# LARGE LANGUAGE MODELS AS EMPIRICAL BAYES PRIORS FOR SPATIAL GENE EXPRESSION PREDICTION

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

## ABSTRACT

Predicting spatial gene expression from histology images typically relies on supervised models trained on paired image-transcriptomics data, which limits accuracy when training samples are scarce. We propose BasilSup (empirical BAyes Spatial Integration of Llm prior with SUPervision), an empirical Bayes framework that uses large language models as priors for spatial gene expression prediction. A multimodal LLM reasons about associations between tissue morphology and gene expression to produce zero-shot predictions that are treated as prior distributions, which are then integrated with a supervised predictor built on pathology foundation model embeddings. The framework automatically balances the two sources of information, relying more on the LLM prior when supervised signal is weak. Across 25 spatial transcriptomics benchmark tasks, BasilSup achieves the highest performance overall, outperforming up to 6 existing state-of-the-art baselines. These results demonstrate that biological knowledge extracted from LLMs provides complementary signal to data-driven histology models, offering a principled way to incorporate domain knowledge into spatial gene expression prediction. Code will be released upon acceptance.

## 1 INTRODUCTION

Spatial transcriptomics technologies measure gene expression while preserving the spatial organization of tissue, enabling the study of how cell types, signaling gradients, and microenvironment interactions are arranged in situ. Because these assays require specialized protocols and remain expensive at scale, a growing line of work aims to predict spatially resolved gene expression directly from routine hematoxylin and eosin (H&E) histology images, which are far more widely available. Since the pioneering work of ST-Net (He et al., 2020), methods for this task have grown to include transformers (Pang et al., 2021; Zeng et al., 2022), graph neural networks (Ganguly et al., 2025; Shi et al., 2025), ranking-based models (Nishimura et al., 2025), and flow matching (Huang et al., 2025), and have been benchmarked on curated collections such as HEST-Bench (Jaume et al., 2024) and STimage-1K4M (Chen et al., 2024a). All of these approaches are purely supervised: they learn a mapping from image features to expression using paired image-transcriptomics training data. This reliance on paired data creates a fundamental bottleneck. For rare tissues, uncommon cancer subtypes, or newly profiled gene panels, training samples may be too few for the model to learn reliable gene-level predictions. The problem is compounded by the high dimensionality of the output space, where hundreds to thousands of genes must each be predicted accurately.

Large language models offer a complementary source of information. Through pretraining on biomedical literature, LLMs absorb extensive knowledge about which genes are expressed in which cell types, how those cell types appear histologically, and which morphological features correlate with specific molecular programs. This is evidenced by work showing that LLMs capture gene-function relationships and biological reasoning from text (Jin et al., 2024; Chen & Zou, 2024; Luo et al., 2022), and that multimodal LLMs can reason about medical images without task-specific training (Li et al., 2023; Tu et al., 2024; Ferber et al., 2024). Recent work demonstrates that language models can make zero-shot predictions about biological phenomena such as mutation effects (Meier et al., 2021), suggesting that LLM-encoded knowledge is directly useful for biological prediction tasks. This knowledge is precisely the kind of prior information that could regularize predictions when supervised signal is scarce, yet the question of how to integrate it with data-driven models in a principled way remains open.



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.

We propose an empirical Bayes framework that addresses this challenge. A multimodal LLM compares pairs of histology patches and predicts which patch expresses a given gene more highly, producing zero-shot spatial predictions without any paired training data. These predictions serve as a prior that is combined with a supervised predictor trained on pathology foundation model embeddings. The framework automatically balances the two sources on a per-gene basis, relying more on the LLM prior when supervised signal is weak and falling back to the data-driven predictor when the LLM contributes no useful information.

Our contributions are as follows:

- To the best of our knowledge, we are the first to propose an empirical Bayes framework for spatial gene expression prediction that leverages LLMs' internal knowledge as a biological prior, which is then updated by data-driven supervised training.
- The framework automatically adapts the weighting between the two complementary components on a per-gene basis, relying more on the LLM prior when supervised signal is weak and on the data-driven predictor when paired training data is sufficient.
- We achieve state-of-the-art performance through comparison with up to six baseline methods over 25 benchmark tasks covering different studies, sequencing technologies, and species, plus a large-panel KPMP evaluation.
- Our work serves as an example of how generalist frontier models (LLMs) and specialist foundation models (pathology encoders) collaborate to achieve stronger performance.
- Our framework is broadly applicable to a wide range of scientific prediction tasks where LLM-encoded domain knowledge complements data-driven approaches.

## 2 Related work

**Spatial gene expression prediction from histology.** Predicting spatially resolved gene expression from hematoxylin-and-eosin-stained histology images was pioneered by ST-Net (He et al., 2020), which trained a DenseNet on paired breast cancer images and spatial transcriptomics measurements. Subsequent work introduced attention-based architectures: HisToGene (Pang et al., 2021) applied a Vision Transformer to predict super-resolution expression in tumors, and Hist2ST (Zeng et al., 2022) combined a Transformer encoder with a graph neural network to capture spatial dependencies among neighboring spots. More recent methods have scaled to whole-slide images and larger gene panels. STFlow (Huang et al., 2025) uses flow matching to model the conditional distribution of expression given histology at the whole-slide level. MERGE (Ganguly et al., 2025) constructs a hierarchical graph over tissue regions and applies a GNN with SimCLR-pretrained features. M2TGLGO (Shi et al., 2025) fuses cell-, patch-, and image-level features with gene similarity information through a multi-modal graph convolutional network. STRank (Nishimura et al., 2025) takes a different approach by learning relative expression rankings rather than absolute values. All of these methods are purely supervised, requiring large sample sizes to learn complex high-dimensional expression patterns and overcome batch effects and slide- or institution-specific artifacts (Howard et al., 2021; Lin et al., 2025), with no mechanism for incorporating external biological knowledge.

**Pathology foundation models.** Self-supervised pretraining on large-scale histopathology image collections has produced general-purpose feature extractors that transfer broadly across downstream tasks. CTransPath (Wang et al., 2022) introduced a hybrid CNN-Transformer backbone pretrained with semantically relevant contrastive learning on 15 million patches. UNI (Chen et al., 2024b) scaled pretraining to over 100 million tissue patches from more than 100,000 diagnostic slides using a Vision Transformer, establishing strong baselines across 34 clinical tasks. Virchow (Vorontsov et al., 2024) further scaled to 1.5 million whole-slide images, enabling pan-cancer detection. CONCH (Lu et al., 2024) extended these ideas to the vision-language setting by pretraining on over one million image-caption pairs, supporting zero-shot classification of tissue types. Our supervised predictor builds on UNI2-h embeddings, the successor to UNI trained on over 200 million pathology images, which provide rich morphological representations without task-specific fine-tuning.

**LLMs for biological reasoning.** Large language models encode substantial biomedical knowledge acquired during pretraining on scientific literature. GeneGPT (Jin et al., 2024) demonstrated that

LLMs augmented with NCBI database APIs can answer genomics questions with high accuracy, while the GeneTuring benchmark (Shang et al., 2025) systematically evaluated LLMs on 16 genomics tasks spanning gene nomenclature, functional analysis, and sequence alignment. GenePT (Chen & Zou, 2024) showed that gene embeddings derived from GPT-generated text descriptions rival purpose-built single-cell foundation models on cell-type classification and gene property prediction. These results suggest that LLMs capture meaningful gene-function relationships from the literature. Our work leverages this capability differently: rather than using LLM embeddings or tool-augmented question answering, we use a multimodal LLM to produce zero-shot spatial expression predictions that serve only as a prior, which is then updated by real paired image-expression data through empirical Bayes inference.

**LLMs as Bayesian priors.** A growing body of work explores extracting prior knowledge from LLMs for use in Bayesian inference. Riegler et al. (2025) showed that LLMs can suggest informative prior distributions for Bayesian regression, while AutoElicit (Capstick et al., 2025) systematically extracts domain knowledge from LLMs to construct priors for predictive modeling, demonstrating substantial error reduction when labeled examples are scarce. Complementary work has studied the implicit priors that LLMs acquire through pretraining: Zhu & Griffiths (2025) used iterated in-context learning to elicit these priors and found alignment with human prior beliefs, and Qiu et al. (2026) showed that Bayesian teaching can improve probabilistic reasoning in LLMs. Our work applies this idea in a new domain: we use a multimodal LLM's zero-shot predictions as an informative prior for spatial gene expression, which is then updated by paired image-expression data through empirical Bayes inference.

## 3 MODEL

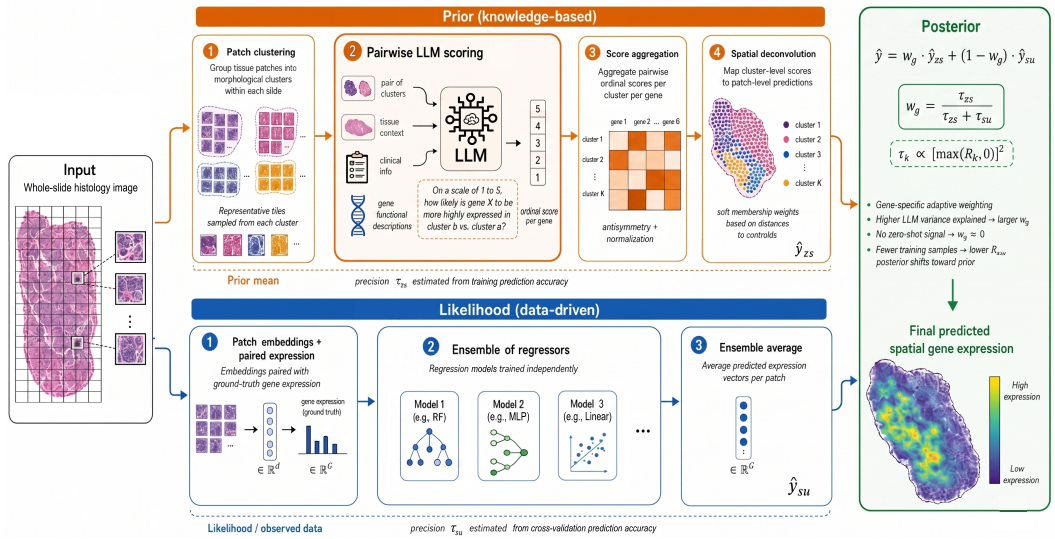


Figure 1: Overview of BasilSup, which combines zero-shot LLM priors with supervised histology predictors using gene-specific empirical Bayes weights.

BasilSup combines a zero-shot multimodal-LLM prior over morphology–gene associations with supervised predictors trained on pathology foundation model embeddings, using gene-specific empirical Bayes weights to form the final spatial expression prediction (Figure 1).

### 3.1 EMPIRICAL BAYES FORMULATION

We combine a supervised predictor trained on pathology foundation model embeddings with a zero-shot predictor derived from pairwise LLM comparisons of morphological clusters. The two sources of information are integrated through an empirical Bayes framework that treats the zero-shot predictions as a prior and the supervised predictions as a likelihood.

Let $\hat{y}_{ig}^{\mathrm{su}}$ and $\hat{y}_{ig}^{\mathrm{zs}}$ denote the supervised and zero-shot predictions for patch $i$ and gene $g$, respectively. We place a Gaussian prior on the true expression centered at the zero-shot prediction with precision $\tau_g^{\mathrm{zs}}$:

$$y_{ig} \;\sim\; \mathcal{N}\big(\hat{y}_{ig}^{\mathrm{zs}},\; (\tau_g^{\mathrm{zs}})^{-1}\big)\,. \tag{1}$$

The supervised prediction provides a noisy observation of the true expression with precision $\tau_g^{\mathrm{su}}$:

$$\hat{y}_{ig}^{\mathrm{su}} \mid y_{ig} \;\sim\; \mathcal{N}\big(y_{ig},\; (\tau_g^{\mathrm{su}})^{-1}\big)\,. \tag{2}$$

By Normal–Normal conjugacy, the posterior mean is a precision-weighted combination of the prior and likelihood:

$$\hat{y}_{ig} = w_g\,\hat{y}_{ig}^{\mathrm{zs}} + (1 - w_g)\,\hat{y}_{ig}^{\mathrm{su}}, \qquad w_g = \frac{\tau_g^{\mathrm{zs}}}{\tau_g^{\mathrm{zs}} + \tau_g^{\mathrm{su}}}. \tag{3}$$

Rather than specifying the precisions a priori, we estimate them empirically from each predictor's accuracy on the training set by setting $\tau_g^{(k)} \propto q_g^{(k)}$ for $k \in \{\mathrm{zs}, \mathrm{su}\}$, where

$$q_g^{(k)} = \big[\max\big(R_g^{(k)},\, 0\big)\big]^2 \tag{4}$$

and $R_g^{(k)}$ is the PCC between predictor $k$'s output and ground truth for gene $g$ on training data. Since the zero-shot predictor requires no training, its quality is evaluated directly on the full training set:

$$R_g^{\mathrm{zs}} = \mathrm{corr}\big(\hat{y}_{\mathrm{train},g}^{\mathrm{zs}},\; y_{\mathrm{train},g}\big)\,. \tag{5}$$

Because the supervised predictor is fit to the same data, its in-sample correlation is optimistic, so we shrink it by the ratio of model dimension to sample size:

$$R_g^{\mathrm{su}} = \big(1 - \min(p/n,\, 0.9)\big)\,\mathrm{corr}\big(\hat{y}_{\mathrm{train},g}^{\mathrm{su}},\; y_{\mathrm{train},g}\big)\,, \tag{6}$$

where $n$ is the number of training patches and $p = 256$ is the number of embedding principal components used by the Ridge model. The $\max(\cdot, 0)$ in Eq. 4 discards negatively correlated predictions, and squaring converts correlation to an $R^2$-like measure of variance explained. If both qualities are zero (neither predictor has positive correlation), the prediction defaults to the supervised estimate ($w_g = 0$). Genes for which the zero-shot predictor explains more variance receive higher $w_g$, while genes for which it has no predictive power receive $w_g \approx 0$. With fewer training slides or lower data quality, $R_g^{\mathrm{su}}$ decreases and the method relies more on the zero-shot prior. For poorly characterized genes, $R_g^{\mathrm{zs}}$ remains low and the method relies on supervised predictions alone.

### 3.2 Zero-shot predictor

The zero-shot predictor estimates per-patch gene expression without paired training data by leveraging a multimodal LLM's ability to reason about morphology-gene associations. The pipeline consists of four stages: patch clustering, pairwise LLM scoring, score aggregation, and spatial deconvolution. Specific hyperparameter values are given in Appendix A.

**Patch clustering.** Consider a collection of $S$ slides, where slide $s$ contains $N_s$ histology patches $\{\mathbf{x}_{s,1}, \ldots, \mathbf{x}_{s,N_s}\}$. A pathology foundation model $\phi\colon \mathcal{X} \to \mathbb{R}^D$ maps each patch to an embedding $\mathbf{e}_{s,i} = \phi(\mathbf{x}_{s,i})$. Within each slide, we partition the embeddings into $K$ clusters via K-Means, producing centroids $\boldsymbol{\mu}_{s,1}, \ldots, \boldsymbol{\mu}_{s,K} \in \mathbb{R}^D$ and assignments $z_{s,i} \in \{1, \ldots, K\}$. Each cluster corresponds to a recurring morphological motif (e.g., tumor epithelium, immune infiltrate, stroma). From each cluster $c$ in slide $s$, we randomly sample $T$ representative tiles, denoted $\mathcal{T}_{s,c}$.

**Pairwise LLM scoring.** Let $\mathcal{C} = \{(s, c) : s = 1, \ldots, S,\; c = 1, \ldots, K\}$ denote the set of all slide-cluster pairs. We construct a set of ordered pairs $\mathcal{P} \subseteq \mathcal{C} \times \mathcal{C}$ with $|\mathcal{P}| \le P$, sampled from both within-slide and cross-slide combinations. For each pair $(a, b) \in \mathcal{P}$, a multimodal LLM $\mathcal{M}$ receives the tiles $\mathcal{T}_a$ and $\mathcal{T}_b$ together with a structured prompt containing tissue context and functional descriptions of target genes. For each gene $g$, the LLM returns an ordinal score

$$v_{ab,g} = \mathcal{M}(\mathcal{T}_a,\, \mathcal{T}_b,\, g) \in \{1, \ldots, L\}, \tag{7}$$

indicating the relative expression of gene $g$ in cluster $b$ compared to cluster $a$: $v = 1$ denotes much lower expression in $b$, $v = \lceil L/2 \rceil$ denotes approximately equal expression, and $v = L$ denotes much higher expression.

**Score aggregation.** By construction, the ordinal scale is antisymmetric: if cluster $b$ scores $v$ relative to $a$, then $a$ scores $L+1-v$ relative to $b$. For each cluster $a \in \mathcal{C}$ and gene $g$, we aggregate across all comparisons involving $a$:

$$\bar{v}_{a,g} = \frac{1}{|\mathcal{N}(a)|}\left(\sum_{b:\,(b,a)\in\mathcal{P}} v_{ba,g} \;+\; \sum_{b:\,(a,b)\in\mathcal{P}} (L+1-v_{ab,g})\right), \tag{8}$$

where $\mathcal{N}(a) = \{b : (b,a) \in \mathcal{P}\} \cup \{b : (a,b) \in \mathcal{P}\}$ is the set of clusters compared against $a$. The raw scores are standardized to zero mean and unit variance per gene within each slide, yielding normalized scores $s_{a,g}$.

**Spatial deconvolution.** To map cluster-level scores to patch-level predictions, we define soft membership weights from the embedding distances. For patch $i$ in slide $s$, let $d_{ic} = \|\mathbf{e}_{s,i} - \boldsymbol{\mu}_{s,c}\|_2$ denote the distance to centroid $c$, and define kernel weights $\tilde{w}_{ic} = \kappa(d_{ic};\, \sigma_s)$ using a distance kernel $\kappa$ with per-slide bandwidth $\sigma_s$. The zero-shot prediction for patch $i$ and gene $g$ is

$$\hat{y}^{\mathrm{zs}}_{ig} = \sum_{c=1}^{K} \frac{\tilde{w}_{ic}}{\sum_{c'=1}^{K} \tilde{w}_{ic'}}\, s_{(s,c),g}. \tag{9}$$

This soft assignment produces predictions that vary smoothly across space, with each patch reflecting its proximity to multiple morphological patterns.

### 3.3 Supervised predictor

The supervised predictor is an ensemble of $M$ regression models trained on pathology foundation model embeddings. Given a training set of patches with paired expression measurements $\{(\mathbf{e}_{s,i},\, \mathbf{y}_{s,i})\}$ where $\mathbf{e}_{s,i} = \phi(\mathbf{x}_{s,i}) \in \mathbb{R}^D$ and $\mathbf{y}_{s,i} \in \mathbb{R}^G$, each model $f_m \colon \mathbb{R}^D \to \mathbb{R}^G$ is trained independently to predict expression from embeddings. The ensemble prediction for patch $i$ and gene $g$ is

$$\hat{y}^{\mathrm{su}}_{ig} = \frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{e}_{s,i})_g. \tag{10}$$

Averaging over diverse model families reduces prediction variance while allowing the ensemble to capture complementary structure in the embedding-to-expression mapping. The specific model architectures are described in Appendix A.

## 4 Experiments

### 4.1 Datasets

We evaluate on three spatial transcriptomics benchmarks spanning diverse tissues, species, and sequencing technologies. **HEST-Bench** provides cancer-specific tasks from the HEST-1k collection (Jaume et al., 2024), each predicting 50 highly variable genes from H&E histology patches. We use 7 of its 10 tasks (PAAD, COAD, IDC, L.IDC, READ, PRAD, CCRCC). **STimage-1K4M** (Chen et al., 2024a) contains over 1,400 slides across multiple species and sequencing technologies. We stratify by species and technology, selecting the top 50 highly variable genes per homogeneous subset. The manuscript analyses use 8 human tasks and 10 mouse tasks. **KPMP** comprises 138 kidney slides from the Kidney Precision Medicine Project. We select the top 1000 most highly variable genes.

### 4.2 Baselines

We compare against six baselines: **STFlow** (Huang et al., 2025), a flow matching model, **MERGE** (Ganguly et al., 2025), a hierarchical graph-based GNN with SimCLR-pretrained ResNet18 features, **M2TGLGO** (Shi et al., 2025), a multi-modal GCN combining cell-, patch-, and image-level features with gene2vec embeddings, **STRank** (Nishimura et al., 2025), a linear model trained with a pairwise ranking loss, and **HisToGene** (Pang et al., 2021) and **Hist2ST** (Zeng et al., 2022), transformer-based architectures operating on raw image patches.

Table 1: HEST-Bench results. Each column is a cancer-specific task. Values are mean per-gene PCC across cross-validation folds (50 genes per task). Supervised only is BasilSup without the zero-shot LLM prior. Bold: best, underline: second best.

| Method | PAAD | COAD | IDC | L.IDC | READ | PRAD | CCRCC | Avg |
|---|---|---|---|---|---|---|---|---|
| BasilSup | **0.495** | **0.346** | **0.599** | **0.232** | **0.220** | 0.384 | **0.267** | **0.363** |
| Supervised only | 0.384 | 0.285 | 0.595 | 0.229 | 0.191 | **0.387** | 0.257 | 0.332 |
| STFlow | 0.454 | 0.288 | 0.569 | 0.213 | 0.210 | 0.233 | 0.226 | 0.313 |
| STRank | 0.412 | 0.219 | 0.558 | 0.190 | 0.146 | 0.349 | 0.246 | 0.303 |
| M2TGLGO | 0.352 | 0.307 | 0.382 | 0.193 | 0.061 | 0.262 | 0.068 | 0.232 |
| MERGE | 0.339 | 0.291 | 0.128 | 0.095 | 0.029 | 0.184 | 0.070 | 0.162 |
| HisToGene | 0.250 | 0.139 | 0.228 | 0.143 | −0.030 | 0.244 | 0.005 | 0.140 |
| Hist2ST | 0.221 | 0.120 | 0.066 | 0.185 | −0.069 | 0.047 | 0.164 | 0.105 |

### 4.3 HEST-BENCH

The 7 HEST-Bench tasks shown cover distinct cancer types, testing whether methods maintain consistent performance across disease contexts (Table 1). BasilSup achieves the highest average and outperforms all existing baselines on every task, while GNN-based methods (MERGE, M2TGLGO) exhibit high variance across tissue types.

### 4.4 STIMAGE-1K4M

We next evaluate on STimage-1K4M, which includes both legacy STv1 and modern Visium platforms as well as human and mouse tissues.

Table 2: STimage-1K4M results (human, technology-stratified). Values are mean per-gene PCC (50 genes). Column headers indicate tissue and sequencing platform. STv1 denotes legacy Spatial Transcriptomics and remaining tasks use 10x Visium. Bold: best, underline: second best.

| Method<br>Technology | SKIN<br>STv1 | BREAST<br>STv1 | SKIN<br>Visium | BREAST<br>Visium | MOUTH<br>Visium | BRAIN<br>Visium | COLON<br>Visium | STOM.<br>Visium | Avg |
|---|---|---|---|---|---|---|---|---|---|
| BasilSup | **0.719** | **0.536** | **0.533** | **0.465** | **0.421** | **0.324** | 0.301 | **0.288** | **0.448** |
| STFlow | 0.622 | NA† | 0.284 | 0.413 | 0.294 | 0.271 | **0.382** | 0.184 | 0.350 |
| STRank | 0.599 | 0.481 | 0.425 | 0.404 | 0.357 | 0.296 | 0.259 | 0.218 | 0.380 |
| HisToGene | 0.328 | 0.198 | 0.128 | 0.367 | 0.095 | 0.247 | 0.128 | 0.040 | 0.191 |
| MERGE | 0.691 | 0.499 | −0.002 | 0.085 | 0.099 | 0.173 | 0.071 | 0.009 | 0.203 |
| Hist2ST | 0.493 | 0.275 | −0.152 | 0.195 | 0.346 | 0.145 | −0.110 | 0.082 | 0.159 |
| M2TGLGO | 0.668 | 0.287 | 0.356 | 0.333 | 0.294 | 0.241 | 0.223 | 0.206 | 0.326 |

† STFlow failed on BREAST (STv1) due to errors caused by the sample size.

On the eight human subsets, which span two sequencing technologies and six tissue types, BasilSup again leads overall (Table 2). It achieves the best result on 7 of 8 analyzed tasks and the highest average PCC, with STRank ranking second overall.

#### 4.4.1 MOUSE CANCER SUBSETS

We further test whether these trends hold across species by evaluating on mouse cancer subsets from the same STimage-1K4M collection. All 10 subsets use Visium and span diverse tumor types.

BasilSup again achieves the highest average PCC (Table 3), leading on 5 of 10 tasks, followed by STFlow and STRank. M2TGLGO is omitted because its nuclear feature extraction produced NaN on most mouse tasks. Predictability varies widely across tumor types, with BREAST and MELANOMA being the most predictable and COLON and LUNG the least.

Table 3: STimage-1K4M results (mouse cancer, Visium only). Values are mean per-gene PCC (50 genes). Each column is a tumor type. Bold: best, underline: second best.

| Method | BRAIN | LIVER | PANC. | B16F10 | BONE | BRST. | COLON | LUNG | LYM.N. | MELAN. | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BasilSup | **0.300** | 0.218 | **0.412** | **0.257** | 0.538 | **0.744** | 0.211 | 0.217 | 0.670 | **0.745** | **0.431** |
| STFlow | 0.277 | 0.217 | 0.382 | 0.190 | **0.562** | 0.704 | 0.136 | **0.269** | **0.725** | 0.673 | 0.414 |
| STRank | 0.267 | 0.221 | 0.361 | 0.220 | 0.507 | 0.692 | **0.229** | 0.186 | 0.636 | 0.669 | 0.399 |
| HisToGene | 0.179 | **0.228** | 0.308 | 0.170 | 0.531 | 0.676 | 0.121 | 0.048 | 0.634 | 0.626 | 0.352 |
| MERGE | −0.005 | 0.222 | 0.117 | −0.013 | 0.355 | 0.161 | 0.032 | 0.004 | 0.470 | 0.163 | 0.151 |
| Hist2ST | 0.291 | 0.133 | 0.144 | 0.112 | 0.005 | 0.340 | 0.105 | −0.127 | 0.439 | 0.262 | 0.170 |

## 4.5 KPMP

Finally, we scale to a larger gene panel to test whether the methods' advantages persist beyond the 50-gene regime. KPMP provides a 1000-gene panel across 138 slides, allowing us to evaluate how performance degrades as less variable (and thus harder to predict) genes are included.

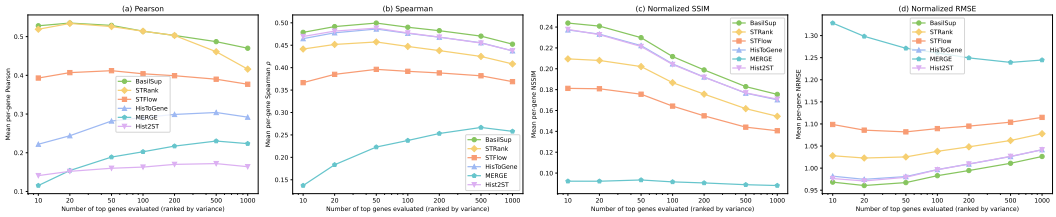


Figure 2: KPMP: performance as a function of gene panel size, measured by (a) Pearson correlation, (b) Spearman correlation, (c) normalized SSIM, and (d) normalized RMSE (lower is better). For (c) and (d), predictions and ground truth are standardized to zero mean and unit variance per gene. SSIM is computed on the Visium array grid, averaging only over observed spot positions.

STFlow, HisToGene, MERGE, and Hist2ST lag substantially behind BasilSup across all gene panel sizes on PCC (Figure 2a). BasilSup and STRank perform comparably on the top 10–200 genes, but BasilSup pulls ahead on larger panels, maintaining the strongest PCC throughout as less variable genes are included. BasilSup also leads on Spearman correlation, normalized RMSE, and normalized SSIM across all panel sizes, confirming that its advantage is robust to the choice of evaluation metric. M2TGLGO is excluded from KPMP because its multi-loss formulation diverged during training at this scale (138 slides, 108+ per fold).

## 4.6 Analysis of zero-shot LLM contribution

### 4.6.1 Per-gene improvement decomposition

To understand where the zero-shot prior helps, we decompose the BasilSup improvement at the per-gene level across all 350 gene-task pairs on HEST-Bench. For each gene-task pair, we compute the improvement as the difference between BasilSup and supervised-only PCC, and examine its relationship with the zero-shot weight $w_g$.

Figure 3a compares the zero-shot and supervised PCC for each gene-task pair, showing that the zero-shot predictor is weaker on average but occasionally competitive or stronger. Across all 350 gene-task pairs, the per-gene zero-shot weight $w_g$ is correlated with the actual improvement from BasilSup integration (Figure 3b, PCC $= 0.62$, $p < 10^{-37}$). Genes in the lowest $w_g$ quartile show minimal change on average, with average improvement $= -0.0001$ and 41% of genes improved (Figure 3c), indicating that the method does not systematically degrade genes where the LLM has little predictive power. In contrast, genes in the highest quartile gain $+0.045$ average PCC, with 99% of genes improved. This demonstrates that the empirical Bayes weighting correctly identifies genes that benefit from the LLM prior.

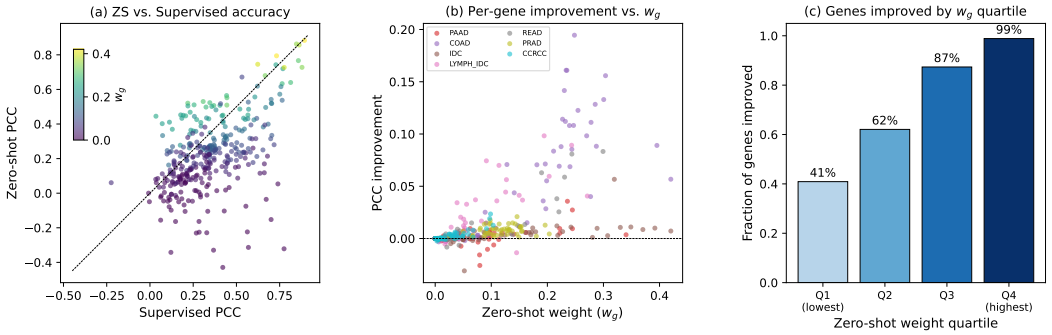


Figure 3: Per-gene decomposition of BasilSup improvement on HEST-Bench. (a) Zero-shot vs. supervised PCC for each gene-task pair, colored by $w_g$. (b) Per-gene PCC improvement over the supervised-only predictor (BasilSup PCC − Supervised PCC) vs. zero-shot weight $w_g$. Each point is one gene-task pair, colored by task. (c) Fraction of genes improved over the supervised-only predictor, binned by $w_g$ quartile.

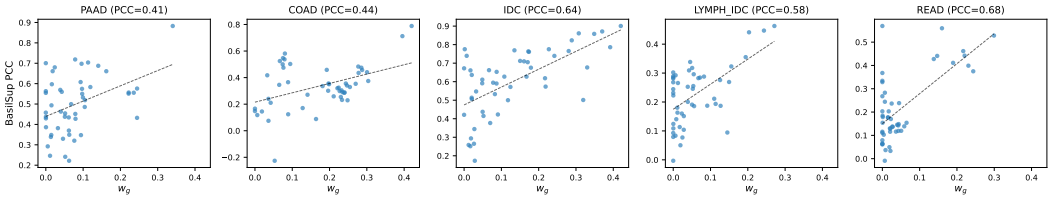


Figure 4: Per-gene zero-shot weight ($w_g$) vs. BasilSup prediction accuracy (PCC) on HEST-Bench (first 5 tasks). Each point is one gene-task pair. PCC is shown per task.

### 4.6.2 Zero-shot weight and prediction accuracy

Beyond showing that higher $w_g$ is associated with improvement, we also ask whether genes assigned higher zero-shot weight are more predictable overall. Figure 4 shows that $w_g$ is positively correlated with BasilSup prediction accuracy (PCC) across all seven HEST tasks (PCC ranging from 0.37 to 0.68), with the full 7-task version shown in Figure 7. The association is strongest in READ (PCC $= 0.68$) and IDC (PCC $= 0.64$), tasks where morphologically distinct cell types drive expression variation, and weakest in CCRCC (PCC $= 0.37$, $p = 0.009$). Genes with higher zero-shot weight also show stronger agreement between the two predictors (Figure 5; full 7-task version in Figure 8), consistent with both capturing the same morphologically driven expression patterns, though the agreement remains moderate (mean PCC $= 0.15$–$0.48$ across tasks, Table 5), indicating that the two sources provide complementary rather than redundant signal. This supports the conclusion that the empirical Bayes weighting scheme selectively upweights genes for which the LLM provides informative morphological cues, and that these genes tend to be the most predictable overall.

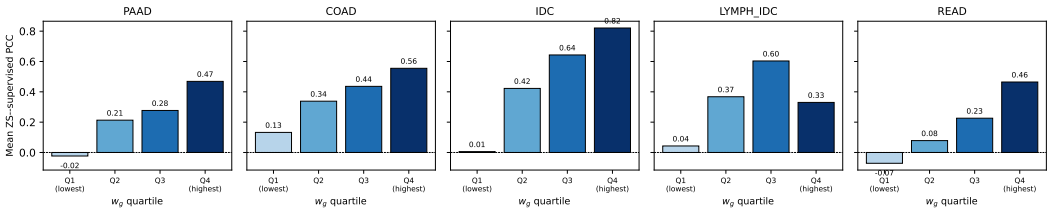


Figure 5: Mean per-gene PCC between zero-shot and supervised predictions, stratified by within-task $w_g$ quartile (first 5 tasks). Genes with higher zero-shot weight show stronger agreement between the two predictors, indicating that BasilSup upweights genes where the zero-shot and supervised signals are most aligned.

#### 4.6.3 FUNCTIONAL CHARACTERIZATION OF HIGH-WEIGHT GENES

Examining the genes that receive the highest zero-shot weights reveals a consistent pattern: they are markers of cell types with distinctive histological morphology (Figure 6). Across tasks, these fall into four functional themes: *proliferation markers* (e.g., MKI67 in PAAD, TOP2A in IDC) whose mitotic figures are recognizable in H&E (Scholzen & Gerdes, 2000), *immune markers* (e.g., PTPRC in IDC, GZMK in PAAD) associated with visible immune infiltrates (Salgado et al., 2015), *epithelial differentiation markers* (e.g., ERBB2, GATA3 in IDC) delineating tumor from stroma (Perou et al., 2000), and *GI secretory markers* (e.g., CEACAM6, CA1 in COAD/READ) marking glandular architecture (Smillie et al., 2019). Conversely, stromal smooth-muscle genes (e.g., MYH11, MYLK in IDC) receive $w_g \approx 0$. The zero-shot weight thus reflects whether the gene's associated cell type exhibits distinctive morphological features for which the LLM can generate informative predictions.

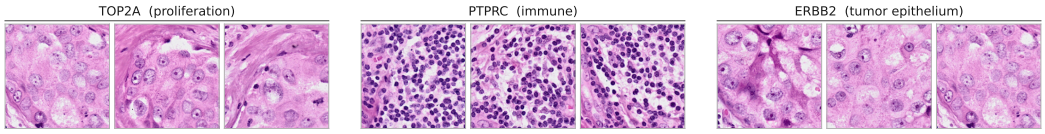


Figure 6: H&E patches from high-expressing regions of three genes with high zero-shot weight in an example cancer slide, illustrating how each gene's functional role corresponds to a distinct morphological pattern.

### 4.7 ABLATION STUDIES

Table 4 compares BasilSup using different LLMs for zero-shot prediction. Gemini 3 Flash achieves the highest average PCC (0.363), followed closely by Gemma 4 31B (0.356), which nearly matches despite being smaller and open-weight. Notably, Gemma 4 31B outperforms the older Gemini 2.5 Flash Lite (0.356 vs. 0.347) at a fraction of the size, suggesting that biomedical knowledge in LLMs is improving rapidly across model generations. The narrow spread across backbones (0.347–0.363) indicates that the relevant biological knowledge is broadly captured across modern LLMs.

Table 4: Ablation of LLM backbone on HEST-Bench. All variants use BasilSup (Ridge+MLP+ZS). Values are mean per-gene PCC (50 genes per task). Bold: best, underline: second best.

| LLM | PAAD | COAD | IDC | L.IDC | READ | PRAD | CCRCC | Avg |
|---|---|---|---|---|---|---|---|---|
| Gemini 3 Flash (default) | <u>0.495</u> | **0.346** | **0.599** | **0.232** | **0.220** | **0.384** | **0.267** | **0.363** |
| Gemini 2.5 Flash Lite | **0.498** | 0.300 | 0.594 | 0.210 | 0.197 | 0.372 | <u>0.261</u> | 0.347 |
| Gemma 4 31B | <u>0.495</u> | <u>0.334</u> | <u>0.596</u> | <u>0.216</u> | <u>0.211</u> | <u>0.381</u> | 0.257 | <u>0.356</u> |

## 5 CONCLUSION

BasilSup shows that biological knowledge encoded in large language models can improve spatial gene expression prediction from histology. Across diverse tissues, species, sequencing technologies, and gene panel sizes, it achieves the highest average performance across all benchmarks, indicating that LLMs provide complementary signal beyond paired training data. More broadly, these results suggest that scientific prediction models can benefit from treating LLM knowledge as a useful source of domain context rather than only as a text-generation tool. By improving gene expression prediction from routine histology, this work could accelerate tissue biology research and reduce dependence on costly spatial transcriptomics assays, though care should be taken before applying such predictions in clinical decision-making. A limitation of the current analysis is that it focuses on STv1 and Visium platforms. Future work could extend this framework to other spatial transcriptomics technologies (e.g., MERFISH, Slide-seq), other spatial omics modalities (e.g., proteomics, metabolomics), and settings where paired training data are limited.

### AI USE STATEMENT

In this work, we used generative AI tools (LLM-based coding assistants) to help implement and debug the data processing, evaluation, and plotting code, to check and reformat bibliography entries against Semantic Scholar and publisher records, and to edit the manuscript text. Multimodal LLMs are also a component of the proposed method, and that use is described in Section 3.2 and Appendix A. We have not used generative AI tools to generate synthetic data sets, formulate mathematical claims or proofs, or propose research hypotheses. The authors reviewed all AI-assisted work: text, figures, and citations were checked manually. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses only publicly available histology and spatial transcriptomics data (HEST-1k, STimage-1K4M, and KPMP) and involves no new data collection from human subjects. Predicted gene expression is intended for research use and should not inform clinical decisions without further validation.

### REPRODUCIBILITY STATEMENT

Section 3.2 and Section 3.3 define the zero-shot and supervised predictors and their empirical Bayes combination. Appendix A lists all hyperparameters, including the foundation model, clustering, LLM, kernel, and supervised training settings. Section 4.1 describes the benchmarks, all of which are publicly available. Code will be released upon acceptance.

## REFERENCES

Alexander Capstick, Rahul G. Krishnan, and Payam Barnaghi. AutoElicit: Using large language models for expert prior elicitation in predictive modelling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6746–6777. PMLR, 2025.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. STimage-1K4M: A histopathology image-gene expression dataset for spatial transcriptomics. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35796–35823, 2024a. doi: 10.52202/079017-1129.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024b. doi: 10.1038/s41591-024-02857-3.

Yiqun T. Chen and James Y. Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2024. doi: 10.1038/s41551-024-01284-6.

Dyke Ferber, Georg Wölflein, Isabella C. Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar S. M. El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, and Jakob Nikolas Kather. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15:10104, 2024. doi: 10.1038/s41467-024-51465-9.

Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis S. Johnson, and Chao Chen. MERGE: Multi-faceted hierarchical graph-based GNN for gene expression prediction from whole slide histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15611–15620, 2025. doi: 10.1109/CVPR52734.2025.01455.

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. doi: 10.1038/s41551-020-0578-x.

Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12:4423, 2021. doi: 10.1038/s41467-021-24698-1.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25550–25565. PMLR, 2025.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53798–53833, 2024. doi: 10.52202/079017-1704.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024. doi: 10.1093/bioinformatics/btae075.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023. doi: 10.52202/075280-1240.

Siyu Lin, Haowen Zhou, Mark Watson, Ramaswamy Govindan, Richard J. Cote, and Changhuei Yang. Impact of stain variation and color normalization for prognostic predictions in pathology. *Scientific Reports*, 15:2369, 2025. doi: 10.1038/s41598-024-83267-w.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg K. Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022. doi: 10.1093/bib/bbac409.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303, 2021.

Kazuya Nishimura, Haruka Hirose, Ryoma Bise, Kaito Shiku, and Yasuhiro Kojima. Learning relative gene expression trends from pathology images in spatial transcriptomics. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96393–96414, 2025. doi: 10.52202/085713-2900.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021. doi: 10.1101/2021.11.28.470212.

Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, Øystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X. Zhu, Per E. Lønning, Anne-Lise Børresen-Dale, Patrick O. Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406 (6797):747–752, 2000. doi: 10.1038/35021093.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, 17: 1238, 2026. doi: 10.1038/s41467-025-67998-6.

Michael A. Riegler, Kristoffer H. Hellton, Vajira Lasantha Thambawita, and Hugo Hammer. Using large language models to suggest informative prior distributions in Bayesian regression analysis. *Scientific Reports*, 15:33386, 2025. doi: 10.1038/s41598-025-18425-9.

Roberto Salgado, Carsten Denkert, Sandra Demaria, Nicolas Sirtaine, Frederick Klauschen, Giancarlo Pruneri, Stephan Wienert, Gert Van den Eynden, Frederick L Baehner, Frederique Penault-Llorca, E. A. Pérez, E. A. Thompson, W. F. Symmans, A. L. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, J. Sparano, Z. Kos, T. Nielsen, D. L. Rimm, K. H. Allison, J. S. Reis-Filho, S. Loibl, C. Sotiriou, G. Viale, S. Badve, S. Adams, K. Willard-Gallo, and S. Loi. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Annals of Oncology*, 26(2):259–271, 2015. doi: 10.1093/annonc/mdu450.

Thomas Scholzen and Johannes Gerdes. The Ki-67 protein: From the known and the unknown. *Journal of Cellular Physiology*, 182(3):311–322, 2000. doi: 10.1002/(SICI)1097-4652(200003)182:3<311::AID-JCP1>3.0.CO;2-9.

Xinyi Shang, Xu Liao, Zhicheng Ji, and Wenpin Hou. Benchmarking large language models for genomic knowledge with GeneTuring. *Briefings in Bioinformatics*, 26(5):bbaf492, 2025. doi: 10.1093/bib/bbaf492.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20810–20819, 2025. doi: 10.1109/CVPR52734.2025.01938.

Christopher S Smillie, Moshe Biton, Jose Ordovas-Montañes, Keri M Sullivan, Grace Burgin, Daniel B Graham, Rebecca H Herbst, Noga Rogel, Michal Slyper, Julia Waldman, Malika Sud, Elizabeth Andrews, Gabriella Velonias, Adam L. Haber, Karthik Jagadeesh, Sanja Vickovic, Junmei Yao, Christine Stevens, Danielle Dionne, Lan T. Nguyen, Alexandra-Chloé Villani, Matan Hofree, Elizabeth A. Creasey, Hailiang Huang, Orit Rozenblatt-Rosen, John J. Garber, Hamed Khalili, A. Nicole Desch, Mark J. Daly, Ashwin N. Ananthakrishnan, Alex K. Shalek, Ramnik J. Xavier, and Aviv Regev. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, 178(3):714–730.e22, 2019. doi: 10.1016/j.cell.2019.06.029.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI. *NEJM AI*, 1(3), 2024. doi: 10.1056/AIoa2300138.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022. doi: 10.1016/j.media.2022.102559.

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. doi: 10.1093/bib/bbac297.

Jian-Qiao Zhu and Thomas L. Griffiths. Eliciting the priors of large language models using iterated In-Context learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2025.

## A Implementation details

This section specifies the hyperparameters and design choices for the zero-shot prediction pipeline (Section 3.2). These choices are fixed across all experiments unless otherwise noted.

**Pathology foundation model.** We use UNI2-h as the embedding model $\phi$, a Vision Transformer pretrained on over 200 million pathology images, producing $D = 1{,}536$-dimensional embeddings per patch.

**Clustering and tile selection.** We set the number of clusters to $K = 10$ and the number of representative tiles per cluster to $T = 4$. Tiles are sampled uniformly at random from each cluster, with a seed derived from the slide identifier for reproducibility.

**LLM and scoring scale.** The multimodal LLM $\mathcal{M}$ is Gemini 3 Flash with temperature set to 0. The ordinal scale has $L = 5$ levels. Up to $P = 1{,}000$ cluster pairs are sampled per task from both within-slide and cross-slide combinations. The structured prompt provides tissue context (organ type and staining protocol) and brief functional descriptions of each target gene, which are generated by the LLM in a separate call and cached per task.

**Distance kernel.** We use a Gaussian kernel $\kappa(d;\, \sigma_s) = \exp\!\left(-d^2/2\sigma_s^2\right)$, with bandwidth set to half the median patch-centroid distance within each slide: $\sigma_s = \frac{1}{2}\,\mathrm{median}_{i,c}\, d_{ic}$.

**Supervised predictor.** We use an ensemble of $M = 2$ models trained on UNI2-h embeddings: (1) a Ridge regression model fit without intercept on the top 256 principal components of the standardized embeddings, with regularization strength $100/(256\,G)$ for $G$ genes, and (2) an MLP with two 256-dimensional hidden layers, each followed by LayerNorm, GELU activation, and dropout (rate 0.1), trained with MSE loss using Adam (learning rate $10^{-4}$, batch size 256, 200 epochs). Both models predict $\log(1 + x)$-transformed expression, and their predictions are averaged.

**Compute resources.** All experiments were run on a single NVIDIA RTX 4090 GPU.

## B Additional results

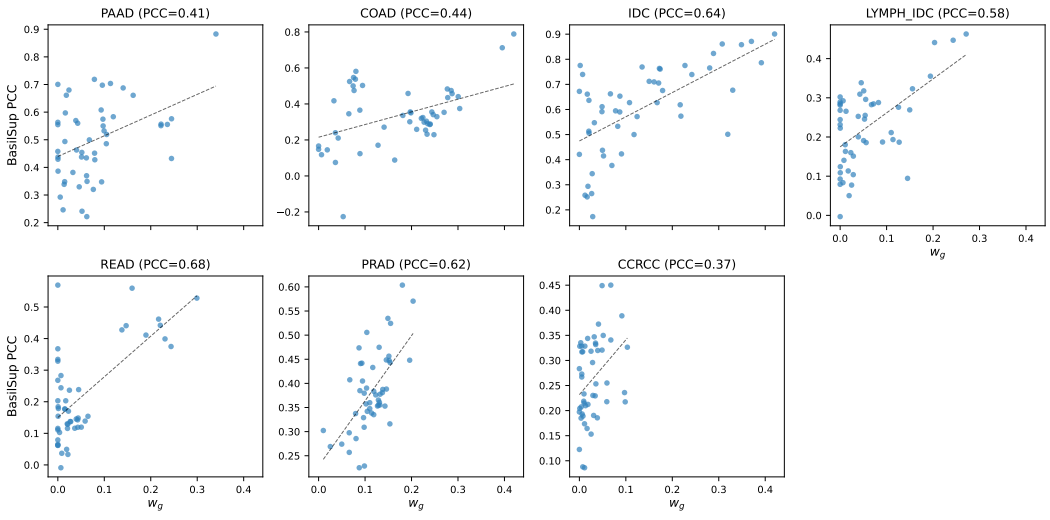


Figure 7: Per-gene zero-shot weight ($w_g$) vs. BasilSup prediction accuracy (PCC) on HEST-Bench. Each point is one gene-task pair. PCC is shown per task.

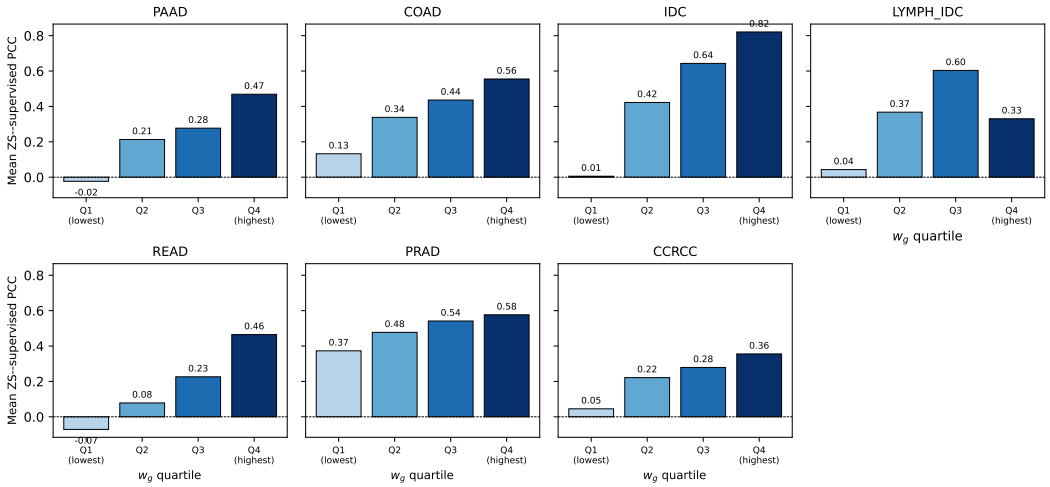


Figure 8: Mean per-gene PCC between zero-shot and supervised predictions, stratified by within-task $w_g$ quartile (all 7 HEST tasks). Genes with higher zero-shot weight show stronger agreement between the two predictors.

Table 5: Per-gene PCC between zero-shot and supervised predictions on HEST-Bench test sets. For each gene, we compute the PCC between the two predictors' spatial patterns across test patches, then average across genes and folds. Low PCC indicates complementary signal.

| Task | Mean PCC | Median PCC | Std | Pos. / 50 |
|---|---|---|---|---|
| READ | 0.146 | 0.156 | 0.260 | 38 |
| CCRCC | 0.238 | 0.247 | 0.177 | 45 |
| PAAD | 0.244 | 0.286 | 0.221 | 43 |
| L.IDC | 0.336 | 0.310 | 0.294 | 37 |
| COAD | 0.408 | 0.431 | 0.235 | 46 |
| IDC | 0.467 | 0.555 | 0.352 | 45 |
| PRAD | 0.476 | 0.505 | 0.116 | 50 |

## C STATISTICAL SIGNIFICANCE TESTS

We assess the statistical significance of BasilSup's performance gains using two complementary approaches: task-level paired tests across benchmark datasets, and per-gene tests within individual benchmarks.

**Task-level paired Wilcoxon signed-rank tests.** For each benchmark, we treat the per-task mean PCC values as paired observations and apply the two-sided Wilcoxon signed-rank test to compare BasilSup against each baseline. Table 6 reports the results. BasilSup significantly outperforms all baselines on both benchmarks ($p < 0.05$), where each baseline is compared only on the benchmarks where it was evaluated. With $n = 7$ and $n = 8$ paired observations, the minimum achievable two-sided Wilcoxon $p$-values are 0.016 and 0.008, respectively.

**Per-gene paired tests: BasilSup vs. supervised-only predictor.** To assess whether integrating the zero-shot LLM prior provides a statistically significant improvement over the supervised predictor alone, we apply the paired Wilcoxon signed-rank test to per-gene PCC values. On HEST-Bench (350 gene-task pairs), BasilSup significantly outperforms the supervised-only predictor ($p = 1.1 \times 10^{-29}$, mean improvement $= +0.014$ PCC, 72.3% of genes improved). The improvement is significant in 5 of 7 individual tasks ($p < 0.05$), with PAAD and READ not reaching significance. On KPMP (1000 genes), the overall improvement is smaller in magnitude but significant ($p = 5.4 \times 10^{-6}$), consistent with the zero-shot prior contributing less when the supervised predictor is trained on a large number of slides.

Table 6: Task-level paired Wilcoxon signed-rank tests comparing BasilSup vs. each baseline. $n$: number of tasks, W/L: wins/losses (tasks where BasilSup has higher/lower mean PCC), $\Delta$: mean difference in per-task PCC. $^{*}p < 0.05$, $^{**}p < 0.01$.

| | HEST-Bench ($n$=7) | | | STimage-1K4M Human ($n$=8) | | |
|---|---|---|---|---|---|---|
| Method | W/L | $\Delta$ | $p$ | W/L | $\Delta$ | $p$ |
| STRank | 7/0 | +0.060 | $0.016^{*}$ | 8/0 | +0.069 | $0.008^{**}$ |
| M2TGLGO | 7/0 | +0.131 | $0.016^{*}$ | 8/0 | +0.122 | $0.008^{**}$ |
| MERGE | 7/0 | +0.201 | $0.016^{*}$ | 8/0 | +0.245 | $0.008^{**}$ |
| HisToGene | 7/0 | +0.223 | $0.016^{*}$ | 8/0 | +0.257 | $0.008^{**}$ |
| Hist2ST | 7/0 | +0.258 | $0.016^{*}$ | 8/0 | +0.289 | $0.008^{**}$ |

**Per-gene unpaired tests: BasilSup vs. baselines.** For baselines where per-gene PCC values are available on HEST-Bench (HisToGene, Hist2ST, MERGE, M2TGLGO), we apply per-task Mann-Whitney $U$ tests comparing BasilSup's 50 per-gene PCC values against each baseline's 50 per-gene PCC values. All four baselines are significantly outperformed in the pooled analysis across all 350 gene-task pairs (HisToGene: $p = 6.7 \times 10^{-47}$, Hist2ST: $p = 2.6 \times 10^{-65}$, MERGE: $p = 1.1 \times 10^{-44}$, M2TGLGO: $p = 2.8 \times 10^{-17}$). At the individual task level, BasilSup achieves significant gains on all 7 tasks against HisToGene and Hist2ST, on 6 of 7 tasks against MERGE (COAD being the exception), and on 5 of 7 tasks against M2TGLGO (COAD and L.IDC being the exceptions).