PROTOTYPE-DRIVEN FUSION OF PATHOLOGY AND SPATIAL TRANSCRIPTOMICS FOR INTERPRETABLE MULTIPLE INSTANCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) measures gene expression in situ, complementing tissue morphology with spatially resolved molecular information. Together, these modalities provide rich morphological and molecular information, but identifying and integrating signals relevant to diagnosis and prognosis remains challenging under limited patient-level supervision. We introduce PathoSpatial, a multimodal multiple instance learning (MIL) framework integrating co-registered H&E images and ST through task-guided prototype learning. By combining supervised adaptation with epoch-level EMA updates over training slides, modality-specific prototype banks capture recurring morphological and molecular patterns, which are then fused through gated attention for prediction. The prototypes make each prediction interpretable: we characterize them biologically and use Owen values to attribute each prediction to prototypes grouped by modality, linking it to the histological patches and transcriptomic spots behind it. Across three cohort-level ST datasets spanning survival prediction, treatment-response classification, and disease-status classification, PathoSpatial provides this interpretability while performing comparably to the strongest multimodal baselines, supporting the investigation of patient-specific patterns for precision medicine.

1 INTRODUCTION

Whole-slide images (WSIs) capture tissue morphology informative of disease progression and patient outcomes (Campanella et al., 2019; Wulczyn et al., 2021; Yu et al., 2016). Their gigapixel resolution and limited availability of detailed annotations motivate multiple instance learning (MIL), which represents each slide as a bag of image patches and aggregates their features for prediction using only slide-level labels (Ilse et al., 2018; Lu et al., 2021). Beyond morphology, multimodal methods integrate histology with transcriptomic profiles to capture complementary molecular information and improve patient outcome prediction (Chen et al., 2022; Jaume et al., 2024; Song et al., 2024b). However, bulk expression measurements average signals across heterogeneous cells, obscuring localized molecular programs and their organization within the tissue.

Advances in spatial transcriptomics (ST) technologies have made spatially resolved molecular profiling increasingly accessible, enabling gene expression measurements within intact tissue sections (Naara et al., 2026; Yan et al., 2026; Ståhl et al., 2016; He et al., 2022; Merritt et al., 2020). Preserving tissue location allows molecular states to be examined alongside local morphology, providing a richer view of tissue heterogeneity than bulk profiles alone. Studies have linked expression programs at the tumor core and leading edge, as well as spatial tumor archetypes, to patient survival and treatment response (Arora et al., 2023; Wang et al., 2024). These findings motivate joint learning from paired H&E images and ST for diagnostic and prognostic prediction. Nevertheless, ST remains expensive and time-consuming, with major cohort-scale studies still limited to a few dozen patients.

This setting presents a joint representation-learning and fusion challenge: each slide provides many paired morphological and molecular measurements, yet patient-level supervision remains limited. The model must compress redundant instances while preserving task-relevant heterogeneity and integrate two bags whose features describe distinct aspects of the same tissue. Co-registration aligns the two modalities in space but not in relevance: which morphological and molecular patterns jointly

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.

inform prediction must be learned from weak patient-level labels. Prototype-based MIL provides a compact intermediate level at which recurring patterns from both modalities can be summarized and fused (Song et al., 2024a; Zhang et al., 2024b; Liu et al., 2025a). However, prototypes learned without labels (e.g., clustering) may favor dominant tissue patterns and miss subtle predictive signals, limiting predictive utility. Conversely, prototypes learned solely from outcome labels may overfit small cohorts and lack clear biological meaning. A hybrid approach could offer both: prototypes that stay tied to recurring patterns in the cohort but are tuned by the labels would remain interpretable while still being predictive.

Motivated by these insights, we introduce PathoSpatial, a multimodal MIL framework for patient-level prediction from co-registered WSIs and ST. PathoSpatial learns task-relevant morphological and molecular prototypes and integrates them into a compact representation for interpretable prediction. Our main contributions are: (1) **Task-guided prototype learning.** We combine unsupervised aggregation with supervised adaptation to capture recurring morphological and molecular patterns while aligning them with the prediction task. Attention-based fusion integrates these patterns with slide-specific weights. (2) **Prototype-level interpretation and attribution.** By summarizing similar instances into a small set of morphological and molecular prototypes, our framework provides a compact basis for Owen-value decomposition of each prediction into prototype and modality contributions. Tissue annotations, cell types, and pathway enrichment provide biological context for the prototypes, supporting hypothesis generation. (3) **Evaluation across modalities and tasks.** We evaluate PathoSpatial on three cohort-level ST datasets spanning treatment-response classification, survival prediction and disease-status classification, where it matches the strongest multimodal baselines while remaining interpretable.

2 Related Work

MIL in computational pathology. MIL aggregators differ in their use of spatial information. **Spatially unaware** methods treat slides as permutation-invariant sets of patch embeddings. Key-instance selectors use max/top-k pooling to capture localized signals but may miss diffuse patterns (Campanella et al., 2019); attention pooling forms weighted bag representations (Ilse et al., 2018; Lu et al., 2021; Li et al., 2021); and cluster-aware variants organize tissue phenotypes before pooling to reduce redundancy and stabilize learning (Yao et al., 2020; 2019). **Spatially aware** methods model slide layout: self-attention with positional encoding captures long-range and multi-scale tissue organization (Shao et al., 2021; Zheng et al., 2023; Zhao et al., 2022), while tile-adjacency graphs propagate local-to-global information, often with attention or hierarchical pooling (Chen et al., 2021a; Li et al., 2018; Zheng et al., 2022; Ding et al., 2023b). These approaches inform our context encoder; our focus is the subsequent compression and multimodal fusion.

Multimodal fusion. Combining histology with genomic and clinical information has shown promise for prognosis prediction (Mobadersany et al., 2018; Chen et al., 2022; Ding et al., 2023a; Volinsky-Fremond et al., 2024). **Late fusion** encodes modalities independently and merges their sample-level embeddings through concatenation or bilinear fusion (Chen et al., 2020; 2022). **Early fusion** models token-level interactions before per-modality pooling, using self-attention over joint token sets (Jaume et al., 2024) or cross-attention between modalities (Chen et al., 2021b; Xu & Chen, 2023; Zhou & Chen, 2023).

Histology and spatial transcriptomics. Paired H&E and ST have mainly been used to predict spot-level gene expression from histology, first with convolutional networks (He et al., 2020) and later with transformer–graph models (Zeng et al., 2022), contrastive image–expression embeddings (Xie et al., 2023), and super-resolution inference (Zhang et al., 2024a); visual–omics foundation models now align the two modalities at scale (Chen et al., 2025). These methods learn spot-level mappings or shared embeddings, whereas PathoSpatial uses paired spots and patches for patient-level prediction under weak supervision.

Prototype-based representations. Prototype methods summarize recurring histological patterns into compact descriptors of tissue heterogeneity. **Unsupervised approaches** derive task-agnostic bag representations through K-means (Vu et al., 2023; Yao et al., 2020; Hou et al., 2023), Gaussian mixture models (Song et al., 2024a), Leiden community detection (Claudio Quiros et al., 2024), or graph models (Hou et al., 2023; Claudio Quiros et al., 2024; Chan et al., 2023). Compositional vectors, aggregated prototype features, or pooled graph features then summarize biological varia-

tion within slides. **Task-conditioned approaches** instead optimize representations using task labels. For example, PIBD (Zhang et al., 2024b) uses information bottlenecks to learn risk-associated prototypes and select relevant features. **Prior-guided approaches** complement data-driven prototypes by incorporating textual or tissue-level semantic priors to guide instance aggregation and representation learning (Liu et al., 2025b; Shi et al., 2024; Wu et al., 2025). Such priors can support semantic interpretability, but their use requires suitable domain concepts and compatible representations. Task-relevant descriptions may be difficult to specify for endpoints such as treatment response, and extending language guidance to ST requires alignment between text and molecular features. PathoSpatial instead learns task-guided morphological and molecular prototypes from patient-level supervision, without requiring predefined textual priors or text-aligned encoders.

3 METHODS

PathoSpatial encodes paired histology and ST, aggregates each modality through task-guided prototypes, and fuses the resulting representations for task-specific prediction (Fig. 1).

3.1 PROBLEM FORMULATION AND FEATURE ENCODING

For slide i, a bag $B_i = \{(\mathbf{x}_{ij}^{(H)}, \mathbf{x}_{ij}^{(S)})\}_{j=1}^{N_i}$ contains N_i co-registered histology patches and ST spots. A bag is associated with an outcome y_i: a time/censoring pair for survival, or a binary label for classification (e.g., treatment response and disease status). Labels are given per patient and shared by all of that patient's slides, without patch- or spot-level task labels; the model predicts per slide, and predictions are averaged over each patient's slides.

We encode histology patches with UNI2 (Chen et al., 2024) and ST expression profiles with scVI (Lopez et al., 2018), yielding 1536- and 64-dimensional embeddings, respectively. Each modality $m \in \{H, S\}$ is projected to $d = 256$ dimensions, producing features $\mathbf{h}_{ij}^{(m)}$. A context encoder then produces contextual features $\mathbf{t}_{ij}^{(m)}$; in PathoSpatial-Attn, it is a TransMIL-inspired global self-attention encoder (Shao et al., 2021) with PPEG between two Transformer layers. We blend these with the projected inputs:

$$\tilde{\mathbf{h}}_{ij}^{(m)} = (1-\alpha)\,\mathrm{LayerNorm}(\mathbf{h}_{ij}^{(m)}) + \alpha\,\mathrm{LayerNorm}(\mathbf{t}_{ij}^{(m)}). \tag{1}$$

Here $\alpha \in (0,1)$ controls contextual refinement. The global self-attention encoder arranges the instance sequence in a padded square grid for PPEG; implementation details are in Appendix C.1.

PathoSpatial-GNN instead uses a local spatial context encoder that mixes each instance with its k_{nn} nearest spatial neighbors, repeated for L rounds of message passing ($k_{\mathrm{nn}} = 8$ and $L = 2$ by default), using each instance's physical coordinates directly. Both context encoders produce $\mathbf{t}_{ij}^{(m)}$ at the same dimension and normalization scale, so every downstream prototype operation (Sec. 3.2) is unaffected by which encoder is used; both variants are evaluated in Table 1, with encoder details in Appendix C.1.

3.2 TASK-GUIDED PROTOTYPE LEARNING

In each modality branch, the shared bank provides a common set of prototype queries across slides, while attention-based aggregation produces a separate representation for each slide. Task gradients adapt the bank to the labels; EMA updates use the aggregated training features to retain a cohort-derived reference. These roles motivate the combined update procedure.

Slide-specific aggregation. Each modality has a learnable prototype bank $\mathbf{P}^{(m)} = \{\mathbf{p}_k^{(m)}\}_{k=1}^{K_m}$, with $\mathbf{p}_k^{(m)} \in \mathbb{R}^d$ shared across slides. For prototype k, let $s_{ikj}^{(m)}$ be the cosine similarity between its projected query and the projected key of instance j (Appendix C.3). We retain the indices $\mathcal{I}_{ik}^{(m)} = \mathrm{TopK}_{k_{im}}(\mathbf{s}_{ik,:}^{(m)})$, where

$$k_{im} = \max\Big(1,\ \min\big(k_{\mathrm{cap}},\ \lfloor N_i \rho / K_m \rfloor,\ N_i\big)\Big). \tag{2}$$

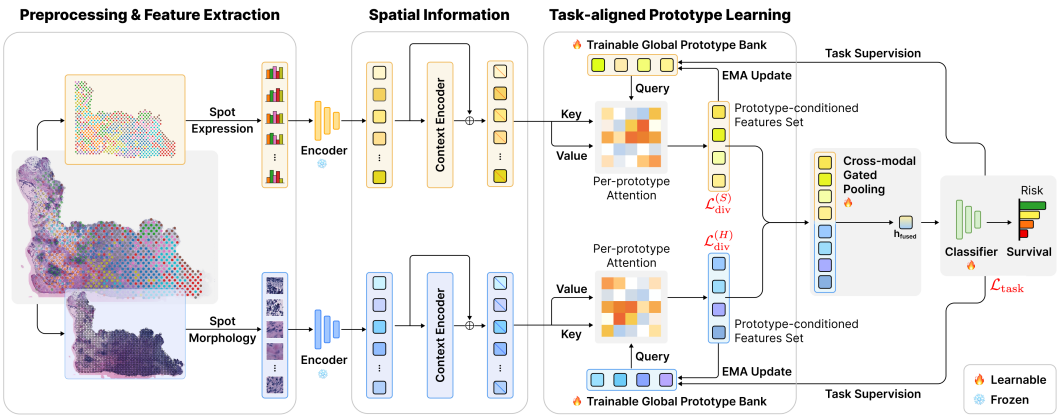

Figure 1: **PathoSpatial Overview.** Modality-specific encoders extract paired histology and ST features, while a context encoder captures spatial information. Prototype bank queries aggregate instances into slide-level representations for gated-attention prediction, with prototypes updated by supervised gradients and training-set EMA. Instance assignments enable post-hoc morphological and molecular characterization and prediction decomposition.

The cap k_{cap} bounds the number of selected instances per prototype and the proportion $\rho \in (0, 1]$ lets this bound scale with bag size; the outer $\max(1, \cdot)$ guarantees every prototype selects at least one instance, avoiding an empty centroid on very small bags. The full model uses $k_{\text{cap}} = 100$ and $\rho = 0.6$. Normalized weights and prototype-conditioned features are

$$\pi_{ikj}^{(m)} = \frac{\exp(s_{ikj}^{(m)})}{\sum_{j' \in \mathcal{I}_{ik}^{(m)}} \exp(s_{ikj'}^{(m)})}, \qquad j \in \mathcal{I}_{ik}^{(m)}, \tag{3}$$

$$\tilde{\mathbf{p}}_{ik}^{(m)} = \sum_{j \in \mathcal{I}_{ik}^{(m)}} \pi_{ikj}^{(m)} \tilde{\mathbf{h}}_{ij}^{(m)}. \tag{4}$$

Stacking these vectors gives $\tilde{\mathbf{P}}_i^{(m)} \in \mathbb{R}^{K_m \times d}$, a compact representation of the slide in modality m.

Shared prototype updates. The banks are randomly initialized and optimized with the network through the supervised training objective below. At the end of each epoch, we additionally update each bank using the mean prototype-conditioned features across training slides:

$$\mathbf{P}^{(m)} \leftarrow \beta \mathbf{P}^{(m)} + (1 - \beta) \bar{\mathbf{P}}_{\text{epoch}}^{(m)}. \tag{5}$$

Here $\beta \in (0, 1)$ is the EMA coefficient. All model parameters, including the feature projections, context encoders, prototype banks, fusion network, and prediction head, are jointly optimized by gradient descent throughout training. The prototype banks additionally receive the EMA update once per epoch using training slides only.

Prototype diversity. We use an assignment-based diversity loss to encourage prototypes within each modality to attend to distinct instances while maintaining prototype usage. For each slide and modality, we form the top-k_{im} masked assignment matrix from $\pi_{ikj}^{(m)}$, setting entries outside each prototype's selected set to zero. The loss combines an overlap penalty on the ℓ_2-normalized assignment rows with a usage-floor penalty:

$$\mathcal{L}_{\text{div}}^{(m)} = \mathcal{L}_{\text{ov}}^{(m)} + w_{\text{usage}} \mathcal{L}_{\text{usage}}^{(m)}. \tag{6}$$

The overlap term discourages prototypes from attending to the same instances. The usage term penalizes prototypes whose share of instances falls below a prescribed fraction of uniform usage, with w_{usage} controlling its contribution. Both terms are computed from the assignment distributions; their full definitions are given in Appendix B.

3.3 Cross-modal Fusion and Task-specific Training

We stack the two sets of slide-specific representations into $\mathcal{H}_i = [\tilde{\mathbf{P}}_i^{(H)}; \tilde{\mathbf{P}}_i^{(S)}] \in \mathbb{R}^{(K_H+K_S)\times d}$. By default, a gated attention network (Ilse et al., 2018) assigns normalized fusion attention weights a_r to its rows $\mathbf{h}_r$, giving

$$\mathbf{h}_{\text{fused}} = \sum_{r=1}^{K_H+K_S} a_r \mathbf{h}_r. \tag{7}$$

A task-specific head maps the fused representation to the prediction. For survival, the head outputs discrete-time hazard probabilities; for treatment-response and disease-status classification, it outputs class probabilities for the binary outcome. The common training objective is

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{task}} + \lambda_{\text{div}}\big(\mathcal{L}_{\text{div}}^{(H)} + \mathcal{L}_{\text{div}}^{(S)}\big). \tag{8}$$

Here $\mathcal{L}_{\text{task}}$ is the task loss: the censoring-aware negative log-likelihood (NLL) (Zadeh & Schmid, 2021) for survival and the class-weighted cross-entropy for classification (Appendix B.3), λ_{div} controls prototype regularization, and $\mathcal{L}_{\text{div}}^{(m)}$ is the assignment-based diversity loss defined above. We set $\lambda_{\text{div}} = 2$ for the full model.

3.4 Model Interpretation

Prototype identity. We characterize each prototype by its top-attended instances. Expression prototypes are described by pathway enrichment of their marker genes. Histology prototypes are matched to the tissue category (pathologist annotations) or cell type (label-transferred from a single-cell reference) most enriched among their top-attended instances relative to the slide background. Because prototype indices differ between cross-validation folds, each interpreted cohort is characterized with one fixed fold's model (Appendix G.1).

Individual prediction attribution. We compute exact Owen values (Owen, 1977) of the slide-specific prototype representations, grouping prototypes by modality. The explained output is the predicted probability of class 1 for classification and a risk score for survival, each relative to the prediction head's bias-only output. For each coalition, we retain its prototypes and renormalize their fusion attention weights before evaluating the output. Owen values sum to the difference between the full prediction and this baseline, and their sums within each modality give its net contribution. Because prototype-level signs partly reflect differences between modalities, we compare individual prototypes only within the same modality. Fusion attention weights describe aggregation and are distinct from these signed attributions; the formulation and its limits are given in Appendix G.2.

4 Experiments

4.1 Datasets

We evaluate PathoSpatial on three public cohorts that provide paired H&E whole-slide images and spatial transcriptomics data from the same tissue section, along with clinical endpoint labels; details are given in Appendix A.1.

TNBC Response (Yan et al., 2026): pre-treatment biopsies from 44 patients with TNBC profiled with Xenium, with labels indicating pathological complete response versus residual disease after neoadjuvant chemotherapy. **TNBC Survival** (Wang et al., 2024): 92 patients with early-stage TNBC (273 ST sections/samples), with five survival endpoints: IDFS, IBCFS, DRFS, RFS, and OS. **Lung Disease** (Vannan et al., 2025): one representative Xenium sample from each of 35 donors (26 with pulmonary fibrosis, 9 unaffected), with disease-status labels.

These are, to our knowledge, the largest publicly available ST cohorts with paired training data (i.e., spatial transcriptomics & high resolution digital histopathology images) and clinical outcome labels.

4.2 Baseline Methods

We evaluate two families of multimodal fusion baselines that differ in molecular representation and fusion strategy. All methods use the same UNI2 histology features, and the H+S baselines and

Table 1: Multimodal prediction benchmark. TNBC Survival: C-index (mean ± SD over five folds); TNBC Response and Lung Disease: AUC and accuracy (ACC) on pooled leave-one-out predictions.

	TNBC Response ($N = 44$)		TNBC Survival ($N = 92$) C-index ↑						Lung Disease ($N = 35$)	
Method	AUC ↑	ACC ↑	IDFS	IBCFS	DRFS	RFS	OS	Mean	AUC ↑	ACC ↑
Customized late fusion (H+S, selected concat/add/dot)										
CLAM	0.684	0.636	0.671 ± 0.162	0.716 ± 0.113	0.659 ± 0.044	0.670 ± 0.090	0.651 ± 0.076	0.673	**1.000**	0.743
TransMIL	0.646	0.636	0.724 ± 0.120	0.715 ± 0.059	0.654 ± 0.110	0.657 ± 0.052	0.669 ± 0.103	0.684	**1.000**	**1.000**
Patch-GCN	0.686	0.523	0.669 ± 0.172	0.685 ± 0.078	0.667 ± 0.133	0.666 ± 0.066	0.633 ± 0.062	0.664	**1.000**	0.800
PANTHER	0.657	0.591	0.718 ± 0.143	0.688 ± 0.090	0.674 ± 0.086	0.646 ± 0.022	0.651 ± 0.027	0.675	**1.000**	0.971
Pseudo-bulk fusion (H+G)										
PORPOISE	0.611	0.568	0.724 ± 0.137	0.717 ± 0.118	0.667 ± 0.080	**0.678 ± 0.053**	0.651 ± 0.034	0.687	**1.000**	0.971
MCAT	0.600	0.614	0.701 ± 0.082	0.722 ± 0.056	0.656 ± 0.081	0.657 ± 0.080	0.640 ± 0.033	0.675	**1.000**	0.971
SurvPath	0.537	0.523	0.720 ± 0.112	0.726 ± 0.079	**0.692 ± 0.082**	0.652 ± 0.083	0.653 ± 0.073	0.689	**1.000**	0.743
PIBD	0.568	0.523	0.670 ± 0.065	0.649 ± 0.101	0.644 ± 0.065	0.672 ± 0.058	0.653 ± 0.043	0.658	**1.000**	0.943
ProSurv	0.589	0.545	0.700 ± 0.158	0.687 ± 0.100	0.672 ± 0.083	0.664 ± 0.082	**0.673 ± 0.024**	0.679	**1.000**	0.743
PathoSpatial-Attn	**0.691**	0.659	0.684 ± 0.131	0.718 ± 0.056	0.659 ± 0.110	0.651 ± 0.061	0.652 ± 0.140	0.673	**1.000**	0.971
PathoSpatial-GNN	0.680	**0.682**	**0.729 ± 0.117**	**0.727 ± 0.103**	0.684 ± 0.104	0.665 ± 0.096	0.659 ± 0.111	**0.693**	**1.000**	0.971

Mean: unweighted average over the five survival endpoints. Bold and underline mark the best and second-best values in each column; ties share highlighting. H: histology; S: spot-level ST; G: pseudo-bulk expression. The evaluation protocol is described in Sec. 4.3.

PathoSpatial also share the same scVI spot features, so differences reflect aggregation and fusion rather than feature extraction:

Customized late fusion (H+S): We encode co-registered histology patches with UNI2 and ST spots with scVI, aggregate each modality with a MIL backbone, and combine the resulting representations using concatenation (concat), addition (add), or dot-product (dot) fusion. The four backbones cover distinct aggregation strategies: CLAM (Lu et al., 2021) uses instance clustering-constrained attention; Patch-GCN (Chen et al., 2021a) models local spatial neighborhoods; TransMIL (Shao et al., 2021) models global correlations; and PANTHER (Song et al., 2024a) uses unsupervised prototypes. For each backbone, Table 1 reports the best of the three operators, selected separately for each cohort and survival endpoint.

Pseudo-bulk fusion (H+G): We pair histology with expression summed across ST spots. PORPOISE (Chen et al., 2022) and ProSurv (Liu et al., 2025a) use profiles of 2,000 genes selected by median absolute deviation, with concatenation fusion. MCAT (Chen et al., 2021b), SurvPath (Jaume et al., 2024), and PIBD (Zhang et al., 2024b) use expression grouped into 50 Hallmark pathways in our implementation. These three represent co-attention, self-attention, and supervised prototype-based integration, respectively. Encoding and implementation details are provided in Appendix B.

4.3 Implementation Details

Data: For TNBC Survival, each ST spot ($100\,\mu$m diameter) is paired with the co-registered, non-overlapping 256×256-pixel H&E patch; the WSI is at $\sim 26\times$ magnification ($0.39\,\mu$m/pixel). For the two Xenium cohorts, we tile the H&E image into 256×256-pixel patches and sum the transcripts in each patch into a pseudo-spot (Appendix A.1). We apply GrandQC (Weng et al., 2024) quality control to the WSIs, then extract patch features with UNI2 (Chen et al., 2024) and spot features with scVI (Lopez et al., 2018). Baselines that require bulk expression use pseudo-bulk expression summed over all spots.

Full model: Unless swept in the sensitivity analysis (Appendix E), we use $K_H = K_S = 12$ prototypes; top-k instance selection with proportion $\rho = 0.6$ and cap $k_{\text{cap}} = 100$; joint gradient and EMA prototype updates with EMA coefficient $\beta = 0.5$; a fixed feature-blend coefficient $\alpha = 0.5$; assignment-based prototype diversity with $\lambda_{\text{div}} = 2$; and joint gated attention over the stacked prototype pool. PathoSpatial-GNN uses the local spatial context encoder, with $k_{\text{nn}} = 8$ neighbors and $L = 2$ rounds (Sec. 3.1); PathoSpatial-Attn uses the global self-attention context encoder (TransMIL/PPEG). We refer to PathoSpatial-GNN with these settings as the *full model*.

Training and evaluation: Following common practice in multimodal survival MIL (Zhang et al., 2024b), we use five-fold cross-validation for TNBC Survival, train all methods for 30 epochs, report the C-index as mean ± SD across folds, and define Kaplan–Meier risk groups by each fold's median

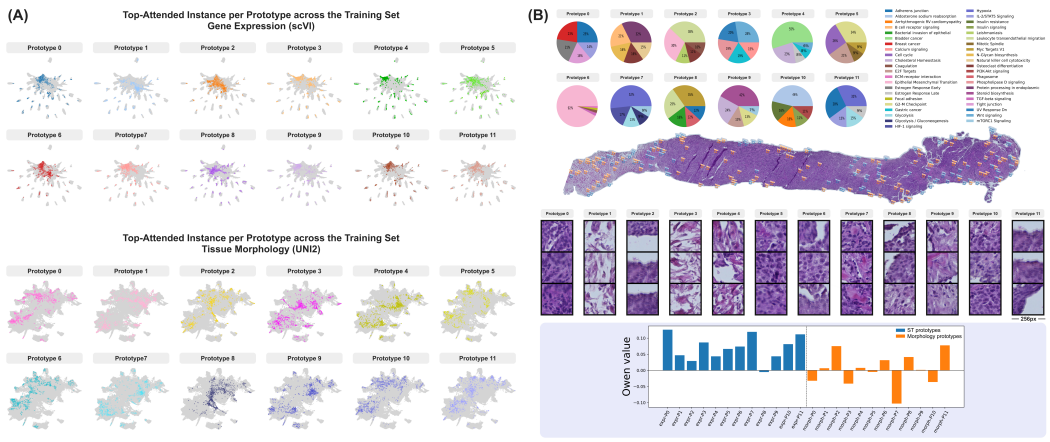

Figure 2: **Multimodal prototype interpretation on TNBC Response.** Histology and ST prototypes are learned jointly within PathoSpatial ($K_H = K_S = 12$). (A) UMAPs of instance features across the training set for gene expression (scVI; top) and tissue morphology (UNI2; bottom); in each small plot, colored points are the instances most attended by that prototype. (B) Held-out patient P87. Top: pathway composition of each ST prototype. Middle: the top 10 attended instances of each histology prototype on the whole-slide image, with representative 256×256-pixel patches for each prototype. Bottom: Owen-value contribution of each ST (expr) and histology (morph) prototype to the prediction. Appendix Fig. 5 shows the full global prototype characterizations in the training set.

predicted risk. Folds are split by patient and stratified by survival-time bin and censoring status, and each fold's best epoch is reported. For TNBC Response and Lung Disease, we use leave-one-out cross-validation and report AUC and accuracy on the pooled held-out predictions at the epoch with the highest AUC. Further details are given in Appendix B.

5 Results

5.1 Survival Prediction

On TNBC survival prediction, PathoSpatial-GNN achieves the highest mean C-index across the five endpoints (0.693), leading on IDFS and IBCFS, while remaining competitive with the strongest baselines on the other three (Table 1).

Local spatial context. PathoSpatial-GNN exceeds PathoSpatial-Attn on all five endpoints (e.g., IDFS 0.729 versus 0.684). We attribute this in part to the local spatial context encoder, which mixes each instance only with its spatial neighbors and so preserves the tissue neighborhoods that the prototypes summarize, whereas global self-attention can mix distant regions before aggregation. The context-encoder ablation shows the same pattern (Sec. 5.4).

Prototype-level fusion. With the same spot-level features, PathoSpatial-GNN exceeds all four late-fusion baselines on IDFS, IBCFS, and DRFS, but trails the best of them on RFS (CLAM, 0.670) and OS (TransMIL, 0.669). Since these baselines differ from PathoSpatial mainly in how instances are aggregated and fused, this is consistent with prototype-level fusion adding to the choice of MIL backbone. Pseudo-bulk methods remain competitive, suggesting that summed expression already carries much of the molecular signal for survival at this cohort size. Because molecular resolution, encoding, and fusion architecture vary together across these families, the comparisons assess complete pipelines and do not isolate the effect of spatial resolution.

Risk stratification. Fig. 3A compares PathoSpatial-GNN with SurvPath, the strongest baseline by mean C-index, with risk groups defined by each fold's median predicted risk. PathoSpatial-GNN has the smaller log-rank p-value on all five endpoints and reaches $p < 0.05$ on all five, versus four for SurvPath (all except RFS). The two methods' fold-averaged C-indices (Table 1) differ by at most 0.013. With 92 patients and correlated endpoints, the power to detect differences between methods with similar C-indices is limited, motivating evaluation on larger cohorts as they become available.

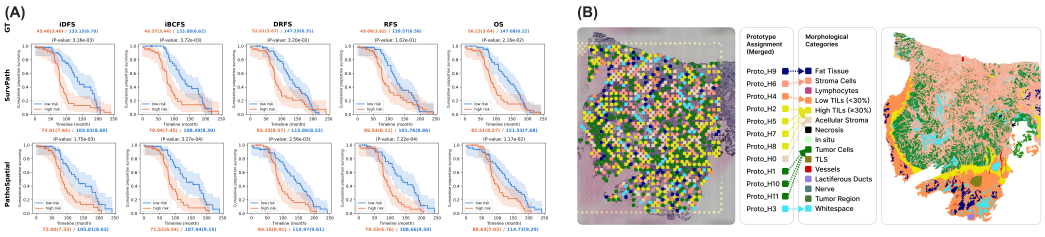

Figure 3: **Risk stratification and prototype matching on TNBC Survival.** (A) Kaplan–Meier curves of high-risk (orange) and low-risk (blue) patients for SurvPath (top) and PathoSpatial-GNN (bottom), split at each fold's median predicted risk, with log-rank p-values. Numbers give each group's mean (SD) observed time in months; the GT row splits patients at their median observed time. (B) Sample TNBC82-CN41-E1: histology prototype assignments on the ST spots (left), with prototypes matched to the pathologist's morphological categories (right) by contrast-weighted enrichment; 6 of 12 matches are significant (Appendix G.1).

5.2 Classification Performance

Overall, PathoSpatial variants give the best AUC and accuracy on TNBC Response, while Lung Disease is at ceiling for all methods. On TNBC Response (44 patients, leave-one-out), PathoSpatial-Attn achieves the highest reported AUC (0.691), followed by Patch-GCN (0.686). PathoSpatial-GNN reports an AUC of 0.680 and the highest accuracy (0.682), followed by PathoSpatial-Attn (0.659). On Lung Disease (35 donors, leave-one-out), all eleven configurations achieve a best-epoch AUC of 1.000. Both PathoSpatial variants report a corresponding accuracy of 0.971. TransMIL achieves the highest accuracy (1.000), followed by both PathoSpatial variants, PANTHER, PORPOISE, and MCAT (all 0.971). With 44 and 35 patients, formal significance testing has limited power; we report the point estimates and defer statistical comparisons to larger cohorts.

5.3 Interpretability

This section focuses on patient-level interpretation: for each TNBC cohort, we interpret one held-out patient that the model predicts correctly and briefly summarize the cohort-level prototype characterization, which is given in full in Appendix G. Lung Disease, where all methods reach an AUC of 1.000, serves as a positive control and is not interpreted. Expression prototypes are characterized by pathway enrichment of their top-attended instances, and histology prototypes by their enrichment for annotated tissue categories or cell types (Appendix G.1).

TNBC Survival. Slide TNBC82-CN41-E1 had an IDFS event at 49.9 months and is ranked at the 84th percentile of predicted risk. Across the training set, the expression prototypes capture EMT and extracellular-matrix stroma (4 of 12), immune programs (antigen presentation, interferon response), metabolic stress (hypoxia, cholesterol homeostasis), and proliferation (Appendix Fig. 6). Six of 12 histology prototypes are significantly enriched for an annotated category: tumor, high-TIL stroma, acellular stroma, or whitespace (Fig. 3B). For this patient, Owen values attribute the high risk to the expression branch (net $+1.95$), led by a lipogenic prototype and two macrophage prototypes, consistent with the association between stromal macrophages and reduced breast cancer survival (Medrek et al., 2012); the histology branch lowers the predicted risk (net -1.21).

TNBC Response. Patient P87, a responder, is predicted to respond with probability 0.72. Across the training set, half of the expression prototypes are dominated by hypoxia and HIF-1 signaling, an axis linked to chemoresistance in TNBC (Xiong et al., 2018) (Appendix Fig. 5). In P87, however, only one prototype retains this identity; the others capture proliferation (G2-M checkpoint, E2F targets) and cholesterol biosynthesis (Fig. 2B), consistent with the association between cell-cycle activity and response reported for this cohort (Yan et al., 2026). The slide is also T-cell rich, in line with the association between tumor-infiltrating lymphocytes and pCR (Denkert et al., 2010), and the histology prototypes separate fibroblast-, tumor-, and B-cell-enriched patches. Fusion attention is close to uniform for this patient, so no single prototype drives the prediction. These analyses are descriptive; details and limitations are given in Appendices G.4 and G.3.

Table 2: Ablation summary. Each row changes one component of the full model (first row).

		TNBC Response ($N = 44$)		TNBC Survival ($N = 92$) C-index ↑					Lung Disease ($N = 35$)	
Axis	Variant	AUC ↑	ACC ↑	IDFS	IBCFS	DRFS	RFS	OS	AUC ↑	ACC ↑
–	Full model (PathoSpatial-GNN)	0.680	**0.682**	**0.729 ± 0.117**	0.727 ± 0.103	**0.684 ± 0.104**	**0.665 ± 0.096**	0.659 ± 0.111	**1.000**	0.971
Prototype updates	Gradient only	0.672	0.545	0.675 ± 0.073	0.684 ± 0.015	0.610 ± 0.130	0.634 ± 0.050	0.641 ± 0.166	**1.000**	0.971
Modality contribution	H&E only	0.646	0.636	0.684 ± 0.083	0.641 ± 0.098	0.720 ± 0.097	0.672 ± 0.151	0.611 ± 0.132	**1.000**	0.943
Modality contribution	ST only	0.656	**0.682**	0.710 ± 0.060	0.671 ± 0.025	0.595 ± 0.110	0.626 ± 0.049	**0.663 ± 0.082**	**1.000**	**1.000**
Aggregation	Within-modality attention	0.587	0.568	0.679 ± 0.148	**0.731 ± 0.113**	0.662 ± 0.124	0.646 ± 0.061	0.611 ± 0.109	**1.000**	**1.000**
Prototype diversity	Cosine	0.653	0.636	0.655 ± 0.115	0.668 ± 0.067	0.598 ± 0.121	0.647 ± 0.056	0.603 ± 0.137	**1.000**	0.886
Context encoder	Global self-attention	**0.691**	0.659	0.684 ± 0.131	0.718 ± 0.056	0.659 ± 0.110	0.651 ± 0.061	0.652 ± 0.140	**1.000**	0.971

Full-model values are as in Table 1; both modalities are used unless stated. Metrics as in Table 1; bold marks the best value in each column, and ties share bold.

5.4 Ablation Study and Sensitivity Analysis

Ablation study. We ablate five design axes of the full model (Sec. 4.3), changing one component at a time: the prototype update rule, modality contribution, cross-modal aggregation, prototype diversity, and context encoder. Table 2 reports representative variants, and Appendix D reports every variant. Every component contributes on most endpoints, and the diversity formulation and prototype update rule have the largest average effects. Removing the EMA update lowers all five survival C-indices (DRFS 0.684 to 0.610), consistent with the epoch-level EMA updates over training slides stabilizing prototypes under limited supervision. Neither modality alone matches the full model: histology alone is weaker on IDFS and OS and ST alone on IBCFS and DRFS, indicating complementary signals. Joint attention exceeds within-modality attention on four endpoints, as weighting prototypes across both modalities lets each slide shift weight between them. Assignment-based diversity exceeds cosine-based diversity on every survival endpoint (DRFS 0.684 versus 0.598); because cluster centroids in this feature space are not mutually orthogonal (Appendix D.4), penalizing overlapping instance assignments constrains the prototypes less than penalizing vector similarity. Replacing the local spatial context encoder with global self-attention lowers all five survival C-indices, in line with Sec. 5.1. Lung Disease AUC remains 1.000 across ablations. **Sensitivity analysis.** In Appendix Table 6, the default configuration performs best on four of the five survival endpoints and on TNBC Response AUC. Changing any of the four hyperparameters (number of prototypes, ρ, β, or α) lowers the survival C-index by similar amounts, whereas Lung Disease AUC stays at or near 1.000. The ablation and sensitivity differences are mostly within the fold-level standard deviations and were not tested for significance.

6 Conclusions and Limitations

We present PathoSpatial, a multimodal and interpretable MIL framework that integrates co-registered WSIs and spatial transcriptomics through task-guided prototypes, linking spatial expression programs with histopathological patterns. Against customized histology–ST late fusion and histology–pseudo-bulk fusion, PathoSpatial matches or exceeds the strongest baselines across tasks: on TNBC Survival, PathoSpatial-GNN achieves the highest mean c-index and separates risk groups on all five endpoints (log-rank p ¡ $p < 0.05$). Its main strength is interpretability at matched performance. Each slide is summarized by a small set of recurring morphological and molecular patterns that align with pathologist annotations, cell types, and known tumor biology. Owen values then trace each prediction to these prototypes and to each modality, making the morphological and molecular evidence behind it inspectable and supporting hypothesis generation from spatial omics. The main limitation is cohort size, reflecting the current scarcity of population-level ST cohorts with paired H&E images. Such cohorts remain scarce and small because ST technologies are recent, while patient-level labels such as treatment response or long-term survival require clinical follow-up. With 92, 44, and 35 patients, the cohorts are too small to distinguish methods, and larger external cohorts are needed to quantify predictive gains more precisely and test generalization beyond the cancer types studied here. Dataset resolution and the number of hyperparameters are further limitations. As labeled ST cohorts grow, PathoSpatial's prototype framework is designed to scale with them, and its interpretability design, including prototype characterization and Owen-value attribution applies without modification to new cohorts and endpoint targets.

AI USE STATEMENT

We used a large language model (LLM) solely for minor editorial assistance, including language polishing and typo correction. The LLM was not used for research ideation, experimental design, data analysis, interpretation, or substantive content generation.

REPRODUCIBILITY STATEMENT

Section 3 describes the model, and Section 4 describes the experimental setup. Appendix A provides dataset and preprocessing details; Appendix B gives implementation details. Further ablation settings and sensitivity analyses are provided in Appendices D and E. All three cohorts are from published studies (Appendix A.1). The code for this project will be released upon acceptance.

ETHICS STATEMENT

This work uses only publicly available data from previously published studies; no new patient data were collected. Each original study was approved by its institutional ethics board, with informed consent from participants: TNBC Survival by the Local Ethical Review Board of the Institut Jules Bordet (reference CE2862) (Wang et al., 2024); TNBC Response by the Institutional Review Board of the University of Texas MD Anderson Cancer Center within the ARTEMIS trial (NCT02276443) (Yan et al., 2026); and Lung Disease by the Vanderbilt University Medical Center (060165, 171657) and Western (20181836) Institutional Review Boards (Vannan et al., 2025).

REFERENCES

Ruchan Arora, Calvin Cao, Maryam Kumar, Sarthak Sinha, Alvin Chanda, Robert McNeil, Davis Samuel, Rajan K Arora, T Wayne Matthews, Shamir Chandarana, and Robert Hart. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14(1):5029, 2023. doi: 10.1038/s41467-023-40806-4.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Tsun-Hin Chan, Fernando J Cendra, Lei Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15661–15670, 2023. doi: 10.1109/CVPR52729.2023.01503.

Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 339–349. Springer International Publishing, 2021a. doi: 10.1007/978-3-030-87237-3_33. URL https://doi.org/10.1007/978-3-030-87237-3_33.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021b.

Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, Zahra Noor, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 2022.

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

Weiqing Chen, Pengzhi Zhang, Tu N. Tran, Yiwei Xiao, Shengyu Li, Vrutant V. Shah, Hao Cheng, Kristopher W. Brannan, Keith Youker, Li Lai, Longhou Fang, Yu Yang, Nhat-Tu Le, Jun-ichi Abe, Shu-Hsia Chen, Qin Ma, Ken Chen, Qianqian Song, John P. Cooke, and Guangyu Wang. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, 2025. doi: 10.1038/s41592-025-02707-1.

Adalberto Claudio Quiros, Nicolas Coudray, Anna Yeaton, Xinyu Yang, Bojing Liu, Hortense Le, Luis Chiriboga, Afreen Karimkhan, Navneet Narula, David A Moore, and Christopher Y Park. Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. *Nature Communications*, 15(1):4596, 2024. doi: 10.1038/s41467-024-48666-7.

Carsten Denkert, Sibylle Loibl, Aurelia Noske, Marc Roller, Berit Maria Müller, Martina Komor, Jan Budczies, Silvia Darb-Esfahani, Ralf Kronenwett, Claus Hanusch, Christian von Törne, Wilko Weichert, Knut Engels, Christine Solbach, Iris Schrader, Manfred Dietel, and Gunter von Minckwitz. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *Journal of Clinical Oncology*, 28(1):105–113, 2010. doi: 10.1200/JCO.2009.23.7370.

Kexin Ding, Mu Zhou, Dimitris N Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In Hayit Greenspan et al. (eds.), *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023*, volume 14225 of *Lecture Notes in Computer Science*, pp. 622–632, Cham, 2023a. Springer. doi: 10.1007/978-3-031-43987-2_60.

Saisai Ding, Juncheng Li, Jun Wang, Shihui Ying, and Jun Shi. Multi-scale efficient graph-transformer for whole slide image classification. *IEEE Journal of Biomedical and Health Informatics*, 27(12):5926–5936, 2023b. doi: 10.1109/JBHI.2023.3317067.

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.

Shanshan He, Ruchir Bhatt, Carl Brown, Emily A Brown, Derek L Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G Garrison, Gary Geiss, and Mark T Gregory. High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology*, 40(12):1794–1806, 2022. doi: 10.1038/s41587-022-01483-z.

Wentai Hou, Yuyao He, Bingzheng Yao, Lequan Yu, Renping Yu, Feng Gao, and Liansheng Wang. Multi-scope analysis driven hierarchical graph transformer for whole slide image based cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 745–754, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43987-2_72.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, Paul Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016. doi: 10.1093/nar/gkw377.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5491–5500. PMLR, 2020. URL https://proceedings.mlr.press/v119/kumar20e.html.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021.

Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 174–182, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-030-00934-2_20.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

Fengchun Liu, Linghan Cai, Zhikang Wang, Zhiyuan Fan, Jin-gang Yu, Hao Chen, and Yongbing Zhang. Prototype-Guided Cross-Modal Knowledge Enhancement for Adaptive Survival Prediction . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15965. Springer Nature Switzerland, September 2025a.

Pei Liu, Luping Ji, Jiaxiang Gou, Bo Fu, and Mao Ye. Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In *International Conference on Learning Representations*, 2025b. URL https://iclr.cc/virtual/2025/poster/28016.

Sherene Loi, Damien Drubay, Sylvia Adams, Giancarlo Pruneri, Prudence A. Francis, Magali Lacroix-Triki, Heikki Joensuu, Maria Vittoria Dieci, Sunil Badve, Sandra Demaria, et al. Tumor-infiltrating lymphocytes and prognosis: a pooled individual patient analysis of early-stage triple-negative breast cancers. *Journal of Clinical Oncology*, 37(7):559–569, 2019. doi: 10.1200/JCO.18.01010.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2. URL https://doi.org/10.1038/s41592-018-0229-2.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

Catharina Medrek, Fredrik Pontén, Karin Jirström, and Karin Leandersson. The presence of tumor associated macrophages in tumor stroma as a prognostic marker for breast cancer patients. *BMC Cancer*, 12:306, 2012. doi: 10.1186/1471-2407-12-306.

Christopher R Merritt, Giang T Ong, Sarah E Church, Kinisha Barker, Patrick Danaher, Gary Geiss, Margaret Hoang, Jane Jung, Yan Liang, Jodi McKay-Fleisch, and Khoa Nguyen. Multiplex digital spatial profiling of proteins and rna in fixed tissue. *Nature Biotechnology*, 38(5):586–599, 2020. doi: 10.1038/s41587-020-0472-9.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018. doi: 10.1073/pnas.1717139115.

Shorook Naara, Veena Kochat, Xiayu Rao, Emre Arslan, Robert Saddawi-Konefka, Suresh Satpati, Jennifer Garbarino, Jennifer L. Anderson, Frederico O. Gleber-Netto, Priyadharsini Nagarajan, Travis D. Kerr, Shamima Akhter, Sophie Li, R'ay Fodor, Shlomo A. Koyfman, Dan Yaniv, Tongxin Xie, Mica Glaun, Michael Bobian, William Israel Padron, Colin Ng, Preethi H. Gunaratne, Shiyanth Thevasagayampillai, Michael R. Migden, Hussein A. Abbas, Patrick K. Reville, K. Y. Tsai, D. J. McGrail, J. Wang, J. N. Myers, N. D. Gross, K. Rai, and M. Amit. Spatial biology reveals altered macrophage states in immunosuppressed non-melanoma skin cancer. *Cell*, 189(19):6170–6188.e9, 2026. doi: 10.1016/j.cell.2026.07.031.

Guillermo Owen. Values of games with a priori unions. In Rudolf Henn and Otto Moeschlin (eds.), *Mathematical Economics and Game Theory*, pp. 76–88. Springer, 1977. doi: 10.1007/978-3-642-45494-3_7. URL https://link.springer.com/chapter/10.1007/978-3-642-45494-3_7.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. ViLa-MIL: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://cvpr.thecvf.com/virtual/2024/poster/31886.

Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag Jayant Vaidya, Alexander Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. In *Forty-first International Conference on Machine Learning*, 2024b.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, and Annelie Mollbrink. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. doi: 10.1126/science.aaf2403.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 2020. URL https://proceedings.mlr.press/v119/sundararajan20b.html.

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

Stephanie Volinsky-Fremond, Nanda Horeweg, Sarah Andani, Jelmar Barkey Wolf, Maxime W Lafarge, Cornelis D de Kroon, Gitte Ørtoft, Estrid Høgdall, Jouke Dijkstra, Jan Jakob Jobsen, et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nature Medicine*, pp. 1–12, 2024. doi: 10.1038/s41591-024-03005-w.

Quoc Dang Vu, Kashif Rajpoot, Shan E Ahmad Raza, and Nasir Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical Image Analysis*, 85:102743, 2023. doi: 10.1016/j.media.2023.102743.

Xiao Wang, David Venet, Floriane Lifrange, Denis Larsimont, Matteo Rediti, Luc Stenbeck, Françoise Dupont, Grégory Rouas, Alexandre J Garcia, Ligia Craciun, and Laurence Buisseret. Spatial transcriptomics reveals substantial heterogeneity in triple-negative breast cancer with potential clinical implications. *Nature Communications*, 15(1):10232, 2024. doi: 10.1038/s41467-024-54445-6.

Zhilong Weng, Alexander Seper, Alexey Pryalukhin, Fabian Mairinger, Claudia Wickenhauser, Marcus Bauer, Lennert Glamann, Hendrik Bläker, Thomas Lingscheidt, Wolfgang Hulla, Danny Jonigk, Simon Schallenberg, Andrey Bychkov, Junya Fukuoka, Martin Braun, Birgid Schömig-Markiefka, Sebastian Klein, Andreas Thiel, Katarzyna Bozek, George J. Netto, Alexander Quaas, Reinhard Büttner, and Yuri Tolkach. Grandqc: A comprehensive solution to quality control problem in digital pathology. *Nature Communications*, pp. 10685, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54769-y. URL https://doi.org/10.1038/s41467-024-54769-y.

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.

Junxian Wu, Minheng Chen, Xinyi Ke, Tianwang Xun, Xiaoming Jiang, Hongyu Zhou, Lizhi Shao, and Youyong Kong. Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://cvpr.thecvf.com/virtual/2025/poster/34433.

Ellery Wulczyn, David F Steiner, Michael Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Thomas Brown, Peter Regitnig, Po-Hsuan Chen, and Niels Hegde. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digital Medicine*, 4 (1):71, 2021. doi: 10.1038/s41746-021-00427-2.

Ronald Xie, Kuan Pang, Sai W. Chung, Catia T. Perciani, Sonya A. MacParland, Bo Wang, and Gary D. Bader. Spatially resolved gene expression prediction from H&E histology images via bimodal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Gaofeng Xiong, Rachel L. Stewart, Jie Chen, Tianyan Gao, Timothy L. Scott, Luis M. Samayoa, Kathleen O'Connor, Andrew N. Lane, and Ren Xu. Collagen prolyl 4-hydroxylase 1 is essential for HIF-1α stabilization and TNBC chemoresistance. *Nature Communications*, 9:4456, 2018. doi: 10.1038/s41467-018-06893-9.

Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21241–21251, October 2023.

Yun Yan, Yiyun Lin, Tapsi Kumar, Shanshan Bai, Aatish Thennavan, Jianzhuo Li, Emi Sei, Tuan Tran, Min Hu, Mitchell Rao, Chenling Tang, Siyuan He, Anna Casasent, Elizabeth Ravenberg, Gaiane Margishvili Rauch, Alyson R. Clayborn, Debu Tripathy, Alastair Thompson, Bora Lim, Lei Huo, Stacy Moulder, Clinton Yam, and Nicholas Navin. Ecotypes of triple-negative breast cancer in response to chemotherapy. *Nature*, 654(8120):1088–1097, 2026. doi: 10.1038/s41586-026-10469-9.

Jiawen Yao, Xinliang Zhu, and Junzhou Huang. Deep multi-instance learning for survival prediction from whole slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I*, pp. 496–504. Springer, 2019. doi: 10.1007/978-3-030-32239-7_55.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, October 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101789. URL http://dx.doi.org/10.1016/j.media.2020.101789.

Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(1):12474, 2016. doi: 10.1038/ncomms12474.

Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3126–3137, 2021. doi: 10.1109/TPAMI.2020.2979450.

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.

Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle Y. Y. Lee, Kyung S. Cho, Katalin Susztak, George X. Xu, Michael D. Feldman, Edward B. Lee, Emma E. Furth, Linghua Wang, and Mingyao Li. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 42(9):1372–1377, 2024a. doi: 10.1038/s41587-023-02019-9.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=otHZ8JAIgh.

Yuxuan Zhao, Zhenrong Lin, Kai Sun, Yi Zhang, Junzhou Huang, Lin Wang, and Jiawen Yao. Setmil: Spatial encoding transformer-based multiple instance learning for pathological image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 66–76, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16434-7_7.

Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama. A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11):3003–3015, 2022. doi: 10.1109/TMI.2022.3176598.

Yushan Zheng, Jun Li, Jun Shi, Fengying Xie, Jianguo Huai, Ming Cao, and Zhiguo Jiang. Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis. *IEEE Transactions on Medical Imaging*, 42(9):2726–2739, 2023. doi: 10.1109/TMI.2023.3264781.

Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21485–21494, 2023.

APPENDIX

A Dataset Description **17**

A.1 Cohorts and Prediction Targets 17

B Implementation Details **18**

B.1 Model Training 18

B.2 Benchmarking 18

B.3 Loss Functions 19

C Methods **20**

C.1 Spatial Context Modeling 20

C.2 Local Spatial Context Encoder (Alternative) 21

C.3 Prototype-Conditioned Feature Set 22

C.4 Cross-Modality Fusion via Gated Attention 22

D Ablation Studies **24**

D.1 Ablation 1: Prototype Update Rule 24

D.2 Ablation 2: Modality Contribution 24

D.3 Ablation 3: Cross-modal Prototype Aggregation 24

D.4 Ablation 4: Prototype Diversity 24

D.5 Ablation 5: Context Encoder 24

E Sensitivity Analysis **25**

F Survival Risk Stratification **25**

G Model Interpretability **26**

G.1 Global Prototype Characterization 26

G.2 Patient-level Attribution 27

G.3 TNBC Response 28

G.4 TNBC Survival 29

A Dataset Description

A.1 Cohorts and Prediction Targets

Table 3 summarizes the three public cohorts. All pair an H&E whole-slide image with spatially resolved expression from the same tissue section.

Table 3: Summary of the three cohorts.

	TNBC Survival	TNBC Response	Lung Disease
Source	Wang et al. (2024)	Yan et al. (2026)	Vannan et al. (2025)
Tissue	Surgical resection, frozen	Pre-treatment core biopsy	Explant/donor lung, FFPE
ST platform	ST array (100 μm spots)	Xenium (5K panel + 100 genes)	Xenium (343 genes)
ST instance	Native spot	Pseudo-spot (summed transcripts)	Pseudo-spot (summed transcripts)
Patients	92	44	35
Task	Survival (5 endpoints)	pCR vs. RD (19/25)	PF vs. unaffected (26/9)

TNBC Survival. This retrospective cohort comprises 92 patients with early-stage TNBC (ER/PR-negative, HER2-non-amplified) treated with primary surgery at the Institut Jules Bordet between 2000 and 2016 (Wang et al., 2024). Frozen surgical tumor samples with $> 15\%$ tumor cellularity were profiled on ST arrays of 1,934 barcoded spots ($100\,\mu$m diameter, $150\,\mu$m center-to-center), using up to three consecutive $16\,\mu$m sections per tumor (273 sections in total). Each section was H&E-stained and imaged at $20\times$ before sequencing, and 14,019 genes detected across all samples were retained. Pathologist annotations of 15 histomorphological categories (e.g., tumor, low/high-TIL stroma, necrosis, tertiary lymphoid structures) are available for one section per patient; we use them only for qualitative interpretation (Appendix G). Outcomes follow STEEP v2.0, with approximately five years of clinical follow-up and censoring at ten years. Table 4 defines the five endpoints by the events they count; all are measured from the date of diagnosis.

Table 4: Survival endpoints in the TNBC Survival cohort (STEEP v2.0).

Endpoint	Death (any cause)	Distant recurrence	Locoregional recurrence	Contralateral BC	Second primary non-BC invasive cancer
Overall survival (OS)	✓				
Distant relapse-free survival (DRFS)	✓	✓			
Recurrence-free survival (RFS)	✓	✓	✓		
Invasive BC-free survival (IBCFS)	✓	✓	✓	✓	
Invasive disease-free survival (IDFS)	✓	✓	✓	✓	✓

✓: an event that ends event-free time for that endpoint. BC: breast cancer.

TNBC Response. Pre-treatment core biopsies were obtained from treatment-naive patients with early-stage TNBC enrolled in the ARTEMIS trial (NCT02276443) (Yan et al., 2026). Patients received neoadjuvant doxorubicin and cyclophosphamide followed by paclitaxel-based chemotherapy. Response was assessed on the resected surgical specimen by residual cancer burden (RCB) and binarized as pathological complete response (pCR; RCB-0) or residual disease (RD; RCB-I–III); the 44 patients used here comprise 19 pCR and 25 RD cases. We use the 44 patients profiled with Xenium in situ RNA (12 fresh-frozen, 32 FFPE sections), measured with the Xenium Human 5K Pan Tissue & Pathways panel plus a custom 100-gene add-on. A post-Xenium H&E image of the same section provides the paired histology.

Lung Disease. This cohort contains 45 FFPE lung tissue samples from 35 donors: nine unaffected lungs declined for transplantation and 26 lungs explanted from patients with pulmonary fibrosis (PF), of which 12 were diagnosed with idiopathic pulmonary fibrosis (Vannan et al., 2025). Diagnoses were made by the treating clinicians and confirmed on explant pathology. Tissue cores were arranged on tissue microarrays and profiled with Xenium using 343 genes (246 from the Xenium human lung panel and 97 custom genes), followed by H&E staining of the same section. We use one representative sample per donor, giving 35 samples, and predict disease status (PF vs. unaffected).

H&E–ST registration. Each cohort requires the H&E image and the ST coordinates to be in the same space. For TNBC Survival, the original authors aligned the ST spots to the H&E images, and

we use their registration directly. For TNBC Response, we register the post-Xenium H&E image to the Xenium morphology image automatically with a pipeline based on PALOM.[1] For Lung Disease, we use the registration released by the original authors (Vannan et al., 2025) where available, and manually register the remaining samples, those on the fifth tissue microarray (TMA5).

Instance construction. The TNBC Survival cohort provides spot-level expression natively, so each ST spot and its co-registered 256×256-pixel H&E patch form one paired instance. Xenium measures individual transcripts at subcellular resolution and has no spots. For TNBC Response and Lung Disease, we therefore tile the registered H&E image into a regular grid of non-overlapping 256×256-pixel patches ($54.4 \times 54.4\ \mu$m at $0.2125\ \mu$m/pixel) and sum the transcript counts falling within each patch to obtain a pseudo-spot expression profile. Each pseudo-spot is then paired with its patch in the same way as a native spot, so all three cohorts share the same instance definition and downstream feature extraction (Sec. 4).

B Implementation Details

B.1 Model Training

All models were developed and trained with Python version 3.10.18 and PyTorch version 2.2.0+cu121 using NVIDIA A100 GPUs, with the batch size set to 1 and the random seed fixed at 1. The Adam optimizer was employed for model optimization with a constant learning rate of 1e-4 and a weight decay of 5e-5. The dropout rate was set to 0.25 for all models whenever feasible. Each model was trained from epoch 0 through epoch 30, and the best epoch was selected.

Evaluation protocol. All data splits are made at the patient level, so no patient appears in both training and test sets. In TNBC Survival, all sections from a patient (273 sections from 92 patients) are assigned to the same fold; in TNBC Response and Lung Disease, each patient or donor contributes one sample. For TNBC Survival, we use five-fold cross-validation stratified jointly by the four discretized survival-time bins and censoring status; the best epoch is selected separately in each fold, and we report the C-index as mean and standard deviation across folds. For TNBC Response and Lung Disease, we use leave-one-out cross-validation, holding out each patient once. Predictions from all held-out patients are pooled, and AUC and accuracy are computed once on the pooled set, so no standard deviation is reported. Every model is trained for a fixed 30 epochs, and the reported epoch is the one with the highest pooled AUC, taking the earliest epoch in case of ties. The predicted class is the one with the larger softmax probability. All experiments use random seed 1.

All PathoSpatial parameters, including the feature projections, context encoders, prototype banks, fusion network, and prediction head, are jointly optimized throughout training using the task-specific objective and prototype diversity penalty. The prototype banks receive gradient updates together with the rest of the model and an additional EMA update at the end of every epoch using training slides only. The gradient-only ablation disables EMA throughout training while retaining joint gradient optimization of all parameters.

B.2 Benchmarking

All benchmark configurations combine histology with ST or pseudo-bulk expression. CLAM, TransMIL, Patch-GCN, and PANTHER serve as modality-specific aggregators within the customized H+S late-fusion pipelines, with the best configuration selected from concatenation, addition, and dot-product fusion reported for each backbone. PORPOISE, MCAT, SurvPath, PIBD, and ProSurv pair histology with pseudo-bulk expression using the encodings below. Hyperparameters largely follow the defaults in the original publications; modified and user-determined settings are listed here:

CLAM: We set *instance_eval* to be *True* to include the instance-level clustering loss.

TransMIL: We added dropout with a rate of 0.25 to the first fully connected layer of the original architecture to enable a fair comparison.

[1] https://pypi.org/project/palom/

Patch-GCN: We set *radius* to 8 for graph construction in both histology and spatial transcriptomics modalities. The number of graph convolutional layers (*num_gcn_layers*) was set to 4 in both modalities.

PANTHER: We set *n_proto* to 16 in each modality, corresponding to 16 mixture components (prototypical distributions) in the Gaussian mixture model. We set *out* to *allcat*, meaning the final slide representation is formed by concatenating all the descriptors for all components.

PORPOISE: We retained 2,000 genes with the highest median absolute deviation (MAD) across samples in the pseudo-bulk expression matrix to represent the genomic feature profile of each sample. We set *size_arg* to *small* to use a smaller version of the genomic SNN.

MCAT: We encode pseudo-bulk expression to 50 Hallmark pathways and construct genomic SNNs with input dimensions determined by the number of genes in each pathway. The updated TNBC survival performance summary labels the selected configuration as concatenation fusion.

SurvPath: We encode pseudo-bulk expression to 50 Hallmark pathways and construct genomic SNNs with input dimensions determined by the number of genes in each pathway.

PIBD: We encode pseudo-bulk expression to 50 Hallmark pathways and construct genomic SNNs with input dimensions determined by the number of genes in each pathway. We adjust *num_patches* to 2000 according to the number of spots/patches per slide in our dataset. We set both *ratio_omics* and *ratio_wsi* to be 0.5 when determining the top-k value in each modality.

ProSurv: We retained 2,000 genes with the highest median absolute deviation (MAD) across samples in the pseudo-bulk expression matrix to represent the genomic feature profile of each sample.

B.3 Loss Functions

All methods share the same task loss for a given task: the NLL survival loss below for TNBC Survival and the cross-entropy loss for TNBC Response and Lung Disease. Auxiliary terms that belong to a baseline's original design, such as CLAM's instance clustering loss, are kept. PathoSpatial adds the assignment-based diversity loss below to the task loss (Eq. 8); the cosine-based alternative is used only in the diversity ablation (Appendix D.4).

Survival Loss: We employ the negative log-likelihood (NLL) survival loss (Zadeh & Schmid, 2021) as the primary prediction loss, following prior multimodal survival models (Chen et al., 2021b; Zhang et al., 2024b). We discretize continuous time into a set of non-overlapping intervals determined by the quartiles of survival times in the training set. For a patient i, let y_i denote the discrete time bin index and $c_i \in \{0, 1\}$ denote the censoring status, where $c_i = 1$ indicates censoring (event not observed) and $c_i = 0$ indicates an observed event.

The model outputs a hazard probability h_j for each time interval j, representing the conditional probability that the event occurs in interval j given that the patient has survived up to interval $j - 1$. The survival probability S_j, representing the probability that the patient survives past interval j, is defined recursively as the cumulative product of non-event probabilities:

$$S_j = \prod_{k=1}^{j} (1 - h_k) \tag{9}$$

where $S_0 = 1$.

The loss function aims to maximize the likelihood of the observed data. For an uncensored patient who experiences the event at time interval y_i, the likelihood is defined as the probability of surviving up to $y_i - 1$ and experiencing the event at y_i. For a censored patient known to survive up to interval y_i, the likelihood is simply the survival probability S_{y_i}. The formulation is defined as:

$$\mathcal{L}_{uncensored} = -\sum_{i:c_i=0} \log(S_{y_i-1} \cdot h_{y_i}) \tag{10}$$

$$\mathcal{L}_{censored} = -\sum_{i:c_i=1} \log(S_{y_i}) \tag{11}$$

Combining these terms, the total negative log-likelihood loss is:

$$\mathcal{L}_{NLL} = (1 - w_{\text{unc}})\mathcal{L}_{total} + w_{\text{unc}}\mathcal{L}_{uncensored} \tag{12}$$

where $\mathcal{L}_{total} = \mathcal{L}_{uncensored} + \mathcal{L}_{censored}$ and w_{unc} is a weighting hyperparameter. In our experiments, we set $w_{\text{unc}} = 0.4$ to encourage the model to focus on learning from the exact event times provided by uncensored samples.

Classification Loss: For the binary tasks, the head outputs logits over the two classes, converted to probabilities $p_{i,c}$ by a softmax. With label $y_i \in \{0, 1\}$, we use class-weighted cross-entropy to offset class imbalance:

$$\mathcal{L}_{CE} = -\sum_{i} \sum_{c\in\{0,1\}} w_c \, \mathbf{1}[y_i = c] \, \log p_{i,c}, \tag{13}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $w_c = N/(C\, n_c)$ is the inverse-frequency class weight, with $C = 2$ classes, n_c the number of training samples in class c, and $N = \sum_c n_c$. The weights average to 1 over the training samples, and a balanced training set gives $w_c = 1$ for all classes.

Assignment-based Diversity Loss: The full model uses assignment-based diversity in each modality, combining an assignment-overlap penalty with a usage-floor penalty. We penalize overlap in the sets of instances different prototypes attend to. Let $\mathbf{\Pi} \in \mathbb{R}^{K\times N}$ be the top-k_{im} masked assignment distribution from Appendix C.3 (each row sums to 1 over its own selected instances). Row-normalizing, $\hat{\mathbf{\Pi}} = \mathbf{\Pi}/\|\mathbf{\Pi}\|_{2,\text{row}}$, the assignment-overlap term is

$$\mathcal{L}_{\text{ov}}(\mathbf{\Pi}) = \frac{1}{K^2} \sum_{u,v} \big((\hat{\mathbf{\Pi}}\hat{\mathbf{\Pi}}^\top)_{uv} - \delta_{uv}\big)^2, \tag{14}$$

which penalizes similarity between assignment rows. This alone cannot detect a *dead* prototype, since $\mathbf{\Pi}$'s rows are each normalized to sum to 1 by construction and so cannot reveal that a given prototype dominates only a negligible share of instances overall. We therefore add a usage-floor hinge computed on the complementary, column-normalized view $\hat{\pi}_{kj} = \Pi_{kj}/\sum_{k'} \Pi_{k'j}$ (the posterior probability of prototype k given instance j): letting $u_k = \frac{1}{N}\sum_j \hat{\pi}_{kj}$ be prototype k's share of instances,

$$\mathcal{L}_{\text{usage}}(\mathbf{\Pi}) = \frac{1}{K} \sum_{k=1}^{K} \Big[\max\Big(0,\ \tfrac{\tau}{K} - u_k\Big)\Big]^2, \tag{15}$$

where $\tau \in (0, 1]$ sets the usage floor relative to a uniform share $1/K$. The assignment-based diversity loss is $\mathcal{L}_{\text{ov}} + w_{\text{usage}}\mathcal{L}_{\text{usage}}$, where w_{usage} controls the relative contribution of the usage penalty and $\lambda_{\text{div}} = 2$ weights the combined diversity loss in the full-model training objective. We use $\tau = 0.25$, i.e., a floor of $0.25/K \approx 0.021$ for $K = 12$ prototypes, and $w_{\text{usage}} = 1$.

C Methods

C.1 Spatial Context Modeling

To capture long-range dependencies while preserving local morphological details, we implement a modified TransMIL encoder. This process is applied independently to each modality $m \in \{H, S\}$. For clarity, we drop the modality superscript (m) and slide index i in this section, describing the operations on a sequence of patches for a single slide. Let $\mathbf{M} \in \mathbb{R}^{N\times D_{input}}$ denote the input bag of features (corresponding to the collection of embeddings $\mathbf{m}_j$), where N is the number of patches.

Feature Projection and Dimension Alignment: Both histology features and spatial transcriptomics features are projected to a common latent dimension $D = 256$ to ensure cross-modal alignment. Specifically, the input features $\mathbf{M}$ are transformed using a fully connected layer followed by a ReLU activation and Dropout ($p = 0.25$). This yields the initial local feature sequence $\mathbf{H} \in \mathbb{R}^{N\times D}$, consisting of the embeddings $\mathbf{h}_j$ described in the main text.

Sequence Padding and Tokenization: The Pyramid Positional Encoding Generator (PPEG) module in TransMIL requires input tokens to be arranged in a 2D grid. Since N varies and is rarely a perfect square, we perform dynamic padding following the TransMIL implementation: We calculate the spatial dimensions of the closest enclosing square grid, $S = \lceil\sqrt{N}\rceil$.

The sequence $\mathbf{H}$ is padded to length S^2 by repeating the initial patches of the sequence until the length requirement is met. This padding strategy ensures that the PPEG module operates on valid feature information rather than zero-tokens.

A learnable classification token ($\mathbf{h}_{cls}$) is prepended to the sequence.

Spatial Encoding Layers: The sequence passes through a stack comprising a first Transformer layer, the PPEG module, and a second Transformer layer. We insert the PPEG between the transformer layers to leverage the inductive bias of image-like structures:

$$\mathbf{Z}_0 = [\mathbf{h}_{cls} \parallel \mathbf{H}_{padded}] \tag{16}$$

$$\mathbf{Z}_1 = \text{TransLayer}_1(\mathbf{Z}_0) \tag{17}$$

$$\mathbf{Z}_{pos} = \text{PPEG}(\mathbf{Z}_1, H = S, W = S) \tag{18}$$

$$\mathbf{Z}_{out} = \text{TransLayer}_2(\mathbf{Z}_{pos}) \tag{19}$$

Here, $\parallel$ denotes concatenation. The PPEG module reshapes the sequence into an $S \times S$ grid, applies depth-wise separable convolutions, and flattens the result back to a sequence.

Feature Recovery and Weighted Residual: Post-encoding, the classification token is discarded, and the padded tail is truncated to restore the original sequence length N, resulting in the context-aware features $\mathbf{T} \in \mathbb{R}^{N \times D}$.

To ensure the final embeddings retain high-fidelity local information while benefiting from global context, we employ the weighted residual connection. Both the original projected features $\mathbf{H}$ and the transformer output $\mathbf{T}$ undergo Layer Normalization. The final embedding feature set $\tilde{\mathbf{H}}$ is computed as:

$$\tilde{\mathbf{H}} = (1 - \alpha) \cdot \text{LayerNorm}(\mathbf{H}) + \alpha \cdot \text{LayerNorm}(\mathbf{T}) \tag{20}$$

The full model uses fixed $\alpha = 0.5$ in both modality branches. α can instead be learned per branch as $\alpha = \sigma(\gamma)$ for a trained scalar logit γ; when learnable, we rescale $\mathbf{H}$ by the ratio of root-mean-square magnitudes, $\mathbf{H} \leftarrow \mathbf{H} \cdot (\mathrm{rms}(\mathbf{T})/\mathrm{rms}(\mathbf{H}))$, before blending, since $\mathrm{LayerNorm}(\mathbf{T})$ sits at RMS ≈ 1.0 while the raw projected term sits at RMS ≈ 0.32 in our features, so an untied α would not otherwise reflect the intended mixture. At $\alpha = 0.5$, the raw and contextual terms receive equal blend coefficients.

C.2 Local Spatial Context Encoder (Alternative)

As an alternative to the global self-attention context encoder above, each modality branch can instead use a local spatial context encoder that mixes every instance only with its k_{nn} nearest spatial neighbors, so the receptive field grows locally across rounds rather than covering the whole bag in one step. This directly encodes physical adjacency, the inductive bias underlying tile-adjacency graph methods for pathology (Sec. 2), whereas the attention path above never uses coordinates at all.

Given per-instance coordinates $\mathbf{c}_j \in \mathbb{R}^2$ and the projected features $\mathbf{H} = \{\mathbf{h}_j\}_{j=1}^N$ from Feature Projection above, we first build a k_{nn}-nearest-neighbor index from pairwise Euclidean distance over $\{\mathbf{c}_j\}$, excluding self-matches. Writing $\mathcal{N}(j)$ for instance j's neighbor set, the encoder applies L rounds of

$$\bar{\mathbf{h}}_j^{(\ell)} = \frac{1}{|\mathcal{N}(j)|} \sum_{j' \in \mathcal{N}(j)} \mathbf{h}_{j'}^{(\ell-1)}, \tag{21}$$

$$\mathbf{h}_j^{(\ell)} = \mathrm{LayerNorm}\Big(\mathbf{h}_j^{(\ell-1)} + \mathrm{Dropout}\big(\mathrm{ReLU}(\mathbf{W}_{\mathrm{self}}^{(\ell)} \mathbf{h}_j^{(\ell-1)} + \mathbf{W}_{\mathrm{nbr}}^{(\ell)} \bar{\mathbf{h}}_j^{(\ell)})\big)\Big), \tag{22}$$

for $\ell = 1, \ldots, L$, with $\mathbf{h}_j^{(0)} = \mathbf{h}_j$, giving the contextual features $\mathbf{t}_j = \mathbf{h}_j^{(L)}$ used in place of $\mathbf{T}$ above. The residual-plus-LayerNorm structure per round keeps the output at the same post-LayerNorm scale (RMS ≈ 1.0) as the attention encoder, so the weighted-residual blend and every downstream prototype operation are shared unmodified between the two encoders; only the padded-grid tokenization used for PPEG is specific to the attention path and is not needed here. This encoder requires per-instance spatial coordinates, which are not required by the alternative attention path. The full model uses $k_{\mathrm{nn}} = 8$ neighbors and $L = 2$ message-passing rounds.

C.3 PROTOTYPE-CONDITIONED FEATURE SET

This section details the generation of the prototype-conditioned feature set described in Sec. 3.2. This process is applied independently to each modality. We drop the modality superscript (m) and slide index i in this section for clarity. Let $\tilde{\mathbf{H}} \in \mathbb{R}^{N\times D}$ be the spatially contextualized feature bag and $\mathbf{P} \in \mathbb{R}^{K\times D}$ be the learnable global prototype bank.

Within each forward pass, we compute the bag-level *prototype-conditioned feature set* $\tilde{\mathbf{P}} \in \mathbb{R}^{K\times D}$ using a specialized Multi-Head Attention mechanism. We treat global prototypes as queries and instance tokens as keys and values.

First, the prototypes and instance features are projected into the attention subspace:

$$\mathbf{Q} = \mathbf{P}\mathbf{W}_Q, \quad \mathbf{K} = \tilde{\mathbf{H}}\mathbf{W}_K \tag{23}$$

where $\mathbf{W}_Q$ and $\mathbf{W}_K$ are learnable projection matrices. To strictly measure directional alignment rather than magnitude, we compute the similarity matrix $\mathbf{S} \in \mathbb{R}^{K\times N}$ using cosine similarity:

$$S_{kj} = \frac{\mathbf{q}_k \cdot \mathbf{k}_j}{\|\mathbf{q}_k\|_2 \|\mathbf{k}_j\|_2} \tag{24}$$

where $\mathbf{q}_k$ and $\mathbf{k}_j$ denote the rows of $\mathbf{Q}$ and $\mathbf{K}$, respectively.

To filter out noise and enforce concept specialization, we employ a *sparse attention* mechanism. For each prototype k, we select only the top subset of instances. Let $\mathcal{I}_k = \mathrm{TopK}_M(\mathbf{S}_{k,:})$ be the indices of the M largest similarity scores for the k-th prototype, with

$$M = \max\left(1,\ \min\left(k_{\mathrm{cap}},\ \left\lfloor \frac{N\rho}{K} \right\rfloor,\ N\right)\right). \tag{25}$$

The cap k_{cap} bounds M directly, while the proportion $\rho \in (0, 1]$ lets M scale with bag size N; $\rho = 1$ recovers a pure-cap rule. The outer $\max(1, \cdot)$ is a correctness floor: without it, a very small bag (few instances N relative to K) can otherwise drive M to zero, leaving a prototype with an empty selection and an undefined centroid; the inner cap at N additionally prevents requesting more instances than the bag contains. The full model uses $k_{\mathrm{cap}} = 100$ and $\rho = 0.6$. This operation is applied row-wise, ensuring that each prototype focuses on its most relevant instances. Note that while the attention is sparse (ignoring most patches), a single highly informative instance may be selected by multiple prototypes if it exhibits compound biological traits.

We then form a *masked* logit matrix $\mathbf{A}$ by assigning $-\infty$ to all unselected entries:

$$A_{kj} = \begin{cases} S_{kj}, & j \in \mathcal{I}_k \\ -\infty, & j \notin \mathcal{I}_k \end{cases} \tag{26}$$

The final attention weights $\mathbf{\Pi} \in \mathbb{R}^{K\times N}$ are obtained via a row-wise softmax, ensuring probability mass is distributed *only* over the selected top-k patches:

$$\Pi_{kj} = \frac{\exp(A_{kj})}{\sum_{j' \in \mathcal{I}_k} \exp(A_{kj'})} \tag{27}$$

Finally, the prototype-conditioned feature $\tilde{\mathbf{p}}_k$ is computed as the weighted average of the *original* spatial features $\tilde{\mathbf{H}}$ (prior to projection or normalization):

$$\tilde{\mathbf{p}}_k = \sum_{j\in\mathcal{I}_k} \Pi_{kj} \tilde{\mathbf{h}}_j \tag{28}$$

Stacking these vectors yields the final modality-specific feature set $\tilde{\mathbf{P}} = [\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_K]^\top$.

C.4 CROSS-MODALITY FUSION VIA GATED ATTENTION

To synthesize a unified patient representation from the distinct biological signals captured by the prototypes, we employ a gated attention mechanism. This approach allows the model to dynamically assign importance to each prototype based on the specific context of the slide, effectively functioning as an expert to softly select risk-relevant features. For clarity, we drop the slide index i in this section.

Feature Stacking: First, we aggregate the prototype-conditioned feature sets from both the Histology (H) and Spatial Transcriptomics (S) branches into a single candidate pool. Let $\tilde{\mathbf{P}}^{(H)} \in \mathbb{R}^{K_H \times D}$ and $\tilde{\mathbf{P}}^{(S)} \in \mathbb{R}^{K_S \times D}$ be the output features from the previous stage (with $D = 256$). We stack these along the instance dimension to form the multimodal bag representation $\mathcal{H} \in \mathbb{R}^{R \times D}$, where $R = K_H + K_S$ is the total number of prototypes from two modalities:

$$\mathcal{H} = \left[\tilde{\mathbf{P}}^{(H)} ; \tilde{\mathbf{P}}^{(S)}\right] \tag{29}$$

where ; denotes vertical concatenation. Let $\mathbf{h}_r \in \mathbb{R}^D$ denote the r-th row of $\mathcal{H}$ (i.e., the feature vector of the r-th item in the joint prototypes).

Gated Attention Mechanism: To compute the attention scores, we map the input features into a latent attention space using two separate linear projections followed by non-linear activations. Specifically, we employ the gated attention mechanism, which uses a sigmoid activation to gate the output of a tanh activation. This allows the model to learn complex non-linear relations among instances.

For each prototype feature $\mathbf{h}_r$, we compute the unnormalized attention score s_r as follows:

$$s_r = \mathbf{w}_c^\top \left(\tanh\left(\mathbf{W}_a \mathbf{h}_r^\top\right) \odot \sigma\left(\mathbf{W}_b \mathbf{h}_r^\top\right)\right) \tag{30}$$

where: $\mathbf{W}_a, \mathbf{W}_b \in \mathbb{R}^{D' \times D}$ are learnable weight matrices projecting the input to the internal attention dimension $D' = 256$.

$\mathbf{w}_c \in \mathbb{R}^{D'}$ is a learnable weight vector projecting the gated features to a scalar score.

$\odot$ denotes element-wise multiplication.

$\sigma(\cdot)$ is the sigmoid activation function.

Dropout ($p = 0.25$) is applied after the $\tanh$ and σ activations during training to prevent overfitting.

Bag Feature Aggregation: The attention scores are normalized using a softmax function to obtain a probability distribution over the prototypes:

$$a_r = \frac{\exp(s_r)}{\sum_{k=1}^{R} \exp(s_k)} \tag{31}$$

The final fused representation $\mathbf{h}_{\text{fused}} \in \mathbb{R}^D$ is computed as the sum of the prototype features weighted by their respective attention scores:

$$\mathbf{h}_{\text{fused}} = \sum_{r=1}^{R} a_r \mathbf{h}_r \tag{32}$$

This vector $\mathbf{h}_{\text{fused}}$ serves as a compact, multimodal summary of the patient's slide, prioritizing the most prognostically relevant morphological and molecular patterns.

Within-modality Attention (Alternative): Instead of one gated attention network over the stacked pool $\mathcal{H}$, we can instantiate a separate gated attention network, of the same form as above, for each modality's prototype set $\tilde{\mathbf{P}}^{(H)}$ and $\tilde{\mathbf{P}}^{(S)}$ independently, giving two pooled vectors $\mathbf{h}_{\text{fused}}^{(H)}, \mathbf{h}_{\text{fused}}^{(S)} \in \mathbb{R}^D$. The fused representation is their unweighted average, $\mathbf{h}_{\text{fused}} = \frac{1}{2}(\mathbf{h}_{\text{fused}}^{(H)} + \mathbf{h}_{\text{fused}}^{(S)})$. This isolates each modality's own most-relevant prototypes before combining, rather than letting the joint attention network trade off prototypes across modalities directly.

Uniform Mean Pooling (Alternative): Fusion weights can also be fixed to a uniform distribution $a_r = 1/R$ rather than learned, giving $\mathbf{h}_{\text{fused}} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{h}_r$. This removes the input-dependent weighting entirely, treating every detected prototype as equally informative.

Single-modality Operation: Either branch can also be run on its own, with the other branch and the fusion step omitted entirely rather than masked at inference; the retained branch's prototype-conditioned feature set is pooled directly (by whichever of the three aggregation rules above is in effect) to produce $\mathbf{h}_{\text{fused}} \in \mathbb{R}^D$, and the diversity loss (Appendix B) is computed from the retained branch's assignments alone. This isolates each modality's standalone contribution to the prediction.

Table 5: Complete ablation results. Each row changes one component of the full model; metrics and highlighting as in Table 2.

		TNBC Response ($N = 44$)		TNBC Survival ($N = 92$) C-index ↑					Lung Disease ($N = 35$)	
Axis	Variant	AUC ↑	ACC ↑	IDFS	IBCFS	DRFS	RFS	OS	AUC ↑	ACC ↑
–	Full model (PathoSpatial-GNN)	0.680	**0.682**	**0.729 ± 0.117**	0.727 ± 0.103	**0.684 ± 0.104**	**0.665 ± 0.096**	0.659 ± 0.111	**1.000**	0.971
Updates	Gradient only	0.672	0.545	0.675 ± 0.073	0.684 ± 0.015	0.610 ± 0.130	0.634 ± 0.050	0.641 ± 0.166	**1.000**	0.971
Modality	H&E only	0.646	0.636	**0.684 ± 0.083**	0.641 ± 0.098	0.720 ± 0.097	0.672 ± 0.151	0.611 ± 0.132	**1.000**	0.943
	ST only	0.656	**0.682**	0.710 ± 0.060	0.671 ± 0.025	0.595 ± 0.110	0.626 ± 0.049	**0.663 ± 0.082**	**1.000**	**1.000**
Pooling	Within-modality attention	0.587	0.568	0.679 ± 0.148	**0.731 ± 0.113**	0.662 ± 0.124	0.646 ± 0.061	0.611 ± 0.109	**1.000**	**1.000**
	Uniform mean pooling	0.672	0.659	0.667 ± 0.164	0.721 ± 0.103	0.678 ± 0.080	0.661 ± 0.050	0.648 ± 0.104	**1.000**	0.971
Diversity	Cosine	0.653	0.636	0.655 ± 0.115	0.668 ± 0.067	0.598 ± 0.121	0.647 ± 0.056	0.603 ± 0.137	**1.000**	0.886
	No diversity term	0.628	0.603	0.702 ± 0.103	0.681 ± 0.072	0.592 ± 0.120	0.655 ± 0.079	0.620 ± 0.141	**1.000**	0.886
Encoder	Global self-attention	**0.691**	0.659	0.684 ± 0.131	0.718 ± 0.056	0.659 ± 0.110	0.651 ± 0.061	0.652 ± 0.140	**1.000**	0.971

D Ablation Studies

This section describes each ablation axis. Each variant changes one component of the full model (Sec. 4.3) and keeps the other settings fixed; Table 5 reports all variants.

D.1 Ablation 1: Prototype Update Rule

The prototype banks can be updated by gradient descent and the epoch-level EMA rule (Sec. 3.2) jointly, or by gradient descent alone with the EMA update disabled throughout training.

D.2 Ablation 2: Modality Contribution

Each modality branch can be run on its own (Appendix C.4), dropping the other branch and the fusion step entirely, to isolate its standalone contribution relative to using both modalities.

D.3 Ablation 3: Cross-modal Prototype Aggregation

We compare joint gated attention over the stacked prototype pool (default), gated attention applied within each modality's prototype set followed by averaging, and uniform mean pooling (Appendix C.4).

D.4 Ablation 4: Prototype Diversity

The full model uses assignment-based diversity (Appendix B) with $\lambda_{\text{div}} = 2$ for all three cohorts. We ablate this choice by replacing the assignment-based loss with a cosine-based penalty on prototype directions, or by disabling diversity regularization entirely ($\lambda_{\text{div}} = 0$).

For the cosine-based ablation, let $\mathbf{P}^{(m)} \in \mathbb{R}^{K_m \times d}$ contain the prototype vectors as rows and let $\hat{\mathbf{P}}^{(m)}$ be its row-wise ℓ_2-normalized form. We substitute

$$\mathcal{L}^{(m)}_{\text{div,cos}} = \frac{1}{K_m^2} \sum_{u,v} \left[\left(\hat{\mathbf{P}}^{(m)} \hat{\mathbf{P}}^{(m)\top} \right)_{uv} - \delta_{uv} \right]^2 \tag{33}$$

for $\mathcal{L}^{(m)}_{\text{div}}$ in the training objective, retaining $\lambda_{\text{div}} = 2$. This penalty encourages mutually orthogonal prototype directions within each modality. However, cluster centroids in our feature spaces are not mutually orthogonal (measured pairwise cosine similarity ≈ 0.64 for k-means centroids), motivating the default loss's focus on distinct instance assignments. The cosine-based formulation is used only in this ablation.

D.5 Ablation 5: Context Encoder

The full model uses the local spatial (GNN) context encoder of Appendix C.2; the alternative uses the global self-attention context encoder (TransMIL/PPEG).

Table 6: Sensitivity analysis. Each row varies one hyperparameter of the full model.

		TNBC Response ($N = 44$)		TNBC Survival ($N = 92$) C-index ↑					Lung Disease ($N = 35$)	
Parameter	Value	AUC ↑	ACC ↑	IDFS	IBCFS	DRFS	RFS	OS	AUC ↑	ACC ↑
–	Full model	**0.680**	**0.682**	**0.729 ± 0.117**	**0.727 ± 0.103**	**0.684 ± 0.104**	0.665 ± 0.096	**0.659 ± 0.111**	**1.000**	0.971
$K_H = K_S$	8	0.653	0.652	0.669 ± 0.104	0.694 ± 0.011	0.606 ± 0.142	0.656 ± 0.050	0.656 ± 0.136	**1.000**	0.914
	16	0.639	0.632	0.675 ± 0.096	0.691 ± 0.012	0.601 ± 0.138	0.656 ± 0.056	0.640 ± 0.141	**1.000**	0.914
ρ	0.2	0.661	0.636	0.668 ± 0.078	0.693 ± 0.033	0.600 ± 0.127	0.664 ± 0.070	0.640 ± 0.144	**1.000**	0.829
	0.4	0.632	0.591	0.675 ± 0.090	0.690 ± 0.011	0.606 ± 0.128	**0.668 ± 0.085**	0.640 ± 0.161	**1.000**	0.886
β	0.3	0.678	0.636	0.670 ± 0.096	0.694 ± 0.017	0.605 ± 0.146	0.652 ± 0.057	0.644 ± 0.164	**1.000**	0.943
	0.7	0.661	0.636	0.682 ± 0.067	0.701 ± 0.012	0.604 ± 0.135	0.658 ± 0.082	0.633 ± 0.164	**1.000**	0.857
	0.8	0.624	**0.682**	0.672 ± 0.106	0.700 ± 0.034	0.610 ± 0.131	0.628 ± 0.067	0.633 ± 0.163	**1.000**	0.829
α	0.0 (raw only)	0.653	0.636	0.680 ± 0.157	0.702 ± 0.111	0.634 ± 0.054	0.611 ± 0.086	0.637 ± 0.070	0.996	0.971
	1.0 (encoder only)	0.678	0.614	0.689 ± 0.097	0.694 ± 0.076	0.616 ± 0.136	0.632 ± 0.056	0.622 ± 0.173	**1.000**	**1.000**
	Learnable	0.653	0.636	0.672 ± 0.077	0.684 ± 0.025	0.608 ± 0.142	0.653 ± 0.051	0.629 ± 0.161	**1.000**	**1.000**

The full model uses $K_H = K_S = 12$, $\rho = 0.6$, $\beta = 0.5$, and fixed $\alpha = 0.5$. Metrics and highlighting as in Table 2.

E Sensitivity Analysis

We vary four hyperparameters of the full model (Sec. 4.3) one at a time: the number of prototypes per modality, the top-k instance-selection proportion ρ (Sec. 3.2), the EMA coefficient β (Sec. 3.2), and the feature-blend coefficient α (Sec. 3.1). Table 6 reports the results. The default values were fixed before this sweep.

Number of Prototypes: We vary the size of the prototype bank, using the same count in each modality, to assess the risk of under-clustering (missing patterns) or over-clustering (redundant patterns).

Top-k Instance-selection Proportion: With the cap fixed at $k_{\text{cap}} = 100$ (Sec. 3.2), we vary the proportion ρ that determines how the number of selected instances per prototype scales with bag size.

EMA Coefficient: Since the EMA update fires once per epoch (Sec. 3.2), β's effective time constant is $1/(1-\beta)$ epochs; we vary β toward the fast end of this range to assess how quickly the bank should track the current epoch's mean prototype-conditioned features versus retain its running average.

Feature-blend Coefficient: We vary the fixed blend coefficient α across its full range, and separately evaluate a learnable variant, $\alpha = \sigma(\gamma)$ learned per modality branch; RMS-matching between the raw and contextual terms is enabled for every arm in this sweep, since α is not otherwise interpretable across arms (Appendix C.1).

F Survival Risk Stratification

We compare the held-out risk scores of PathoSpatial-GNN and SurvPath on TNBC Survival. For each fold, risk scores come from the same best epoch as the C-index in Table 1, and slides from the same patient are averaged to one score per patient.

Per-fold median split. The five fold models are trained independently and do not share a risk scale. On IDFS, for example, the spread of PathoSpatial-GNN's fold medians is 1.66 times the mean within-fold standard deviation of its risk scores, and 21 of 92 patients change group depending on whether a single pooled median or per-fold medians are used. We therefore split each fold's held-out patients at that fold's own median, as in Zhang et al. (2024b), and combine the resulting groups for the Kaplan–Meier curves and log-rank test (Fig. 3A). Table 7 reports the results.

Per-fold splitting lowers the log-rank p-values of both methods on most endpoints. PathoSpatial-GNN has the larger χ^2 on every endpoint under both schemes. Ranking performance is also similar: fold-averaged C-indices differ by at most 0.013 (Table 1). On predictions pooled across folds, which assumes a shared risk scale, SurvPath has a slightly higher C-index on four endpoints (mean difference 0.017). The two risk scores are only weakly correlated (median Spearman $\rho = 0.295$), so the methods disagree on which patients are high risk. Individual risk calls also vary across endpoints:

Table 7: Risk stratification on TNBC Survival: log-rank tests under per-fold and pooled median splits.

	Per-fold median split				Pooled median split	
	Log-rank χ^2		Log-rank p		Log-rank p	
Endpoint	PS	SP	PS	SP	PS	SP
IDFS	9.80	8.71	1.8e-3	3.2e-3	3.3e-2	5.3e-2
IBCFS	12.91	8.41	3.3e-4	3.7e-3	8.5e-5	2.6e-2
DRFS	9.10	4.60	2.6e-3	3.2e-2	2.6e-3	5.0e-2
RFS	11.43	2.68	7.2e-4	1.0e-1	1.2e-1	6.0e-1
OS	6.36	5.28	1.2e-2	2.2e-2	3.0e-2	8.0e-2

PS: PathoSpatial-GNN; SP: SurvPath. The per-fold split corresponds to Fig. 3A; the pooled split (Fig. 4) is a sensitivity analysis.

for the example slide in Fig. 3, whose outcome is the same for all five endpoints, its risk percentile ranges from 11% (OS) to 84% (IDFS) across the five separately trained models, so patient-level attributions on this cohort are illustrative.

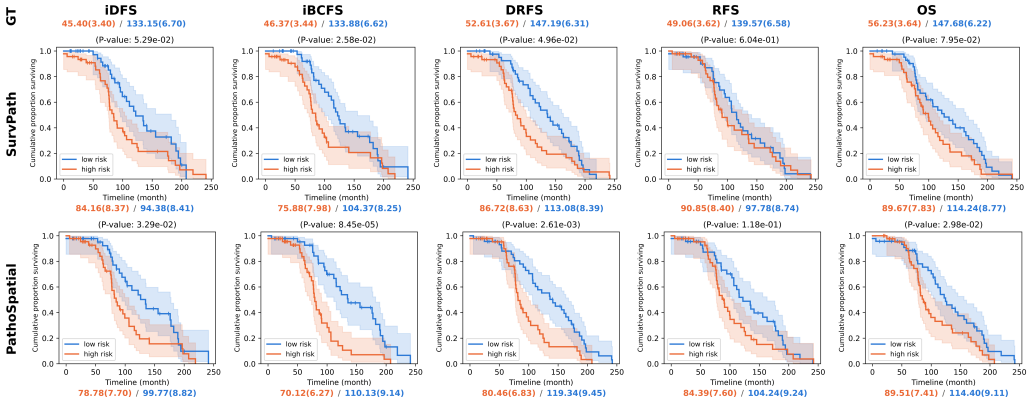

Figure 4: **Risk stratification with a pooled-median split.** Same risk scores as Fig. 3A, but all 92 patients are split at a single median predicted risk. Top: SurvPath; bottom: PathoSpatial-GNN; the GT row is the same as in Fig. 3A.

G Model Interpretability

We interpret PathoSpatial at two levels. *Global* interpretation characterizes what each learned prototype represents across the training set. *Patient-level* interpretation explains a single prediction by attributing it to the slide-specific prototype representations. Interpretation covers the two TNBC cohorts; Lung Disease, on which all methods reach an AUC of 1.000, serves as a positive control. The main text (Figs. 3 and 2) focuses on patient-level interpretation with a brief cohort-level summary; this appendix gives the protocols (Appendices G.1 and G.2) and the full cohort-level characterization and additional patient-level results for each cohort (Appendices G.3 and G.4).

G.1 Global Prototype Characterization

Prototype indices are specific to a trained model and cannot be pooled across cross-validation folds. For each TNBC cohort, we therefore fix one fold's model and run it over that fold's training slides for cohort-level analysis and over one held-out slide for patient-level analysis. For each prototype, we collect its top-attended instances and characterize them as follows.

Feature space. We project all instance features to two dimensions with UMAP, separately for gene expression and tissue morphology, and highlight the top-attended instances of each prototype.

Pathway enrichment. For each expression prototype, we take the top one-vs-rest marker genes of its top-attended instances (75 genes for TNBC Survival) and test them for over-representation with Enrichr (Kuleshov et al., 2016) against the MSigDB Hallmark 2020 (Liberzon et al., 2015) and KEGG 2021 (Kanehisa & Goto, 2000) gene sets, keeping terms with Benjamini–Hochberg-adjusted $q < 0.05$ and at least four overlapping genes. Significance follows Fisher's exact test: with N background genes, K genes in a pathway, and n marker genes of which k fall in the pathway,

$$p = \sum_{i=k}^{\min(n,K)} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}. \tag{34}$$

Several KEGG infectious-disease terms (e.g., Influenza A, Allograft rejection) are dense in ribosomal and MHC genes; we read them as translation- or antigen-presentation-heavy programs rather than as infection biology. For TNBC Response, up- and down-regulated marker genes are tested separately, and each enriched pathway is labeled accordingly; for TNBC Survival, only up-regulated markers are used.

Tissue and cell-type labels. For TNBC Survival, each spot carries the fraction of its area in each pathologist-annotated tissue category; unannotated area is dropped and the remaining fractions are renormalized. For TNBC Response, we use the cell segmentation provided by the original study and annotate each cell by transferring labels from that study's single-cell RNA-seq atlas (Yan et al., 2026) with `scanpy.tl.ingest` (Wolf et al., 2018). Each cell is assigned to the 256×256-pixel pseudo-spot patch that contains it (Appendix A.1), giving a cell-type composition for every patch.

Matching histology prototypes to categories. For each prototype p and instance i, let $\bar{\pi}_{pi}$ be the prototype's attention over all instances of the slide, i.e., the weights of Sec. 3.2 before top-k truncation. We take $\bar{\pi}_{pi}$ in excess of the mean over all K prototypes, which removes the component shared by all prototypes, and normalize it over instances:

$$w_{pi} \propto \max\Big(0,\ \bar{\pi}_{pi} - \tfrac{1}{K}\textstyle\sum_{p'} \bar{\pi}_{p'i}\Big), \qquad \textstyle\sum_i w_{pi} = 1. \tag{35}$$

With A_{ic} the fraction of instance i in category c and b_c its slide-wide mean, the composition of prototype p is $\mathrm{comp}_{pc} = \sum_i w_{pi} A_{ic}$ and its enrichment over the slide background is $\log_2(\mathrm{comp}_{pc}/b_c)$. Each prototype is matched to its most enriched category among categories with $\mathrm{comp}_{pc} \geq 0.02$ and $b_c \geq 0.02$. For TNBC Survival, each match is tested by permuting the annotations across spots with the weights fixed (2,000 permutations), using the maximum enrichment over categories as the null statistic and Benjamini–Hochberg correction across prototypes. The permutation ignores spatial autocorrelation between spots, so these p-values are optimistic.

Visualization. In Fig. 3B, each spot is colored by the matched category of the prototype with the largest contrast weight at that spot, and the 70% of spots with the largest winning weight are shown. This filter uses only the model's weights, not the annotation. Colors therefore show each prototype's matched category rather than a per-spot classification.

Sample selection. For each TNBC cohort, we interpret one held-out patient that the model predicts correctly and whose saved checkpoint reproduces its recorded prediction.

G.2 Patient-level Attribution

Coalition value and baseline. For a fixed slide, let $\mathcal{N} = B_H \cup B_S$ index the slide-specific prototype representations $\mathbf{x}_r$ entering fusion, partitioned into histology and ST groups. Let a_r be their full-model fusion attention weights. For a nonempty coalition $C \subseteq \mathcal{N}$, we retain the selected representations and renormalize their weights:

$$\mathbf{z}(C) = \frac{\sum_{r\in C} a_r \mathbf{x}_r}{\sum_{r\in C} a_r}. \tag{36}$$

The coalition value $v(C)$ is the model output computed from $\mathbf{z}(C)$ by the prediction head with parameters $W, \mathbf{b}$:

$$v_{\mathrm{cls}}(C) = \mathrm{softmax}\big(W\mathbf{z}(C) + \mathbf{b}\big)_1, \qquad v_{\mathrm{surv}}(C) = -\sum_{t=1}^{T}\prod_{j=1}^{t}\big[1 - \sigma\big((W\mathbf{z}(C) + \mathbf{b})_j\big)\big]. \tag{37}$$

For binary classification, v_{cls} is the predicted probability of class 1, so a positive attribution moves the prediction toward class 1 and a negative one toward class 0. For TNBC Response, class 1 is response (pCR). For survival, v_{surv} is the negative sum of predicted survival probabilities over the T outputs of the head, so larger values indicate higher predicted risk. The head has $T = 8$ outputs, of which only the first four, one per time bin, are supervised by the NLL loss; the remaining four stay at zero weight (hazard 0.5) and add a fixed $0.9375\, S_4$ to the sum, where S_4 is the survival probability after the last supervised bin. We keep $T = 8$ so that v_{surv} equals the risk score used for all reported results; using only the four supervised outputs gives nearly identical attributions (correlation 0.999 for the example slide, with the same five top-ranked prototypes). Prototype representations are held fixed; absent prototypes are omitted from fusion. Since gated attention scores depend on each prototype independently, weight renormalization exactly reproduces fusion over the retained prototypes. We define $\mathbf{z}(\emptyset) = \mathbf{0}$, giving the bias-only baseline $v(\emptyset)$. This is a model-defined reference, not an empirical population-average risk. Attribution depends on the chosen coalition value function and reference (Sundararajan & Najmi, 2020).

Exact modality-grouped Owen values. For $r \in B_m$, let $B_{\bar{m}}$ denote the other modality's group and define $q_m(A) = |A|!(|B_m| - |A| - 1)!/|B_m|!$. With two modality groups, the Owen value (Owen, 1977) is

$$\phi_r = \frac{1}{2} \sum_{A \subseteq B_m \setminus \{r\}} q_m(A) \sum_{Q \in \{\emptyset, B_{\bar{m}}\}} \left[v(Q \cup A \cup \{r\}) - v(Q \cup A) \right]. \tag{38}$$

This averages marginal contributions over both modality orderings and all within-modality prototype orderings. We enumerate the subsets exactly in double precision, using exported prototype representations, attention weights, and head parameters. With 12 prototypes per modality, there are 4096 marginal-contribution pairs per prototype; no Monte Carlo sampling is required. We numerically check the completeness identity $\sum_{r \in \mathcal{N}} \phi_r = v(\mathcal{N}) - v(\emptyset)$ for each slide. Positive and negative values indicate contributions toward class 1 and class 0 for classification, and toward higher and lower risk for survival. Summing ϕ_r within each group gives its net modality contribution.

Interpretation. The two prototype banks occupy different regions of the fused feature space, so the sign of a prototype's Owen value partly reflects its modality. We therefore read the modality totals as the net contribution of each branch and compare individual prototypes only within the same bank. Fusion attention sums are reported separately as a description of fusion, not as attributions. Coalitions differ from the complete prototype pool used during training, so attributions can reflect extrapolation, particularly when an entire modality is absent, consistent with the extrapolation concerns discussed by Kumar et al. (2020). These are explanations of model predictions, not causal biological effects.

G.3 TNBC Response

We interpret the full model (PathoSpatial-GNN). Cohort-level analyses use one fixed leave-one-out fold (best epoch 13), run over its 43 training patients.

Global interpretation. The expression bank is dominated by hypoxia (Fig. 5B): six of 12 prototypes (P0, P3, P5, P7, P9, P11) lead with Hypoxia, four of them together with HIF-1 signaling or glycolysis. HIF-1α stabilization promotes chemoresistance in TNBC models and has been associated with shorter relapse-free survival after chemotherapy (Xiong et al., 2018). The remaining prototypes capture myeloid and inflammatory programs (P1, Fcγ receptor-mediated phagocytosis; P8, IL-6/JAK/STAT3), epithelial junctions (P4, P10), and mesenchymal programs (P2, P6, EMT), including leukocyte transendothelial migration (P2).

Patient-level interpretation. Patient P87 is a responder, held out in the corresponding fold and predicted to respond with probability 0.724; this is the most confident correct prediction among folds whose checkpoints reproduce their recorded predictions. The patient's expression prototypes (Fig. 2B) add programs absent at the cohort level: proliferation (P5, G2-M checkpoint and E2F targets), cholesterol biosynthesis (P9), Myc targets (P8), and Wnt signaling (P3). Only P7 keeps the hypoxia identity. The slide is T-cell rich (56% T cells, 27% tumor cells, 6% fibroblasts), in line with the association between tumor-infiltrating lymphocytes and pCR (Denkert et al., 2010). After background correction, six histology prototypes are enriched for fibroblasts, three for tumor cells, two for B cells, and one for T cells, so the histology bank separates stromal, tumor, and lymphoid-aggregate patches. Fusion attention is close to uniform for this patient (0.032–0.058 per prototype;

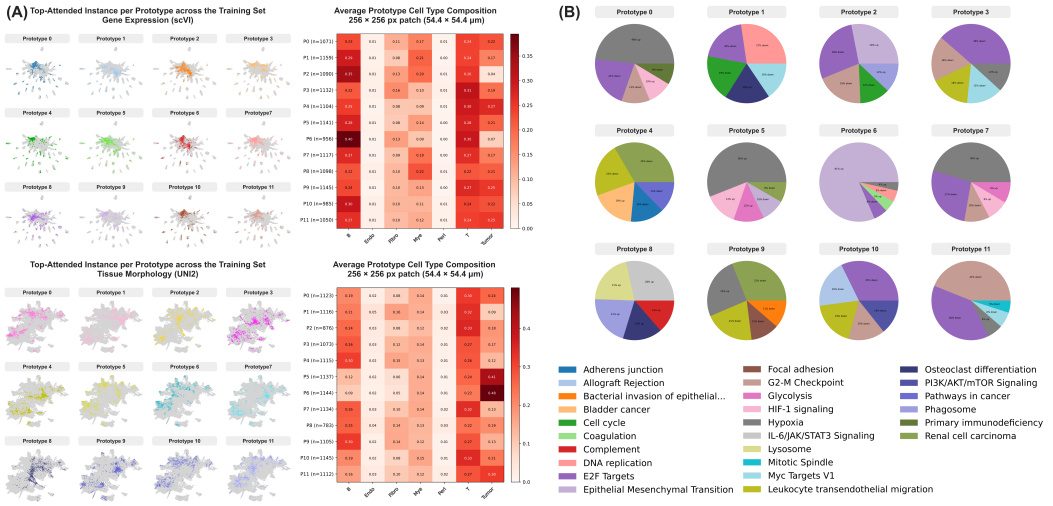

Figure 5: **Global prototype characterization on TNBC Response.** (A) Top-attended instances of each prototype across the training set, shown on UMAPs of gene expression (scVI; top) and tissue morphology (UNI2; bottom), with the average cell-type composition of each prototype's 256×256-pixel patches ($54.4 \times 54.4\,\mu$m) on the right. B: B cells; Endo: endothelial cells; Fibro: fibroblasts; Mye: myeloid cells; Peri: perivascular cells; T: T cells. (B) Pathways enriched in each expression prototype across the training set, labeled by whether they come from up- or down-regulated marker genes.

61% on the expression bank), so no single prototype drives the prediction. These enrichments come from one slide and were not tested for significance.

G.4 TNBC Survival

We interpret the full model (PathoSpatial-GNN) trained for IDFS, using the fold-1 model at its best epoch (15). Cohort-level analyses run this model over the fold's 215 training slides. This is the only endpoint and fold whose best epoch also ranks the example slide correctly.

Global interpretation. The expression prototypes fall into four groups (Fig. 6B): EMT and extracellular-matrix stroma (P1, P3, P6, P9); immune programs, namely antigen presentation and allograft rejection (P7, P11) and interferon response (P8); metabolic stress, namely hypoxia and HIF-1 signaling (P0), cholesterol homeostasis (P2), and lysosome and glutathione metabolism (P4); and translation or proliferation, namely ribosome and p53 (P5) and Myc targets (P10).

Patient-level interpretation. Slide CN41-E1 (patient TNBC82, 1,355 spots) had an uncensored IDFS event at 49.9 months and is ranked at the 84th percentile of predicted risk; 6 of its 7 comparable patient pairs are ordered correctly. Within-slide marker genes resolve the expression prototypes into specific cell states: hypoxic glycolytic carcinoma (P0; *SLC2A1*, *NDRG1*), collagen-producing fibroblasts (P3; *COL1A1*, *SPARC*), three macrophage states (P4, C1Q$^+$; P7, MHC-II$^+$; P11, MMP9$^+$), B and plasma cells (P6; *IGHM*, *IGKC*), and lipogenic cells (P2; *FASN*). The Owen values attribute the high predicted risk to the expression branch (net $+1.95$), led by P2 ($+0.51$), P7 ($+0.28$), P9 ($+0.23$), and P4 ($+0.19$); stromal CD68$^+$ macrophages have been associated with reduced breast cancer-specific survival (Medrek et al., 2012). The histology branch receives 10.4% of the fusion weight and lowers the predicted risk (net -1.21).

Six of 12 histology prototypes are significantly enriched for an annotated category ($q < 0.05$): high-TIL stroma (P2, P7), tumor (P10, P11), acellular stroma (P0), and whitespace (P3). Stromal TILs are prognostic in early-stage TNBC (Loi et al., 2019). The enrichments are modest (at most about $1.8\times$), and per-spot agreement with the annotation does not exceed a majority-class baseline at any coverage (at most 35.5%, at 15% coverage, versus 39.4%), so the histology prototypes describe tissue composition rather than a segmentation.

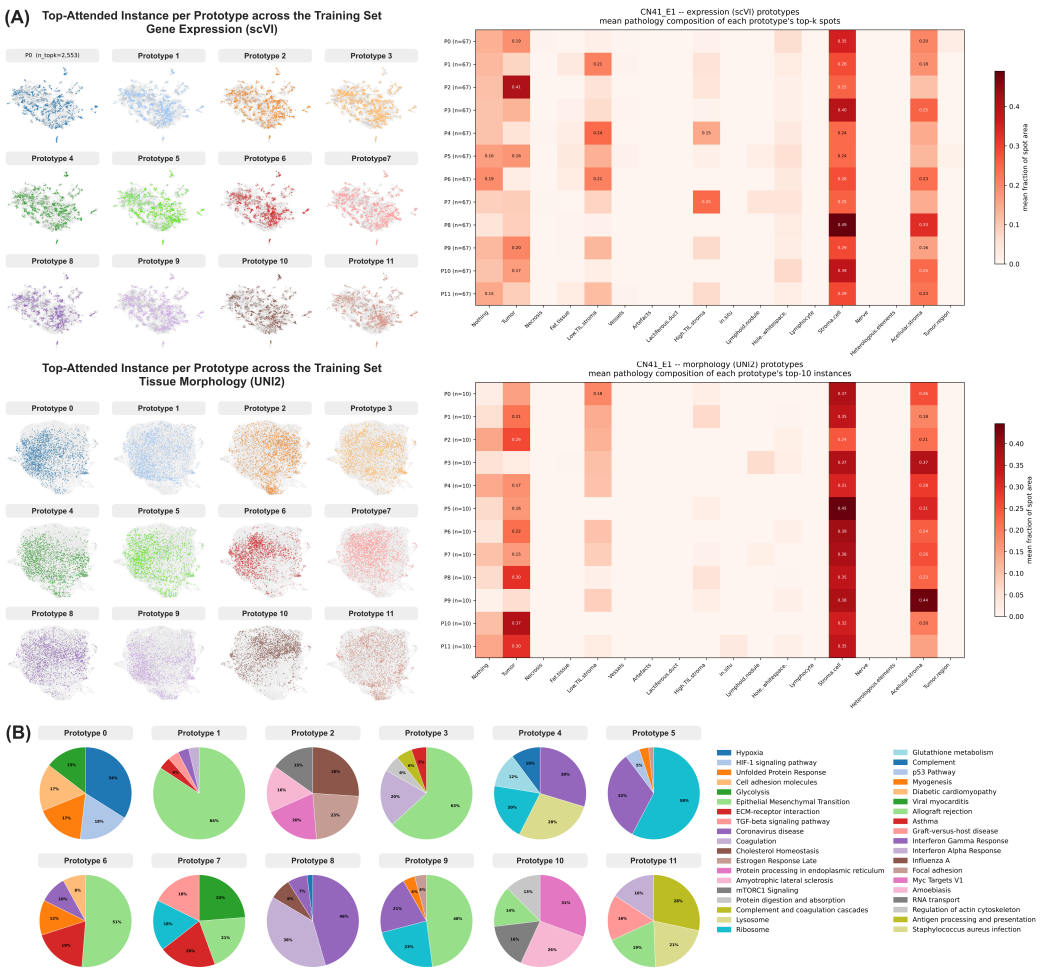

Figure 6: **Global prototype characterization on TNBC Survival.** (A) Top-attended instances of each prototype across the training set, shown on UMAPs of gene expression (scVI; top) and tissue morphology (UNI2; bottom). Heatmaps show the mean pathology-annotation composition of each prototype's top-attended spots on slide CN41-E1 (67 spots per expression prototype and 10 per histology prototype). (B) Pathways enriched in each expression prototype across the training set.