

## Suppl. A. GLMP Implementation Details

**Foreground tissue detection.** Each whole-slide image (WSI) is tiled into non-overlapping square patches of a fixed physical size (e.g.,  $128 \times 128 \mu\text{m}$ ). We derive an adaptive tissue mask by converting a downsampled WSI thumbnail to the Hue, Saturation, and Value (HSV) color space and applying Otsu’s method to the non-zero values of each channel [62]. Patches are retained for subsequent analysis only if their masked tissue area exceeds a minimum threshold of 1%.

**Patch clustering for efficient MLLM prompting.** We use a cluster-based strategy to ensure efficiency of the MLLM API calls. For each WSI, foreground patches are first clustered into  $k$  histology patch groups using  $k$ -means based on their similarity in the feature space of a vision encoder backbone. Since within-slide clustering analysis by definition is invariant to slide-level (and thus TSI-level) batch effects, which are constant for any single tissue slide, this clustering step is commonly used for efficiently organizing within-slide patches into histopathologically representative groups [19, 64, 74]. Existing works find this single-slide clustering process to be stable with respect to model choice and the number of clusters when  $k$  is between 10 and 25 per WSI [3, 14, 26, 64, 83]. We thus set the number of clusters to  $k = 10$  in our experiments to control MLLM context window usage, since long contexts can lead to performance degradation [4, 38, 45]. The clustering serves only as a preprocessing step for efficient API usage and does not contribute to the final embeddings, which are generated entirely through the MLLM-text encoder pipeline. After clustering, a soft probabilistic membership score is then computed for each patch to every cluster via a temperature-scaled softmax over their cosine similarities:

$$p_{ij} = \frac{\exp(\langle \hat{\mathbf{x}}_i, \hat{\boldsymbol{\mu}}_j \rangle / \tau)}{\sum_{\ell=1}^k \exp(\langle \hat{\mathbf{x}}_i, \hat{\boldsymbol{\mu}}_{\ell} \rangle / \tau)} \quad (\text{S1})$$

where  $\langle \cdot, \cdot \rangle$  denotes cosine similarity and  $\tau$  is the softmax temperature, which is set to  $\tau = 0.1$  as recommended in temperature-tuning literature [2, 23]. These probabilities are used to select  $m = 25$  representative patches per cluster for the MLLM API call.

**Semantic description generation.** Biologically focused semantic descriptions are generated for all histology patch groups in an API call to the MLLM (default: Gemini 2.5 Pro [15]). The request consists of a structured prompt (Suppl. J.1) followed by representative patches from each group, with images placed immediately after their textual identifiers. To induce a deterministic and structured output, the generation temperature is set to 0.0, and the model is instructed to return a single JSON object covering all groups. The process is repeated 10 times to form an ensemble of outputs, allowing minor textual variations from the inherent stochasticity of the MLLM to be averaged out in the subsequent embedding step.

**Patch-level embedding generation.** The structured MLLM descriptions for the  $k$  histology patch groups are encoded into  $k$  corresponding semantic embeddings using a text embedding model (Gemini Embedding [43]). To reduce randomness in outputs due to the inherent stochastic properties of MLLMs, the embeddings from 10 repeated MLLM runs for each group are then averaged to yield a single embedding  $\mathbf{s}_j$  for each group  $j$ . The final GLMP embedding for an individual patch,  $\mathbf{z}_i$ , is then computed as a weighted sum of these  $k$  group embeddings using the patch’s soft probabilistic membership score  $p_{ij}$  from (Equation S1) as the weight:

$$\mathbf{z}_i = \sum_{j=1}^k p_{ij} \mathbf{s}_j \quad (\text{S2})$$

## Suppl. B. Experimental details

**Data preprocessing.** For all datasets, image patches are extracted to represent a physical area of  $128 \times 128 \mu\text{m}$  at an effective 20x magnification ( $\approx 0.5 \mu\text{m}/\text{pixel}$ ). The exception is the TumSeg dataset, which is available only at a low resolution (1.25x), from which we extract  $256\mu\text{m} \times 256\mu\text{m}$  patches. For CAMELYON16 and TCGA-LUSC, we select foreground patches located entirely within an annotated region ("Tumor") or entirely outside of any annotated region ("Normal").

**Linear probe and training.** We train a single linear layer on frozen embeddings, using the Adam optimizer with a learning rate of  $10^{-4}$  and cross-entropy loss. Training is performed with a batch size of 256 patches for 20 epochs, and we use the weights from the final epoch for evaluation.

### Evaluation protocols.

- **Cross-TSI testing:** We evaluate generalization across TSIs by training and testing on slides from different TSIs. For this task, patch-level datasets are constructed by randomly sampling up to 2,000 tumor or normal patches from each slide. On the CAMELYON16 dataset, which includes 2 TSIs, models are trained on one TSI and tested on the other, and vice-versa. For the TCGA-LUSC dataset, we employ a leave-one-TSI-out testing, training on two TSIs and testing on the third, with the held-out TSI in rotation.
- **Within-TSI testing:** We pool slides from all TSIs and use 5-fold splitting for creating the training and testing sets. Splits are created at the slide level so that no slide appears in more than one fold, and stratification maintains a balanced distribution of TSIs across folds. The patch sampling procedure (up to 2,000 patches per slide) is identical to that used in the cross-TSI experiments.
- **TSI Confounding:** We adopt the protocol from Kömen et al. [40] using a 56-slide subcohort from CAMELYON16. From each slide, 200 patches are sampled, with tumor patches drawn from metastatic slides and normal patches only from non-metastatic slides. This design yields three training splits with increasing levels of label-site correlation (50/50 no bias, 75/25 low bias, and 100/0 high bias), as summarized in Table 1. A linear probe is then trained on each biased split and evaluated on a test set constructed with the opposite correlation.
- **Neighborhood-based Robustness Evaluation:** Following [41], we evaluated the Robustness Index (RI) on a strictly balanced  $2 \times 2$  subcohort of CAMELYON16. We selected 10 slides for each combination of tissue class (Normal vs. Tumor) and TSI (Radboud UMC vs. UMC Utrecht), totaling 40 slides, and sampled 200 patches per slide. We computed the  $k$ -nearest neighbors using cosine similarity on  $L_2$ -normalized embeddings, dynamically masking patches from the same WSI as the query patch to ensure cross-slide retrieval. The neighborhood size  $k$  was selected by first determining the optimal  $k$  for each model via leave-one-slide-out cross-validation accuracy on the biological class, and then adopting the global median  $k$  across all models for the final RI calculation.
- **TSI Prediction:** We quantify site-specific artifacts by training a linear probe to predict the acquisition site across four datasets: CAMELYON16, TCGA-LUSC, AI4SKIN, and TumSeg. For each dataset, we sample up to 2,000 patches per slide and perform a 5-fold combined-TSI cross-testing. We report the mean prediction accuracy and standard deviation across folds.
- **k-NN on Principal Components:** Following the protocol of Kömen et al. [40], we evaluate site-specific signal in the TCGA-LUSC and AI4SKIN embeddings. Using 5-fold cross-testing at the slide level, we train a  $k$ -nearest neighbors classifier ( $k = 5$ , cosine similarity) on the top  $l$  principal components, with  $l$  varying from 1 to 50. This assesses how much site information is captured in the dimensions of highest variance.

## Suppl. C. Cross-TSI generalization

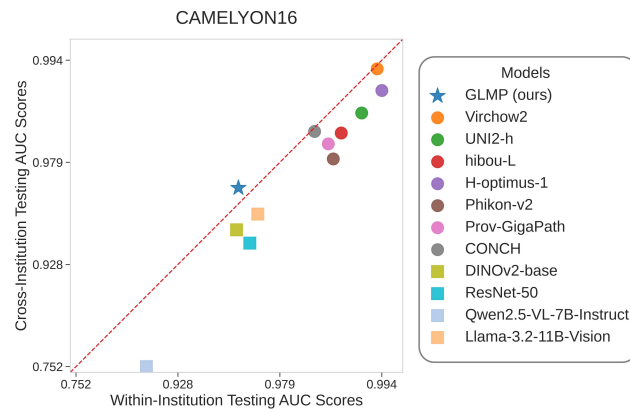


Figure S6. Accuracy of within-TSI testing (i.e. training data includes some WSIs from the testing TSIs) vs. cross-TSI testing (i.e. training does not include any WSIs from the testing TSIs) on CAMELYON16 tissue classification task. Patches in training and testing sets are from disjoint WSIs, ensuring that patches from the same WSI do not appear in both sets. Points falling below the  $y = x$  line indicate a performance drop in model generalization when new TSIs (not included during training) are used for testing. Accuracy is measured by AUC, with the axes transformed using  $-\log(1-\text{AUC})$  for clearer visualization scale. PFMs are marked with circles while general-purpose vision models are marked with squares.

Table S4. AUC results for tissue classification under within-TSI and cross-TSI evaluation on TCGA-LUSC and CAMELYON16, corresponding to the correlation plots in Figure 4 and Figure S6, including baseline PFMs and variants with Macenko normalization and DANN alignment.

Model	CAMELYON16		TCGA-LUSC	
	Within-TSI AUC	Cross-TSI AUC	Within-TSI AUC	Cross-TSI AUC
GLMP (ours)	0.9653	0.9714	0.8819	0.8913
Virchow2	0.9935	0.9932	0.8968	0.8548
UNI2-h	0.9922	0.9885	0.8878	0.8364
hibou-L	0.9900	0.9853	0.9066	0.8405
H-optimus-1	0.9939	0.9912	0.9106	0.8805
Phikon-v2	0.9890	0.9799	0.9187	0.8580
Prov-GigaPath	0.9883	0.9832	0.8853	0.8540
CONCH	0.9862	0.9856	0.8154	0.7988
DINOv2-base	0.9645	0.9526	0.8157	0.7981
ResNet-50	0.9698	0.9443	0.8724	0.8564
Qwen2.5-VL-7B-Instruct	0.8943	0.7517	0.8330	0.7984
Llama-3.2-11B-Vision	0.9726	0.9607	0.7998	0.7895
<b>Macenko Normalization</b>				
Virchow2	0.9872	0.9886	0.8928	0.8699
UNI2-h	0.9916	0.9894	0.8899	0.8685
hibou-L	0.9873	0.9859	0.8958	0.8600
H-optimus-1	0.9927	0.9910	0.8978	0.8767
Phikon-v2	0.9862	0.9811	0.9020	0.8477
Prov-GigaPath	0.9845	0.9815	0.8843	0.8650
<b>DANN</b>				
Virchow2	0.9921	0.9861	0.8858	0.8442
UNI2-h	0.9912	0.9821	0.8834	0.8296
hibou-L	0.9892	0.9869	0.9088	0.8554
H-optimus-1	0.9930	0.9901	0.8935	0.8550
Phikon-v2	0.9868	0.9759	0.9130	0.8500
Prov-GigaPath	0.9851	0.9737	0.8744	0.8549

## Suppl. D. Baseline model configurations

Our comparative analysis involved 12 foundation models, encompassing pathology-specific vision encoders, general-purpose vision models, and MLLMs (Table S5). All pre-trained models were obtained from the Hugging Face Hub, except for the ResNet-50 baseline from torchvision. For each model, the feature extraction procedure was tailored to its specific architecture and the authors’ recommendations. For Virchow2, we concatenated the classification token with the mean-pooled representation of all patch tokens. For hibou-L, we utilized the model’s `pooler_output`, while for Phikon-v2 and the DINOv2 vision encoder, we extracted the final hidden state of the classification token. For UNI2-h, H-optimus-1, and the Prov-GigaPath tile encoder, we used the single feature vector returned directly by a forward pass. For the contrastive vision-language model CONCH, we obtained features from its vision encoder prior to the final contrastive projection layer. The ResNet-50 baseline was represented by its final global average pooling layer. For the MLLMs, we investigated two types of semantic embeddings: features from the vision projection layer, and contextualized features from the full model’s last hidden layer. Accordingly, for Qwen2.5-VL-7B-Instruct, we obtained the embeddings by mean-pooling the output of its `get_image_features` function. For Llama-3.2-11B-Vision, we obtained the contextualized embeddings by invoking the model’s standard forward pass with a minimal, non-informative text prompt and then mean-pooling the representations from the last hidden layer.

Table S5. Overview of the foundation models benchmarked in this study.

Model Repository ID	Architecture (Param.)	Pre-training Data	Dim.	Reference
<b>Pathology-Specific Vision Encoders</b>				
paige-ai/Virchow2	ViT-H/14 (632M)	3.1M WSIs	2560	[86]
MahmoodLab/UNI2-h	ViT-H/14 (681M)	350k WSIs	1536	[12]
histai/hibou-L	ViT-L/14 (304M)	1.14M WSIs	1024	[52]
bioptimus/H-optimus-1	ViT-g/14 (1.1B)	1M WSIs	1536	[8]
owkin/phikon-v2	ViT-L/16 (304M)	58.4K WSIs	1024	[24]
prov-gigapath/prov-gigapath	ViT-g/14 (1.1B)	171k WSIs	1536	[78]
MahmoodLab/CONCH	ViT-B/16 (86M)	1.17M image-text pairs	512	[47]
<b>General-Purpose Vision Encoders</b>				
facebook/dinov2-base	ViT-B/14 (86M)	LVD-142M	768	[58]
torchvision/resnet50	ResNet-50 (25.6M)	ImageNet-1K	2048	[29]
<b>Multi-Modal Language Models (MLLMs)</b>				
Qwen/Qwen2.5-VL-7B-Instruct	MLLM (7B)	Image-text/documents/video	3584	[73]
meta-llama/Llama-3.2-11B-Vision	MLLM (11B)	6B image-text pairs	4096	[27]

**Suppl. E. Additional experimental results**  
**Suppl. E.1. Multi-TSI Embedding Clustering**

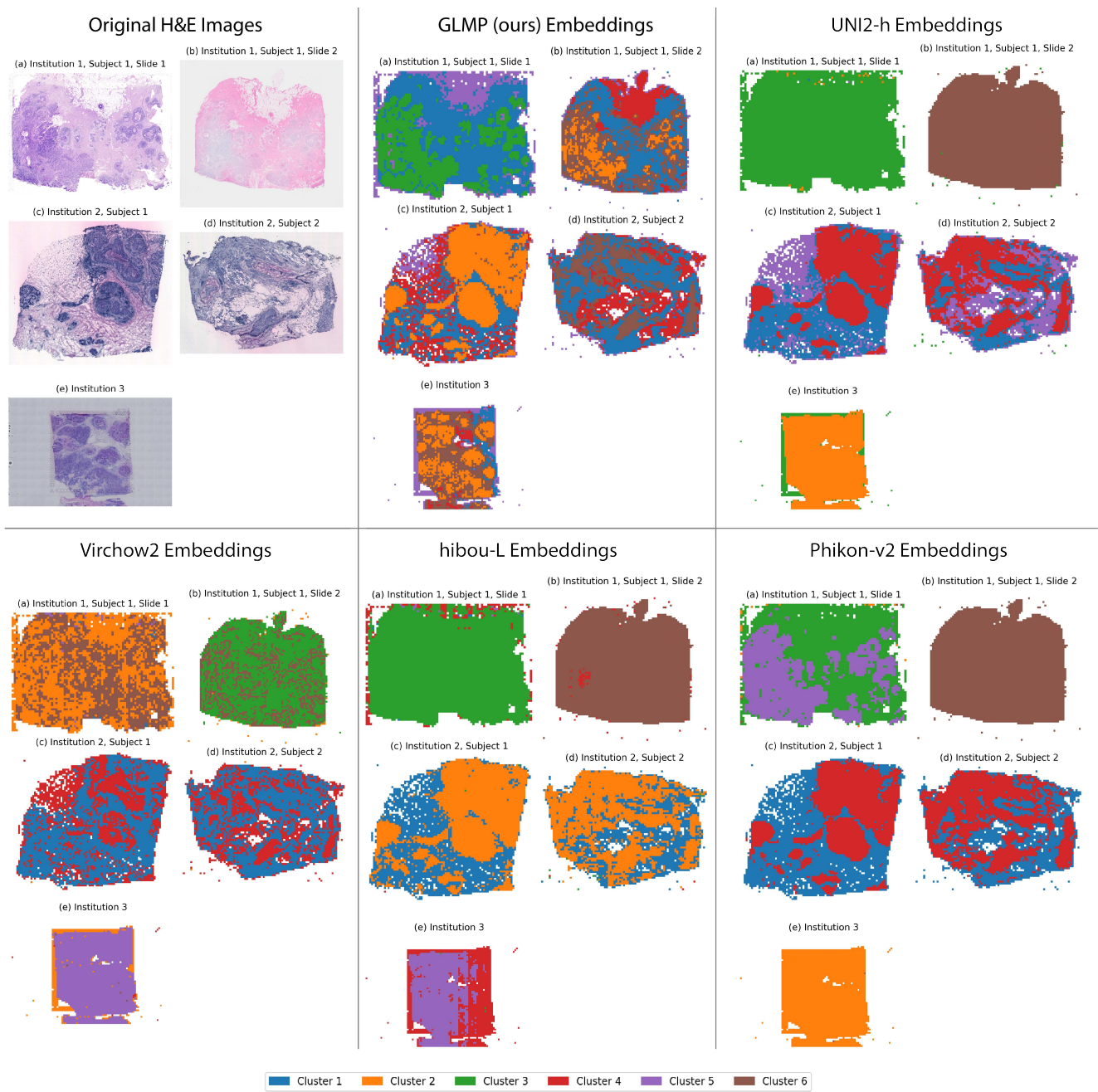


Figure S7. Clustering results on the MSBCD dataset using  $k$ -means on histology image patch embeddings generated by different models.

## Suppl. E.2. Visualization of embedding spaces

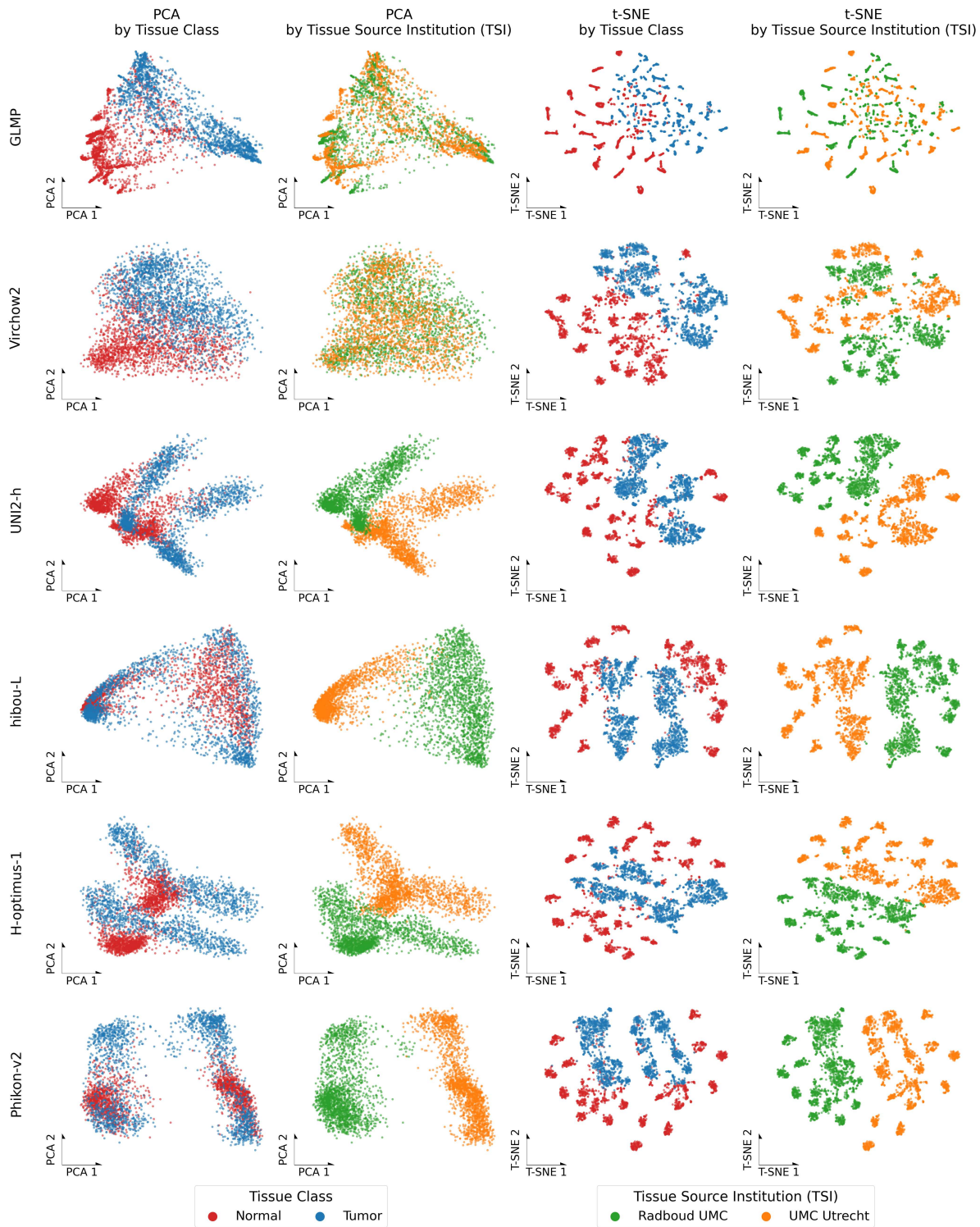


Figure S8. Dimension reduction of pathology model embeddings on the CAMELYON16 Dataset.

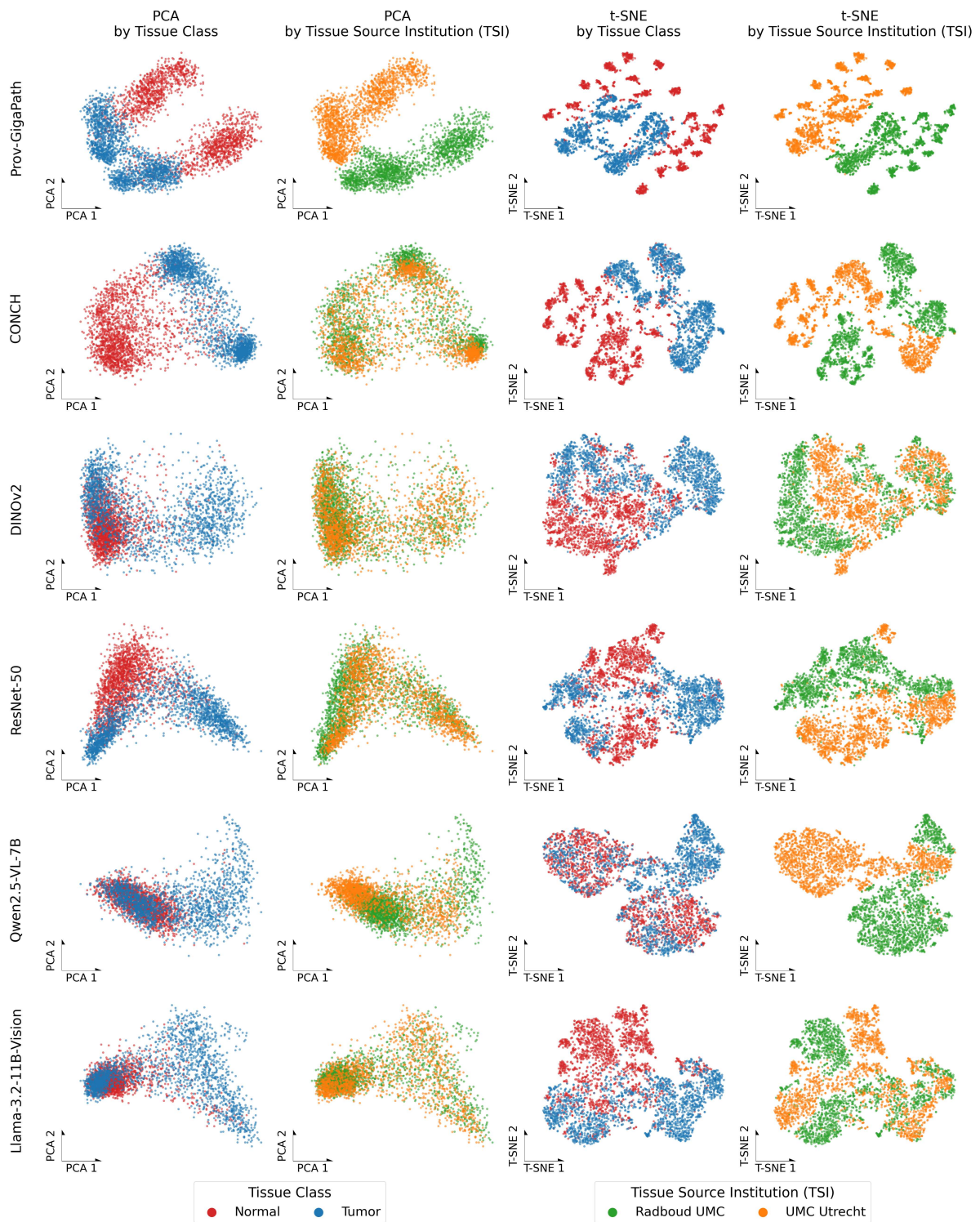


Figure S8. (Continued) Dimension reduction of pathology model embeddings on the CAMELYON16 Dataset.

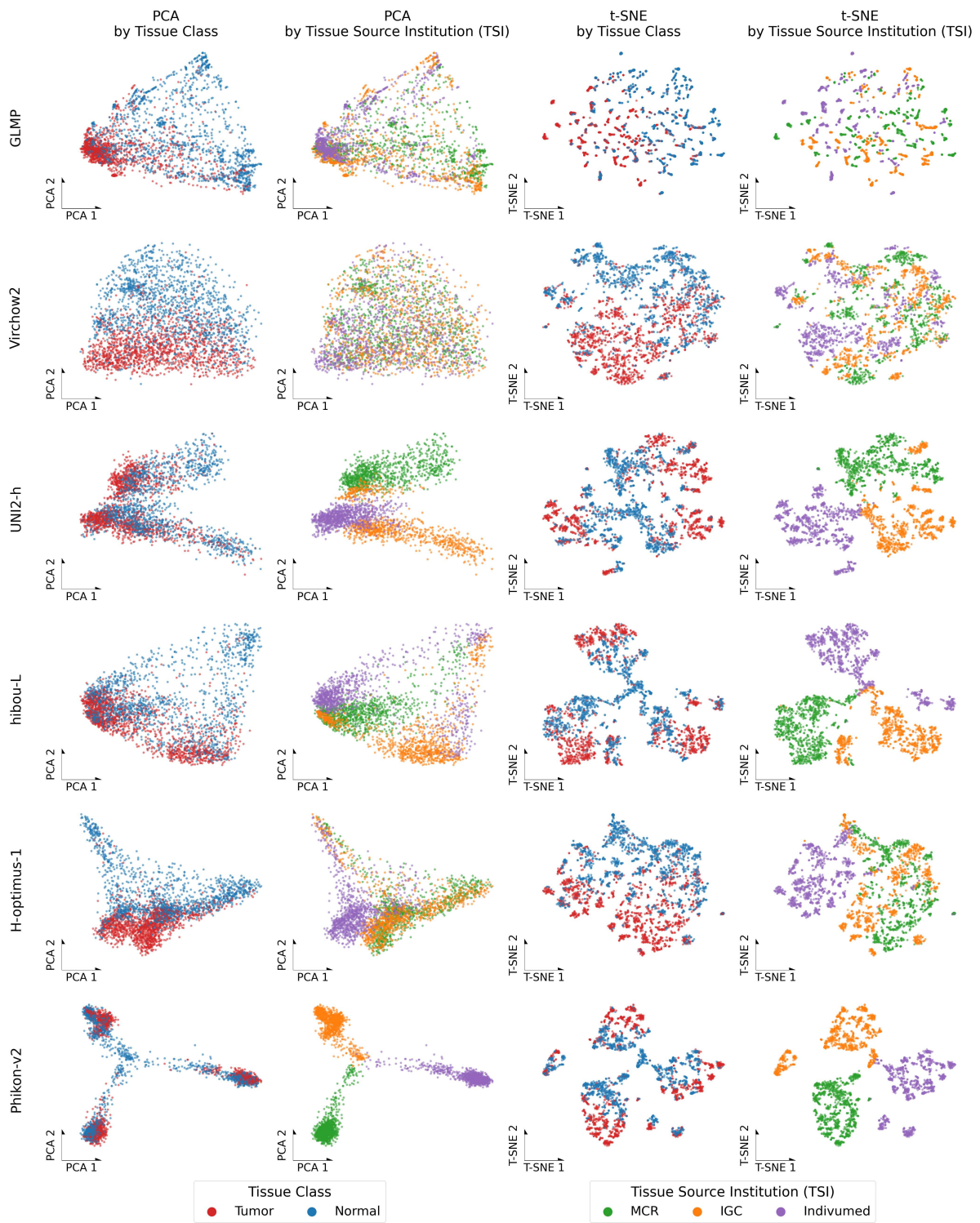


Figure S9. Dimension reduction of pathology model embeddings on the TCGA-LUSC Dataset.

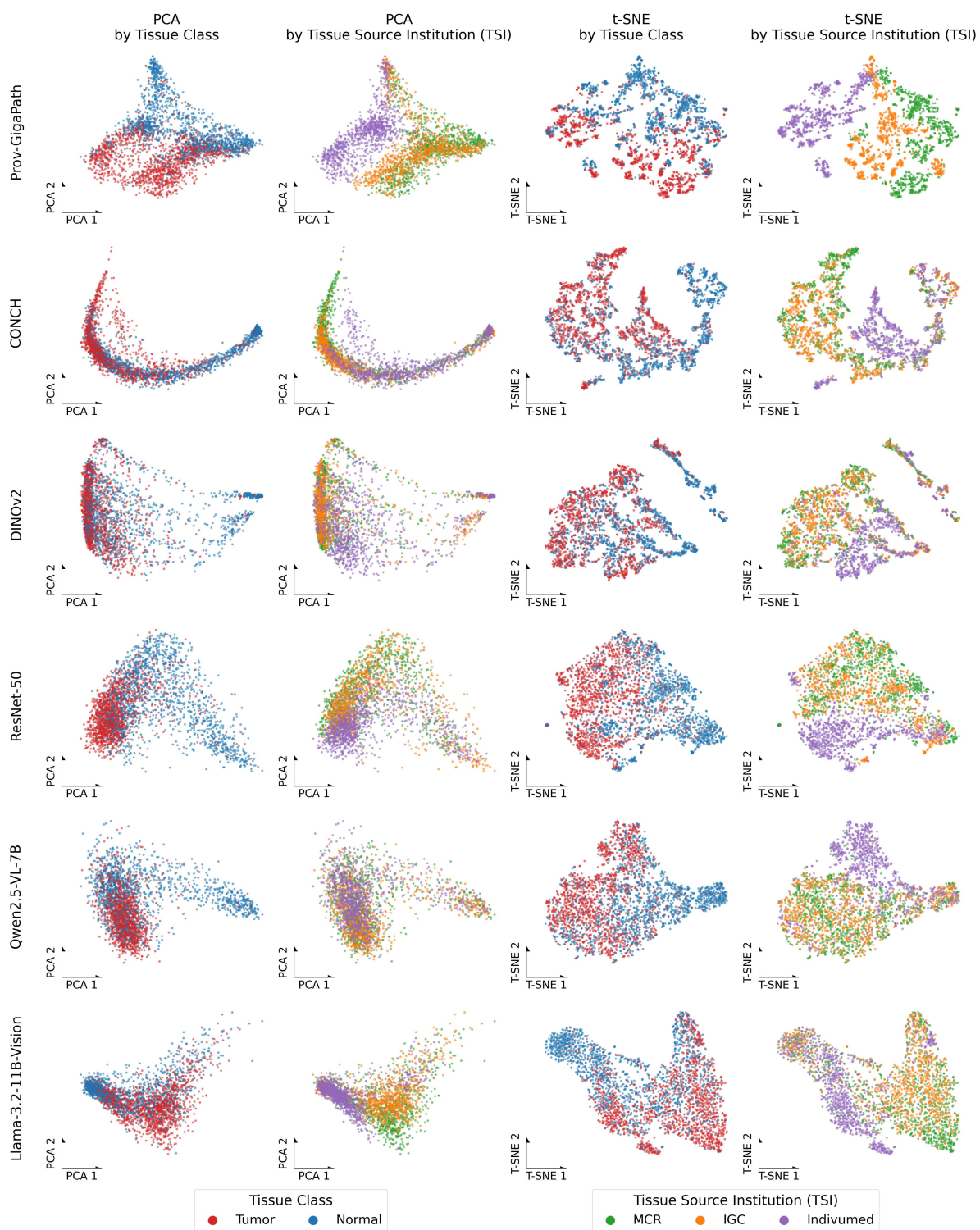


Figure S9. (Continued) Dimension reduction of pathology model embeddings on the TCGA-LUSC Dataset.

## Suppl. F. TSI confounding

### Suppl. F.1. Details for TCGA-LUSC TSI confounding experiments

To evaluate the models on the TCGA-LUSC cohort, we employed a 3-fold cross-validation strategy using samples from three TSIs. In each fold, we grouped two sites together and separated the third to construct spurious correlations between the TSI and the tissue class.

For instance, in Fold 1, TSI-66 was assigned to represent one grouping, while TSI-22 and TSI-56 represented the other. We partitioned the data into four distinct splits with increasing bias levels, sampling exactly 200 patches per WSI. At the 100/0 completely confounded split, all normal training patches originated exclusively from TSI-66, while all tumor training patches originated from TSI-22 and TSI-56. The corresponding test set perfectly reversed this correlation. The precise patch distribution for Fold 1 is shown in [Table S6](#).

Table S6. Representative dataset split (Fold 1) for the 3-Fold TSI confounding experiment on TCGA-LUSC. Patch counts are based on sampling 200 patches per WSI. The training sets introduce spurious correlations between the tissue class and TSI, while the testing set features a completely reversed correlation.

Split	Set	TSS 66		TSS 22 & TSS 56		Total Patches
		# Normal	# Tumor	# Normal	# Tumor	
50/50 (No Bias)	Training	600	600	600	600	2,400
67/33 (Low Bias)	Training	800	400	400	800	2,400
83/17 (Med Bias)	Training	1,000	200	200	1,000	2,400
100/0 (High Bias)	Training	1,200	0	0	1,200	2,400
	Testing	0	800	800	0	1,600

## Suppl. F.2. Details for TCGA-LUSC TSI confounding experiments

In the main text, we compare the unnormalized GLMP to baseline PFMs utilizing Macenko stain normalization. To ensure a comprehensive evaluation of stain normalization techniques, we also evaluated the baseline PFMs using Reinhard normalization [60] on CAMELYON16.

As shown in Figure S10, Reinhard normalization proved less effective than Macenko normalization at baseline PFMs from shortcut learning, consistent with findings in [41]. The performance of several baseline PFMs stagnates near an AUC of zero under the 100/0 high-bias condition. This indicates that the models still heavily relied on residual TSI-specific color statistics that Reinhard normalization failed to remove.

Reinhard Normalization for Baseline PFMs

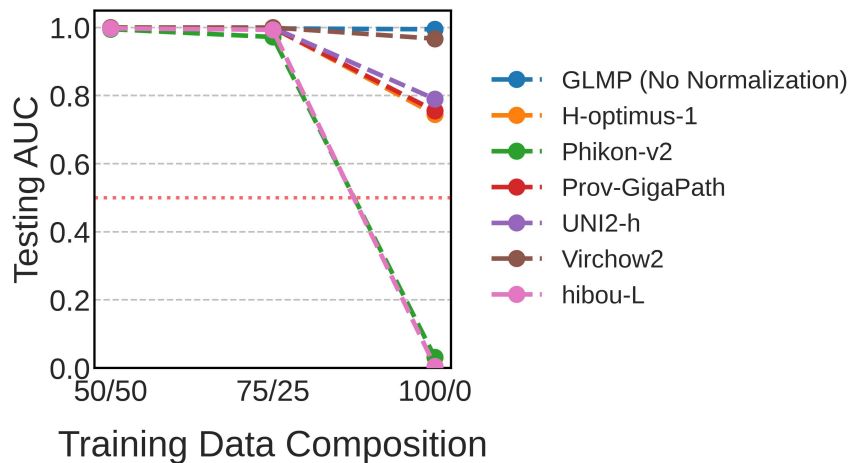


Figure S10. Tissue classification performance under increasing correlations between class label and TSI in CAMELYON16, using Reinhard stain normalization for the baseline PFMs.

## Suppl. G. Neighborhood-based Robustness Evaluation

We evaluated the Robustness Index (RI) on a strictly balanced CAMELYON16 subcohort to quantify the influence of technical artifacts on local embedding geometries. Higher RI values confirm that embedding neighborhoods are driven by true biological similarity rather than site-specific confounders.

Table S7. Robustness Index (RI) evaluated on a balanced CAMELYON16 subcohort. Higher values indicate that a patch’s local neighborhood in the embedding space is driven by biological similarity rather than confounding TSI-specific artifacts.

<b>Model</b>	<b>Robustness Index (RI)</b>
GLMP (Ours)	<u>0.789</u>
Virchow2	0.773
H-optimus-1	0.697
Prov-GigaPath	0.405
UNI2-h	0.202
hibou-L	0.114
Phikon-v2	0.024

## Suppl. H. TSI Prediction

### Suppl. H.1. Linear probe TSI prediction

Table 2 expands upon the TSI prediction results presented in the main text by including the standard deviations across 5-fold cross-validation. Accuracies near random chance indicate strong invariance to non-biological batch effects.

Table S8. Model susceptibility to batch effects, measured by accuracy on the TSI prediction task using a linear probe. Lower accuracy indicates less potential influence by TSI-specific signatures and higher robustness to technical artifacts, with a truly TSI-agnostic model expected to perform no better than random chance. Standard deviations were computed across five cross-validation folds using slide-level splits.

Model	CAMELYON16	TCGA-LUSC	AI4SKIN	TumSeg
Completely random (ideal performance)	0.5000	0.3333	0.5000	0.5000
GLMP (ours)	$0.6212 \pm 0.0678$	$0.4750 \pm 0.1053$	$0.5287 \pm 0.0757$	$0.6154 \pm 0.1564$
Virchow2	$0.9998 \pm 0.0002$	$0.9600 \pm 0.0589$	$0.9926 \pm 0.0136$	$0.9209 \pm 0.0305$
UNI2-h	$0.9999 \pm 0.0001$	$0.9690 \pm 0.0440$	$0.9965 \pm 0.0045$	$0.9201 \pm 0.0365$
hibou-L	$0.9995 \pm 0.0008$	$0.9593 \pm 0.0572$	$0.9983 \pm 0.0016$	$0.8962 \pm 0.0512$
Midnight-12k	$0.9972 \pm 0.0061$	$0.9921 \pm 0.0083$	$0.9700 \pm 0.0425$	$0.8906 \pm 0.0509$
H-optimus-1	$0.9999 \pm 0.0001$	$0.9794 \pm 0.0288$	$0.9914 \pm 0.0058$	$0.8896 \pm 0.0579$
Phikon-v2	$0.9999 \pm 0.0001$	$0.9973 \pm 0.0035$	$0.9894 \pm 0.0156$	$0.9027 \pm 0.0435$
Prov-GigaPath	$0.9997 \pm 0.0002$	$0.9560 \pm 0.0498$	$0.9893 \pm 0.0203$	$0.9278 \pm 0.0326$
CONCH	$0.9985 \pm 0.0004$	$0.7222 \pm 0.0859$	$0.9536 \pm 0.0205$	$0.8559 \pm 0.0401$
DINOv2-base	$0.9966 \pm 0.0010$	$0.7335 \pm 0.0942$	$0.9640 \pm 0.0258$	$0.8395 \pm 0.0635$
ResNet-50	$0.9877 \pm 0.0098$	$0.8316 \pm 0.0779$	$0.9690 \pm 0.0265$	$0.8030 \pm 0.0832$
Qwen2.5-VL-7B-Instruct	$0.9990 \pm 0.0004$	$0.8124 \pm 0.0852$	$0.9907 \pm 0.0063$	$0.8825 \pm 0.0734$
Llama-3.2-11B-Vision	$0.9976 \pm 0.0012$	$0.6665 \pm 0.0644$	$0.9813 \pm 0.0140$	$0.8467 \pm 0.0546$

## Suppl. H.2. Details for PC-based TSI prediction experiments

TSI prediction using top principal components (PCs). A KNN classifier is trained on the top PCs of the embeddings based on a protocol adapted from [40].

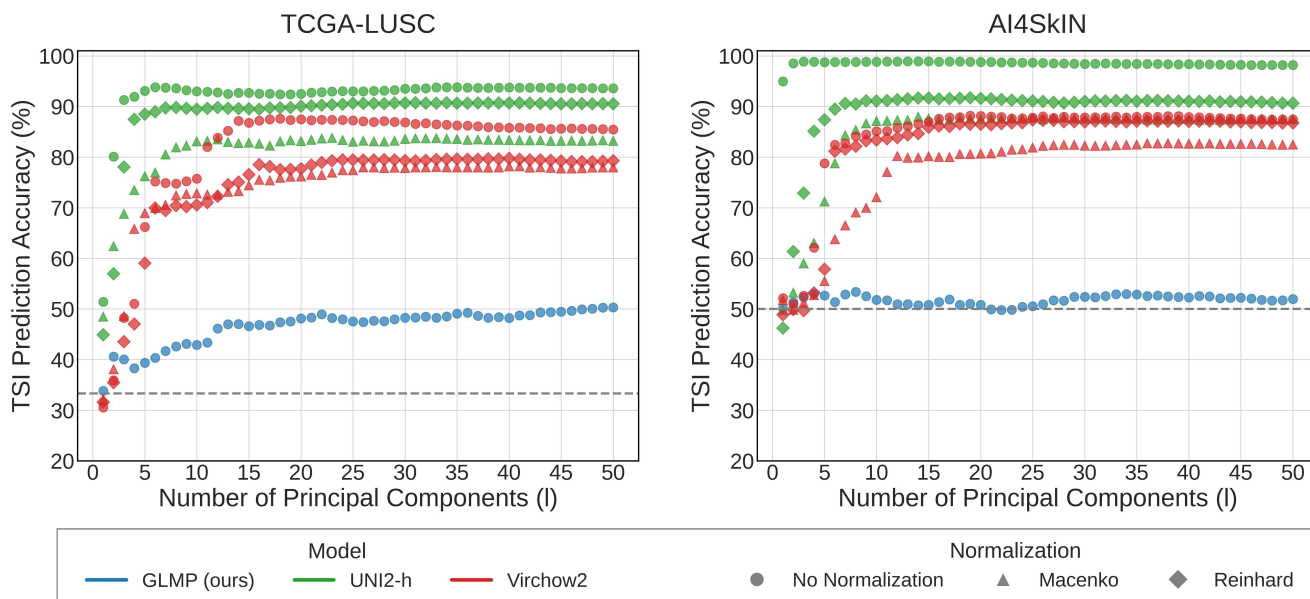


Figure S11. Accuracy for predicting TSI in TCGA-LUSC and AI4SKIN using KNN based on the top principal components (PCs) of the embeddings generated by different models. The horizontal dashed line indicates random chance.

## Suppl. I. Results of ablation studies

### Suppl. I.1. Effect of MLLM prompt focus

We modified the MLLM prompt to explicitly describe non-biological characteristics (e.g., stain color, scanner focus) to test its filtering capability.

Table S9. Effects of MLLM prompt focus on TSI prediction on CAMELYON16. Lower scores indicate better robustness against batch effects.

MLLM Prompt Focus	TSI prediction accuracy
Biological features only	0.62
Biological + non-biological features	0.73

### Suppl. I.2. Sensitivity to clustering and text encoders

To address the sensitivity of the model to the initial grouping and embedding components, we evaluate alternative configurations for the clustering and text encoding models. While our default implementation uses  $k$ -means ( $k = 10$ ) and the Gemini text encoder, we test variations including alternative clustering algorithms (Spectral Clustering [54], Mini-Batch KMeans [63]), varying levels of cluster granularity ( $k \in \{8, 12, 15\}$ ), and alternative text encoders ranging from large general-purpose models (Qwen3 [84]) to smaller domain-specific ones (PubMedBERT [28]).

Table S10. Sensitivity analysis of GLMP components on tissue classification (AUC). We compare the default configuration (KMeans  $k = 10$ , Gemini Embedding) against alternative clustering strategies and text encoders.

Model Variation	CAMELYON16		TCGA-LUSC	
	Within-TSI	Cross-TSI	Within-TSI	Cross-TSI
<b>GLMP (default)</b>	0.9653	0.9714	0.8819	0.8913
<i>Clustering Strategy</i>				
KMeans ( $k = 8$ )	0.9650	0.9706	0.8675	0.8745
KMeans ( $k = 12$ )	0.9806	0.9772	0.8858	0.8973
KMeans ( $k = 15$ )	0.9748	0.9681	0.8888	0.8902
Mini-Batch KMeans ( $k = 10$ )	0.9668	0.9641	0.8793	0.8796
Spectral Clustering ( $k = 10$ )	0.9541	0.9486	0.8671	0.8606
<i>Text Encoder</i>				
Qwen3-Embedding-8B	0.9665	0.9707	0.8788	0.8863
Pubmedbert-Base-Embeddings	0.9589	0.9669	0.8845	0.8853

## Suppl. J. MLLM prompt for GLMP

### Suppl. J.1. Standard prompt

The following prompt guides the MLLM to generate biology-focused, artifact-free semantic descriptions for histology images, with `__ORGAN__` and `__SPECIES__` adapted to the specific dataset.

You are a board-certified `__ORGAN__` pathologist.

#### Input Structure & Order

You will receive this entire set of instructions first. Following these instructions, a sequence of content will be provided:

1. A text line identifying a group, for example: “The subsequent image patches pertain to Group 1.”
2. Immediately following this text, all image patches belonging to Group 1 will be provided.
3. This pattern will repeat for subsequent groups (e.g., “The subsequent image patches pertain to Group 2.”, followed by its images, and so on for all available groups).

Your task is to process all groups and then generate a single JSON output summarizing each one.

#### Context for Image Analysis

The H&E image patches you will receive are from a single whole-slide image (WSI) of a `__ORGAN__` tissue from `__SPECIES__`.

- **Group Delineation:** You must strictly use the textual group identifiers provided in the input stream (e.g., “The subsequent image patches pertain to Group 1.”) to define which images belong to which group, the order of the groups, and for keying your final JSON output.
- **Intra-Group Similarity:** All patches within a specific group (as defined by its preceding textual identifier) are expected to share similar histologic features, though minor variations may exist.
- **Expected Content:** Depending on the specific tissue type indicated above, you may encounter a wide range of histologic components. The specific features will be pertinent to the organ/tissue system being examined.

#### Tasks

1. **Within-Group Synthesis:** For each group of images presented (e.g., those following “The subsequent image patches pertain to Group 1.”), carefully examine **all** provided image patches belonging to that specific group. Synthesize a representative description that captures the **predominant, consistent, and defining** histologic features observed across these patches. For each group, your synthesis should specifically address the following aspects, which will directly correspond to the fields in the structured summary:
  - **Architectural Pattern:** Describe the predominant tissue arrangement (e.g., infiltrative growth, preserved native architecture, glandular formation, solid sheets) and structure density. If acellular or non-architectural, describe that arrangement.
  - **Cellular Morphology & Cytologic Grade:** Describe predominant cell features: relative cell size, cell density, nuclear pleomorphism, chromatin, nucleoli, mitotic activity, and cytoplasm. Assign an overall nuclear grade (low, intermediate, high) if applicable. If acellular, state ‘Not applicable’.
  - **Key Structural Interface:** Describe the most significant structural boundary or architectural interface observed. Focus on the relationship between the main lesion and surrounding tissue, such as the status of a capsule, basement membrane, or the nature of a tumor-stroma interface. If not applicable, state so.
  - **Stromal Response & Inflammation:** Describe the stroma and the type, density, and location of any inflammatory infiltrate. If the group is predominantly stroma or inflammation, describe it here.
  - **Necrosis & Other Key Features:** Describe the presence and type of any necrosis. Report other key features only if they are diagnostically significant AND truly widespread across the majority of patches, thereby defining the group’s overall character. Omit minor, focal, or incidental findings.
2. **Internal Comparison (Mental Step Only - Do Not Output):** Mentally compare the synthesized features of each group against the others. This step is for you to refine your within-group synthesis (Task 1). Observing significant differences in key features between groups confirms they are distinct entities. Use this mental differentiation to ensure that the description for each group accurately and uniquely captures *its own* predominant characteristics. **Absolutely do NOT mention these mental comparisons or reference any other group in your final written output.** Each group’s description must be entirely self-contained.

3. **Structured Summary (Final Output):** After all groups and their images have been presented and analyzed, produce a **single JSON object**. This object will contain one top-level key for each group processed. The keys in the JSON (e.g., “Group 1”, “Group 2”) must exactly match the group numbering specified in the textual identifiers that introduced each image set (e.g., “Group 1” from “The subsequent image patches pertain to Group 1.”). Base the descriptions for each field **strictly** on your synthesis from Task 1 for that specific group.

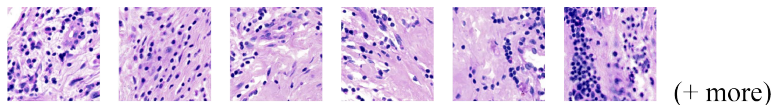
Example JSON structure:

```
{
  "Group 1": {
    "Architectural Pattern": "<description>",
    "Cellular Morphology & Cytologic Grade": "<description>",
    "Key Structural Interface": "<description>",
    "Stromal Response & Inflammation": "<description>",
    "Necrosis & Other Key Features": "<description>"
  },
  "Group 2": {
    // ... same structure ...
  }
}
```

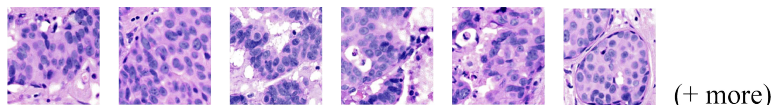
**Style & Constraints for Output**

- **Focus on Predominant, Defining, and Widespread Features:** Your descriptions **MUST** reflect features that are predominant, consistent, and diagnostically significant for the entire group of images. Base descriptions on observations from the clear majority of patches. **AVOID** over-reporting minor or focal variations.
- **Absolutely NO Cross-Group References:** Under **NO circumstances** should the description for one group (e.g., Group 8) mention, compare itself to, or reference any other group. Each group’s description **MUST** be entirely self-contained.
- **Do Not Mention Metadata in Descriptions:** Avoid mentioning patch size, number of patches, magnification, or the concept of clustering itself within the descriptive string values of the JSON.
- **Conciseness and Terminology:** Keep each field’s description to 1-2 sentences. Be precise and use standard histopathology terminology.
- **Strictly Avoid Artifact Description:** Focus exclusively on biological features. Omit descriptions of technical artifacts unless they pervasively obscure the biological assessment of a group.
- **Strict Output Format:** Output **only** the JSON object—no extra introductory text, commentary, explanations, or apologies. The output must start directly with ‘{’ and end directly with ‘}’.

The subsequent image patches pertain to Group 1.

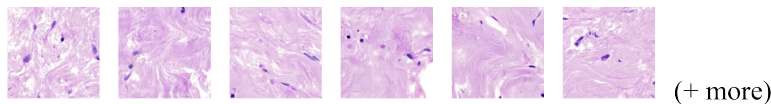


The subsequent image patches pertain to Group 2.



⋮

The subsequent image patches pertain to Group 10.



## Suppl. J.2. Prompt used in the ablation study

This prompt is identical to the biological-only prompt in Appendix [Suppl. J.1](#), except for the following additions and edits to request descriptions of non-biological artifacts in the ablation study.

### Addition to Task 1 fields

- **Non-Biological Characteristics:** Describe consistent non-biological features that characterize the group, including stain color balance or strength, scanner focus or stitching artifacts, and site or batch cues. Report only if these characteristics are widespread within the group.

**Edits to Style and Constraints:** Replace “Strictly Avoid Artifact Description” with: “Report non-biological characteristics only when they are consistent and pervasive within the group. Do not over-interpret minor or focal artifacts.”

### JSON Output

```
{
  "Group 1": {
    "Architectural Pattern": "<description>",
    "Cellular Morphology & Cytologic Grade": "<description>",
    "Key Structural Interface": "<description>",
    "Stromal Response & Inflammation": "<description>",
    "Necrosis & Other Key Features": "<description>",
    "Non-Biological Characteristics": "<description>"
  },
  "Group 2": {
    // ... same structure ...
  }
}
```

## Suppl. J.3. Ablation: the MLLM-only approach for tissue classification

### Prompt for the MLLM-only approach for CAMELYON16 tissue classification

You are a board-certified lymph node pathologist. Your task is to analyze a single H&E stained image patch provided to you. This patch is from a larger whole-slide image (WSI) of a lymph node.

Your analysis must be strictly confined to the features visible within this single patch. Do not make a diagnosis for the entire slide or the patient.

Based on your expert assessment of the morphology, classify the patch. Your response must be a single, valid JSON object and nothing else. The JSON object must contain a single key, “classification”, and its value must be one of two exact strings: “Tumor” or “Normal”.

For example, if you classify the patch as Normal, your **entire response** must be:

```
{
  "classification": "Normal"
}
```

## Suppl. J.4. Ablation: the MLLM-only approach for cluster-level tissue classification

You are a board-certified lymph node pathologist.

### Input Structure & Order

1. You will receive this entire set of instructions first.
2. Following these instructions, a sequence of content will be provided:
  - A text line identifying a group (e.g., “The subsequent image patches pertain to Group 1.”).
  - Immediately following this text, all image patches belonging to that group will be provided.
  - This pattern repeats for all subsequent groups.

**Context for Image Analysis** The H&E image patches you will receive are from a single whole-slide image (WSI) of a lymph node.

- **Group Delineation:** Use the textual group identifiers strictly to identify image-group membership.
- **Intra-Group Homogeneity:** All patches within a group represent the same tissue type.

### Tasks

1. **Diagnostic Assessment:** For each group, examine histologic features (morphology, architecture, nuclear grade) to determine malignancy.
  - **Criteria for “Tumor”:** Classify as “Tumor” if patches show metastatic carcinoma, melanoma, lymphoma, or neoplastic cells (high N:C ratio, pleomorphism, hyperchromasia, abnormal mitoses).
  - **Criteria for “Normal”:** Classify as “Normal” if patches show benign components (lymphocytes, germinal centers, sinuses), capsule, adipose tissue, fibrosis, or reactive changes.
2. **Structured Summary:** Produce a **single JSON object** classifying every group.

### Example JSON structure:

```
{
  "Group 1": {
    "Classification": "Tumor"
  },
  "Group 2": {
    "Classification": "Normal"
  }
}
```

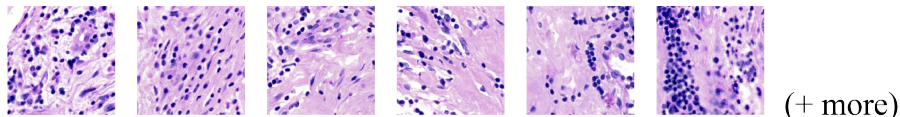
### Style & Constraints for Output

- **Binary Option Only:** Force a decision between “Tumor” or “Normal”. Do not use terms like “Suspicious” or “Atypical”.
- **No Explanations:** Do not include descriptions, reasoning, or confidence scores.
- **Strict Output Format:** Output **only** the JSON object. No introductory text or markdown formatting. The output must start with { and end with }.

## Suppl. K. Examples of MLLM output

Below are some examples of the structured descriptions generated by the MLLM (Gemini 2.5 Pro) for two histology patch groups from a WSI in MSBCD, using the biologically focused prompt in Appendix [Suppl. J.1](#).

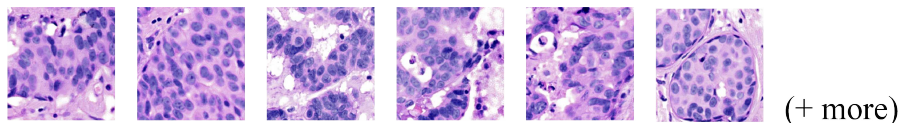
Input Image (Group 1):



Generated Description (Group 1):

```
"Group 1": {  
  "Architectural Pattern": "Disorganized, infiltrative pattern of mixed inflammatory cells within a fibrous and edematous stroma. No discrete epithelial or glandular structures are identified.",  
  "Cellular Morphology & Cytologic Grade": "The cellular population is composed of small, mature lymphocytes with dark nuclei and scant cytoplasm, intermixed with larger histiocytes showing vesicular nuclei and more abundant cytoplasm. Bland spindle cells are also present. Not applicable for nuclear grade.",  
  "Key Structural Interface": "The inflammatory infiltrate diffusely permeates the stromal connective tissue without forming a distinct interface with a neoplastic or epithelial component.",  
  "Stromal Response & Inflammation": "The stroma is fibrotic and edematous, hosting a prominent, diffuse, mixed chronic inflammatory infiltrate of lymphocytes and histiocytes.",  
  "Necrosis & Other Key Features": "No overt necrosis is seen. Scattered apoptotic bodies and cellular debris are present within the inflammatory milieu."  
}
```

Input Image (Group 2):



Generated Description (Group 2):

```
"Group 2": {  
  "Architectural Pattern": "Cohesive nests of neoplastic epithelial cells arranged in solid and cribriform patterns, consistent with ductal carcinoma in situ. The nests fill and expand duct-like spaces.",  
  "Cellular Morphology & Cytologic Grade": "The neoplastic cells exhibit moderate to marked pleomorphism, with enlarged, hyperchromatic to vesicular nuclei, irregular nuclear contours, and prominent nucleoli. The overall features correspond to an intermediate to high nuclear grade (Grade 2-3).",  
  "Key Structural Interface": "The neoplastic epithelial nests are sharply demarcated from the surrounding stroma, suggesting confinement within a ductal-lobular unit.",  
  "Stromal Response & Inflammation": "The periductal stroma is fibrous and contains a mild, patchy lymphoplasmacytic infiltrate.",  
  "Necrosis & Other Key Features": "Focal luminal necrosis is present, characterized by eosinophilic, granular debris. Apoptotic cells are also noted within the epithelial nests."  
}
```