955
	MSE	0.7696$\pm$0.0228	0.4990	0.7318$\pm$0.0297	0.4937
	Obs.*	0.7698$\pm$0.0387	0.5112	**0.7403**$\pm$0.0469	0.5033
	Obs.	**0.8071**$\pm$0.0173	0.5138	0.7356$\pm$0.0266	0.4896
STORM†	MSE*	0.7515$\pm$0.0380	0.5041	0.6942$\pm$0.0308	0.5100
	MSE	0.7564$\pm$0.0180	0.4935	0.7074$\pm$0.0307	0.4952
	Obs.*	0.7770$\pm$0.0235	0.5006	0.7370$\pm$0.0307	0.5038
	Obs.	**0.7987**$\pm$0.0159	0.5023	**0.7528**$\pm$0.0162	0.4892

whereas the 12 DLPFC sections are labeled by cortical layer, L1 to L6, and the white matter. The first two are defined by lesion and the third by morphology, so this section also reports the three cohorts separately.

Table 26 gives the four regimes on the three cohorts. We read it mainly as a comparison of the two objectives trained on the same pool. With both trained on the pooled data, the observation model has the higher AMI in 13 of the 15 backbone and cohort cells and the higher homogeneity in 12. On AMI it leads on every backbone in the prostate and breast cohorts. The DLPFC sections are the exception: there it leads on three of the five backbones, and on STPath† and STORM† it sits below the squared-error model, at 0.257 against 0.275 and at 0.301 against 0.311. Under squared error, pooling itself has no consistent sign: the pooled model's AMI is above the expert's on 9 of the 15 cells, but on the prostate cohort on only one of the five, by less than 0.001.

The same contrast, read slide by slide on AMI, agrees with this. Comparing the pooled observation model with the pooled squared-error model on the same slide, it is ahead on 23 of the 27 slides on each of Probe, Simple-TF and Probe+LoRA, the counts §5.2 quotes, and on 14 and 12 on STPath† and STORM†. Per cohort they follow the split above: on Probe the contrast is won on all 7 prostate slides and all 8 breast slides but on 8 of the 12 DLPFC sections, and on STPath† and STORM† the DLPFC counts fall to 3 and to 2 of the 12. Those are the two backbones whose pooled observation model sits below squared error on that cohort in the table above, and the three where it leads take 8, 11 and 9 of the 12.

G.3 A NEW PLATFORM

Atera measures the whole transcriptome and reads it out by imaging, and each property fixes one choice this evaluation cannot make freely. The first fixes which expert it can be asked of, the Visium one, because an expert trained on Xenium alone predicts its own panel and would leave most of this slide's genes unfitted. That expert is told to assume Visium at inference. Its offset then stays at zero, the value it was trained with; the dispersion and the spread of the slide term are read from the

Table 26: Spatial clustering on the three cohorts, scored against the annotation by adjusted mutual information and homogeneity; mean$\pm$sd over five seeds. *: the Visium expert. Bold: best per backbone and column.

	Prostate (7)		Breast (8)		DLPFC (12)	
Obj.	AMI	Homogeneity	AMI	Homogeneity	AMI	Homogeneity
Probe						
MSE*	0.3187±0.0079	0.5320±0.0106	0.1983±0.0032	0.2818±0.0041	0.1201±0.0041	0.1473±0.0049
MSE	0.3191±0.0064	0.5304±0.0090	0.2276±0.0038	0.3185±0.0073	0.1268±0.0062	0.1551±0.0070
Obs.*	0.3237±0.0034	0.5492±0.0041	0.1967±0.0093	0.2826±0.0115	0.1180±0.0037	0.1434±0.0049
Obs.	**0.3456**±0.0067	**0.5853**±0.0091	**0.2504**±0.0026	**0.3563**±0.0051	**0.1307**±0.0072	**0.1607**±0.0100
Probe+LoRA						
MSE*	0.3297±0.0150	0.5573±0.0186	0.1910±0.0128	0.2721±0.0135	0.1231±0.0111	0.1513±0.0136
MSE	0.3210±0.0060	0.5424±0.0074	0.2187±0.0091	0.3152±0.0156	0.1309±0.0093	0.1593±0.0100
Obs.*	0.3424±0.0081	0.5835±0.0087	0.1981±0.0128	0.2791±0.0164	0.1132±0.0071	0.1389±0.0091
Obs.	**0.3540**±0.0068	**0.6076**±0.0094	**0.2402**±0.0066	**0.3458**±0.0067	**0.1373**±0.0079	**0.1689**±0.0095
Simple-TF						
MSE*	0.3023±0.0115	0.5248±0.0080	0.1811±0.0093	0.2788±0.0090	0.1163±0.0133	0.1467±0.0162
MSE	0.2883±0.0366	0.5375±0.0316	0.2305±0.0332	0.3607±0.0329	0.1222±0.0425	0.1635±0.0457
Obs.*	0.2833±0.0093	0.5109±0.0106	0.1737±0.0078	0.2749±0.0105	0.0858±0.0077	0.1115±0.0086
Obs.	**0.3092**±0.0157	**0.5522**±0.0326	**0.2641**±0.0146	**0.3989**±0.0220	**0.1482**±0.0309	**0.1924**±0.0391
STPath†						
MSE*	0.2978±0.0393	0.5432±0.0546	0.2023±0.0211	0.2950±0.0201	**0.2791**±0.0217	0.3528±0.0201
MSE	0.2742±0.0301	0.5352±0.0269	0.2234±0.0202	**0.3464**±0.0160	0.2752±0.0293	**0.3823**±0.0361
Obs.*	0.2750±0.0349	0.5276±0.0413	0.1705±0.0146	0.2705±0.0146	0.2582±0.0305	0.3530±0.0442
Obs.	**0.3085**±0.0138	**0.5653**±0.0150	**0.2246**±0.0114	0.3403±0.0170	0.2572±0.0192	0.3476±0.0269
STORM†						
MSE*	**0.2715**±0.0166	0.5181±0.0092	0.2718±0.0116	0.4248±0.0155	0.2836±0.0189	0.4005±0.0401
MSE	0.2533±0.0034	0.5048±0.0107	0.2671±0.0104	0.4346±0.0221	**0.3109**±0.0121	**0.4749**±0.0282
Obs.*	0.2506±0.0090	0.5018±0.0151	0.2615±0.0101	0.4199±0.0129	0.2810±0.0156	0.4149±0.0277
Obs.	0.2558±0.0076	**0.5244**±0.0086	**0.2788**±0.0072	**0.4647**±0.0069	0.3014±0.0058	0.4515±0.0086

Visium rows; and on the two backbones that take a platform token it receives the padding token, as in every other evaluation (§D.3). The area term alone is set from this slide rather than from Visium: a bin here is the $100\,\mu\mathrm{m}$ square of §C.3, 4.2 times the area of a Visium spot, and the area term takes the log of that ratio (§B.1). It could be told Xenium instead, since the offset table is estimated apart from any model, but it never fitted a Xenium dispersion, and assuming Visium keeps it within what it saw in training. The second property fixes the platform the pooled model is told to assume, Xenium, so that c_{gp} and the other platform terms of Equation 1 are read as the Xenium ones, and the pooled Simple-TF and STPath† receive the Xenium token they were trained with. Wu et al. (2026) reports Atera to be more sensitive than Xenium, so the offset that ought to be used may be larger than any we estimated. Estimating it would take measurements from Atera itself, which this evaluation withholds. We therefore use the largest of the offsets we did estimate, which is also the one belonging to the imaging platform of the same family. That offset is estimated for only $1{,}539$ of the slide's $18{,}023$ genes (§B.2); the other $16{,}484$ are read with an offset of zero and the area term alone, the same term the expert receives on every gene.

Tables 27 and 28 give the four regimes on the two slides. On the correlation the pooled observation model is the highest of the four in every cell, over all genes and over the 50 most variable alike. Pooling raises the all-gene correlation over the expert on all twenty cells under either objective, by less than 0.001 on one of them. The observation model is above squared error on each sample separately, by 10 to 28% on the breast sample and by 11 to 25% on the cervical one. The squared error separates the regimes less cleanly. Over the 50 most variable genes the pooled observation model is again the lowest on eight of the ten cells, and the pooled squared-error model on the other two, Simple-TF and STPath† on the breast slide. Over all genes the lowest is the pooled squared-error model on five cells, the pooled observation model on three and the expert observation model

Table 27: The breast slide of Atera. PCC and squared error over the slide's 18,023 measured genes and over its 50 most variable; mean±sd over five seeds. *: the Visium expert, read as Visium where the pooled model is read as Xenium, both at the area of the Atera bin. Bold: best per backbone and column.

		All genes		Top 50 by variance	
Backbone	Obj.	PCC	MSE	PCC	MSE
Probe	MSE*	0.4942±0.0226	2.0235±0.0867	0.7298±0.0609	10.8854±0.4183
	MSE	0.4948±0.0561	1.6282±0.0805	0.8132±0.0209	7.3062±0.1116
	Obs.*	0.5419±0.0214	**1.2683**±0.0475	0.7518±0.0556	7.6248±0.3740
	Obs.	**0.6349**±0.0054	1.3228±0.0589	**0.8667**±0.0123	**6.1729**±0.3865
Probe+LoRA	MSE*	0.5383±0.0130	2.1153±0.0511	0.7549±0.0278	11.3823±0.1799
	MSE	0.6071±0.0232	1.2900±0.0873	0.8793±0.0259	6.4039±0.3481
	Obs.*	0.5624±0.0212	1.3260±0.0741	0.7549±0.0461	7.7271±0.4280
	Obs.	**0.6692**±0.0032	**1.2554**±0.0313	**0.8895**±0.0068	**5.5747**±0.2927
Simple-TF	MSE*	0.5188±0.0381	1.8248±0.1508	0.8033±0.0272	8.2731±0.8571
	MSE	0.6120±0.0139	**0.9063**±0.1094	0.9041±0.0137	**4.0679**±0.2634
	Obs.*	0.6093±0.0193	1.0948±0.0909	0.8241±0.0193	5.0335±0.4248
	Obs.	**0.6928**±0.0051	1.0477±0.0354	**0.9173**±0.0038	4.1020±0.1780
STPath†	MSE*	0.4336±0.0624	2.4786±0.1025	0.7611±0.0234	12.0728±0.5630
	MSE	0.5943±0.0179	**0.8264**±0.0889	0.8977±0.0074	**4.2000**±0.4952
	Obs.*	0.5876±0.0252	1.9323±0.1424	0.8183±0.0396	9.0828±0.8123
	Obs.	**0.6666**±0.0128	1.1931±0.1242	**0.8986**±0.0205	4.7868±0.4337
STORM†	MSE*	0.4381±0.0394	2.2855±0.0757	0.7115±0.0318	11.7442±0.2322
	MSE	0.5483±0.0168	**1.3083**±0.1361	0.8005±0.0268	6.0146±0.5045
	Obs.*	0.5425±0.0397	1.4289±0.2741	0.7407±0.0407	8.1228±1.1826
	Obs.	**0.6255**±0.0089	1.3147±0.0940	**0.8700**±0.0096	**5.5020**±0.3688

on two. The two squared-error columns are not on one scale and should be read down rather than across, since the most variable genes carry the larger counts.

H Ablation, Full Results

Table 3 removes the terms of Equation 1 in two groups. The *platform group* is everything indexed by the platform of a spot: the aggregation area A_s, the platform offset c_{gp}, and the platform axis of the dispersion δ_p. The *slide group* is the slide intercept b_{sg} alone. Table 3 reports only the ten-task mean, for reasons of space; this section gives the ten tasks behind that mean, for every backbone.

H.1 The count-likelihood arms, task by task

Table 29 gives the ten tasks behind plain MSE, the reference, and each of the four count-likelihood rows of Table 3. Against plain MSE the full model is higher on 44 of the 50 task-backbone pairs, 23 of the 25 on Visium and 21 of the 25 on Xenium, and on every task on Probe and on STPath†. The likelihood alone accounts for little of this. NB is above MSE on only 32 of the 50, and on Simple-TF it is below MSE on all five Xenium tasks. The terms carry the rest. The full model is above NB on all ten tasks on Probe, nine on STPath† and on STORM†, eight on Simple-TF, and seven on Probe+LoRA, 43 of the 50 in all. Six of the seven cells where it falls short fall short by 0.012 or less; the one sizable exception is PRAD on Simple-TF, where NB reads 0.322 and the full model 0.243.

Removing either group costs the full model across tasks rather than on a few. The full model is above the arm without the platform group on 34 of the 50 pairs, and above the arm without the slide group on 30; the slide group counts for more on Xenium, 18 of the 25, than on Visium, 12 of the 25. Two backbones depart from this. On Probe+LoRA the platform group adds nothing on Xenium, and the arm without it is above the full model on all five Xenium tasks. On STPath† the slide group adds

Table 28: The cervix slide of Atera. PCC and squared error over the slide's 18,023 measured genes and over its 50 most variable; mean±sd over five seeds. *: the Visium expert, read as Visium where the pooled model is read as Xenium, both at the area of the Atera bin. Bold: best per backbone and column.

		All genes		Top 50 by variance	
Backbone	Obj.	PCC	MSE	PCC	MSE
Probe	MSE*	0.3411±0.0151	2.4516±0.1448	0.5894±0.0343	11.2880±0.3684
	MSE	0.4041±0.0252	1.7532±0.0816	0.7202±0.0278	7.3682±0.2677
	Obs.*	0.4095±0.0184	1.3298±0.0756	0.6204±0.0435	8.4165±0.3035
	Obs.	**0.5065**±0.0039	**1.3046**±0.0294	**0.8021**±0.0089	**5.6201**±0.2355
Probe+LoRA	MSE*	0.3887±0.0082	2.5322±0.0576	0.6027±0.0636	11.5848±0.3069
	MSE	0.4765±0.0138	1.2993±0.1125	0.7603±0.0334	7.1428±0.4826
	Obs.*	0.4184±0.0088	1.3499±0.0284	0.6529±0.0396	8.3491±0.1474
	Obs.	**0.5293**±0.0045	**1.2176**±0.0166	**0.8249**±0.0075	**5.3173**±0.4169
Simple-TF	MSE*	0.3634±0.0238	2.3089±0.2829	0.6642±0.0351	8.5043±1.0252
	MSE	0.4742±0.0279	**1.1861**±0.1040	0.8194±0.0184	4.5480±0.4023
	Obs.*	0.3805±0.0159	1.4012±0.1267	0.6660±0.0904	5.9348±0.7960
	Obs.	**0.5263**±0.0075	1.3507±0.0351	**0.8408**±0.0061	**3.7697**±0.4167
STPath†	MSE*	0.2736±0.0448	3.2465±0.1371	0.5827±0.0801	13.6573±0.6153
	MSE	0.4256±0.0255	**1.2356**±0.3219	0.7406±0.0352	7.4514±1.3916
	Obs.*	0.3093±0.0801	2.7866±0.0578	0.6451±0.0492	11.6437±0.4983
	Obs.	**0.4869**±0.0105	1.3828±0.1297	**0.8158**±0.0220	**5.6455**±1.3266
STORM†	MSE*	0.2772±0.0273	2.5719±0.0993	0.4776±0.1315	12.0820±0.2099
	MSE	0.3993±0.0469	1.4385±0.1777	0.7193±0.0431	6.3111±0.5765
	Obs.*	0.3347±0.0329	**1.4191**±0.1575	0.6223±0.0344	8.5665±0.6882
	Obs.	**0.4519**±0.0021	1.4396±0.0518	**0.7527**±0.0143	**5.6119**±0.5621

nothing anywhere, and the arm without it is above the full model on all ten tasks. Both margins are small: on the ten-task mean they are 0.002 and 0.008, within one standard deviation as §5.3 notes.

H.2 The same terms under a squared error

The observation model is written for a count likelihood. We therefore also asked whether its terms can be built into ordinary MSE training, the setting of the reference row in Table 3, and still help there. The slide intercept b_{sg} is added to the head's output, which under this loss is a value in $\log(1+y)$ space rather than a log mean, so the term moves the prediction by what it would have moved $\log \mu$. The platform offset c_{gp} and the aggregation area A_s cannot sit as an offset on a log mean the loss does not have, so they are applied to the *target*: each spot's measured expression is put on the common scale before the error is taken, and the prediction is returned to the platform's scale before it is scored. The platform axis of the dispersion δ_p has no counterpart, since a squared error has no shape parameter, and it is the one term that does not cross over. The squared-error arms therefore carry the same terms as the count-likelihood arms of §H.1, except δ_p.

Table 30 gives the ten tasks for the five squared-error arms. Inside MSE training the slide term helps, the platform terms help on Visium but not on Xenium on four of the five backbones, and neither brings MSE up to the observation model. The slide term carries over intact: it raises the ten-task mean on every backbone, from 0.344 to 0.357 on Probe and from 0.306 to 0.328 on STPath†, and it raises seven to nine of the ten tasks on each. MSE with the slide term is the best squared-error arm on all five backbones, yet it stays below the full observation model on all five (Table 29). Applied to the target, the platform terms raise four or five of the five Visium tasks on four backbones, but on three of those they raise none of the five Xenium tasks, and on those three they lower the ten-task mean. What differs from the count side is where the terms are placed. Applied to the target, they change what each arm is fitted to, while every arm is scored on the same measured counts; in the observation model the same terms enter as an offset on the mean and leave the target untouched. Normalizing the target is no substitute for either term: MSE (CPM) is below plain MSE on four of

Table 29: Per-task Pearson correlation for plain MSE and the count-likelihood arms of Table 3, with all five backbones in one table. Rows are MSE, the reference, and the four count-likelihood arms; bold marks the best of the five (higher is better). Means over five seeds; the standard deviations are omitted for space (median 0.016 over the cells here). Tasks are named as the benchmark names them, except that LYMPH_IDC is written LYMPH.

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE	0.182	0.139	0.062	0.195	0.245	0.433	0.572	0.575	0.443	0.596	0.344
NB	0.203	0.162	0.076	0.197	0.229	0.480	0.577	0.585	0.455	0.620	0.358
Obs. w/o platform	**0.228**	0.174	0.070	**0.218**	0.266	**0.494**	0.590	0.596	0.454	**0.629**	0.372
Obs. w/o slide	0.227	0.170	**0.087**	0.203	0.259	0.485	0.596	0.594	**0.474**	0.624	0.372
Obs.	0.227	**0.181**	0.086	0.213	**0.287**	0.490	**0.602**	**0.605**	0.463	0.622	**0.378**
Probe+LoRA											
MSE	0.229	0.176	0.080	0.224	0.260	0.515	0.611	0.622	0.497	0.650	0.386
NB	0.235	0.183	0.083	0.222	0.285	0.519	0.615	0.625	**0.499**	0.640	0.391
Obs. w/o platform	**0.251**	0.189	0.087	0.227	0.306	**0.528**	**0.623**	**0.627**	0.488	**0.654**	**0.398**
Obs. w/o slide	0.232	**0.192**	**0.094**	**0.241**	**0.317**	0.515	0.612	0.620	0.486	0.639	0.395
Obs.	0.244	0.190	0.093	0.238	0.309	0.525	0.613	0.620	0.487	0.641	0.396
Simple-TF											
MSE	0.214	0.184	0.088	0.236	0.305	0.490	0.570	0.585	0.427	0.557	0.366
NB	0.176	**0.192**	0.089	0.226	**0.322**	0.448	0.523	0.530	0.375	0.498	0.338
Obs. w/o platform	**0.260**	0.184	0.095	0.225	0.288	0.507	0.592	0.605	0.465	0.630	0.385
Obs. w/o slide	0.237	0.187	0.100	**0.247**	0.303	0.485	0.577	0.587	0.441	0.613	0.378
Obs.	0.258	0.189	**0.117**	0.233	0.243	**0.532**	**0.625**	**0.629**	**0.486**	**0.647**	**0.396**
STPath†											
MSE	0.158	0.125	0.053	0.152	0.170	0.428	0.536	0.521	0.365	0.551	0.306
NB	0.202	0.122	0.047	0.197	0.265	0.465	0.541	0.525	0.395	0.590	0.335
Obs. w/o platform	0.213	0.142	0.054	**0.227**	**0.281**	0.466	0.552	0.516	0.403	0.591	0.344
Obs. w/o slide	**0.250**	**0.167**	**0.076**	**0.227**	0.260	**0.511**	**0.597**	**0.583**	**0.447**	**0.632**	**0.375**
Obs.	0.244	0.147	0.064	0.215	0.259	0.505	0.595	0.581	0.439	0.625	0.367
STORM†											
MSE	0.174	0.148	0.060	0.176	0.255	0.459	0.518	0.530	**0.422**	0.525	0.327
NB	0.208	0.135	0.063	0.176	0.254	0.471	0.511	0.518	0.396	0.546	0.328
Obs. w/o platform	0.223	0.143	0.066	0.191	0.274	0.481	0.526	0.532	0.398	**0.556**	0.339
Obs. w/o slide	0.221	0.137	0.068	0.210	0.269	0.482	0.537	0.545	0.380	0.535	0.338
Obs.	**0.224**	**0.160**	**0.069**	**0.213**	**0.285**	**0.488**	**0.556**	**0.558**	0.385	0.548	**0.349**

the five backbones, and on those four it is below on seven or more of the ten tasks, which is the statement §5.3 makes on the ten-task mean.

I Extended Related Work

I.1 Predicting expression from histology

Predicting expression from H&E has grown into a broad literature, and the directions it has taken can be separated from one another. ST-Net (He et al., 2020) set the initial form of the task: a convolutional network takes the patch under a spot and regresses that spot's expression, one spot at a time. Most of what followed widens the context a spot is allowed to see. One line lays a graph over the spots of a section and propagates along it, as in MERGE (Ganguly et al., 2025), which builds a hierarchical graph of spots, and M2TGLGO (Shi et al., 2025). A second line relates the spots of a section through attention, as in HisToGene (Pang et al., 2021) and FEAST (Jeong et al., 2026), and hybrids of the two have followed as well, as in Hist2ST (Zeng et al., 2022), which passes each patch through a Convmixer, a Transformer and a graph network in turn. A third line changes the scale at which features are read rather than their support, fusing representations taken at

Table 30: The same for the squared-error arms. MSE (CPM) normalizes the target by the spot's own library size and sits below the plain MSE it is usually put in place of on four of the five backbones (§H.2). Bold marks the best of the five (higher is better). Means over five seeds; the standard deviations are omitted for space (median 0.017 over the cells here).

	Visium					Xenium					Mean
	READ	CCRCC	HCC	LYMPH	PRAD	COAD	IDC	LUNG	PAAD	SKCM	All
Probe											
MSE	0.182	0.139	0.062	0.195	0.245	0.433	0.572	0.575	0.443	0.596	0.344
MSE (CPM)	**0.217**	0.116	0.067	0.180	0.238	0.418	0.529	0.477	0.400	0.605	0.325
MSE + slide	0.173	0.149	0.073	0.191	0.261	**0.448**	**0.587**	**0.597**	**0.463**	**0.628**	**0.357**
MSE + platform	0.198	0.159	0.081	0.191	0.284	0.408	0.533	0.514	0.389	0.577	0.334
MSE + both	0.202	**0.167**	**0.083**	**0.199**	**0.285**	0.399	0.535	0.529	0.383	0.585	0.337
Probe+LoRA											
MSE	0.229	0.176	0.080	0.224	0.260	0.515	0.611	0.622	**0.497**	**0.650**	0.386
MSE (CPM)	0.244	0.165	0.081	0.214	0.261	0.497	0.571	0.573	0.479	0.635	0.372
MSE + slide	0.231	0.190	0.082	0.238	0.280	**0.520**	**0.616**	**0.630**	0.496	0.640	**0.392**
MSE + platform	0.246	0.198	0.093	**0.251**	0.310	0.439	0.560	0.560	0.430	0.594	0.368
MSE + both	**0.256**	**0.202**	**0.095**	0.249	**0.324**	0.448	0.563	0.565	0.433	0.573	0.371
Simple-TF											
MSE	0.214	0.184	0.088	**0.236**	**0.305**	0.490	0.570	0.585	0.427	0.557	0.366
MSE (CPM)	0.166	0.175	0.081	0.224	0.303	0.417	0.476	0.472	0.339	0.438	0.309
MSE + slide	0.235	0.184	0.101	0.215	0.290	**0.507**	**0.609**	**0.608**	**0.470**	**0.624**	**0.384**
MSE + platform	0.252	0.181	**0.113**	0.219	0.261	0.464	0.581	0.561	0.448	0.586	0.367
MSE + both	**0.259**	**0.191**	0.105	0.224	0.244	0.463	0.586	0.559	0.442	0.589	0.366
STPath†											
MSE	0.158	0.125	0.053	0.152	0.170	0.428	0.536	0.521	0.365	0.551	0.306
MSE (CPM)	0.187	**0.152**	**0.080**	0.177	**0.270**	0.402	0.511	0.487	0.370	0.536	0.317
MSE + slide	0.174	0.114	0.055	0.197	0.214	**0.465**	**0.550**	**0.532**	**0.390**	**0.588**	**0.328**
MSE + platform	0.226	0.135	0.055	0.197	0.197	0.394	0.533	0.476	0.368	0.545	0.313
MSE + both	**0.238**	0.129	0.057	**0.209**	0.228	0.394	0.547	0.492	0.384	0.543	0.322
STORM†											
MSE	0.174	0.148	0.060	0.176	0.255	0.459	0.518	0.530	**0.422**	0.525	0.327
MSE (CPM)	0.178	0.136	**0.070**	0.137	0.254	0.431	0.446	0.376	0.327	0.500	0.286
MSE + slide	**0.194**	0.169	0.060	0.187	0.272	**0.468**	**0.532**	**0.548**	0.411	**0.562**	**0.340**
MSE + platform	**0.194**	0.159	0.060	0.184	**0.282**	0.413	0.506	0.513	0.387	0.489	0.319
MSE + both	0.188	**0.177**	0.067	**0.207**	0.277	0.405	0.510	0.507	0.372	0.514	0.323

several magnifications, as in TRIPLEX (Chung et al., 2024) and M2OST (Wang et al., 2025), which regresses one spot from several magnifications at once.

Other work departs from regression itself. BLEEP (Xie et al., 2023) aligns patch and expression embeddings by contrastive learning and answers a query spot by retrieval rather than by a regressed value, and mclSTExp (Min et al., 2024) predicts a query spot by aggregating the measured expression of its nearest neighbors in that shared space. A further group recasts the task as imputation or reconstruction, recovering the expression of unmeasured locations from the measured ones together with the image, as in PH2ST (Niu et al., 2026) and CAMMST (Kim et al., 2026). Generative formulations replace the regression loss with a generative objective: STEM (Zhu et al., 2025) with a diffusion model, STFlow (Huang et al., 2025a) with flow matching, and DriftST (Yang et al., 2026) with a distribution-matching loss. A last line changes the grain of the output, moving from the spot down to the single cell, as in GHIST (Fu et al., 2025) and GHIST+ (Zhang et al., 2026).

What these directions share is where they intervene. They change the input, the architecture, or the form the prediction takes. The setting in which a model is trained has moved far less: one dataset at a time, measured on a single platform and usually drawn from a single study, with a fixed set of genes chosen within that dataset, most often between fifty and a few hundred. Where studies are pooled, as in mclSTExp, STimage and GHIST+, the pool stays within one platform, and a second platform, when it appears, appears only at test time. The objective has not been left untouched either. Hist2ST

fits a zero-inflated negative binomial to the counts in place of a squared error on their logarithm, and STimage (Tan et al., 2026) returns a negative binomial per spot in place of a point estimate, reading its spread, together with an ensemble, as the uncertainty of the prediction. But a count likelihood models how the target is distributed, not the slide and platform on which it was read, and a term for the latter has largely stayed absent. XFuse (Bergenstråhle et al., 2022) comes closest: its generative model of image and counts admits covariates per sample that shift the rate of its negative binomial gene by gene, which its authors note can control for batch effects, but it is fitted to the sections of one study rather than pooled across studies and platforms. In a single-platform setting there is only one platform, so no platform term is called for, and the slides are few and mostly drawn from one study, so the spread between them can be small. The measurement process could be left out of the objective not because it does not matter, but because such a setting tends to keep it out of sight.

I.2 Foundation models

The field has been moving from a model per cohort toward foundation models: single models pretrained across many cohorts and meant to serve them all at once. STPath (Huang et al., 2025b) is a transformer over the spots of a slide, pretrained on 928 slides from four platforms by reconstructing masked expression, so that one model both predicts from H&E alone without fine-tuning and, given the measured expression of some spots, fills in the rest. STORM (Xiang et al., 2026) pretrains a hierarchical encoder on 632 slides by masked reconstruction over neighborhoods of spots, reading expression through the frozen spatial foundation model Novae (Blampey et al., 2025), and predicts from H&E with a head trained on its histology branch. MoLF (Hu & Speidel, 2026) is trained on HEST-1k (Jaume et al., 2024) slides spanning 29 cancer types and learns a flow into the latent space of an expression autoencoder, conditioned on the image and the cancer type, with a sparse mixture of experts as its velocity field. PAST (Yang et al., 2025) works at single-cell resolution, aligning a cell-centered image encoder with an expression encoder by contrastive learning on about nineteen million Xenium cell pairs from fifteen cancer types. Phoenix (Tran et al., 2026) generates single-cell expression by latent flow matching and draws on a curated collection of 22.2 million Xenium cell pairs from sixteen organ systems. SciCore-Omics (Xiao et al., 2026) widens the set of modalities further by placing histology and expression in the token space of a language model.

Two things widen together in this shift. The first is biological, and it is the stated goal: these models are pan-cancer or pan-organ by design, so one set of weights is asked to cover tissues whose expression programs differ. The second is technical, and it follows from the first without being asked for. Assembling enough paired slides to pretrain at this scale means drawing on what has been collected publicly, and those collections were measured on different platforms. Breadth of tissue and breadth of platform arrive in the same step.

How the pooled data is met differs by model, and the responses fall broadly into three groups, which one model may combine. The first names a source of variation as an input: STPath adds a learned embedding of the platform and of the organ to every spot token, withholding the platform in most training steps so that the model can also run without it, and MoLF conditions on the cancer type. The second removes the variation from the data before the model sees it. Phoenix sets Visium aside, citing its lower sensitivity and sparser transcripts, and trains on Xenium alone, as PAST does; PAST also converts expression into gene ranks to damp differences in sequencing depth and batch, and STORM reads expression through Novae, which was pretrained with slide-aware batch correction. The third leaves the variation implicit: MoLF carries no indicator of platform although HEST-1k spans several, and STORM none of platform or organ, so what sets the sources apart is carried by the representation rather than named.

The three differ, but they agree on where the remedy sits: on the input side or in the choice of data. The objective is left as it is, and the prediction target is fitted as it was measured. The slide is the clearest case. None of these foundation models has a parameter indexed by the slide or a term for it in its own loss, and where differences between slides are addressed, it is through the expression input or the data, not in how the prediction target is fitted. Yet the slide is the unit at which tissue is sectioned, stained and read, and even within one platform and one organ it accounts for a share of the variance in the counts comparable to that of the platform (§2.2).

I.3 Batch correction and data harmonization

That samples processed in different batches differ systematically for technical rather than biological reasons, and that leaving this unaddressed can move the conclusions drawn from the data, has been discussed for a long time (Leek et al., 2010). The usual response in transcriptomics is to repair the expression matrix before anything downstream is fitted to it, and tools for exactly that already exist.

The lineage begins in bulk and single-cell transcriptomics. ComBat (Johnson et al., 2007) is the most widely used: for each gene it estimates a location and a scale per batch and shifts the values onto a common one, with the per-batch estimates shrunk by empirical Bayes toward a prior shared across genes, which keeps them stable when a batch holds few samples. ComBat-seq (Zhang et al., 2020) carries the same idea to raw counts, replacing the Gaussian model with a negative binomial one so that the corrected values remain integers. A second group aligns in a latent space rather than in the matrix: Harmony (Korsunsky et al., 2019) iteratively clusters cells and removes the batch component within each cluster, Seurat (Butler et al., 2018) finds the sources of variation that datasets share by canonical correlation analysis and aligns them in that shared space, and scVI (Lopez et al., 2018) folds the batch in as a covariate of a generative latent variable model. These methods were written for dissociated cells and ignore spatial coordinates, which a third group brings back: STAligner (Zhou et al., 2023) draws mutually nearest spots across slices together, SPIRAL (Guo et al., 2023) solves representation learning and batch removal jointly in one model, and LLOKI (Haber et al., 2025) routes slices through single-cell foundation model embeddings so that panels need not overlap for them to be aligned.

A natural way to apply these tools to prediction is to correct the target matrix first and then fit the usual objective to the corrected values. This has been done where the data permit: BLEEP corrects four consecutive Visium slices of human liver with Harmony before holding one of them out. A collection pooled across cohorts and platforms is less amenable to this approach, for three reasons. The first concerns the shape of the matrix. ComBat and ComBat-seq estimate a location and a scale for each gene in each batch, and therefore require every gene to be measured in every batch. Pooled slides do not share their genes. Visium measures the whole transcriptome, whereas Xenium measures a targeted panel, and panels differ even within Xenium: of the 1,082 genes on the Xenium slides analyzed in Appendix A.1, only 277 were measured in more than one organ. Across imaging platforms, the overlap is smaller still. LLOKI, which was designed for non-overlapping panels, reports that five such platforms share only nineteen genes, although the smallest panel contains 248. Restricting the correction to the genes shared by all slides would exclude most of the prediction targets, and a gene measured on only some slides has no reference against which to be aligned. In addition, the spatial tools have been demonstrated mainly by integrating a few slices of one tissue at a time, and LLOKI and STAligner output integrated embeddings for joint analysis rather than a corrected expression matrix to which a regression could be fitted.

The second reason arises when the correction is restricted to subsets with a complete matrix, such as a single cohort or panel. Slides within each subset are then placed on a common scale, but the subsets are not aligned with one another, and the differences between them remain in the targets as differences between slides. These differences constitute the slide share of §2.2, which is comparable to the platform share even within one platform and one organ. The third concerns the study design. ComBat protects biological variation correlated with batch only if it is supplied as a covariate, and Zhang et al. (2020) note that ComBat-seq may not perform well under severely or completely confounded designs. A pooled collection is confounded in this manner: the platforms were not applied to the same mixture of organs. Among the organs with at least three slides on each platform, only breast and lung were profiled on both Visium and Xenium (Appendix A.1). For an organ measured on a single platform, platform and organ effects cannot be separated. Where an organ is measured on both platforms, including it as a covariate still leaves out tissue composition, which varies from slide to slide. Moreover, each slide within a subset forms its own batch and carries its own tissue, so using the slide as the batch removes its biology along with its measurement level.