From WSIs to Bulk, Spatial and Single-Cell Transcriptomics via Multi-Scale Graph Neural Networks

Miguel Andrade[1], José Almeida[2], Carlos Santiago[1], and Catarina Barata[1]

[1]Departamento de Engenharia Electrotécnica e de Computadores, Instituto Superior Técnico, Lisboa, Portugal
[2]Champalimaud Foundation, Lisbon, Portugal

arXiv:2610.41692v1 [cs.LG] 30 Oct 2026

Abstract

Understanding the spatial distribution of gene expression is key to studying complex diseases like cancer. While Spatial Transcriptomics (ST) and Single-Cell Transcriptomics (scT) provide high-resolution insights, they remain expensive and scarce. In contrast, standard histopathology Whole-Slide Images (WSIs) paired with bulk RNA sequencing are widely available. This work introduces a Multi-Scale Graph Neural Network (GNN) capable of simultaneously predicting bulk, spatial, and single-cell transcriptomics directly from WSIs. By extracting morphological features using a Pathology Foundation Model (PFM) (UNI2-h) and constructing spatial graphs at both patch and cell levels, the model learns the hierarchical relationships within the tissue. The network is trained on a combined dataset of The Cancer Genome Atlas Program (TCGA), Visium, and Xenium samples across lung, kidney, and breast tissues. A hybrid training strategy is employed, leveraging direct supervision alongside weakly supervised consistency losses to ensure predictions align across biological scales. Results show the model accurately predicts gene expression at the slide level and moderately at the patch level (strongest on Xenium), and achieves moderate correlation at the challenging single-cell level. Furthermore, downstream tasks explore the clinical relevance of these predictions, demonstrating their utility in survival analysis, spatial hazard mapping, and tumor classification via molecular signatures.

1 Introduction

Understanding genes is key to grasping how our body works and evolves, as well as making sense of complex diseases such as cancer. That is why ST and scT may provide critical insights for understanding and effectively treating tumors. The rapid evolution of ST and scT has transformed our understanding of tissue biology by enabling the systematic measurement of expression levels for all or most genes throughout tissue space [1]. Despite their revolutionary potential, these methods remain constrained by high experimental costs (thousands of dollars per slide), and are not yet routine in most hospitals. On the other hand, histopathology is already a common practice in cancer diagnosis, producing hematoxylin and eosin-stained WSIs. WSIs are high-resolution digital representations of histopathology slides that capture the complete morphological context of the tissue. They preserve cellular structures, tissue architecture and spatial organization of cells.

Although ST provides a good understanding of the spatial distribution of genes, it does not offer the finest granularity available. In one patch there are often dozens to hundreds of cells, so the expression of each spot is actually the mixture of the gene expressions of each cell within each patch. scT achieves the finest granularity available, providing the gene expression for each detected cell in a WSI. Additionally, WSIs are often used as a conclusive diagnostic tool because many diseases, including cancer, are characterized by abnormal spatial organization within tissues [2]. By mapping mRNA within tissues, we improve our ability to visualize and interpret biological processes, since abnormal mRNA expression is frequently associated with human tumor initiation, development, progression and metastasis [3].

The relationship between tissue morphology and gene expression arises from the fact that tissue architecture and histological appearance are a direct manifestation of underlying genes at play. Differences

in cell type, microenvironmental context and the presence of a tumor can result in characteristic morphological signatures. These morphological features can share a high amount of information with gene expression [4]. Therefore, using WSIs to predict ST and scT represents the next logical step, that not only assists doctors in diagnosing tumor slides but also enables a wide range of downstream tasks such as biomarker discovery and survival prediction.

ST and scT have advanced rapidly in both experimental and computational methods, yet there remains a critical limitation in accurately predicting transcriptomic profiles. Moreover, current deep learning models can infer partial gene expression patterns from histopathology slides. However, their predictions often show limited gene coverage, weak generalization across different tissues, and low accuracy in reconstructing the full transcriptome. While the acquisition of spatial data is expensive and time-consuming, bulk gene expression data, which provides the average gene expression across the whole tissue, is far more accessible and abundant. However, the translation from bulk-level to spatial gene expression profiles remains under-explored.

To bridge this gap, this work leverages three types of data: (1) WSIs paired with bulk gene expression; (2) WSIs paired with ST; and (3) WSIs paired with scT. WSIs are typically divided into smaller, fixed-size patches to enable efficient processing by deep learning models. ST requires that patches align with the spots where mRNA was sequenced (arranged across the WSI in a predefined grid pattern to enable spatially resolved transcriptomic profiling). In addition, in cases where scT data is available, each detected cell has an associated gene expression profile, although these profiles can be very sparse, since each cell may have counts for only a few of the genes in the panel. On the other hand, if only bulk gene expression is available, the spatial information is lost and the gene expression corresponds to the mixture of the whole tissue. This poses a challenge, as the bulk gene expression does not correspond exactly to the WSI and the spatial relationship between the WSI and the genomic data is unknown [5]. Despite this, the bulk gene expression can be reasonably approximated as representative of the average gene expression across the WSI, given that the WSI and the sampled tissue originate from the same primary tumor.

The results obtained by recent studies [6–9] support the hypothesis that, in most cases, morphology encodes sufficient information to approximate gene expression patterns. Consequently, deep learning models trained on paired data WSI-transcriptomics can exploit these correlations to reconstruct spatial gene profile maps directly from WSIs.

Most existing models that have the capability of predicting at the single-cell level employ the same training strategy: treating each spot as a bag of cells and using Multiple Instance Learning (MIL) to predict the spot's gene expression. This way, the model learns how the aggregation of the cell's gene expression should look like through weakly supervised learning. This strategy was employed in a state of the art model, DeepSpot2Cell [10], which leverages a PFM and spatial multi-level context to predict scT using spot-level supervision. A similar approach was proposed in [11]. The methodology is the same, as the spot is represented as a bag of cells, and the cells' gene expression is learned through MIL.

In this work, a Multi-Scale GNN was developed to predict gene expression from WSIs at three levels: bulk, spatial and single-cell. The consistency losses encourage the predictions at these levels to agree.

2 Results

2.1 Overall pipeline architecture

The overall pipeline can be seen in Figure 1, which shows how an unseen WSI would be processed and the inference pipeline of our model. By extracting morphological features using a PFM (UNI2-h) and constructing spatial graphs at both patch and cell levels, the model learns the hierarchical relationships within the tissue. Cell information is passed to the patch branch, and patch information is aggregated to obtain the slide-level prediction. The network is trained on a combined dataset of TCGA, Visium, and Xenium samples across lung, kidney, and breast tissues. A hybrid training strategy is employed, leveraging direct supervision alongside weakly supervised consistency losses to ensure predictions align across biological scales.

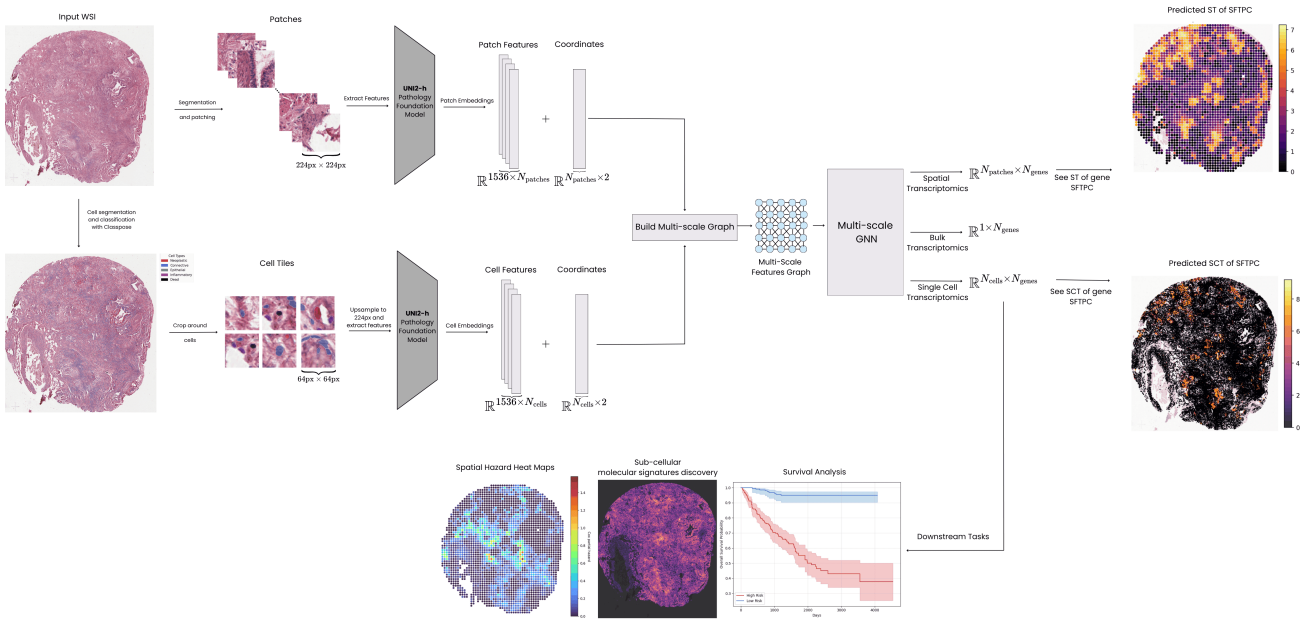


Figure 1: Proposed model's overall pipeline architecture.

2.2 Visual features correlate with gene expression

This study relies on a fundamental hypothesis: gene expression is predictable, to some degree, from the visual features extracted from WSIs.

Visium slides contain a high amount of batch effects both in the feature space and gene expression space. This creates a slide-level confound, where patches from the same slide cluster together in both spaces not necessarily due to shared biology, but due to these batch effects. This phenomenon has been documented for PFMs, which have been shown to encode strong site-specific signatures in their principal components [12].

The patch features, extracted using UNI2-h, were compressed to 100 components using PCA. The 17,038-dimensional gene expression vectors were compressed into 300 principal components. To assess whether the correspondence between embeddings and transcriptomics reflects a biological signal, Harmony [13] was applied to both spaces to correct for *slide_id* as a batch covariate. The number of clusters was computed using the Leiden algorithm, with 22 clusters in the patch feature space and 20 clusters in the gene expression space.

The t-SNE plot of the gene expressions was then colored using the cluster assignment from the patch feature space. The goal was to investigate whether clusters in the patch embedding space correspond to clusters in the gene expression space. As shown in Supplementary Figure S1, this correspondence does indeed occur. The cluster purity was 55%, the Adjusted Rand Index (ARI) was 0.1251, and the Normalized Mutual Information (NMI) was 0.2773. While Harmony correction cannot guarantee the complete removal of every possible batch effect, the fact that a substantial correlation persists after explicit slide-level correction indicates that the observed relationship is not solely attributable to batch effects. This is further supported by computing the PCC between gene expressions in the same cluster, which was found to be 0.3007. While moderate, it suggests that patches with similar features tend to exhibit similar gene expression profiles. Furthermore, the PFM used (UNI2-h) is not "gene-aware", that is, it was only pre-trained on general histopathology images without exposure to ST data. This general-purpose encoder, which was pre-trained without exposure to ST data, recovers patterns that align with the gene expression space. Some of this alignment is expected, since cell types that are distinguishable under the microscope are themselves defined by the expression of specific genes, so the observed correlation likely reflects, at least in part, cell-type composition rather than finer morphological cues. Nonetheless, it indicates that the visual features capture information that is informative of the underlying transcriptomic state.

To further study this relationship, Classpose [23] classifies each cell into one of the five available

Table 1: Proposed model's metrics across datasets. Metrics are reported as the mean and standard deviations across five cross-validation folds.

Metric	TCGA $N = 1724$	Visium $N = 105$	Xenium $N = 47$
Bulk PCC	0.881 ± 0.010	0.774 ± 0.146	0.797 ± 0.057
Patch PCC	–	0.509 ± 0.026	0.786 ± 0.041
Cell PCC	–	–	0.396 ± 0.021
Patch AUROC	–	0.554 ± 0.004	0.703 ± 0.011
Cell AUROC	–	–	0.649 ± 0.011

labels: neoplastic, connective, inflammatory, epithelial, and dead. In Supplementary Figure S2, it is clear that certain gene profiles are related to the percentage of the cell type present in the corresponding patch. For example, for neoplastic cells (tumor-related), there are only certain clusters that contain a high percentage of this cell, meaning certain gene expression profiles are associated with this cell type.

2.3 Gene expression prediction at bulk, spatial and single-cell levels

The results are organized by dataset (TCGA, Visium or Xenium) and by tissue type (lung, kidney, breast and overall cohort). This enables a more detailed analysis, uncovering the performance of each cohort in the dataset. The metrics are reported as the mean and standard deviations across five cross-validation folds. Table 1 presents the main performance metrics obtained by the proposed model across datasets. The bulk, patch and cell PCC values in this table compare expression across genes within a sample. The gene-wise mean PCC curves in Figure 2A and B compare expression across samples for a given gene and report the mean for the selected top-predictive genes. The patch and cell AUROC values are computed per gene within each slide: spots (or cells) whose true expression of the gene is in the top quartile are labelled as high, those in the bottom quartile as low, the middle 50% are discarded, and the AUROC measures how well the predicted values separate the two groups. The AUROC is then averaged over genes and over the test slides of each fold (see Section 4.9).

Analyzing the Visium results in Supplementary Table S1, a very distinct performance gap emerges between bulk-level and patch-level predictions. The model achieves a strong bulk PCC (overall: 0.774 ± 0.146), but the patch-level predictions drop to 0.509 ± 0.026, which shows a moderate linear relationship between the predictions and the ground-truth. This drop in performance is expected, since it reflects the inherent challenge of predicting ST data. Furthermore, lung samples exhibit the highest patch-level AUROC, although the spot-wise patch PCC is lower, the patch AUROC is the highest among all organs (0.586). Figure 2A (chart A) corroborates this: lung exhibits the highest gene-wise PCC. This demonstrates that the model is able to capture distinct spatial profiles, consistently predicting higher expression values in regions that genuinely correspond to the top quartile of gene expression. It is worth noting that breast does not have a valid FDR value, this outcome is attributable to the small cohort size ($N = 7$), where each test split contains at most two breast samples, severely limiting statistical power in hypothesis testing.

In the case of the Xenium dataset, the model maintains a strong predictive power at both the slide level (overall: $r = 0.797$) and the patch level ($r = 0.786$). On the other hand, the cell-level PCC is smaller at 0.396, nonetheless this result still shows that the model is able to predict at single-cell level with a moderate positive linear relationship with the ground-truth values. Stratifying by cell type, the model exhibits a higher correlation in epithelial ($r = 0.406$) and neoplastic ($r = 0.456$) cells compared to dead ($r = 0.330$), inflammatory ($r = 0.340$) and connective ($r = 0.367$) cells. Single-cell predictions could only be directly evaluated for the genes measured by Xenium.

Finally, for the TCGA cohort, the model shows high reliability in predicting the bulk gene expression across all tissue types. This is confirmed by the very strong correlations (overall: $r = 0.881$) in Supplementary Table S3, along with the high gene-wise correlations in Figure 2 (A) (chart C). The fraction of significantly predicted bulk genes is also high ($72.6\% \pm 16.0\%$), confirming the model's capacity to predict global slide expression.

Figure 2B compares the proposed model with DeepSpot2Cell and STFlow. DeepSpot2Cell achieves higher spot-level gene-wise PCC values than the proposed model. On Visium the difference is within the fold variability, while on Xenium it is larger than the fold variability. DeepSpot2Cell also attains a higher patch AUROC on both datasets (Visium 0.590 vs 0.554; Xenium 0.764 vs 0.703) (Supplementary Tables S6 and S7). Conversely, at cell level, the proposed model is clearly better than DeepSpot2Cell (cell PCC 0.396 vs 0.349; cell AUROC 0.649 vs 0.523). Differences in the supervision available to each model limit these comparisons: the proposed model has direct supervision at cell level, while DeepSpot2Cell uses a weakly supervised loss function. STFlow has lower gene-wise PCC in the displayed comparisons, although on Visium its patch-level metrics are comparable to the proposed model's.

2.4 Spatial patterns in Visium and Xenium

To complement the quantitative metrics reported above, this section presents qualitative visual comparisons between predicted and ground-truth gene expression for individual slides from the Visium and Xenium test sets. For each slide, the displayed gene was selected as the one with the highest spatial expression variance within that specific slide, as high-variance genes tend to exhibit the clearest spatial patterns and therefore best illustrate whether the model is capturing meaningful morphological structure.

Figure 2 (C) shows a Visium spot-level evaluation for slide INT22, with *IGKC* as the selected gene. The predicted spatial expression map closely reproduces the main structural pattern of the ground truth.

Figure 2 (D) presents the corresponding evaluation on the Xenium slide TENX200, at spot and single-cell resolution, for the gene *FBLN1*. The predicted spot-level map visually reproduces the tumor morphology visible in the WSI, matching the true expression pattern with a spot-level PCC of $r = 0.775$. At the single-cell level, the predicted expression is visibly noisier and the correlation drops accordingly ($r = 0.458$), which is expected given the substantially higher spatial resolution and consequent sparsity of single-cell transcript counts.

2.5 Predicted bulk expression supports survival analysis

In order to validate the quality of the bulk gene expression predictions made by the model, a survival analysis task was performed on both the prediction and the ground truth bulk gene expression, and the results were compared. Clinical outcome variables were obtained from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR), a curated and standardized clinical annotation dataset for TCGA cohorts [14]. More specifically, for TCGA-KIRC, TCGA-LUAD and TCGA-BRCA the Disease-Specific Survival (DSS) time was extracted for each patient. For each cohort, three Cox models [15] were trained: one on the 50 PCs (obtained through PCA on the 17,038 dimensional gene expression) of the true bulk gene expression, another on the 50 PCs of the predicted bulk gene expression and the last on the 50 PCs of the visual features embedding. The Cox models were trained and tested on all patients from each TCGA cohort using five-fold cross validation. Table 2 compares the C-index obtained using ground-truth bulk expression, predicted bulk expression, and patch visual features (mean-pooled UNI2-h embeddings) for each patient. Using the predicted instead of the ground-truth bulk expression lowers the C-index in all three cohorts, by 0.047 in TCGA-LUAD, 0.083 in TCGA-KIRC, and 0.029 in TCGA-BRCA. The results also reveal that the gene expression is not a good predictor for the patient outcome in the TCGA-LUAD cohort, since the C-index is close to 0.5, which means that the Cox model's risk scores have no meaningful relationship with the survival times. In contrast, for TCGA-KIRC and TCGA-BRCA the scores reveal an acceptable model, with C-index scores above 0.6, across the three input types.

2.6 Molecular signatures and spatial hazard heat maps

The analysis made in the previous Section is important, but it has a limitation: since the 17,038 genes were compressed into 50 principal components, the contribution of individual genes to the patient's outcome can only be assessed indirectly. Each principal component is a weighted combination of all

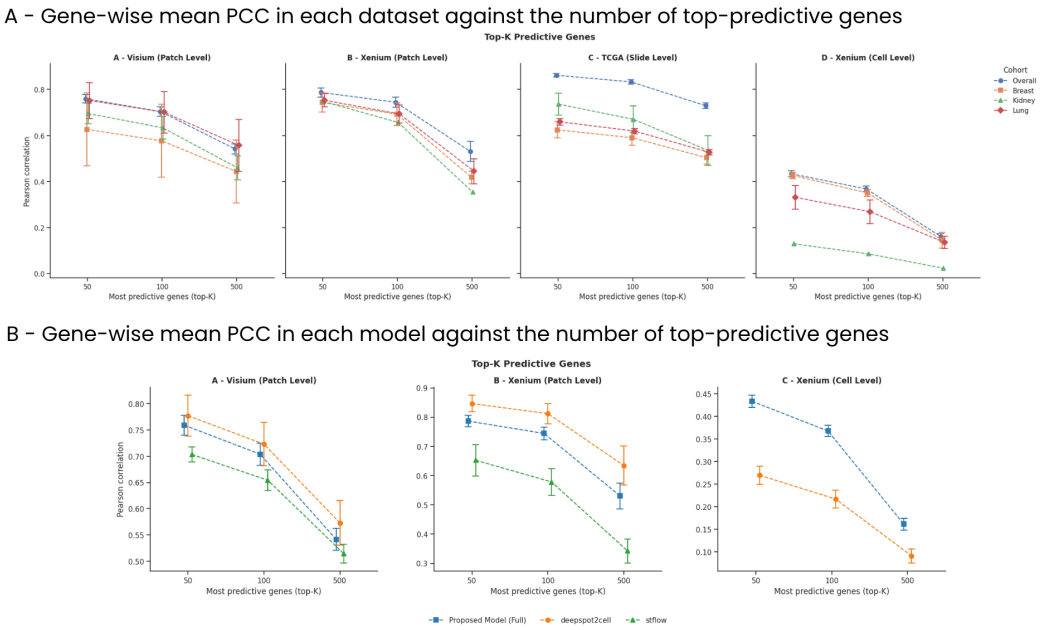


Figure 2: **Multi-Scale GNN results.** **A**) Gene-wise mean PCC in each dataset against the number of top-predictive genes in the proposed model. **B**) Gene-wise mean PCC in each model against the number of top-predictive genes. **C**) Visium spot-level evaluation for slide INT22 (kidney), gene *IGKC*. Left to right: WSI, true spot expression, predicted spot expression, predicted cell expression. **D**) Xenium spot-level evaluation for slide TENX200 (breast), gene *FBLN1*. Left to right: WSI, true spot expression, predicted spot expression, true cell expression, predicted cell expression.

Table 2: Comparison of Bulk C-index metrics across cohorts.

C-index	TCGA-LUAD $N = 393$	TCGA-KIRC $N = 413$	TCGA-BRCA $N = 834$
Real bulk	0.544	0.748	0.639
Predicted bulk	0.497	0.665	0.610
Visual features	0.555	0.688	0.730

genes, so the genes that drive it can be identified from their weights in that component (also known as loadings). The Cox model coefficients can then be mapped back to the genes. However, every gene contributes to every component, and each component often mixes several biological processes, which makes it difficult to attribute an increase or decrease in risk to specific genes or pathways. To tackle this, Cox models were trained directly on the mean expression of established molecular signatures. A molecular signature or Hallmark gene set is a collection of genes that are associated to certain biological processes.

The gene sets used were extracted from the Molecular Signatures Database (MSigDB) [16], and the ones analyzed are: (1) proliferation, this molecular signature contains 200 genes that are required for cellular division and cell cycle progression. Tumors over-express this signature to enable uncontrolled proliferation and rapid expansion of the tumor; (2) apoptosis, this gene set, with 161 genes, is responsible for programmed cell death which eliminates damaged or abnormal cells. Tumor cells frequently evolve to suppress this pathway, enabling them to evade self-destruction; (3) angiogenesis, this hallmark, containing 36 genes, represents the biological process of generating new blood vessels to supply tissues with oxygen and nutrients. Tumors trigger this pathway to build blood supplies, preventing necrosis and sustaining growth; and (4) hypoxia, this signature, with 200 genes, captures the cellular survival and metabolic responses triggered by environments with low oxygen levels. As tumors outgrow their vascular supply, tumor cells activate this pathway to adapt to oxygen deprivation.

The Cox models were trained on the expression of each gene set (the average expression across the genes within each molecular signature) to predict the DSS time. Supplementary Figure S6 displays the log hazard ratios/risk weights and 95% confidence intervals assigned to each gene set, along with their *p-values*. The proliferation molecular signature is the one that contributes more to the risk score especially in TCGA-KIRC. The remaining signatures did not reach statistical significance, so no conclusions are drawn about their association with survival.

Furthermore, to take advantage of the spatial predictions at spot level, the Cox model trained with the bulk gene expression can be applied individually to each spot's predicted gene expression. This analysis should only be done in slides where the predicted pathways yielded a statistically significant *p-value*, in this case TCGA-KIRC samples with the proliferation signature. This way the spatial correlation between the gene expression of each molecular signature and the risk score can be seen. Figure 4 (A) shows one of these examples, on the left the WSI along with the tumor annotation are shown, followed by the spatial expression of the proliferation signature, and on the right the spatial hazard heat map is displayed. The spatial risk scores align with the predicted spatial expression of the proliferation associated genes ($r = 0.978$). These risk scores are derived from the predicted signature expression using the bulk-trained Cox model. This correlation therefore does not independently validate local risk.

2.7 Cell-level molecular signature discovery

At the cellular level, model predictions were evaluated by mapping the spatial activity of MSigDB molecular signatures against subcellular ground truth in Xenium samples. Visualizing these cell-level molecular signatures might prove important in a clinical setup, for example, for the proliferation signature, the previous Sections showed that it is highly correlated with tumor related risk. Visualizing the continuous expression of this signature might help identify the cells that are "driving" the tumor expansion.

Figure 4 (B) shows one example of the comparison between the model's prediction of a signature's

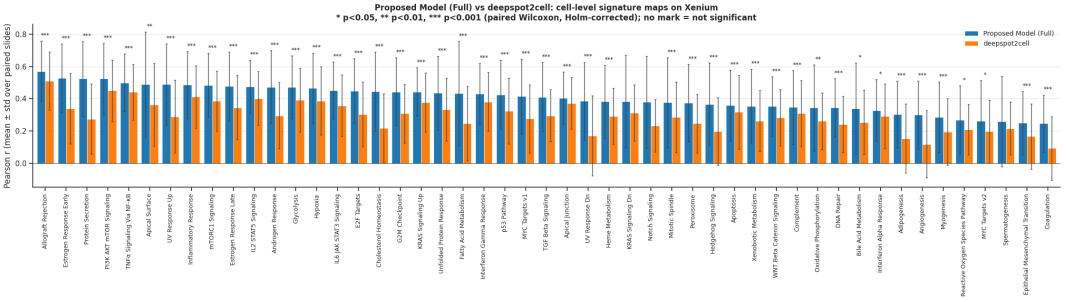


Figure 3: Spatial PCC and standard deviation across paired Xenium slides of the cell-level signature maps for all hallmark gene sets, for the proposed model and DeepSpot2Cell.

Table 3: Average spatial PCC (mean ± standard deviation across slides) across the analyzed signatures in the proposed model and in DeepSpot2Cell.

Signature	Proposed Model	DeepSpot2Cell
Proliferation	0.443 ± 0.211	0.309 ± 0.181
Hypoxia	0.465 ± 0.218	0.387 ± 0.211
Apoptosis	0.358 ± 0.219	0.318 ± 0.229
Angiogenesis	0.299 ± 0.210	0.118 ± 0.208

expression and the ground truth. These signature maps were compared after spatial binning and Gaussian smoothing (Section 4.10). The spatial PCC therefore evaluates the agreement between these signature maps, rather than the gene expression of individual cells. As with the single-cell evaluation, these comparisons are limited to the genes measured by Xenium.

Table 3 reports the average spatial PCC and standard deviation across slides for each signature across all Xenium slides ($N = 47$) in the dataset. The model achieves a moderate spatial agreement across signatures, ranging from $r = 0.299$ for Angiogenesis to $r = 0.465$ for Hypoxia. Hypoxia and proliferation reach higher correlations (0.465 and 0.443) than the gene-level cell PCC of 0.396 (Table 1). The two metrics are not directly comparable, since signature maps are binned and smoothed, but this is consistent with the sparsity of single-cell counts: aggregating the signal over many genes and spatial bins reduces noise and makes it easier to recover. This task shows the model's capability of predicting the gene expression at cell level and enabling molecular signatures discovery. When compared against DeepSpot2Cell under the same evaluation protocol, the proposed model achieves a higher mean correlation across all four signatures, with the largest gains in Angiogenesis (0.299 vs 0.118) and Proliferation (0.443 vs 0.309). Since the same slides are evaluated by both models, the differences were assessed with a paired Wilcoxon signed-rank test, which accounts for the slide-level variability. Figure 3 extends this comparison to all hallmark gene sets in MSigDB, the proposed model achieves a significantly higher correlation in all but three of them, for which the difference is not significant, and none of the hallmarks favors DeepSpot2Cell.

2.8 Molecular signatures localize tumor-associated regions

This task evaluates whether patch-level predictions for TCGA slides are effective for classifying tissue as tumor or non-tumor. To do this, a publicly available dataset [17] containing tumor annotations for TCGA was used. It's important to mention that these annotations were produced by trained annotators rather than pathologists, based on the visual appearance of the tissue, and are not "pixel-perfect", meaning that the patch-level annotations may be coarse. Nonetheless, they provide a valuable benchmark to evaluate spot-level predictions on TCGA slides, which lack ground-truth ST.

The initial analysis involves measuring the spatial correlation between predicted molecular signatures (proliferation, apoptosis, angiogenesis and hypoxia) and tumor annotations. Table 4 reports the average AUROC and the tumor prevalence ratio across the different TCGA cohorts. These metrics were

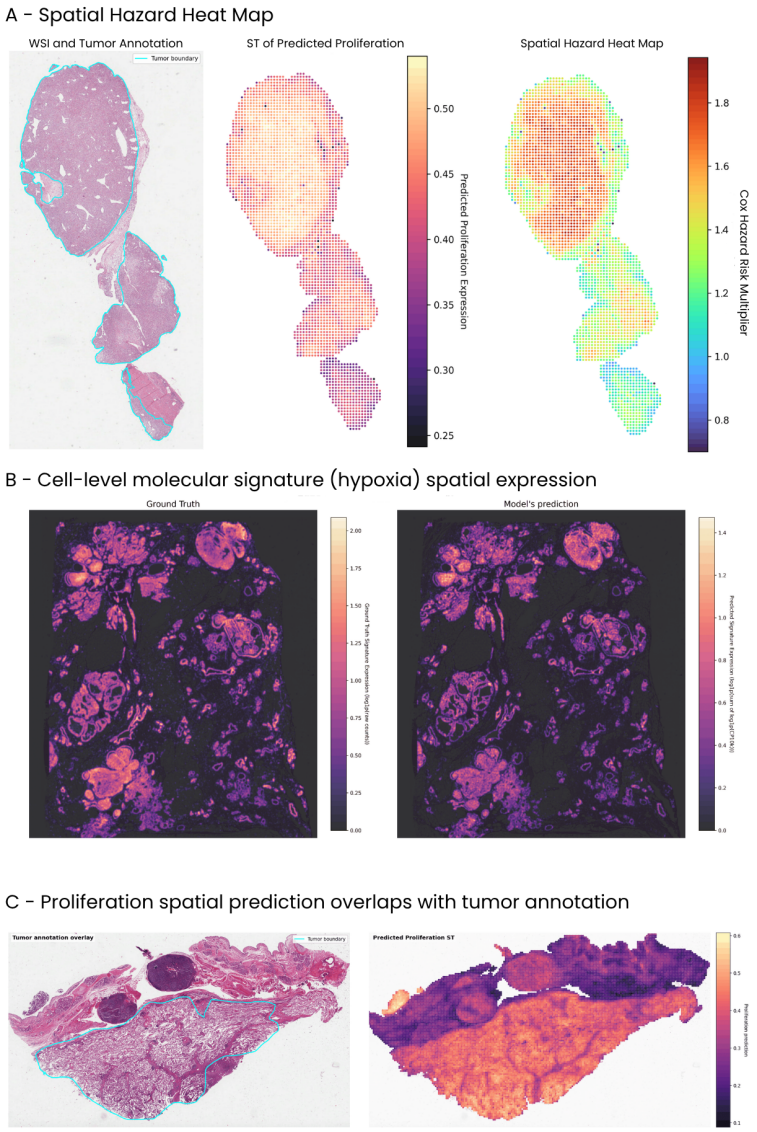


Figure 4: **Downstream tasks using predicted gene expression. A)** WSI with tumor annotation, predicted proliferation expression, and spatial hazard heat map in a TCGA-KIRC sample. **B)** Predicted and ground-truth hypoxia signature maps in a Xenium sample, after spatial binning and Gaussian smoothing ($r = 0.840$). **C)** Tumor annotation and predicted spatial proliferation expression (AUROC = 0.921).

computed for each slide and averaged across all slides within each cohort. Tumor prevalence ratio measures whether patches where a molecular signature is highest are more tumor annotated compared with the cohort's baseline tumor fraction. For example, in TCGA-LUAD, the Proliferation gene set has a tumor prevalence ratio of 1.412, indicating that patches within these hotspots are 1.412 times more likely to be annotated as tumor than patches on that cohort overall. Moreover, in TCGA-LUAD, mainly proliferation related genes co-occur with tumor annotated patches, yielding a high AUROC (0.711). Additionally, for TCGA-KIRC, both hypoxia and proliferation signatures exhibit high co-occurrence with tumor annotations, mainly indicated by the high AUROC values (0.776 and 0.648, respectively). Finally, in TCGA-BRCA, both proliferation and hypoxia showed the strongest AUROC values, indicating that the model successfully associates regions of tumor morphology with these gene sets.

To further illustrate these associations, Figure 4 (C) displays an example of how these molecular signatures co-occur in tumor annotated regions, supported by a high AUROC (0.921). Overall, these results indicate that tumor-associated gene sets (proliferation and hypoxia) consistently align with tumor annotations, showing that spot-level predictions retain tumor-related information on TCGA cohorts.

Table 4: Tumor annotation and predicted molecular signature metrics by cohort. LUAD, lung adenocarcinoma; KIRC, kidney renal clear cell carcinoma; BRCA, breast invasive carcinoma.

Cohort / signature	*N*	AUROC	TPR
LUAD: Prolif.	431	0.711	1.412
LUAD: Hypoxia	431	0.593	1.131
LUAD: Apopt.	431	0.525	0.987
LUAD: Angio.	431	0.517	0.992
KIRC: Prolif.	416	0.648	1.193
KIRC: Hypoxia	416	0.776	1.444
KIRC: Apopt.	416	0.608	1.136
KIRC: Angio.	416	0.579	1.099
BRCA: Prolif.	850	0.810	1.910
BRCA: Hypoxia	850	0.728	1.596
BRCA: Apopt.	850	0.688	1.485
BRCA: Angio.	850	0.631	1.344

TPR: tumor prevalence ratio; Prolif.: proliferation; Angio.: angiogenesis; Apopt.: apoptosis.

Table 5: Tumor classification performance by input type and TCGA cohort.

Cohort	Input	Bal. Acc.	F1-score	AUROC
LUAD	50 signatures	0.707	0.707	0.771
	Full transcript.	0.684	0.679	0.742
	Patch embeddings	0.741	0.740	0.822
KIRC	50 signatures	0.857	0.861	0.932
	Full transcript.	0.851	0.850	0.912
	Patch embeddings	0.862	0.861	0.942
BRCA	50 signatures	0.841	0.838	0.909
	Full transcript.	0.805	0.803	0.886
	Patch embeddings	0.860	0.857	0.925

Bal. Acc.: balanced accuracy. LUAD: lung adenocarcinoma; KIRC: kidney renal clear cell carcinoma; BRCA: breast invasive carcinoma.

2.9 Tumor classification using predicted molecular signatures

Furthermore, a three-layer Multilayer Perceptron (MLP) was trained to classify each patch as tumor or non-tumor based on predicted gene expression. The classifier was trained using three different input feature sets: (1) predicted expression for the 50 MSigDB Hallmark signatures; (2) predicted full transcriptome expression (17,038 genes); and (3) original UNI2-h patch embeddings. Classifiers were trained and evaluated with five-fold cross-validation with splits at slide level, to avoid cross-patient contamination. Table 5 shows that both predicted transcriptomics and patch embeddings provide discriminative features for tumor classification, as reflected by high AUROC and F1-scores. Notably, the classifier performs better when using the 50 molecular signatures rather than the full transcriptome. This may happen because the individual expressions of the 17,038 genes have high variance and noise, the molecular signatures act as a biologically grounded dimensionality reduction that filters out prediction noise. Using the UNI2-h embeddings outperforms using the predicted gene expression in almost all cases. This advantage should be interpreted with care: the tumor annotations were made by non-expert annotators based on the visual appearance of the tissue, which favors features that directly encode morphology, such as the PFM embeddings. Nonetheless, the 50 signatures achieve similar results, while compressing the space into 50 interpretable biological values.

2.10 Predictions retain tissue-specific patterns

This task aims to assess whether the tissue-specific gene expression patterns are maintained and learned by the model. To conduct this analysis, a simple two-layer MLP was trained to classify tissue origin (lung, kidney or breast) from spot gene expression. The classifier was evaluated with five-fold cross-validation,

using 105 Visium slides (210,246 patches). The MLP was optimized on true spot expressions using cross-entropy loss and subsequently evaluated on both measured and predicted test spot expressions. Of particular interest is the difference between the MLP's predictions using the ground truth and the predicted gene expression: high similarity indicates that spot predictions retain tissue-specific patterns. To evaluate this difference, the Kullback-Leibler (KL) divergence is computed between the MLP's predicted probability vectors derived from ground-truth and the predicted expression. Specifically, for each spot, two predictions are obtained: (1) one using the true gene expression; and (2) one using the predicted gene expression by the proposed model. The MLP outputs a logit vector with three entries (one for each class: lung, kidney and breast), and a softmax function is applied to obtain class probabilities. Across all test spots, the average Kullback-Leibler (KL) divergence between the two distributions is 1.595.

The model is further tested on 100 randomly selected TCGA slides in each fold, using the tissue origin annotated in the TCGA metadata as ground truth.

Supplementary Table S4 shows the accuracy, F1-score and AUROC averaged across the five folds; for example, the average accuracy is 0.980 on the true Visium spot expression, 0.859 on the predicted Visium spot expression and 0.815 on the predicted TCGA spot expression.

In addition, the quality of the reconstructed embeddings is evaluated by assessing whether they retain tissue-specific features. The model is evaluated using five-fold cross-validation in the full dataset (TCGA, Visium and Xenium), totaling 21,456,735 patches. The MLP is trained to classify the patch tissue origin using the original UNI2-h embeddings.

The mean KL divergence is 0.5001. The results in Supplementary Table S5 showing comparable accuracy, F1-score, and AUROC across both sets, confirm that the model reconstructs features from predicted ST while preserving tissue-specific characteristics.

2.11 Ablation studies

To assess the contribution of key design choices in our model, four ablated variants of the model were evaluated: (1) without feature reconstruction; (2) without consistency losses; (3) without TCGA samples; and (4) without TCGA cell level weak supervision. Supplementary Tables S6, S7, and S8 report the performance metrics across benchmark models and ablated variants for the Visium, Xenium, and TCGA cohorts, respectively.

Removing the feature reconstruction objective leads to a consistent drop across most metrics on the datasets. On Visium, patch-level AUROC falls from 0.554 to 0.548 and the fraction of significant genes drops from 37.7% to 34.6%. On Xenium, cell-level AUROC drops slightly (0.649 to 0.637), suggesting that reconstructing the original patch embeddings encourages the model to retain morphological information that also benefits downstream expression.

Removing the consistency losses degrades the structural alignment of the model's predictions. The most drastic impact is observed in Cell-to-Patch (CP) consistency on Visium, which collapses from 0.896 to 0.222, demonstrating that without this explicit constraint, the model fails to align both prediction levels. Patch-to-Slide (PS) consistency also drops across all datasets, including a substantial decrease on TCGA (0.893 to 0.691). Consequently, this lack of structural regularization harms overall bulk predictive performance on the spatially resolved datasets, dropping bulk PCC from 0.774 to 0.726 on Visium and from 0.797 to 0.779 on Xenium.

Removing TCGA from training has a more mixed effect: bulk PCC improves substantially on both Visium (0.774 to 0.875) and Xenium (0.797 to 0.831), likely because the model is trained exclusively, at slide-level, with pseudo-bulk labels from these cohorts. However, this comes at the cost of reduced feature reconstruction quality (Visium: 0.634 to 0.522; Xenium: 0.566 to 0.453), indicating that TCGA's scale and diversity still contribute to feature reconstruction.

Finally, ablating the cell-level weak supervision from TCGA reveals its important role in fine-grained spatial localization. In Xenium, patch-level AUROC drops (0.703 to 0.674), while cell-level AUROC also drops slightly (0.649 to 0.640). This demonstrates that the cell-level information extracted from TCGA cohort, although it offers only weak supervision to the model, seems beneficial for guiding the model to accurately resolve expression patterns at the patch and cell-levels, even when evaluating on

entirely different spatial platforms.

3 Discussion

In this work, a Multi-Scale GNN was developed to predict gene expression from WSIs at three levels: bulk, spatial and single-cell. The model combines data from TCGA, Visium and Xenium samples, using direct supervision when gene expression measurements are available and weak supervision through consistency losses. This allows the model to predict 17,038 genes across lung, kidney and breast tissues, while learning from datasets with different levels of data granularity.

The results show that the model performs best at slide level, reaching an average PCC of 0.881 on TCGA. At patch level, the average PCC was 0.509 on Visium and 0.786 on Xenium. Predicting cells was more challenging, with an average PCC of 0.396 on Xenium. Compared with DeepSpot2Cell, our model obtained higher cell-level PCC and AUROC, while DeepSpot2Cell was better at patch level, in both PCC and AUROC. These results support the use of a shared model for the three prediction tasks, although performance remains uneven across genes and tissues.

The downstream tasks showed that the predictions retain tissue-specific and tumor-related information. Predicted molecular signatures supported tumor localization and classification, while the cell-derived signature maps partially reproduced spatial patterns observed in Xenium. Using 50 molecular signatures for tumor classification performed better than using the full predicted gene expression, showing the value of grouping genes into interpretable biological processes. Predicted bulk expression also supported survival analysis in kidney and breast cancer, although performance was lower than with measured expression in all three cohorts.

The ablation studies showed that the consistency losses are important for keeping predictions aligned across scales. Feature reconstruction and weak supervision from TCGA cells also helped several prediction metrics. However, adding TCGA did not improve every task: removing these samples increased the bulk PCC on Visium and Xenium. This shows that combining datasets provides useful information, but also introduces cohort specific differences that the model must accommodate.

These findings have some limitations. Single-cell predictions could only be directly evaluated for the genes measured by Xenium, and TCGA spatial predictions could not be compared with ground truth. Batch effects, small cohorts, cell segmentation errors and memory constraints also affected the study. Furthermore, differences in the supervision available to each model limit the conclusions drawn from the benchmark comparisons. Nevertheless, this work provides a framework for combining bulk, spatial and single-cell data to obtain molecular predictions from histopathology images.

Future work should evaluate the model on larger and more balanced datasets, including samples from independent institutions. This would help determine how well the model handles changes in tissue preparation, imaging and sequencing platforms. Broader single-cell measurements would also allow more of the predicted genes to be validated.

Another area of improvement is addressing batch effects directly during training. While algorithms like Harmony can align representations in low-dimensional spaces, they cannot be easily applied to the full transcriptome during end-to-end learning. Future iterations could incorporate methods that explicitly penalize the model for learning hospital or scanner-specific features, encouraging the model to focus strictly on biological patterns rather than technical differences between cohorts.

On the architectural level, the current pipeline only passes information bottom-up, from cells to patches and then to the slide. Incorporating top-down information flow would allow global tissue context to guide local patch and cell predictions, which could help resolve ambiguous local features. Additionally, the cell feature extraction pipeline could be refined. Instead of using a frozen patch-level foundation model on upsampled cell crops, fine-tuning the visual backbone or using encoders pre-trained specifically on single-cell morphology could provide more informative input embeddings for the cell graph.

4 Methods

4.1 Dataset and experimental design

The multi-modal dataset created comprises several individual datasets, spanning across three different organs: lung, kidney and breast. The dataset aggregates bulk RNA-seq data from TCGA studies, and ST and scT data from several different studies all aggregated by the HEST-1k dataset [18]. For this study, from HEST-1k, 152 cases were selected, lung had a total of 60 cases (36 Visium and 24 Xenium), kidney had 64 cases (62 Visium and 2 Xenium), while breast had 28 cases (7 Visium and 21 Xenium). The ST data comes from the Visium samples, on the other hand, scT data is present in Xenium samples that were acquired using the Xenium In Situ Platform which is commercialized by 10x Genomics.

When it comes to the bulk gene expression data, there is much more data available. Publicly available TCGA datasets for lung adenocarcinoma (TCGA-LUAD) [19], clear cell renal cell carcinoma (TCGA-KIRC) [20] and breast invasive carcinoma (TCGA-BRCA) [21], were analyzed and obtained via the NCI Genomic Data Commons portal. In the TCGA-LUAD project we used 465 cases, this is, 465 pairs of DX (diagnostic) slides and bulk gene expression, in TCGA-KIRC 510 cases were used and in TCGA-BRCA there were 1052 cases used. In TCGA projects the number of sequenced genes is 60,660. Out of the 2027 TCGA samples, 268 of them were randomly selected to have their cells detected, segmented and classified (with Classpose) and then their features extracted with UNI2-h. For these slides, during training, the model performs weak supervision at cell level through the consistency losses. At inference time, the model is able to also predict at cell-level for these slides, effectively generating novel virtual scT data.

In terms of training strategy, the model was trained using five-fold cross-validation with splits at slide level, to avoid cross-patient contamination (there is only one WSI for each patient), meaning that the test WSIs/patients were never seen during training. Moreover, due to memory constraints of the hardware used, a hard limit of 300,000 cells per WSI at test time was used. For slides exceeding this limit, 300,000 cells were randomly sampled for inference, and the remaining cells were not processed. The evaluated datasets contained 1,724 TCGA, 105 Visium and 47 Xenium slides.

4.2 WSI segmentation and feature extraction

For the HEST-1k slides with ST, each patch is centered around a spot; consequently, the patch's gene expression is approximated by the gene expression in the spot. Throughout this work, the terms spot and patch are used interchangeably.

Tissue segmentation involves deciding which parts of the image are tissue or background and removing the background. Trident [22] was the segmentation tool used. Trident also allows automatic patching of the segmented tissue: it extracts the coordinates of each patch based on two specified parameters, patch size and magnification (here, 224×224 px at $20\times$ magnification). With the coordinates of each patch in the WSI, the patch images can be extracted and fed into the PFM (UNI2-h in this work) for feature extraction.

Following WSI segmentation and patching, the next step in the pipeline is the segmentation and classification of individual cells. The model used for this task was Classpose [23], which is designed to perform simultaneous instance segmentation and classification of cells in histopathology images. A Classpose model was trained on the PanNuke dataset [29, 30] to predict five cell classes: neoplastic, inflammatory, connective, dead and epithelial. The data was randomly split into 80% for training and 20% for testing, and training used the default Classpose parameters with uncertainty weighting, a batch size of 4 and 200 epochs. To compare with CellViT [31], the currently available model trained on PanNuke, separate models were also trained on the three original PanNuke training folds. The average over the three validation folds shows that Classpose outperforms CellViT in panoptic quality (PQ) for all cell types and matches or exceeds it in detection quality (DQ) (Supplementary Table S9). Test performance on the 20% hold-out set closely matches the 3-fold average: the PQ/DQ for neoplastic, inflammatory, connective, dead and epithelial cells is 0.601/0.711, 0.546/0.652, 0.444/0.557, 0.319/0.423 and 0.595/0.738, respectively.

For each detected cell, the 64×64px (at the maximum magnification, typically $20\times$) crops around each detected cell are extracted. These 64px $\times$ 64px crops are then upsampled to 224px $\times$ 224px, and their features are extracted with UNI2-h, producing cell embeddings.

A particularly relevant PFM that was extensively used is *UNI2-h* [25], a Vision Transformer (ViT) designed specifically for large-scale histopathology representation learning. This model takes as input a 224×224px patch image and outputs a 1536×1 embedding.

4.3 Transcriptomic measurements and normalization

Single-Cell Transcriptomics (scT) reveals how individual cells express genes. The data used in the study was acquired using the Xenium In Situ Platform, which is commercialized by 10x Genomics. Contrary to the Visium, also commercialized by 10x Genomics, the Xenium is an imaging-based technology which provides subcellular spatial resolution. This technology enables the detection of mRNA transcripts with their exact 2D position within the tissue, which, when paired with the cell segmentation from Classpose, allows a gene expression to be assigned to each cell. This is done by counting the mRNA transcripts that intersect each cell (the area of the cells is defined by the segmentation mask inferred by Classpose), essentially assigning which mRNA transcripts belong to each cell.

Despite its high reliability, the main drawback of using Xenium samples is their limited gene panel, the samples used from the HEST-1k dataset only contain 300 to 600 genes, depending on the sample, which is fairly limited when compared to the 10,000 to 33,000 genes sequenced by Visium. Furthermore, it is worth mentioning that scT can be converted into ST data for compatibility. The HEST-1k dataset already handles this by creating "pseudo-Visium" spots derived from the sub-cellular mRNA transcripts detected in each $55 \times 55\,\mu$m spot.

Raw count data can vary from gene to gene, from zero up to millions of counts. Such high variance is problematic for deep learning models to learn from. These problems introduce the need for a normalization strategy that is invariant to the library size (total number of mRNA transcripts detected) and that compresses the variance. The strategy used was to apply the $\log(1 + \text{CP10k})$ normalization, which compresses the variance to the range of $0 - 8$ and makes gene expression comparable between different samples.

4.4 Construction of the graphs

The principle used for constructing the graphs is that neighboring patches/cells can share important information and by giving each node the spatial context from its neighboring nodes, the prediction made for that node will be better when compared to not giving spatial information about the neighboring nodes. To this end, an algorithm was designed to build graphs based on the neighborhood and distance between the nodes. The algorithm starts by computing the k-Nearest Neighbors (k-NN) for each node.

Patch-level graphs were built using a dynamic-radius neighborhood with $k = 6$ (radius factor 1.5) for HEST slides and $k = 8$ for TCGA slides, while cell-level graphs used a pruned $k = 6$ neighborhood with a distance multiplier of 1.5.

4.5 Multi-Scale GNN architecture

Supplementary Figure S10 depicts the architecture of the proposed model. The architecture consists of four main components that work together: (1) the cell prediction branch, which includes the *Cell GAT Backbone*, *Cell Head*, *Cell Attention* and *Fusion Gate*; (2) the patch prediction branch, which includes the *Patch GAT Backbone*, *Patch Head* and *Patch Attention* ; (3) the slide prediction branch, that is comprised by the *Slide Head*; and (4) the reconstruction branch which is only composed by the *Reconstruction Head*.

In this model's architecture the information always flows from the smaller scale to the higher scale, information from the cell branch flows to the patch branch and information from the patch branch flows to the slide branch.

The cell branch is responsible for receiving the cells' feature graph and predicting each cell's gene expression. The cell feature graph contains one node for each detected cell, whose initial embedding

was extracted using UNI2-h. The *Cell GAT backbone* is responsible for performing the attention mechanism between graph nodes and essentially modeling the cell-cell interactions. This produces a refined feature graph where each cell is represented by an updated embedding of size 1×1024 enriched by the neighborhood nodes through the attention mechanism. Each cell's refined embedding is provided to the *Cell Head*, which regresses the gene expression vector.

In order to enrich the patch branch with cell-level information, the *Cell Attention* and *Fusion Gate* are employed. The *Cell Attention* layer computes attention scores across the refined cell embeddings, which are normalized using the *Per Patch Softmax* layer to weight the importance of each cell to its parent patch. Consequently, each patch will be assigned an embedding that represents the information gathered by the cell branch. The *Fusion Gate* is used to map this embedding from the cell feature space into the patch feature space, where it is added to the patch's refined embedding in the patch branch.

During training, only 25 patches per WSI are randomly selected to retain the underlying cell graph. Since most patches in the WSI will not have the cell's graph, the embedding that results from the *Per Patch Softmax* layer will be all zeros, to preserve the all zeros embedding the first layer of the fusion gate has bias $= 0$. This results in the cell's information embedding being all zeros and not affecting the patch's refined embedding.

The *Patch GAT Backbone* performs two *GATConv* operations, meaning that the message passing mechanism performs a two neighbor hop attention. A skip connection is added to prevent oversmoothing and preserve the original features of each node. The refined patch embeddings are enriched with the cell's information embedding, after which the *Patch Head* regresses a gene expression vector for each node.

The *Patch Attention* layer is responsible for weighing the importance of each patch to create an embedding that represents the slide embedding. The slide branch is composed just by the *Slide Head*, which takes as input the slide embedding output by the *Patch Attention* layer and regresses the bulk gene expression vector.

The reconstruction branch is composed by the *Reconstruction Head*, this layer is responsible for reconstructing the original patch features from the predicted patch gene expressions. This is basically a parallel objective to generate patch-level feature representations from ST data.

4.6 Multi-Scale aggregation

Mathematically, normalized ($\log(1 + \text{CP10k})$) transcriptomic distributions cannot be linearly aggregated in the normalized space, as most state-of-the-art models suggest [9,10]. Summing or averaging normalized transcriptomics profiles would severely distort the normalization properties. In order to maintain biological consistency across all prediction levels, in the training stage, the proposed hierarchical aggregation algorithm implements the following pipeline: (1) It reverts the predicted multi-level gene expressions out of their logarithmic domains back into the library-size corrected (CP10k) space; (2) It aggregates these library-size corrected expressions across different levels using library-size weighted pooling, effectively going back into the raw count space and summing the gene expressions; and (3) It maps the resulting raw count profiles back into the normalized space.

In cases where the true cells' library sizes are unavailable (TCGA and Visium samples), the model approximates the aggregation by simply averaging the CP10k values of the cell within each patch. This approximation assumes that all cells have the same library size, which is biologically incorrect, but necessary when cell's library sizes are unobserved.

For training samples that do not have patch-level ground-truth (TCGA) and thus the patch library sizes are unknown, the model performs an average of the patch-level CP10k values.

4.7 Loss function and optimization

In order to train the Multi-Scale GNN across bulk (TCGA), spot-level (Visium) and single-cell (Xenium) datasets, a hybrid weakly and fully supervised co-training strategy was developed. To ensure balanced learning across all modalities, the training loop dynamically mixes data from all three datasets.

Because training is conducted on a multi-modal dataset, available ground-truth labels differ across cohorts, resulting in sample-dependent loss formulations. For TCGA samples only the slide-level ground truth is available, so only the *Slide Head* gets direct supervision, on the other hand, the *Patch Head* and *Cell Head* receive weak supervision through the consistency losses. For Visium samples, the ground truth is available at spot-level, and the slide-level ground truth is inferred by the spot-level labels (by summing the raw counts of all genes in all spots), creating the pseudo-bulk gene expression. For these training samples there is direct supervision at slide and patch-level, while also having all the consistency losses. Finally, for Xenium samples, the ground truth is available at sub-cellular level, the cell ground truth is computed by counting the transcripts inside each detected cell, as discussed in Section 4.3, there are also pseudo-spots and pseudo-bulk gene expression ground truths available. For these samples, all the regression heads get direct supervision, along with the weak supervision from the consistency terms.

A hybrid Pearson-MSE loss was used, mixing squared-error minimization with MSE and structural pattern alignment with PCC:

$$\mathcal{L}_{\text{Pearson-MSE}}(P, T) = \delta \cdot \text{MSE}(P, T) + (1 - \delta) \cdot (1 - \text{PCC}(P, T)). \tag{1}$$

In this formulation, P and T represent the predicted and ground-truth gene expression profiles, and $\delta = 0.3$ balances the two components.

For the direct supervision tasks, the loss components were dynamically balanced using an uncertainty weighting scheme proposed in [26]. The total loss function integrates the direct supervision targets along with the cross-level consistency constraints:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \frac{1}{\sigma_{\text{slide}}^2}\mathcal{L}_{\text{slide}} + \frac{1}{\sigma_{\text{patch}}^2}\mathcal{L}_{\text{patch}} + \frac{1}{\sigma_{\text{cell}}^2}\mathcal{L}_{\text{cell}} \\ + \lambda_{\text{CP}}(t)\mathcal{L}_{\text{CP}} + \lambda_{\text{PS}}(t)\mathcal{L}_{\text{PS}} + \omega(t)\mathcal{L}_{\text{recon}} + \log(\sigma_{\text{slide}}\sigma_{\text{patch}}\sigma_{\text{cell}}) \end{aligned} \tag{2}$$

Here, σ_{slide}, σ_{patch} and σ_{cell} are learnable parameters. The learned weights should not be interpreted as the task's uncertainty/variance, but as coarse task-level balancing parameters that improve optimization by balancing supervised tasks with different noise levels and different learning dynamics.

The direct supervision terms use the Pearson-MSE loss. $\mathcal{L}_{\text{CP}}$ ensures consistency between pooled single-cell predictions and their corresponding patches, while $\mathcal{L}_{\text{PS}}$ aligns patch-level predictions with global slide predictions; both also use the Pearson-MSE loss. $\mathcal{L}_{\text{recon}}$ drives the reconstruction of the original patch features using a pure MSE loss. The regularization weights scale linearly from zero up to their target values of 0.01 for both λ_{CP} and λ_{PS}, and 0.1 for ω, over the first 50 epochs.

Optimization used AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with a `ReduceLROnPlateau` scheduler (factor 0.5, patience of 5 epochs, minimum learning rate 1×10^{-6}), mixed-precision training with gradient scaling, and gradient-norm clipping at 1.0. Training proceeded for a maximum of 50 epochs, along with a batch size of 3 WSIs and gradient accumulation of 16 steps, with an effective batch size of 48 WSIs.

4.8 Baseline models and benchmarks

In order to establish a baseline, two other models (DeepSpot2Cell and STFlow) were trained and tested on our dataset. Due to the architecture of these models, which only have supervision at the patch-level, they can only be trained with WSIs which have ground-truth at spot-level (Visium and Xenium). For this reason DeepSpot2Cell and STFlow were only trained on our Visium and Xenium data, TCGA samples were not considered. Nonetheless, for direct comparison, these models were trained on the exact same splits as our model, excluding the TCGA samples. Both baselines used the same UNI2-h embeddings, common gene panel, and random seed (42) as our model.

4.9 Evaluation metrics

The Pearson Correlation Coefficient (PCC) is used throughout this work as the primary metric for assessing how well the predicted expression profiles align with the ground truth, both across genes within

a sample and across samples for a given gene. Since gene expression prediction can also be framed as a discrimination problem, that is, distinguishing spatial regions or cells of high versus low expression, the Area Under the Receiver Operating Characteristic (AUROC) is used to complement the PCC by evaluating this property directly.

This metric was designed for classification tasks, but as suggested in [27] it can be applied to continuous gene expressions as we have in ST data. To leverage AUROC for evaluating the model's ability to capture distinct spatial profiles, the continuous ground-truth expression values are binarized into "high" or "low" expression.

Specifically, in the evaluation pipeline, for each gene across all spots in each WSI, the 25^{th} percentile (p_{25}) and the 75^{th} percentile (p_{75}) of the gene's true spatial expression distribution are computed. For each gene, the spots are then partitioned into high expression (H) and low expression (L) subsets. H contains spots where the spot's gene expression is greater or equal to p_{75}, while L contains spots where the gene expression is lower or equal to p_{25}. Intermediate spots between p_{25} and p_{75} are masked and excluded from the computation.

The AUROC is then calculated using these discrete subsets H and L against the model's continuous predicted expression values for the unmasked spots. Finally, the individual gene AUROC scores are averaged to yield a per WSI spatial discrimination metric. A high AUROC in this context means that the model consistently predicts higher expression values in regions that genuinely correspond to the top quartile of gene expression.

To assess the statistical reliability of these correlations across thousands of genes simultaneously, the False Discovery Rate (FDR) is used to control for false discoveries. In practice, the FDR is implemented using the Benjamini-Hochberg (BH) procedure [28]. First, the PCC between the true and predicted spatial expressions is calculated for each gene on every test slide. A one-sided binomial test is then applied to the signs of these slide-level correlations, producing one p-value per gene. Genes for which the correlation cannot be computed are excluded (happens in cases where the measured or predicted expression is constant). Genes satisfying $p^{\text{BH}} < 0.05$ are classified as having a significantly predicted spatial expression. To allow direct comparison, FDR values are reported as percentages of valid genes per fold.

4.10 Cell-level molecular signature maps

To compare predicted and ground-truth signature distributions, each WSI was partitioned into a grid of 2,000 × 2,000 spatial bins. Within each bin, the gene expression of the signature was computed by summing the predicted expression of each cell within the bin. On the other hand, for the ground truth, the signature expression associated with each bin is computed by summing the number of raw transcripts associated to the signature inside the bin. Moreover, since single-cell data is extremely sparse and noisy, spatial smoothing was applied using a Gaussian filter. The comparison between the proposed model and DeepSpot2Cell was assessed with a paired Wilcoxon signed-rank test across slides.

5 Conclusion

This work presented a Multi-Scale GNN that predicts bulk, spatial and single-cell transcriptomics directly from WSIs. By combining TCGA, Visium and Xenium samples in a hybrid training strategy, with direct supervision where gene expression measurements are available and weak supervision through consistency losses, a single model learns from datasets with different levels of granularity while keeping its predictions aligned across scales.

The model performs best at the slide level and moderately at the patch level, while single-cell prediction remains the most challenging task; at this level, it achieved higher correlation and AUROC than DeepSpot2Cell. Performance remains uneven across genes and tissues, and single-cell predictions could only be directly evaluated for the genes measured by Xenium. Nevertheless, the downstream tasks showed that the predicted profiles retain tissue-specific and tumor-related information, supporting survival analysis, tumor localization and classification, and cell-level molecular signature maps.

Although spatial and single-cell transcriptomics do not yet have an established clinical application, this work helps advance computational tools that make them accessible from routine histopathology images, which may prove valuable as their practical medical utility develops.

References

[1] A. Rao, D. Barkley, G. S. França, and I. Yanai, "Exploring tissue architecture using spatial transcriptomics," Nature, vol. 596, no. 7871, pp. 211–220, Aug. 2021, doi:10.1038/s41586-021- 03634-9. [Online]. Available: https://doi.org/10.1038/s41586-021-03634-9.

[2] I. Damjanov, Histopathology: A Color Atlas and Textbook. Williams & Wilkins, 1996. [Online]. Available: https://books.google.pt/books?id=IgAW0AEACAAJ.

[3] X. Li and C.-Y. Wang, "From bulk, single-cell to spatial rna sequencing," International Journal of Oral Science, vol. 13, no. 1, p. 36, 2021. [Online]. Available: https://www.nature.com/articles/s41368-021-00146-0.

[4] E. Chelebian, C. Avenel, and C. Wählby, "Combining spatial transcriptomics with tissue morphology," Nature Communications, vol. 16, no. 1, p. 4452, 2025. [Online]. Available: https://doi.org/10.1038/s41467-025-58989-8.

[5] L. A. D. Cooper, E. G. Demicco, J. H. Saltz, R. T. Powell, A. Rao, and A. J. Lazar, "Pancancer insights from the cancer genome atlas: the pathologist's perspective," The Journal of Pathology, vol. 244, no. 5, pp. 512–524, 2018.

[6] T. Huang, T. Liu, M. Babadi, W. Jin, and R. Ying, "Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching," 2025. [Online]. Available: https://arxiv.org/abs/2506.05361.

[7] J. Zou, K. Xiao, Z. Chen, J. Pei, J. Xu, T. Chen, L. Hou, C. Wu, Y. She, Z. Yuan, and L. Chen, "Predicting spatial transcriptomics from h&e image by pretrained contrastive alignment learning," bioRxiv, 2025, preprint. [Online]. Available: https://www.biorxiv.org/content/10.1101/2025.06.15.659438v1.

[8] F. Dashtiahangar and J.-S. Yuan, "Bridging morphology and molecular signatures: Multi-task deep learning for multi-omics prediction from histopathology," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025, pp. 1–9.

[9] Y. Şenbabaoğlu, V. Prabhakar, A. Khormali, J. Eastham, E. Liu, E. Warner, B. Nabet, M. Srivastava, M. Ballinger, and K. Liu, "Mosby enables multi-omic inference and spatial biomarker discovery from whole slide images," Scientific Reports, vol. 14, p. 18271, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-69198-6.

[10] K. Nonchev, G. Manaiev, V. H. Koelzer, and G. Rätsch, "DeepSpot2Cell: Predicting virtual single-cell spatial transcriptomics from H&E images using spot-level supervision," bioRxiv preprint, 2025, doi: 10.1101/2025.09.23.678121.

[11] L. Chadoutaud, M. Lerousseau, D. Herrero-Saboya, J. Ostermaier, J. Fontugne, E. Barillot, and T. Walter, "sCellST predicts single-cell gene expression from H&E images," Nature Communications, vol. 17, no. 1, p. 1194, Jan. 2026, doi: 10.1038/s41467-025-67965-1.

[12] J. Kömen, H. Marienwald, J. Dippel, and J. Hense, "Do histopathological foundation models eliminate batch effects? a comparative study," 2024. [Online]. Available: https://arxiv.org/abs/2411.05489.

[13] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," Nature Methods, vol. 16, no. 12, pp. 1289–1296, 2019. [Online]. Available: https://doi.org/10.1038/s41592-019-0619-0.

[14] J. Liu, T. Lichtenberg, K. A. Hoadley et al., "An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics," Cell, vol. 173, no. 2, pp. 400–416.e11, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867418302290.

[15] D. R. Cox, "Regression models and life-tables," Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 2, pp. 187–202, 1972. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x.

[16] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. Mesirov, and P. Tamayo, "The molecular signatures database hallmark gene set collection," Cell Systems, vol. 1, no. 6, pp. 417–425, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405471215002185.

[17] C. Loeffler and J. N. Kather, "Manual tumor annotations in tcga," Aug. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5320076.

[18] G. Jaume, P. Doucet, A. H. Song, M. Y. Lu, C. Almagro-Pérez, S. J. Wagner, A. J. Vaidya, R. J. Chen, D. F. K. Williamson, A. Kim, and F. Mahmood, "Hest-1k: A dataset for spatial transcriptomics and histology image analysis," 2024. [Online]. Available: https://arxiv.org/abs/2406.16192.

[19] T. C. G. A. R. Network, "Comprehensive molecular profiling of lung adenocarcinoma," Nature, vol. 511, no. 7511, pp. 543–550, 2014. [Online]. Available: https://www.nature.com/articles/nature13385.

[20] T. C. G. A. R. Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," Nature, vol. 499, no. 7456, pp. 43–49, 2013. [Online]. Available: https://www.nature.com/articles/nature12222.

[21] D. C. Koboldt, R. S. Fulton, M. D. McLellan et al., "Comprehensive molecular portraits of human breast tumours," Nature, vol. 490, no. 7418, pp. 61–70, oct 2012. [Online]. Available: https://doi.org/10.1038/nature11412.

[22] A. Zhang, G. Jaume, A. Vaidya, T. Ding, and F. Mahmood, "Accelerating data processing and benchmarking of ai models for pathology," arXiv preprint arXiv:2502.06750, 2025.

[23] S. Mandal, J. G. de Almeida, K. Bräutigam, N. Papanikolaou, and T. A. Graham, "Classpose drives the discovery of colorectal cancer phenotypes in clinical grade whole slide images," bioRxiv, 2026. [Online]. Available: https://www.biorxiv.org/content/early/2026/07/10/2025.12.18.695211.

[24] M. Pachitariu, M. Rariden, and C. Stringer, "Cellpose-sam: superhuman generalization for cellular segmentation," bioRxiv, 2025. [Online]. Available: https://www.biorxiv.org/content/early/2025/05/01/2025.04.28.651001.

[25] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood, "Towards a general-purpose foundation model for computational pathology," Nature Medicine, 2024.

[26] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," CoRR, vol. abs/1705.07115, 2017. [Online]. Available: http://arxiv.org/abs/1705.07115.

[27] B. He, L. Bergenstråhle, L. Stenbeck, A. Abid, A. Andersson, Å. Borg, J. Maaskola, J. Lundeberg, and J. Zou, "Integrating spatial gene expression and breast tumour morphology via deep learning," Nature Biomedical Engineering, vol. 4, no. 8, pp. 827–834, Aug 2020. [Online]. Available: https://doi.org/10.1038/s41551-020-0578-x.

[28] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," Journal of the Royal Statistical Society: Series B (Methodological), vol. 57, no. 1, pp. 289–300, 01 1995. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[29] J. Gamper, N. Alemi Koohbanani, K. Benes, A. Khuram, and N. Rajpoot, "PanNuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification," in *Proc. Eur. Congr. Digit. Pathol. (ECDP)*, ser. Lecture Notes in Computer Science, vol. 11435. Springer, 2019, pp. 11–19, doi: 10.1007/978-3-030-23937-4_2.

[30] J. Gamper *et al.*, "PanNuke dataset extension, insights and baselines," *arXiv preprint arXiv:2003.10778*, 2020.

[31] F. Hörst *et al.*, "CellViT: Vision transformers for precise cell segmentation and classification," *Med. Image Anal.*, vol. 94, p. 103143, 2024, doi: 10.1016/j.media.2024.103143.

Supplementary material

TCGA spatial predictions and tumor classification

In Figure S8, the model produces spatially coherent expression patterns that align with tissue morphology visible in the corresponding H&E image, with the finer-grained cell-level predictions recovering the same regional expression trends identified at the spot level.

In the single-slide example in Figure S9, predictions based on the 50 molecular signatures achieve higher discrimination (AUROC = 0.989) than the original patch embeddings (AUROC = 0.811).

Visual features correlate with gene expression

The clustering analysis is described in Section 2.2. The 100 principal components of the patch features explain 69.82% of the variance, while the 300 principal components of the gene expression vectors explain 25.76% of the variance.

The cluster purity measures, among patches in the same patch-feature cluster, the percentage that map to the same gene expression cluster. ARI measures pair-wise consistency, adjusted for agreements that would happen purely by chance. A score of 0 indicates random clustering, while 1 indicates identical clustering. Finally, NMI measures information overlap between the two cluster assignments, ranging from 0 (no relationship) to 1 (perfect predictability).

Cell type encodes gene expression

In the t-SNE of the Harmony-corrected patches' gene expression of Visium slides, each point is colored by the percentage of cell types present in the corresponding patch (Figure S2). Connective cells, the most common cell type, appear frequently in most clusters.

The same analysis can be done at the cell level on the Xenium slides available in the HEST-1k dataset. To do this, a gene expression is assigned to each cell, as described in Section 4.3. Out of the available cells, 250,000 were selected randomly for the analysis. In Figure S3, it is possible to see the t-SNEs of the Harmony-corrected cells' gene expression, highlighting each cell type. There is not a clear cluster for each cell type: for connective cells, for example, it appears that there is not a specific gene expression profile. On the other hand, for neoplastic and inflammatory cells, there is clearly a certain gene expression profile associated with them.

Multi-Scale GNN architecture

Figure S10 depicts the architecture described in Section 4.5. Each WSI can have up to millions of detected cells; even with a small batch size of WSIs, it would be unfeasible to hold every cell's embedding in GPU memory. Considering that each of the 25 patches selected during training might have 5–50 cells, each training sample will have 125–1,250 cells. This makes each WSI have a considerably smaller memory footprint and enables much faster training.

Multi-Scale aggregation

The aggregation strategy and the approximations used when library sizes are unavailable are described in Section 4.6. The following equations detail the cell-to-patch and patch-to-slide aggregation.

Cell-level to patch-level aggregation. Let $\mathbf{Z}_{\text{cell}} \in \mathbb{R}^{N_{cells} \times N_{genes}}$ represent the predicted cell gene expression matrix generated by the cell-level head, where N_{cells} is the total number of detected cells and N_{genes} is the number of target genes. Every individual cell index $i \in \{1, \ldots, N_{cells}\}$ belongs to a parent patch index $j \in \{1, \ldots, N_{patches}\}$.

In order to aggregate the cell predictions into their corresponding patch, the model first reverts the normalized prediction ($\mathbf{z}_{\text{cell},i}$) to raw counts ($\text{Raw}_{\text{cell},i}$). To do this, the model first reverts the logarithm

to obtain the CP10k value, and then un-scales it using the cell's true library size ($L_{\text{cell},i}$):

$$\text{Raw}_{\text{cell},i} = (\exp(\mathbf{z}_{\text{cell},i}) - 1) \times \frac{L_{\text{cell},i}}{10^4} \tag{3}$$

Subsequently, the raw counts of all cells mapped to patch j are summed to compute the total raw gene expression of the patch ($\text{Raw}_{\text{patch},j}$). The aggregated library size for the patch is the sum of its underlying cells' library sizes ($L_{\text{patch},j} = \sum L_{\text{cell},i}$).

Finally, the original normalization is re-applied to the raw expression to yield the predicted cell-aggregated patch prediction:

$$\mathbf{z}_{\text{patch_from_cells},j} = \log\left(1 + \frac{\text{Raw}_{\text{patch},j}}{L_{\text{patch},j}} \times 10^4\right) \tag{4}$$

Patch-level to slide-level aggregation. The predicted patch level gene expressions ($\mathbf{z}_{\text{patch},j}$) are converted to raw counts using their respective spot library sizes ($L_{\text{patch},j}$):

$$\text{Raw}_{\text{patch},j} = (\exp(\mathbf{z}_{\text{patch},j}) - 1) \times \frac{L_{\text{patch},j}}{10^4} \tag{5}$$

The raw counts from all patches are summed to approximate the bulk gene expression at slide-level ($\text{Raw}_{\text{slide}}$). The original normalization is re-applied using the slide library size to produce the final patch derived bulk gene expression:

$$\mathbf{z}_{\text{slide_from_patches}} = \log\left(1 + \frac{\text{Raw}_{\text{slide}}}{L_{\text{slide}}} \times 10^4\right) \tag{6}$$

Loss weighting

The loss function and optimization settings are described in Section 4.7. The $\log(\sigma_{\text{slide}}\sigma_{\text{patch}}\sigma_{\text{cell}})$ term in Equation 2 is a regularization term to prevent the model from driving the weights to zero in order to minimize the loss.

The reason why the consistency and reconstruction terms were not weighted using the learned weights was because, as experimentally observed, the model becomes over-confident on these tasks, assigning them higher weights and neglecting the direct supervision losses. This happens because these tasks are "easier" and the model just maintains the consistency between the predictions instead of trying to learn the directly supervised tasks. For this reason, these weights were kept small and deterministic.

Baseline models and benchmarks

The benchmark design is described in Section 4.8. Both baseline models were trained using their respective published architectures and default hyperparameters. DeepSpot2Cell used an ensemble of 8 Phi/Rho sub-networks (hidden size 512, dropout 0.3), each aggregating information from a spot's $k = 8$ nearest spatial neighbors. It was trained with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-6}) for up to 50 epochs, batch size 256 spots, with early stopping after 5 epochs without improvement. STFlow was configured as a 4-layer spatial transformer denoiser (hidden dimension 128, edge dimension 128, 4 attention heads, dropout 0.2) trained under a conditional-flow-matching objective with a zero-inflated negative-binomial prior, using AdamW (learning rate 5×10^{-4}, weight decay 1×10^{-6}) for up to 100 epochs with early stopping (patience of 20 epochs), gradient-norm clipping at 1.0, a batch size of 1 slide.

Ablation studies

The four ablated variants are described in Section 2.11. Removing feature reconstruction tests whether this self-supervised signal improves the quality of the learned representations used for gene expression prediction. Removing the patch to slide and cell to patch consistency terms tests whether this constraint is helpful for the model. Removing TCGA slides from training tests whether the additional scale

and tissue diversity provided by this large, spatially-unresolved dataset benefits performance on the spatially-resolved Visium and Xenium benchmarks. Removing TCGA cell level weak supervision tests whether the extra information at cell-level from TCGA is beneficial to the model.

Table S1: Proposed model's metrics for the Visium Dataset

Cohort	**Total** N	**Bulk PCC**	**Patch PCC**	**Feat. Recon. Cos. Sim.**	**PS Cons. PCC**	**CP Cons. PCC**	**Patch AUROC**	**Signif. Genes (Patch) (FDR < 0.05)**
Overall	105	0.774 ± 0.146	0.509 ± 0.026	0.634 ± 0.009	0.777 ± 0.142	0.896 ± 0.016	0.554 ± 0.004	$37.7\% \pm 5.9\%$
Breast	7	0.752 ± 0.080	0.583 ± 0.053	0.481 ± 0.068	0.816 ± 0.082	0.892 ± 0.028	0.567 ± 0.014	–
Kidney	62	0.742 ± 0.216	0.544 ± 0.024	0.672 ± 0.040	0.753 ± 0.216	0.907 ± 0.014	0.536 ± 0.005	$14.0\% \pm 9.7\%$
Lung	36	0.844 ± 0.071	0.430 ± 0.035	0.609 ± 0.020	0.819 ± 0.042	0.874 ± 0.031	0.586 ± 0.018	$13.8\% \pm 13.8\%$

Table S2: Proposed model's metrics for the Xenium Dataset

Cohort / Subtype	**Total** N	**Bulk PCC**	**Patch PCC**	**Cell PCC**	**Feat. Recon. Cos. Sim.**	**PS Cons. PCC**	**CP Cons. PCC**	**Patch AUROC**	**Signif. Genes (Patch) (FDR < 0.05)**	**Cell AUROC**
Overall	47	0.797 ± 0.057	0.786 ± 0.041	0.396 ± 0.021	0.566 ± 0.044	0.858 ± 0.044	0.779 ± 0.029	0.703 ± 0.011	$8.0\% \pm 14.0\%$	0.649 ± 0.011
Connective	-	-	-	0.367 ± 0.019	-	-	-	-	-	0.616 ± 0.003
Dead	-	-	-	0.330 ± 0.019	-	-	-	-	-	0.613 ± 0.018
Epithelial	-	-	-	0.406 ± 0.023	-	-	-	-	-	0.626 ± 0.022
Inflammatory	-	-	-	0.340 ± 0.017	-	-	-	-	-	0.596 ± 0.012
Neoplastic	-	-	-	0.456 ± 0.036	-	-	-	-	-	0.614 ± 0.016
Breast	21	0.785 ± 0.040	0.775 ± 0.019	0.399 ± 0.026	0.503 ± 0.024	0.868 ± 0.018	0.781 ± 0.010	0.736 ± 0.010	–	0.678 ± 0.027
Connective	-	-	-	0.332 ± 0.031	-	-	-	-	-	0.622 ± 0.009
Dead	-	-	-	0.275 ± 0.063	-	-	-	-	-	0.643 ± 0.027
Epithelial	-	-	-	0.402 ± 0.032	-	-	-	-	-	0.626 ± 0.029
Inflammatory	-	-	-	0.308 ± 0.042	-	-	-	-	-	0.607 ± 0.033
Neoplastic	-	-	-	0.475 ± 0.023	-	-	-	-	-	0.592 ± 0.018
Kidney	2	0.572 ± 0.000	0.517 ± 0.000	0.169 ± 0.000	0.504 ± 0.000	0.778 ± 0.000	0.602 ± 0.000	0.567 ± 0.000	–	0.578 ± 0.000
Connective	-	-	-	0.161 ± 0.000	-	-	-	-	-	0.572 ± 0.000
Epithelial	-	-	-	0.164 ± 0.000	-	-	-	-	-	0.440 ± 0.000
Inflammatory	-	-	-	0.168 ± 0.000	-	-	-	-	-	0.537 ± 0.000
Neoplastic	-	-	-	0.156 ± 0.000	-	-	-	-	-	0.513 ± 0.000
Lung	24	0.828 ± 0.057	0.823 ± 0.042	0.416 ± 0.024	0.622 ± 0.057	0.851 ± 0.062	0.796 ± 0.028	0.685 ± 0.015	$15.3\% \pm 34.1\%$	0.629 ± 0.020
Connective	-	-	-	0.417 ± 0.027	-	-	-	-	-	0.617 ± 0.006
Dead	-	-	-	0.368 ± 0.030	-	-	-	-	-	0.600 ± 0.018
Epithelial	-	-	-	0.433 ± 0.023	-	-	-	-	-	0.623 ± 0.028
Inflammatory	-	-	-	0.380 ± 0.027	-	-	-	-	-	0.595 ± 0.014
Neoplastic	-	-	-	0.464 ± 0.045	-	-	-	-	-	0.639 ± 0.013

Table S3: Proposed model's metrics for the TCGA Dataset

Cohort	**Total** N	**Bulk PCC**	**Feat. Recon. Cos. Sim.**	**PS Cons. PCC**	**Bulk AUROC**	**Signif. Genes (Bulk) (FDR < 0.05)**
Overall	1,724	0.881 ± 0.010	0.637 ± 0.014	0.893 ± 0.015	0.661 ± 0.003	$72.6\% \pm 16.0\%$
Breast	866	0.875 ± 0.012	0.615 ± 0.015	0.891 ± 0.018	0.608 ± 0.007	$53.1\% \pm 12.3\%$
Kidney	423	0.886 ± 0.025	0.682 ± 0.014	0.881 ± 0.030	0.586 ± 0.015	$31.0\% \pm 20.0\%$
Lung	435	0.888 ± 0.004	0.637 ± 0.016	0.907 ± 0.008	0.602 ± 0.005	$42.2\% \pm 4.8\%$

Table S4: MLP's classification metrics across the true and predicted spot's gene expression.

Test Set	**Average Accuracy**	**Average F1-score**	**Average AUROC**
Visium true spot's gene expression	0.980 ± 0.025	0.975 ± 0.028	0.994 ± 0.009
Visium predicted spot's gene expression	0.859 ± 0.068	0.804 ± 0.101	0.942 ± 0.056
TCGA predicted spot's gene expression	0.815 ± 0.029	0.774 ± 0.038	0.910 ± 0.022

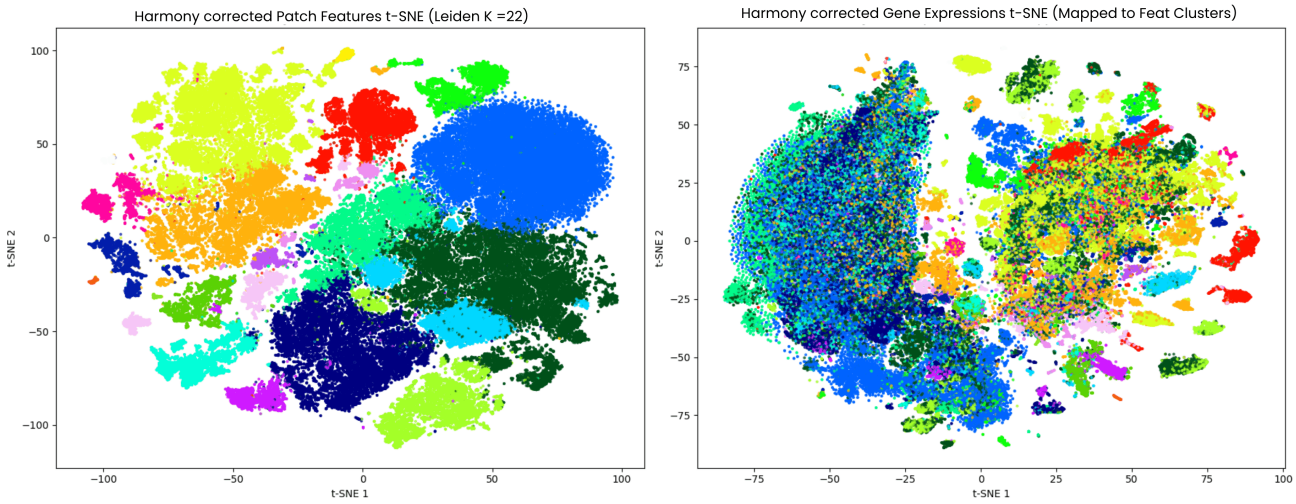


Figure S1: Left: Harmony corrected t-SNE of clustered patch features. Right: Harmony corrected t-SNE of gene expression profiles, colored according to the cluster assignment of the corresponding patch features (across 150,000 randomly selected patches).

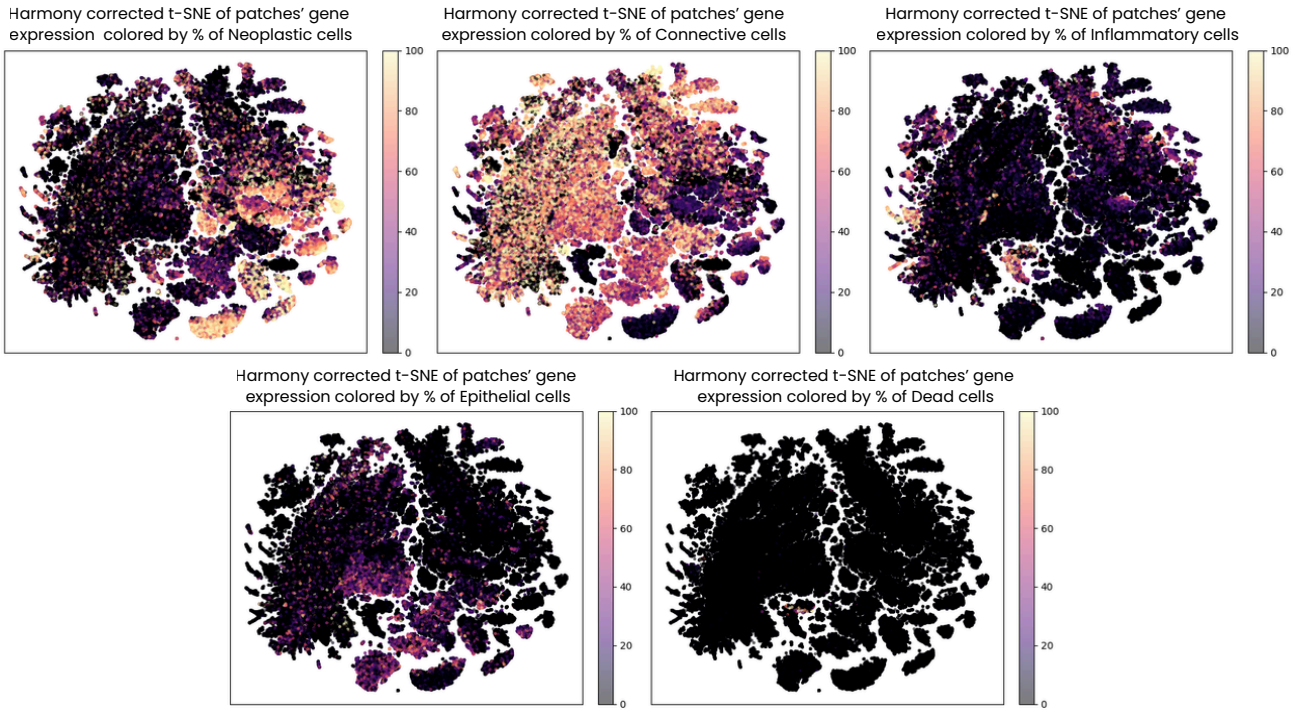


Figure S2: Harmony corrected t-SNEs of patches' gene expression profiles colored by percentage of each cell type (neoplastic, connective, inflammatory, epithelial, dead) (across 250,000 randomly selected patches).

Table S5: MLP's classification metrics across the original and reconstructed spot's embedding.

Test Set	**Accuracy**	**F1-score**	**AUROC**
Original spot's embedding	0.924 ± 0.007	0.915 ± 0.008	0.984 ± 0.005
Reconstructed spot's embedding	0.902 ± 0.009	0.893 ± 0.011	0.981 ± 0.005

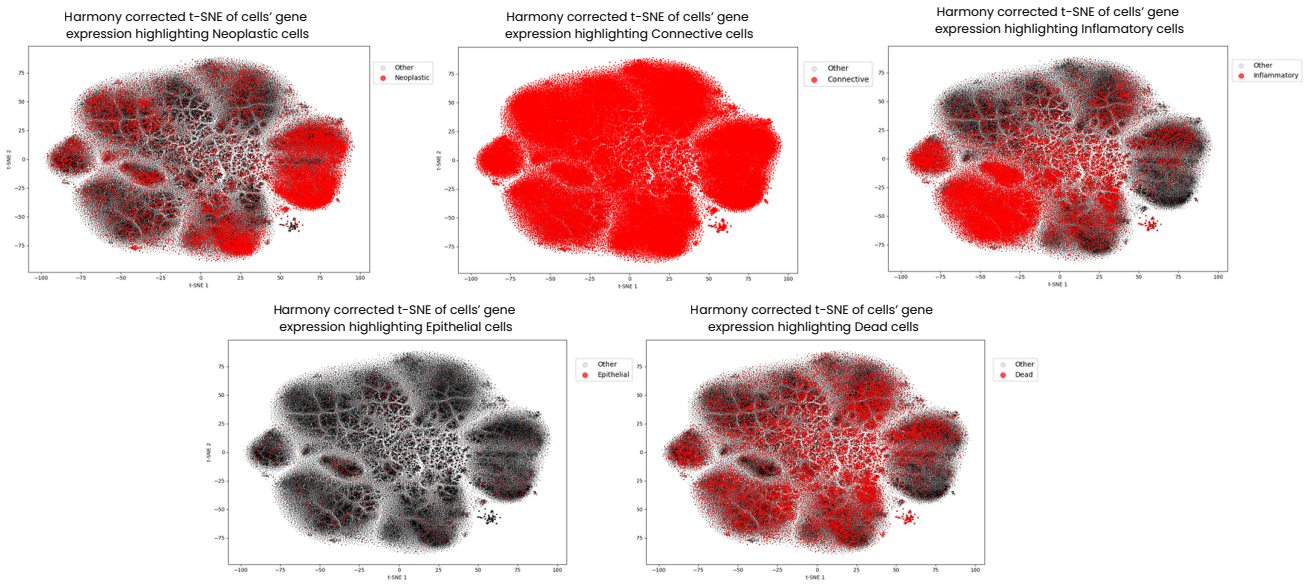


Figure S3: Harmony corrected t-SNEs of cells' gene expression profiles highlighting cell type (neoplastic, connective, inflammatory, epithelial, dead) (across 250,000 randomly selected cells).

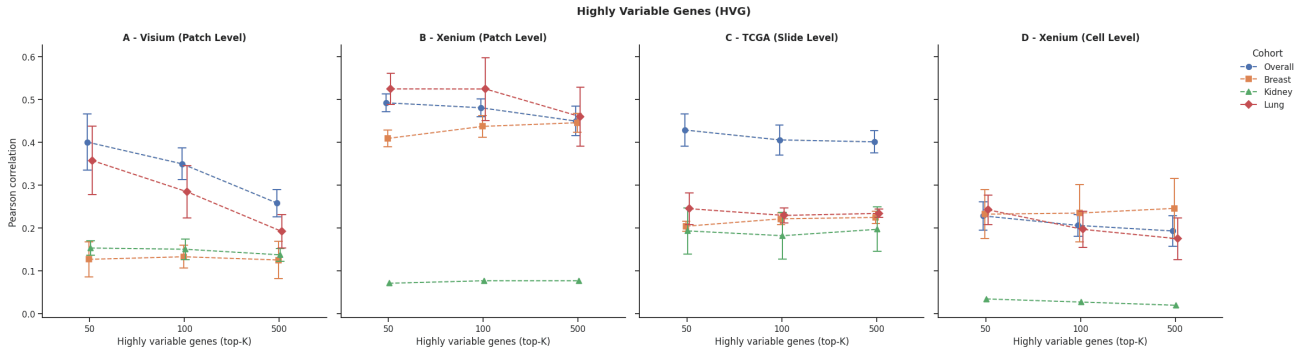


Figure S4: Gene-wise mean PCC in each dataset against the number of top Highly Variable Genes (HVGs) for the proposed model.

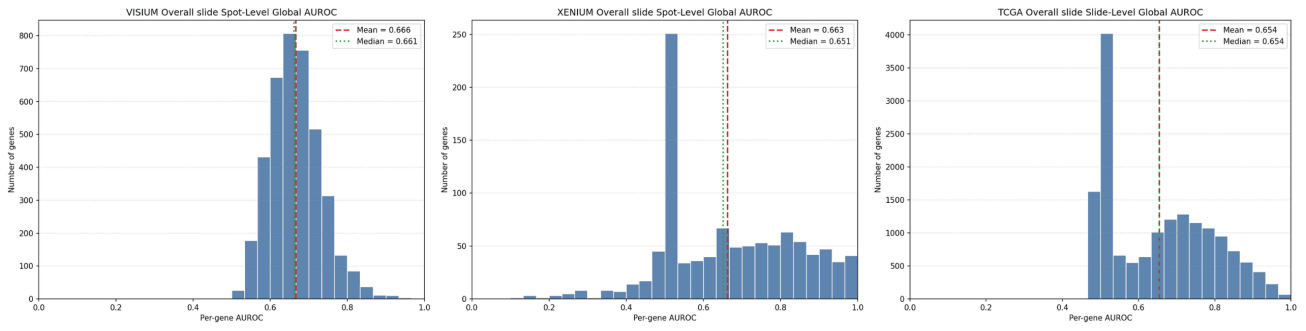
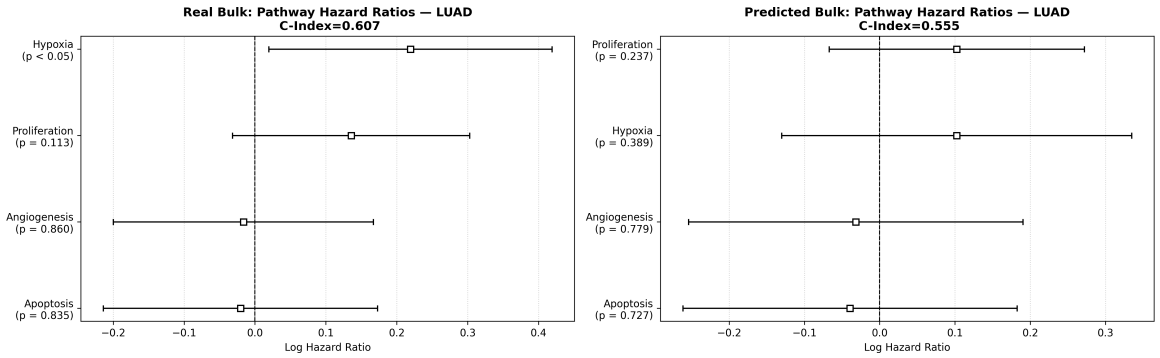
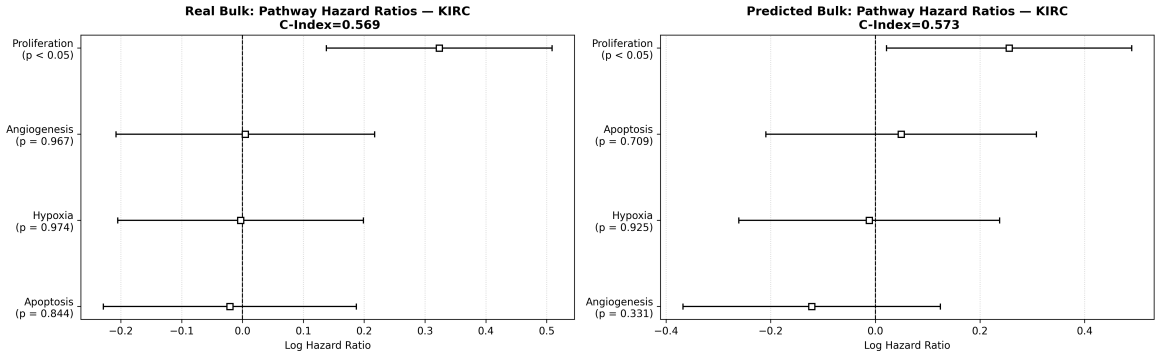


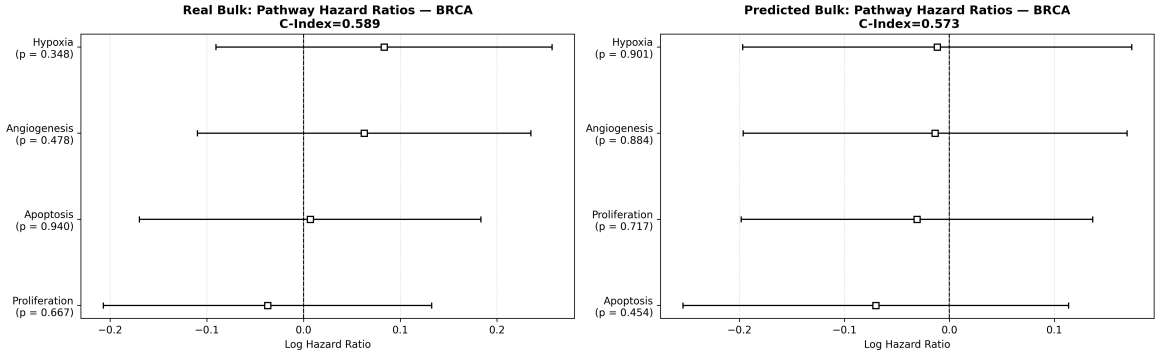
Figure S5: Histograms of gene-wise AUROC in each dataset for the proposed model.



(a) Log hazard ratios for each molecular signature to predict DSS time in TCGA-LUAD.



(b) Log hazard ratios for each molecular signature to predict DSS time in TCGA-KIRC.



(c) Log hazard ratios for each molecular signature to predict DSS time in TCGA-BRCA.

Figure S6: Log hazard ratios for each molecular signature to predict DSS time in TCGA cohorts.

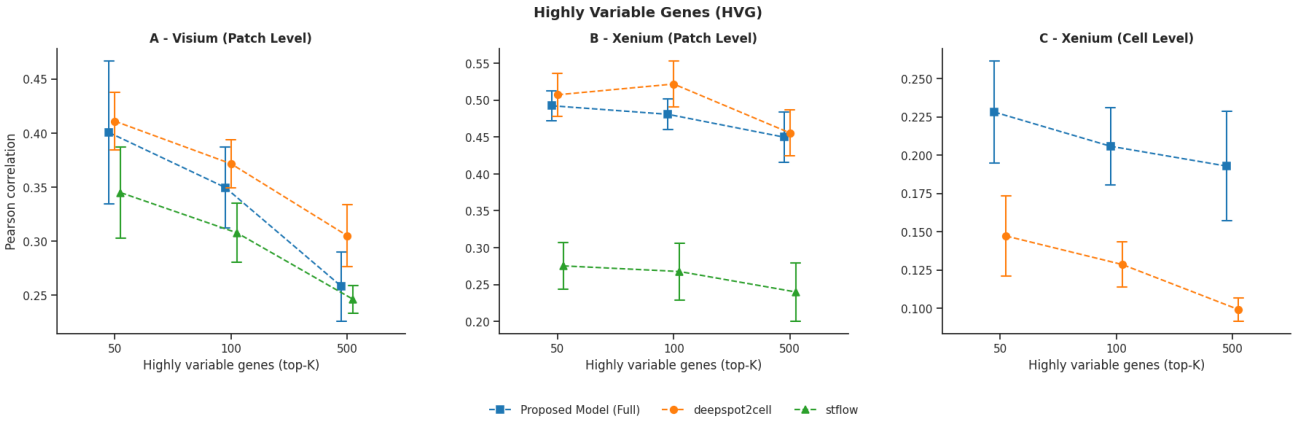


Figure S7: Gene-wise mean PCC in each model against the number of top HVGs.

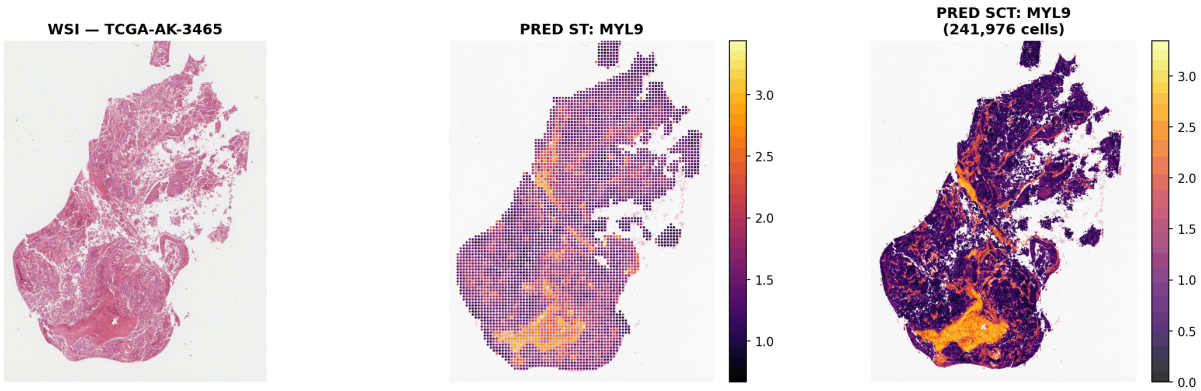


Figure S8: ST and scT predictions for TCGA slide AK-3465, gene *MYL9*. Left to right: WSI, predicted spot expression, predicted cell expression.

Table S6: Metrics for the Visium Dataset (Overall) across different models.

Model	**Total** N	**Bulk PCC**	**Patch PCC**	**Feat. Recon. Cos. Sim.**	**PS Cons. PCC**	**CP Cons. PCC**	**Patch AUROC**	**Signif. Genes (Patch) (FDR < 0.05)**
Proposed Model (Full)	105	0.774 ± 0.146	0.509 ± 0.026	0.634 ± 0.009	0.777 ± 0.142	0.896 ± 0.016	0.554 ± 0.004	$37.7\% \pm 5.9\%$
Proposed Model (w/o Consistency)	105	0.726 ± 0.178	0.512 ± 0.027	0.606 ± 0.036	0.740 ± 0.183	0.222 ± 0.007	0.555 ± 0.004	$43.5\% \pm 9.3\%$
Proposed Model (w/o Feat. Recon.)	105	0.744 ± 0.159	0.501 ± 0.030	–	0.757 ± 0.168	0.881 ± 0.029	0.548 ± 0.002	$34.6\% \pm 12.5\%$
Proposed Model (w/o TCGA)	105	0.875 ± 0.017	0.510 ± 0.025	0.522 ± 0.033	0.924 ± 0.019	0.867 ± 0.030	0.549 ± 0.011	$36.4\% \pm 16.0\%$
Proposed Model (w/o TCGA cells)	105	0.794 ± 0.085	0.506 ± 0.028	0.559 ± 0.063	0.814 ± 0.090	0.850 ± 0.027	0.548 ± 0.005	$37.3\% \pm 4.7\%$
DeepSpot2Cell	105	-	0.535 ± 0.028	–	-	–	0.590 ± 0.004	$55.1\% \pm 4.4\%$
STFlow	105	-	0.519 ± 0.032	–	-	–	0.575 ± 0.009	$48.7\% \pm 6.6\%$

Table S7: Metrics for the Xenium Dataset (Overall) across different models.

Model	**Total** N	**Bulk PCC**	**Patch PCC**	**Cell PCC**	**Feat. Recon. Cos. Sim.**	**PS Cons. PCC**	**CP Cons. PCC**	**Patch AUROC**	**Signif. Genes (Patch)**	**Cell AUROC**
Proposed Model (Full)	47	0.797 ± 0.057	0.786 ± 0.041	0.396 ± 0.021	0.566 ± 0.044	0.858 ± 0.044	0.779 ± 0.029	0.703 ± 0.011	$8.0\% \pm 14.0\%$	0.649 ± 0.011
Proposed Model (w/o Consistency)	47	0.779 ± 0.022	0.770 ± 0.054	0.390 ± 0.029	0.564 ± 0.078	0.837 ± 0.032	0.734 ± 0.032	0.691 ± 0.027	$8.7\% \pm 15.5\%$	0.638 ± 0.014
Proposed Model (w/o Feat. Recon.)	47	0.810 ± 0.045	0.777 ± 0.049	0.384 ± 0.032	–	0.863 ± 0.044	0.741 ± 0.056	0.686 ± 0.015	$8.0\% \pm 15.1\%$	0.637 ± 0.013
Proposed Model (w/o TCGA)	47	0.831 ± 0.036	0.771 ± 0.045	0.398 ± 0.028	0.453 ± 0.056	0.924 ± 0.009	0.780 ± 0.042	0.669 ± 0.031	$7.8\% \pm 15.1\%$	0.648 ± 0.015
Proposed Model (w/o TCGA cells)	47	0.764 ± 0.079	0.759 ± 0.041	0.391 ± 0.020	0.519 ± 0.096	0.817 ± 0.115	0.739 ± 0.035	0.674 ± 0.022	$7.3\% \pm 13.0\%$	0.640 ± 0.009
DeepSpot2Cell	47	-	0.760 ± 0.055	0.349 ± 0.034	–	0.984 ± 0.011	0.726 ± 0.028	0.764 ± 0.010	$8.8\% \pm 16.1\%$	0.523 ± 0.013
STFlow	47	-	0.679 ± 0.088	–	–	1.000 ± 0.000	–	0.634 ± 0.019	$4.5\% \pm 10.1\%$	–

Table S8: Metrics for the TCGA Dataset (Overall) across different models.

Model	**Total** N	**Bulk PCC**	**Feat. Recon. Cos. Sim.**	**PS Cons. PCC**	**Bulk AUROC**	**Signif. Genes (Bulk) (FDR < 0.05)**
Proposed Model (Full)	1,724	0.881 ± 0.010	0.637 ± 0.014	0.893 ± 0.015	0.661 ± 0.003	$72.6\% \pm 16.0\%$
Proposed Model (w/o Consistency)	1,724	0.883 ± 0.014	0.586 ± 0.047	0.691 ± 0.034	0.651 ± 0.007	$69.6\% \pm 14.9\%$
Proposed Model (w/o Feat. Recon.)	1,724	0.888 ± 0.010	–	0.881 ± 0.007	0.649 ± 0.008	$83.4\% \pm 15.0\%$
Proposed Model (w/o TCGA)	–	–	–	–	–	–
Proposed Model (w/o TCGA cells)	1,724	0.882 ± 0.017	0.541 ± 0.061	0.865 ± 0.026	0.647 ± 0.006	$79.3\% \pm 12.6\%$
DeepSpot2Cell	–	–	–	–	–	–
STFlow	–	–	–	–	–	–

Table S9: Average 3-fold performance (panoptic and detection quality; PQ and DQ, respectively) comparison between Classpose and CellViT models trained on PanNuke*. The best performing metrics are in bold.

	CellViT	**Classpose**
Cell type	PQ / DQ	PQ / DQ
Neoplastic	0.581 / 0.71	0.600 / 0.71
Inflammatory	0.417 / 0.59	0.544 / 0.65
Connective	0.423 / 0.53	0.441 / 0.55
Dead	0.149 / 0.39	0.325 / 0.428
Epithelial	0.583 / 0.73	0.586 / 0.73

*Available at https://huggingface.co/datasets/RationAI/PanNuke, accessed on May 18, 2026.

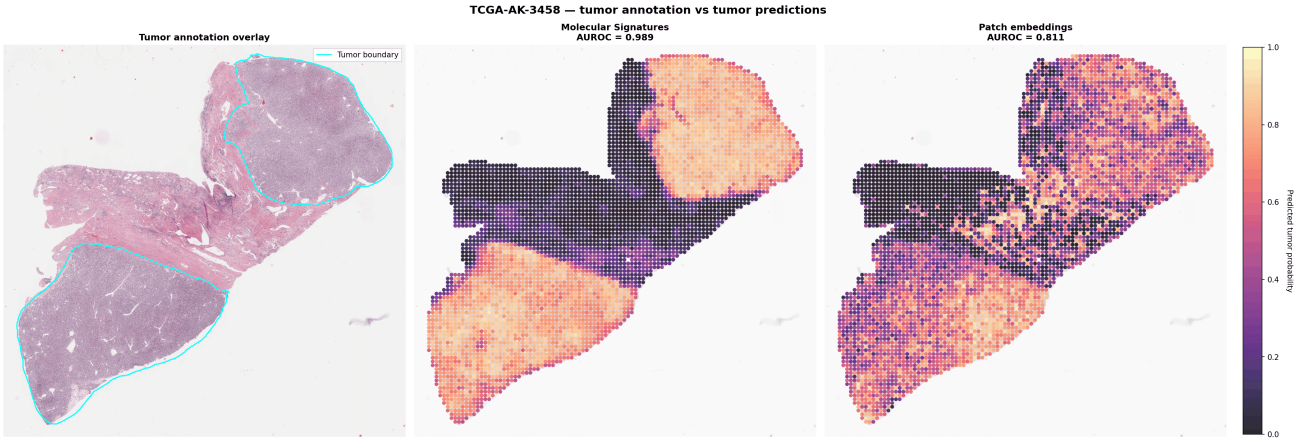


Figure S9: Comparison between the tumor annotations and the predicted tumor probability by each feature set in TCGA slide AK-3458.

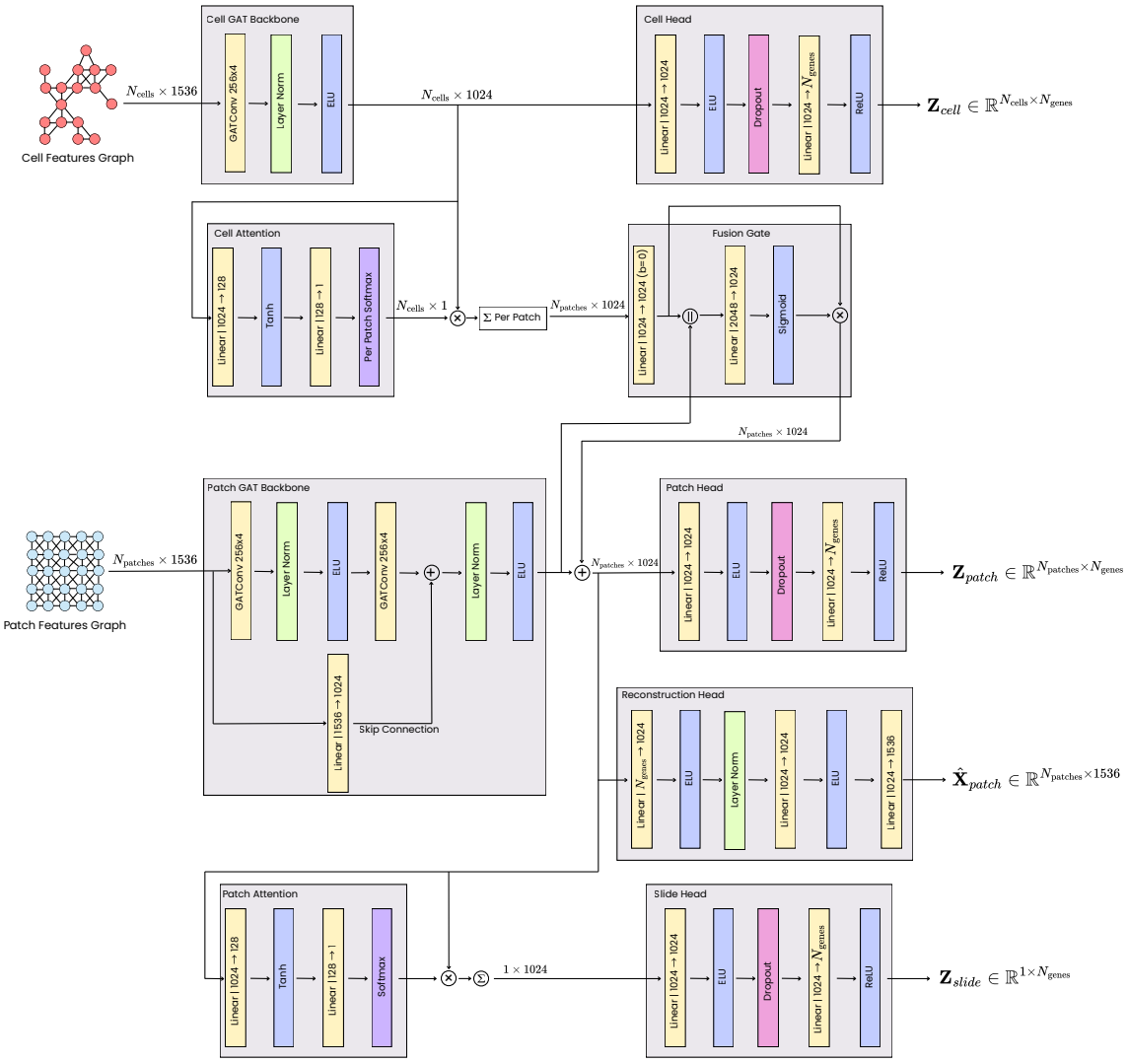


Figure S10: Architecture of the Multi-Scale Graph Neural Network (GNN).