This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material for: TumorGraphs: Patient-Level Fusion of Multi-WSI Histopathology

1 Overview

This supplementary material provides additional dataset and endpoint details, implementation and model-selection information, baseline descriptions, scalability analysis, and extended experimental results omitted from the main paper due to space constraints.

2 Dataset and Endpoint Details

2.1 Specimen Type

The datasets used in this study contain primary tumour WSIs; secondary or metastatic tumour specimens are not included. Thus, cross-slide relations are constructed exclusively among primary tumour WSIs from the same patient.

2.2 Single-WSI Patients

Patients represented by a single WSI were retained. Since no cross-slide relations can be constructed in these cases, TumorGraphs naturally reduces to the corresponding intra-WSI graph followed by patient-level pooling and prediction. Single-WSI cases accounted for 7/618 (1.1%) patients in ccRCC and 25/1,222 (2.0%) patients in the Lung cohort.

2.3 Survival Endpoint

We selected 12-month survival prediction as a clinically interpretable fixed-horizon prognostic task for which integrating information across multiple tumour WSIs is particularly relevant. Prognostic morphological evidence may be distributed across different tissue regions and sections, motivating the integration of all available WSIs at the patient level rather than relying on independently represented slides. This setting therefore provides a clinically motivated task for evaluating multi-WSI patient-level modelling.

The binary 12-month survival endpoint was defined as death within 365 days versus survival beyond 365 days. Patients alive with less than 365 days of follow-up were excluded because their 12-month survival status could not be established.

3 WSI Preprocessing and Quality Control

WSI preprocessing and patch-level feature extraction followed the reference UNI-v2 implementation. Tissue regions were segmented prior to patch extraction, and foreground regions were tiled into non-overlapping 256×256 patches at $20\times$ magnification. Retained patches were subsequently processed using the standard UNI-v2 input transformations and encoded with the pretrained UNI-v2 foundation model.

We did not introduce additional task-specific patch selection or artifact-rejection procedures beyond those applied by the reference preprocessing pipeline. Consequently, graph nodes correspond directly to the tissue patches retained by this preprocessing procedure. The same preprocessing was applied to all WSIs and all compared methods, ensuring that differences between models are not attributable to different tissue masks or patch-selection strategies.

Slides with limited usable tissue were not artificially expanded or balanced by additional patch sampling; the resulting number of retained patches was allowed to vary across WSIs. This variability motivates the patient-specific connectivity budget used by TumorGraphs, which is defined relative to the smallest WSI graph of each patient.

4 Implementation and Model Selection

All models use UNI-v2 patch embeddings and identical patient-level five-fold splits. Models were trained for up to 100 epochs with AdamW (learning rate 10^{-4}, weight decay 5×10^{-4}) and class-weighted binary cross-entropy, with early stopping on validation BACC (patience of 10 epochs).

Predefined GCN and GAT configurations with attention-head configurations 111, 211, and 221 were evaluated. For each model family, the configuration with the highest mean BACC across the five validation folds was selected for reporting. The same five folds were therefore used for configuration selection and performance reporting; no separate held-out test set was used. Identical patient-level folds were used across all methods.

When threshold optimisation was enabled, the classification threshold was determined from validation predictions.

Table 1 summarizes the explored TumorGraphs hyperparameter space.

Table 1. Hyperparameters explored for TumorGraphs. Bold indicates the selected value where a single value was used across the reported configurations.

Parameter	**Role**	**Tested Values**
Budget α	Cross-slide density	5, 10, 20, 100%
Clusters C	Hyperedge granularity	2, 3, **4**, 5, 6, 7
Layers L	GNN depth	2, **3**, 4
Hidden dim.	Embedding size	128, 256, **512**

4.1 GNN Notation

The notation GAT-*a*-*b*-*c* denotes a three-layer GAT whose successive layers use a, b, and c attention heads, respectively. GCN denotes the corresponding graph convolutional backbone.

5 Baseline Architectures

All baselines use identical UNI-v2 patch embeddings and patient-level splits. They differ primarily in the stage and mechanism through which information from multiple WSIs is combined. Importantly, none of the post-hoc aggregation baselines explicitly constructs patch-level relations between different WSIs.

5.1 Feature-Level Aggregation

AvgPool and MaxPool aggregate UNI-v2 patch representations across the patient without explicit relational modelling. ABMIL+AvgPool and TransMIL+AvgPool first encode information using the corresponding MIL architecture and subsequently apply average pooling for patient-level aggregation.

Figure 1 illustrates the feature-pooling baseline.

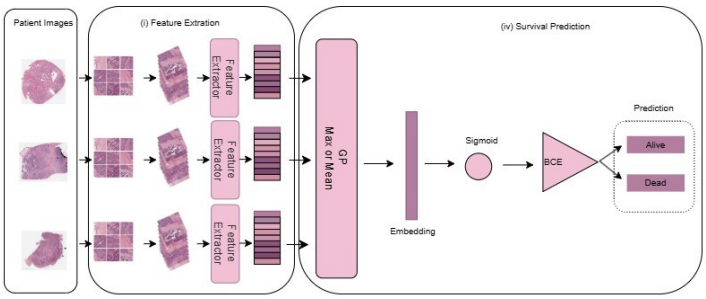


Fig. 1. Overview of the feature-pooling baseline.

PatchGNN represents each WSI independently as a spatial patch graph. Message passing is restricted to patches belonging to the same WSI, after which slide representations are aggregated across WSIs using mean or max pooling.

Figure 2 illustrates this strategy.

5.2 Patient-Bag Models

ABMIL (Patient Bag) and TransMIL (Patient Bag) treat patches from all WSIs belonging to a patient as a single patient-level bag and directly produce a patient representation. These models do not explicitly construct cross-slide patch relations.

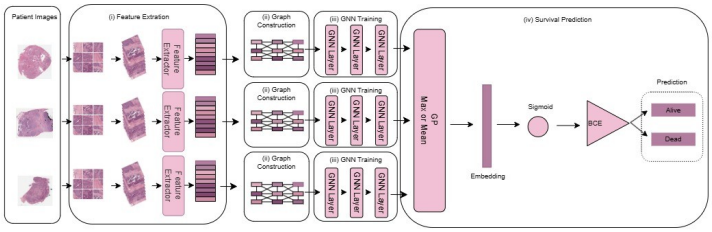


Fig. 2. Overview of PatchGNN feature-level aggregation. Each WSI is independently represented as a spatial graph before patient-level aggregation.

5.3 Prediction-Level Aggregation

PatchGNN-1D and PatchGNN-MV independently process each WSI and combine slide-level predictions using one-dominance (1D) and majority voting (MV), respectively.

Figure 3 illustrates this prediction-level aggregation.

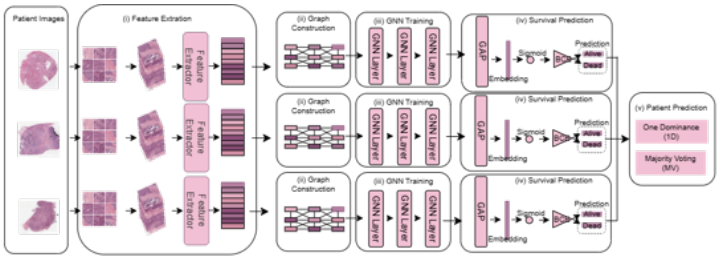


Fig. 3. Overview of PatchGNN prediction-level aggregation using one-dominance (1D) or majority voting (MV).

5.4 Hierarchical Transformer Aggregation

Hierarchical models first encode each WSI independently and subsequently aggregate slide representations at the patient level. P&SrE uses ABMIL for slide encoding followed by Transformer-based patient aggregation. TransMIL+P&SrE and PatchGNN+P&SrE replace the slide encoder while retaining hierarchical patient-level aggregation.

These approaches enable learned aggregation across WSIs after slide-level encoding, but do not explicitly construct patch-level relations between different WSIs.

5.5 Patient-Graph Size

To further characterize the computational scale of TumorGraphs, Table 2 reports typical patient-level graph and hypergraph sizes.

Table 2. Patient-level graph and hypergraph statistics, reported as median (IQR). Cross-slide statistics correspond to a connectivity budget of $\alpha = 5\%$.

Statistic	**ccRCC**	**Lung**
WSIs/patient	4 (4–5)	3 (2–4)
Nodes/patient	28,966 (12,562–39,309)	15,987 (8,190–25,764)
Intra-WSI edges/patient	138,896 (68,171–208,220)	37,268 (19,493–80,724)
Cross-WSI edges/patient (Hetero)	591 (304–1,019)	255 (162–458)
Hyperedges/patient	3 (2–3)	3 (1–4)
Nodes/hyperedge	158 (93–257)	95 (57–167)

6 Cross-Slide Connectivity Budget

The connectivity budget controls the amount of cross-slide information exchange introduced into each patient graph. Because patients have different numbers and sizes of WSIs, using a fixed number of cross-slide connections could result in substantially different graph densities across patients. Moreover, excessive cross-slide connectivity may increase computational cost and promote over-smoothing.

We therefore define a patient-specific connectivity cap as

$$B(\alpha, p) = \left\lfloor \alpha \cdot \min_{m \in \mathcal{S}(p)} |V^{(m)}| \right\rfloor,$$

where $\mathcal{S}(p)$ denotes the set of WSIs belonging to patient p, $|V^{(m)}|$ is the number of patch nodes in WSI m, and α is the connectivity-budget fraction. Thus, the maximum number of cross-slide connections is scaled relative to the number of patches in the smallest WSI of each patient, providing a patient-specific rather than cohort-wide connectivity cap.

For pairwise TumorGraphs (homogeneous and heterogeneous), each patch retrieves its $k' = 3$ nearest neighbours among patches belonging to other WSIs of the same patient in UNI-v2 embedding space. Candidate cross-slide edges are retained up to the patient-specific connectivity cap $B(\alpha, p)$. The homogeneous and heterogeneous variants use the same cross-slide edge construction; they differ only in whether intra- and inter-WSI edges share a common relation type or are represented as distinct relations during message passing.

For hypergraph TumorGraphs, the same parameter α controls hyperedge membership rather than the number of pairwise edges. After embedding-space clustering, hyperedge membership is capped proportionally to the smallest WSI

of the patient. Thus, α controls the extent of cross-slide integration in all TumorGraph variants, although the constrained quantity differs between pairwise graphs and hypergraphs.

We evaluate $\alpha \in \{5, 10, 20, 100\}\%$, spanning sparse to increasingly dense cross-slide connectivity. Importantly, $\alpha = 100\%$ does not correspond to a fully connected patient graph; rather, the connectivity cap equals the number of patch nodes in the smallest WSI of that patient. This sensitivity analysis evaluates whether effective patient-level integration requires dense cross-slide coupling or can be achieved with a limited number of cross-slide relations. For single-WSI patients, no cross-slide relations are constructed.

6.1 Resulting Patient-Graph Sizes

Table **??** reports typical graph sizes for the analytical cohorts. Because graph-size distributions are skewed, statistics are reported as median (IQR). Cross-WSI edge and hyperedge statistics are reported at $\alpha = 5\%$ to characterize the sparse connectivity setting.

Table 3 presents patient-level balanced accuracy (BACC, mean $\pm$ std) for the full set of baselines and TumorGraphs configurations.

7 Scalability Analysis

Experiments were conducted on NVIDIA RTX 6000 Ada Generation GPUs with 48 GB of GPU memory. The system used NVIDIA driver 570.133.07 and CUDA 12.8. Model training and inference were performed on a single GPU per run. Reported values correspond to forward-pass inference per patient.

Table 4 reports average forward-pass inference time per patient. TumorGraphs introduce moderate computational overhead compared with post-hoc aggregation baselines. Homogeneous TumorGraphs show inference time comparable to PatchGNN, while the hypergraph formulation incurs additional cost due to higher-order relation processing.

8 Best-Performing TumorGraph Configurations

The main paper reports the selected configurations for each TumorGraph family. Here, we provide sensitivity and specificity in addition to BACC to characterize the performance of the strongest configurations.

8.1 ccRCC

Table 5 reports the selected configuration within each TumorGraph family for the ccRCC cohort.

Table 3. Patient-level balanced accuracy (BACC, mean ± std) for baselines and TumorGraphs. † Mean pooling is widely used.

	ccRCC	**Lung**
Cross-Slide Aggregation		
Feature-level aggregation		
MaxPool	0.560 ± 0.039	0.588 ± 0.026
AvgPool†	0.731 ± 0.038	0.566 ± 0.025
ABMIL + AvgPool†	0.684 ± 0.087	0.639 ± 0.048
TransMIL + AvgPool†	0.736 ± 0.040	0.629 ± 0.042
PatchGNN + MaxPool	0.766 ± 0.036	0.661 ± 0.031
PatchGNN + AvgPool†	0.767 ± 0.031	0.650 ± 0.029
Patient-bag models		
ABMIL (Patient Bag)	0.689 ± 0.065	0.629 ± 0.026
TransMIL (Patient Bag)	0.723 ± 0.048	0.614 ± 0.027
Prediction-level aggregation		
PatchGNN-1D	0.757 ± 0.032	0.609 ± 0.037
PatchGNN-MV	0.763 ± 0.050	0.610 ± 0.039
Hierarchical Transformer aggregation		
P&SrE (ABMIL + Transformer)	0.725 ± 0.042	0.617 ± 0.029
PatchGNN + P&SrE	0.732 ± 0.047	0.626 ± 0.038
TransMIL + P&SrE	0.786 ± 0.038	0.622 ± 0.026
Cross-Slide Relational Learning (TumorGraphs)		
Homo (GAT 2-1-1 | GCN)	0.789 ± 0.041	0.671 ± 0.022
Hetero (GAT 2-1-1 | GAT 2-1-1)	0.773 ± 0.030	$\mathbf{0.672 \pm 0.034}$
Hyper (GAT 2-2-1 | GAT 2-2-1)	$\mathbf{0.792 \pm 0.029}$	0.670 ± 0.040

8.2 Lung

Table 6 reports the selected configuration within each TumorGraph family for the Lung cohort.

Table 6. Selected TumorGraph configurations on the Lung cohort. Metrics are mean ± std across five folds.

Family	**Configuration**	**BACC**	**Sensitivity**	**Specificity**
Homogeneous	GCN (20%)	$\mathbf{0.672 \pm 0.022}$	0.595 ± 0.158	$\mathbf{0.748 \pm 0.187}$
Heterogeneous	GAT-211 (100%)	0.672 ± 0.034	$\mathbf{0.621 \pm 0.194}$	0.723 ± 0.237
Hypergraph	GAT-221 (5%)	0.670 ± 0.040	0.594 ± 0.148	0.747 ± 0.189

Table 4. Average forward-pass inference time per patient. Lower is better.

Method	Time (ms)
AvgPool	0.6
P&SrE (ABMIL+Transformer)	4.3
PatchGNN-1D	6.5
PatchGNN	11.6
TumorGraphs (Homo)	11.7
TumorGraphs (Hetero)	17.6
TumorGraphs (Hyper)	16.0

Table 5. Selected TumorGraph configurations on MMIST-ccRCC. Metrics are mean ± std across five folds.

Family	Configuration	BACC	Sensitivity	Specificity
Homogeneous	GAT-211 (5%)	0.7886 ± 0.0409	0.8305 ± 0.0274	$\mathbf{0.7467 \pm 0.0980}$
Heterogeneous	GAT-111 (5%)	0.7785 ± 0.0426	0.8102 ± 0.0452	0.7467 ± 0.1147
Hypergraph	GAT-221 (20%)	$\mathbf{0.792 \pm 0.029}$	0.838 ± 0.052	0.747 ± 0.050
Hypergraph	GCN (5%)	0.7834 ± 0.0185	$\mathbf{0.8501 \pm 0.0640}$	0.7167 ± 0.0866

9 CONCH-Based Semantic Audit

CONCH was used as an external pretrained vision–language model to assess the morphological coherence of cross-slide relations. CONCH does not provide predefined tissue-class labels; instead, zero-shot classification is performed by comparing image embeddings with text embeddings of histopathology concepts.

Following the zero-shot classification procedure adopted in the original CONCH work, where ChatGPT was used to generate task-specific class names and descriptions for downstream zero-shot evaluation, we used ChatGPT to define cancer-specific histopathology labels and corresponding text descriptions. Labels were defined separately for ccRCC and lung cancer to reflect relevant histopathological patterns for each cancer type.

For each patch, its CONCH image embedding was compared with the text embeddings of the candidate descriptions, and the category with the highest similarity was assigned. The resulting labels were subsequently grouped into the coarse categories *tumor*, *normal*, and *other* used in the semantic audit.

These labels were not used for model training, graph construction, edge selection, or survival prediction. They were generated only after model training to evaluate whether patches connected across different WSIs showed greater semantic coherence than matched random within-patient groups. Semantic coherence was quantified using purity, normalized entropy, and cross-WSI label agreement.

Importantly, the CONCH-derived labels are zero-shot semantic assignments rather than pathologist-annotated ground truth. Accordingly, this analysis is intended as a semantic consistency audit, not as independent histopathological validation. Its purpose is to assess whether the constructed cross-slide relations

preferentially connect morphologically coherent tissue regions according to an external histology foundation model.

10 Summary of Supplementary Findings

Across both datasets, TumorGraphs achieve competitive or improved performance relative to post-hoc aggregation approaches while maintaining manageable computational cost. Performance is similar across the three relational formulations, and the semantic audit provides evidence that cross-slide relations connect morphologically coherent tissue regions.