Morphological Prototyping for Unsupervised Slide Representation Learning in Computational Pathology

Supplementary Material

1. Mathematical derivations of EM algorithm

Given the Gaussian mixture model (GMM) for a generative model for each individual patch embedding, we provide a detailed derivation for estimation of 1) the posterior probability for the prototype assignment $q(c|\mathbf{z}_n^j;\theta_j)$ and 2) the GMM parameters $\theta^j = \{\pi_c^j, \boldsymbol{\mu}_c^j, \Sigma_c^j\}$. Given the GMM specification,

$$\begin{aligned} p(\mathbf{z}_n^j;\theta^j) &= \sum_{c=1}^{C} p(c_n^j = c;\theta^j)\cdot p(\mathbf{z}_n^j|c_n^j = c;\theta^j) \\ &= \sum_{c=1}^{C} \pi_c^j \cdot \mathcal{N}(\mathbf{z}_n^j;\boldsymbol{\mu}_c^j,\Sigma_c^j),\ \ s.t.\sum_{c=1}^{C}\pi_c^j = 1, \end{aligned} \tag{1}$$

the goal is to estimate θ^j that maximizes the log-likelihood

$$\max_{\theta^j} \log p(\mathbf{Z}^j;\theta^j) = \max_{\theta^j} \sum_{n=1}^{N_j} \log p(\mathbf{z}_n^j;\theta^j). \tag{2}$$

Using Jensen's inequality, we obtain the following lower bound for the log-likelihood,

$$\begin{aligned} &\sum_{n=1}^{N_j} \log p(\mathbf{z}_n^j;\theta^j) \\ &= \sum_{n=1}^{N_j} \log \sum_{c=1}^{C} p(\mathbf{z}_n^j, c_n^j = c;\theta^j) \\ &= \sum_{n=1}^{N_j} \log \sum_{c=1}^{C} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)\cdot \frac{p(\mathbf{z}_n^j, c_n^j = c;\theta^j)}{q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)} \\ &\geq \sum_{n=1}^{N_j}\sum_{c=1}^{C} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)\log \frac{p(\mathbf{z}_n^j, c_n^j = c;\theta^j)}{q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)} \\ &= \sum_{n=1}^{N_j} \underbrace{E_{q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)}\left[\log p(\mathbf{z}_n^j, c_n^j = c;\theta^j)\right]}_{Q(\theta^j;\theta_{\text{old}}^j)} \\ &\quad - \sum_{n=1}^{N_j} \underbrace{E_{q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)}\left[q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)\right]}_{-H(C;\theta_{\text{old}}^j)}. \end{aligned} \tag{3}$$

This allows us to substitute the problem of maximizing the log-likelihood with that of maximizing a surrogate function, which in our case is the lower bound given by Jensen's inequality. It can be shown that increasing the lower bound with respect to θ^j leads to monotonically increasing log-likelihood [27, 49]. The optimization of the surrogate function towards maximizing the log-likelihood is often referred to as the Expectation-Maximization (EM) algorithm, which iteratively alternates between the E-step and the M-step.

The surrogate function is comprised of the two terms, $Q(\theta^j;\theta_{\text{old}}^j)$ and $H(C;\theta_{\text{old}}^j)$, both of which are expectations with respect to the posterior probability of prototype assignment, *i.e.,* $q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)$. In the E-step, we use Bayes' rule to compute the posterior probability and consequently the expectations,

$$\begin{aligned} &q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j) \\ &= \frac{q(\mathbf{z}_n^j|c_n^j = c;\theta_{\text{old}}^j)\cdot q(c_n^j = c;\theta_{\text{old}}^j)}{q(\mathbf{z}_n^j;\theta_{\text{old}}^j)} \\ &= \frac{q(\mathbf{z}_n^j|c_n^j = c;\theta_{\text{old}}^j)\cdot q(c_n^j = c;\theta_{\text{old}}^j)}{\sum_{c=1}^{C} q(\mathbf{z}_n^j|c_n^j = c;\theta_{\text{old}}^j)\cdot q(c_n^j = c;\theta_{\text{old}}^j)} \\ &= \frac{\pi_c^j \cdot \mathcal{N}(\mathbf{z}_n^j;\boldsymbol{\mu}_c^j,\Sigma_c^j)}{\sum_{c=1}^{C}\pi_c^j \cdot \mathcal{N}(\mathbf{z}_n^j;\boldsymbol{\mu}_c^j,\Sigma_c^j)}. \end{aligned} \tag{4}$$

In the M-step, we find θ_{new}^j that maximizes the surrogate function. Since the entropy term $H(C;\theta_{\text{old}}^j)$ is not a function of θ_j (it is a function of θ_{old}^j), we only need to optimize $Q(\theta^j;\theta_{\text{old}}^j)$ by taking the derivative with respect to θ_j,

$$\begin{aligned} &\sum_{n=1}^{N_j} \frac{\partial Q(\theta^j;\theta_{\text{old}}^j)}{\partial \pi_c^j} = 0 \\ &\Rightarrow \pi_c^{j,\text{new}} = \frac{\sum_{n=1}^{N_j} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)}{N_j} \\ &\sum_{n=1}^{N_j} \frac{\partial Q(\theta^j;\theta_{\text{old}}^j)}{\partial \boldsymbol{\mu}_c^j} = 0 \\ &\Rightarrow \boldsymbol{\mu}_c^{j,\text{new}} = \frac{\sum_{n=1}^{N_j} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)\cdot \mathbf{z}_n^j}{\sum_{n=1}^{N_j} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)} \\ &\sum_{n=1}^{N_j} \frac{\partial Q(\theta^j;\theta_{\text{old}}^j)}{\partial \Sigma_c^j} = 0 \\ &\Rightarrow \Sigma_c^{j,\text{new}} = \frac{\sum_{n=1}^{N_j} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)\cdot (\mathbf{z}_n^j - \boldsymbol{\mu}_c^{j,\text{new}})^2}{\sum_{n=1}^{N_j} q(c_n^j = c|\mathbf{z}_n^j;\theta_{\text{old}}^j)}. \end{aligned} \tag{5}$$

2. Training details

For training, we use weight decay of 1×10^{-5} and AdamW optimizer with a learning rate of 1×10^{-4} with the co-

sine decay scheduler. For *slide classification* experiments, we use cross-entropy loss and a maximum of 20 epochs with early stopping if the validation loss does not decrease for 10 epochs. For the supervised baselines, due to the variable-length WSI set, we use a batch size of 1 and a gradient accumulation of 32 steps. For unsupervised baselines (including PANTHER), we use a batch size of 32. For *survival prediction* experiments, we use negative log-likelihood loss (NLL) [92] with a batch size of one patient over 20 epochs for supervised baselines. For unsupervised baselines (including PANTHER), we use Cox proportional hazards loss [48] with a batch size of 64 patients over 50 epochs.

3. Computational considerations

Two NVIDIA 3090 GPUs were used for training PANTHER. PANTHER pre-extracts 32,784-dim slide features (16 prototypes $\times$ 2,049-dim for concatenated π_c, $\boldsymbol{\mu}_c$, Σ_c) for linear or MLP probing, **468**$\times$ smaller than $[15\text{K} \times 1024]$-dim patch patch embeddings used for MIL training. We pre-extract PANTHER features with batch size of 1 (10 WSIs/sec), and can compress 11K TCGA slides (4 TB) with $\sim$ 1.4 GB. While more prototypes imply more features to concatenate, training a linear classifier with PANTHER features still has less number of parameters (32,784) than ABMIL ($\sim$500K).

4. Datasets

We provide brief explanations for the datasets that were used for the evaluation of PANTHER.

4.1. Slide classification

EBRAINS [70]: For fine-grained (30 classes) and coarse-grained (12 classes) brain tumor subtyping tasks, we used Hematoxylin and Eosin (H&E) Formalin-fixed and paraffin-embedded (FFPE) WSIs ($n = 2,319$) collected from the University of Vienna. We label-stratify the dataset into train/val/test fold of 50:25:25 and use the same fold for both the fine-grained and coarse-grained subtyping tasks. Performance was evaluated using balanced accuracy and weighted F1.

NSCLC: For the non-small cell lung carcinoma (NSCLC) subtyping task, we use H&E WSIs from TCGA and CPTAC for classifying lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) cases. The TCGA cohort contains a total of 1,041 slides (LUAD: 529, LUSC: 512) and the CPTAC cohort consists a total of 1,091 slides (LUAD: 578, LUSC: 513). We label-stratify the TCGA cohort into train/val/test fold of 80:10:10, with CPTAC used for external validation. Performance was evaluated using balanced accuracy and weighted F1.

PANDA [8, 9]: For the ISUP grading task, we used prostate cancer core needle biopsies (n=10,616) from the Prostate Cancer Grade Assessment (PANDA) challenge. Each biopsy is given an ISUP grade, making this a 6-class classification task. These biopsies are collected from Karolinska Institute (KRLS) and Radboud University Medical Center (RUMC). We label-stratify the PANDA dataset into train/val/test fold of 80:10:10, with the evaluation performed on KRLS and RUMC cohorts separately. Performance was evaluated using Cohen's quadratic weighted Kappa κ^2 metric.

4.2. Survival prediction

TCGA: We perform site-stratified 5-fold CV [38] evaluation on the following cancer types from TCGA: Breast Invasive Carcinoma (BRCA, $n = 1,041$, WSI $= 1,111$), Colon and Rectum Adenocarcinoma (CRC, $n = 566$, WSI $= 575$), Bladder Urothelial Carcinoma (BLCA, $n = 373$, WSI $= 437$), Uterine corpus endometrial carcinoma (UCEC, $n = 504$, WSI $= 565$), Kidney renal clear cell carcinoma (KIRC, $n = 511$, WSI $= 517$), and Lung adenocarcinoma (LUAD, $n = 456$, WSI $= 1,024$). The train/val split is performed on the patient level.

External dataset (CPTAC, NLST): Using the models trained on TCGA cohort, we perform external validation on KIRC (CPTAC: $n = 180$, WSI $= 341$) and LUAD (CPTAC: $n = 185$, WSI $= 486$, NLST: $n = 244$, WSI $= 686$). We note that evaluation on CPTAC and NLST is much more difficult due to dataset shifts in image acquisition (differences in H&E stain variability), geographic location and demographics (social determinants of health affecting access to healthcare), and other potential biases (differences in follow-up procedures between TCGA and CPTAC/NLST).

5. Additional experiments

Ablation over different feature encoders: We evaluate PANTHER, which relied on features extracted with ViT-L/16 DINOv2 pre-trained on a large internal histology dataset (UNI) [18], with other baselines using features extracted from 1) CTransPath encoder [83], which is a Swin Transformer pretrained on 29,753 WSIs from TCGA and 2,457 WSIs from the Pathology AI Platform (PAIP), and 2) ResNet50 encoder pretrained on natural images (ImageNet) [60]. The results can be found in Table S1.

Ablation over a different number of clusters C: We evaluate how PANTHER and other baselines (AttnMISL, ProtoCounts, H2T, and OT) that depend on the number of prototypes C perform across different choices. We report both the classification and survival prediction results for $C = \{8, 16, 32\}$ in Table S2.

Ablation over different survival loss functions: For survival prediction tasks, we also train our unsupervised baselines with 1) the negative log-likelihood (NLL) loss [92] and 2) the ranking loss [62]. These survival loss functions have been frequently used as alternatives to the Cox loss

in survival analysis problems, especially in medical imaging literature. To maintain consistency with the Cox loss experimental setting, we use a batch size of 64 for training with the NLL and ranking loss. The results can be found in Table S3, where we include the supervised baseline results with NLL loss for completeness.

Evaluation was performed on several representative classification and survival tasks: EBRAINS (challenging difficulty), PANDA (depends on understanding mixture proportions of tissue patterns), CRC survival prediction (tissue can be annotated using CRC-100K [47]), and LUAD survival prediction (assessing out-of-domain generalization).

6. Results, interpretation, and insights

Stronger feature encoders improve supervised MIL baselines: We observe consistent trends that stronger feature encoders improve slide-level tasks, with models trained using UNI reaching the best performance (Table S1). Across all MIL architectures and in all classification tasks (except for DSMIL on RUMC evaluation in PANDA), UNI consistently outperforms ResNet-50 and CTransPath in head-to-head comparisons. On survival tasks, we note that CTransPath was additionally pretrained on TCGA, which may produce optimistic bias in evaluation. However, we find that UNI still outperforms CTransPath on TCGA-CRC and TCGA-LUAD survival prediction across many architectures. Interestingly, AttnMISL, which generally underperformed against ABMIL and TransMIL using ResNet-50 features, becomes one of the top-ranked MIL models in survival tasks when using UNI features (second-highest c-index in CRC survival prediction, and the highest c-index in LUAD survival prediction on the TCGA and NLST cohorts). This can be attributed to the prototypical formulation of AttnMISL, which depends on the representation quality of the data centroids for prototypical pooling, which would improve with stronger feature encoders.

Stronger feature encoders enable unsupervised baselines to compete with MIL: Similar to MIL methods, PANTHER and other unsupervised slide representation methods have consistent improvement using UNI over ResNet-50 and CTransPath (aside from evaluation on PANDA and evaluation using DeepSets/ProtoCounts, Table S1). Interestingly, we note that OT and PANTHER using CTransPath features significantly underperforms against many weakly-supervised baselines (0.377 / 0.369 / 0.518 balanced accuracies comparing OT, PANTHER, ABMIL respectively on EBRAINS; 0.677 / 0.782 / 0.901 κ^2 comparing OT, PANTHER, ABMIL respectively on PANDA-KRLS). When using a stronger feature encoder such as UNI, unsupervised slide representation methods have significant gains and consequently outperform ABMIL and other MIL baselines. As in the case of AttnMISL, this can be attributed to the need for strong pretrained encoders that are able to retrieve similar patch embeddings for prototypical pooling.

PANTHER trains stable survival models: Across all evaluation settings (different number of prototypes C and survival loss functions), we find that PANTHER is able to develop high-performing survival models with out-of-domain generalization (Table S2 and S3). In CRC survival prediction, $\text{PANTHER}_{\text{All}}$+MLP consistently outperforms all cluster-based methods within each setting. In comparison to MIL, though $\text{PANTHER}_{\text{All}}$+MLP with C=16 has lower c-index than TransMIL (0.684), we note that $\text{PANTHER}_{\text{All}}$+MLP with C=8 reaches higher performance (0.691). In LUAD survival prediction, $\text{PANTHER}_{\text{All}}$+MLP is consistently the second best-performing model on TCGA evaluation, behind OT (best c-index of 0.715 of with C=8). Though many baselines such as OT, H2T, and even DeepSets can reach strong performance on TCGA, we note that almost all of these methods have unstable performance on external cohorts, with c-index falling under 0.5 on CPTAC, NLST, or both.

PANTHER prototypes capture distinct tumor morphologies: In **Fig.** S1, S2, and S3, we visualize prototypical assignment maps and heatmap visualizations of various cancer types. Consistent with our findings in **Fig.** 3, PANTHER is able to map the spatial organization of histologic visual concepts. In particular, PANTHER finds several unique tumor populations, delineating: tumor-invading muscle and tumor with immune infiltration in BLCA, nested tumor and tumor-associated connective tissue in BRCA, and clear cell RCC with and without presentation of poorly-differentiated glands in KIRC. Furthermore, we also show concordance of our visualizations using a supervised classifier (developed using patch-level tumor annotations in the TCGA Uniform Tumor dataset [51]) for tumor tissue classification. Visualizing the posterior probability heatmap of the tumor prototype with the highest mixture probability $\widehat{\pi}_c$ (greatest presence), we find that our tumor heatmap visualizations have strong concordance with those generated based on the results from supervised classifiers.

PANTHER prototypes capture distribution of tissue classes in CRC-100K: In **Fig.** S4, we visualize prototypical assignment maps and their correspondence to diverse tissue annotations in COADREAD tissue. Using the CRC-100K dataset (containing 9 tissue classes) [47], we developed a supervised patch-level classifier to predict tissue assignments for all patches in TCGA-COADREAD slides. To match the prototypical assignment maps from PANTHER

with the label distribution in CRC-100K, we applied the previous classifier to the learned prototypes of PANTHER, to predict CRC-100K tissue labels. Overall, we find that the learned prototypes of PANTHER have strong concordance with morphologically-relevant and diverse histopathology tissue patterns annotated by supervised classifiers.

Table S1. **Varying pretrained feature extractors.** We compare the performance of supervised (**top**) and unsupervised (**bottom**) methods with different pretrained encoders, ResNet50 with ImageNet transfer (RN50), CTransPath (CTP), and UNI, on classification tasks (EBRAINS and PANDA) and survival tasks (CRC and LUAD).

	Train on	EBRAINS (fine, 30 classes)		PANDA (grading, 6 classes)		CRC (survival)	LUAD (survival)		
	Test on	EBRAINS (Bal. acc.)	EBRAINS (F1)	KRLS (κ^2)	RUMC (κ^2)	TCGA (C-Index)	TCGA (C-Index)	CPTAC (C-Index)	NLST (C-Index)
Sup. (RN50)	ABMIL [40]	0.197	0.267	0.792	0.789	0.540 ± 0.08	0.537 ± 0.18	0.684 ± 0.01	0.482 ± 0.03
	TransMIL [73]	0.516	0.614	0.841	0.854	0.501 ± 0.05	0.628 ± 0.11	0.555 ± 0.06	0.462 ± 0.01
	DSMIL [55]	0.401	0.510	0.681	0.728	0.494 ± 0.01	0.487 ± 0.03	0.524 ± 0.02	0.519 ± 0.02
	ILRA [84]	0.509	0.599	0.860	0.880	0.588 ± 0.09	0.522 ± 0.17	0.559 ± 0.07	0.443 ± 0.06
	AttnMISL [88]	0.033	0.073	0.128	0.005	0.493 ± 0.04	0.519 ± 0.16	0.666 ± 0.01	0.535 ± 0.01
Sup. (CTP)	ABMIL [40]	0.518	0.594	0.901	0.908	0.642 ± 0.10	0.607 ± 0.03	0.545 ± 0.05	0.560 ± 0.02
	TransMIL [73]	0.642	0.71	0.911	0.918	0.611 ± 0.12	0.614 ± 0.06	0.479 ± 0.05	0.508 ± 0.04
	DSMIL [55]	0.515	0.584	0.890	0.916	0.499 ± 0.03	0.552 ± 0.05	0.466 ± 0.04	0.438 ± 0.01
	ILRA [84]	0.580	0.655	0.917	0.920	0.590 ± 0.09	0.602 ± 0.05	0.427 ± 0.05	0.456 ± 0.03
	AttnMISL [88]	0.033	0.073	0.402	0.837	0.627 ± 0.12	0.602 ± 0.05	0.427 ± 0.05	0.456 ± 0.03
Sup. (UNI)	ABMIL [40]	0.674	0.744	0.935	0.918	0.608 ± 0.09	0.654 ± 0.06	0.572 ± 0.03	0.519 ± 0.04
	TransMIL [73]	0.701	0.758	0.942	0.922	0.684 ± 0.06	0.665 ± 0.10	0.555 ± 0.03	0.484 ± 0.05
	DSMIL [55]	0.648	0.698	0.909	0.911	0.500 ± 0.00	0.501 ± 0.00	0.502 ± 0.00	0.500 ± 0.00
	ILRA [84]	0.618	0.695	0.931	0.925	0.555 ± 0.10	0.586 ± 0.06	0.651 ± 0.05	0.482 ± 0.01
	AttnMISL [88]	0.534	0.636	0.882	0.894	0.639 ± 0.10	0.673 ± 0.10	0.632 ± 0.03	0.577 ± 0.04
Unsup. (RN50)	DeepSets [93]	0.033	0.073	< 0	< 0	0.574 ± 0.08	0.565 ± 0.12	0.715 ± 0.02	0.591 ± 0.01
	ProtoCounts [69, 91]	0.045	0.077	0.016	0.183	0.516 ± 0.05	0.546 ± 0.04	0.499 ± 0.08	0.502 ± 0.07
	H2T [81]	0.047	0.087	0.262	0.329	0.501 ± 0.10	0.585 ± 0.14	0.512 ± 0.07	0.545 ± 0.04
	OT [64]	0.063	0.088	0.211	0.540	0.578 ± 0.08	0.575 ± 0.14	0.581 ± 0.05	0.544 ± 0.02
	$\text{PANTHER}_{\text{WA}}$ + lin.	0.033	0.073	0.150	0.057	0.525 ± 0.09	0.586 ± 0.11	0.552 ± 0.04	0.491 ± 0.01
	$\text{PANTHER}_{\text{All}}$ + lin.	0.063	0.112	0.207	0.535	0.554 ± 0.07	0.586 ± 0.12	0.548 ± 0.04	0.505 ± 0.02
	$\text{PANTHER}_{\text{All}}$ + MLP	0.142	0.216	0.550	0.665	0.585 ± 0.07	0.601 ± 0.07	0.396 ± 0.04	0.465 ± 0.03
Unsup. (CTP)	DeepSets [93]	0.033	0.073	< 0	< 0	0.522 ± 0.10	0.629 ± 0.06	0.463 ± 0.03	0.554 ± 0.05
	ProtoCounts [69, 91]	0.057	0.037	0.059	0.539	0.521 ± 0.05	0.479 ± 0.11	0.537 ± 0.10	0.522 ± 0.11
	H2T [81]	0.053	0.124	0.333	0.704	0.545 ± 0.11	0.586 ± 0.05	0.552 ± 0.07	0.562 ± 0.02
	OT [64]	0.377	0.482	0.677	0.738	0.607 ± 0.09	0.690 ± 0.06	0.466 ± 0.02	0.547 ± 0.06
	$\text{PANTHER}_{\text{WA}}$ + lin.	0.033	0.073	0.203	0.586	0.533 ± 0.11	0.673 ± 0.05	0.474 ± 0.03	0.515 ± 0.05
	$\text{PANTHER}_{\text{All}}$ + lin.	0.398	0.493	0.661	0.757	0.614 ± 0.09	0.672 ± 0.05	0.491 ± 0.04	0.540 ± 0.06
	$\text{PANTHER}_{\text{All}}$ + MLP	0.369	0.483	0.782	0.869	0.661 ± 0.11	0.655 ± 0.08	0.584 ± 0.04	0.511 ± 0.03
Unsup. (UNI)	DeepSets [93]	0.033	0.073	< 0	< 0	0.563 ± 0.10	0.652 ± 0.05	0.550 ± 0.01	0.509 ± 0.04
	ProtoCounts [69, 91]	0.038	0.018	< 0	0.13	0.552 ± 0.06	0.460 ± 0.11	0.577 ± 0.11	0.500 ± 0.01
	H2T [81]	0.117	0.223	0.457	0.755	0.639 ± 0.11	0.662 ± 0.09	0.583 ± 0.03	0.603 ± 0.04
	OT [64]	0.700	0.756	0.817	0.883	0.622 ± 0.09	0.687 ± 0.08	0.641 ± 0.02	0.495 ± 0.04
	$\text{PANTHER}_{\text{WA}}$ + lin.	0.497	0.598	0.663	0.787	0.647 ± 0.12	0.654 ± 0.07	0.461 ± 0.01	0.482 ± 0.06
	$\text{PANTHER}_{\text{All}}$ + lin.	0.691	0.756	0.866	0.909	0.645 ± 0.07	0.672 ± 0.06	0.568 ± 0.05	0.623 ± 0.07
	$\text{PANTHER}_{\text{All}}$ + MLP	0.693	0.760	0.923	0.931	0.665 ± 0.10	0.685 ± 0.06	0.653 ± 0.04	0.634 ± 0.04

Table S2. **Varying C in cluster-based methods.** We compare the performance of cluster-based methods with $C = \{8, 16, 32\}$ on classification tasks (EBRAINS and PANDA) and survival tasks (CRC and LUAD). **Top.** MIL baselines with no clustering (NC). **Bottom.** Cluster-based methods with $C = \{8, 16, 32\}$, which include weakly-supervised MIL (AttnMISL) and unsupervised slide representation learning approaches (ProtoCounts, H2T, OT, PANTHER).

	Train on	EBRAINS (fine, 30 classes)		PANDA (grading, 6 classes)		CRC (survival)	LUAD (survival)		
	Test on	EBRAINS (Bal. acc.)	EBRAINS (F1)	KRLS (κ^2)	RUMC (κ^2)	TCGA (C-Index)	TCGA (C-Index)	CPTAC (C-Index)	NLST (C-Index)
NC	ABMIL [40]	0.674	0.744	0.935	0.918	0.608 ± 0.09	0.654 ± 0.06	0.572 ± 0.03	0.519 ± 0.04
	TransMIL [73]	0.701	0.758	0.942	0.922	0.684 ± 0.06	0.665 ± 0.10	0.555 ± 0.03	0.484 ± 0.05
	DSMIL [55]	0.648	0.698	0.909	0.911	0.500 ± 0.00	0.501 ± 0.00	0.502 ± 0.00	0.50 ± 0.00
	ILRA [84]	0.618	0.695	0.931	0.925	0.555 ± 0.10	0.586 ± 0.06	0.651 ± 0.05	0.482 ± 0.01
	DeepSets [93]	0.033	0.073	< 0	< 0	0.563 ± 0.10	0.652 ± 0.05	0.550 ± 0.01	0.509 ± 0.04
C=8	AttnMISL [88]	0.560	0.655	0.857	0.874	0.595 ± 0.08	0.643 ± 0.07	0.644 ± 0.04	0.545 ± 0.01
	ProtoCounts [69, 91]	0.022	0.03	0.0	0.284	0.479 ± 0.11	0.561 ± 0.06	0.625 ± 0.13	0.562 ± 0.15
	H2T [81]	0.045	0.105	0.286	0.767	0.622 ± 0.09	0.638 ± 0.08	0.525 ± 0.04	0.563 ± 0.03
	OT [64]	0.689	0.756	0.803	0.869	0.626 ± 0.09	0.715 ± 0.10	0.609 ± 0.03	0.523 ± 0.05
	PANTHER$_{\text{WA}}$ + lin.	0.490	0.593	0.671	0.787	0.600 ± 0.10	0.670 ± 0.05	0.526 ± 0.01	0.502 ± 0.07
	PANTHER$_{\text{All}}$ + lin.	0.668	0.742	0.843	0.9	0.678 ± 0.12	0.643 ± 0.06	0.660 ± 0.02	0.615 ± 0.03
	PANTHER$_{\text{All}}$ + MLP	0.674	0.753	0.918	0.936	0.691 ± 0.11	0.648 ± 0.06	0.669 ± 0.03	0.603 ± 0.03
C=16	AttnMISL [88]	0.534	0.636	0.882	0.894	0.639 ± 0.10	0.673 ± 0.10	0.632 ± 0.03	0.577 ± 0.04
	ProtoCounts [69, 91]	0.038	0.018	< 0	0.13	0.552 ± 0.06	0.460 ± 0.11	0.577 ± 0.11	0.500 ± 0.01
	H2T [81]	0.117	0.223	0.457	0.755	0.639 ± 0.11	0.662 ± 0.09	0.583 ± 0.03	0.603 ± 0.04
	OT [64]	0.700	0.756	0.817	0.883	0.622 ± 0.09	0.687 ± 0.08	0.641 ± 0.02	0.495 ± 0.04
	PANTHER$_{\text{WA}}$ + lin.	0.497	0.598	0.663	0.787	0.647 ± 0.12	0.654 ± 0.07	0.461 ± 0.01	0.482 ± 0.06
	PANTHER$_{\text{All}}$ + lin.	0.691	0.756	0.866	0.909	0.645 ± 0.07	0.672 ± 0.06	0.568 ± 0.05	0.623 ± 0.07
	PANTHER$_{\text{All}}$ + MLP	0.693	0.760	0.923	0.931	0.665 ± 0.10	0.685 ± 0.06	0.653 ± 0.04	0.634 ± 0.04
C=32	AttnMISL [88]	0.492	0.598	0.901	0.889	0.572 ± 0.10	0.666 ± 0.06	0.591 ± 0.03	0.587 ± 0.02
	ProtoCounts [69, 91]	0.073	0.105	0.301	0.54	0.578 ± 0.09	0.498 ± 0.14	0.566 ± 0.12	0.529 ± 0.09
	H2T [81]	0.244	0.363	0.626	0.779	0.621 ± 0.12	0.665 ± 0.05	0.599 ± 0.04	0.650 ± 0.03
	OT [64]	0.687	0.746	0.841	0.898	0.605 ± 0.00	0.689 ± 0.08	0.664 ± 0.02	0.518 ± 0.04
	PANTHER$_{\text{WA}}$ + lin.	0.489	0.593	0.670	0.782	0.606 ± 0.11	0.677 ± 0.06	0.522 ± 0.01	0.469 ± 0.06
	PANTHER$_{\text{All}}$ + lin.	0.676	0.751	0.883	0.896	0.649 ± 0.07	0.677 ± 0.06	0.583 ± 0.06	0.594 ± 0.05
	PANTHER$_{\text{All}}$ + MLP	0.674	0.741	0.935	0.931	0.656 ± 0.13	0.676 ± 0.04	0.665 ± 0.06	0.614 ± 0.05

Table S3. **Varying loss function in survival tasks.** We compare the performance of all methods with different loss functions, NLL (**top**), ranking (**middle**), and Cox loss (**bottom**), on survival outcome prediction in CRC and LUAD.

	Train on	CRC	LUAD		
	Test on	TCGA	TCGA	CPTAC	NLST
Sup. (NLL)	ABMIL [40]	0.608 ± 0.09	0.654 ± 0.06	0.572 ± 0.03	0.519 ± 0.04
	TransMIL [73]	0.684 ± 0.06	0.665 ± 0.10	0.555 ± 0.03	0.484 ± 0.05
	DSMIL [55]	0.500 ± 0.00	0.501 ± 0.00	0.502 ± 0.00	0.500 ± 0.00
	ILRA [84]	0.555 ± 0.10	0.586 ± 0.06	0.651 ± 0.05	0.482 ± 0.01
	AttnMISL [88]	0.639 ± 0.10	0.673 ± 0.10	0.632 ± 0.03	0.577 ± 0.04
Unsup. (NLL)	DeepSets [93]	0.559 ± 0.11	0.560 ± 0.19	0.659 ± 0.02	0.587 ± 0.02
	ProtoCounts [69, 91]	0.517 ± 0.03	0.493 ± 0.04	0.496 ± 0.15	0.591 ± 0.05
	H2T [81]	0.563 ± 0.08	0.498 ± 0.17	0.547 ± 0.01	0.520 ± 0.02
	OT [64]	0.626 ± 0.12	0.681 ± 0.09	0.615 ± 0.03	0.462 ± 0.02
	PANTHER$_{\text{WA}}$ + lin.	0.508 ± 0.10	0.647 ± 0.09	0.643 ± 0.02	0.433 ± 0.04
	PANTHER$_{\text{All}}$ + lin.	0.647 ± 0.11	0.670 ± 0.08	0.651 ± 0.04	0.614 ± 0.07
	PANTHER$_{\text{All}}$ + MLP	0.649 ± 0.11	0.668 ± 0.08	0.638 ± 0.08	0.607 ± 0.05
Unsup. (Rank)	DeepSets [93]	0.608 ± 0.11	0.614 ± 0.05	0.556 ± 0.04	0.538 ± 0.04
	ProtoCounts [69, 91]	0.476 ± 0.05	0.503 ± 0.05	0.482 ± 0.14	0.507 ± 0.08
	H2T [81]	0.598 ± 0.11	0.661 ± 0.11	0.558 ± 0.02	0.620 ± 0.04
	OT [64]	0.670 ± 0.11	0.643 ± 0.03	0.595 ± 0.03	0.488 ± 0.05
	PANTHER$_{\text{WA}}$ + lin.	0.626 ± 0.13	0.637 ± 0.06	0.445 ± 0.02	0.518 ± 0.07
	PANTHER$_{\text{All}}$ + lin.	0.661 ± 0.07	0.677 ± 0.06	0.575 ± 0.05	0.625 ± 0.06
	PANTHER$_{\text{All}}$ + MLP	0.671 ± 0.09	0.684 ± 0.06	0.651 ± 0.03	0.628 ± 0.05
Unsup. (Cox)	DeepSets [93]	0.563 ± 0.10	0.652 ± 0.05	0.550 ± 0.01	0.509 ± 0.04
	ProtoCounts [69, 91]	0.552 ± 0.06	0.460 ± 0.11	0.577 ± 0.11	0.500 ± 0.01
	H2T [81]	0.639 ± 0.11	0.662 ± 0.09	0.583 ± 0.03	0.603 ± 0.04
	OT [64]	0.622 ± 0.09	0.687 ± 0.08	0.641 ± 0.02	0.495 ± 0.04
	PANTHER$_{\text{WA}}$ + lin.	0.647 ± 0.12	0.654 ± 0.07	0.461 ± 0.01	0.482 ± 0.06
	PANTHER$_{\text{All}}$ + lin.	0.645 ± 0.07	0.672 ± 0.06	0.568 ± 0.05	0.623 ± 0.07
	PANTHER$_{\text{All}}$ + MLP	0.665 ± 0.10	0.685 ± 0.06	0.653 ± 0.04	0.634 ± 0.04

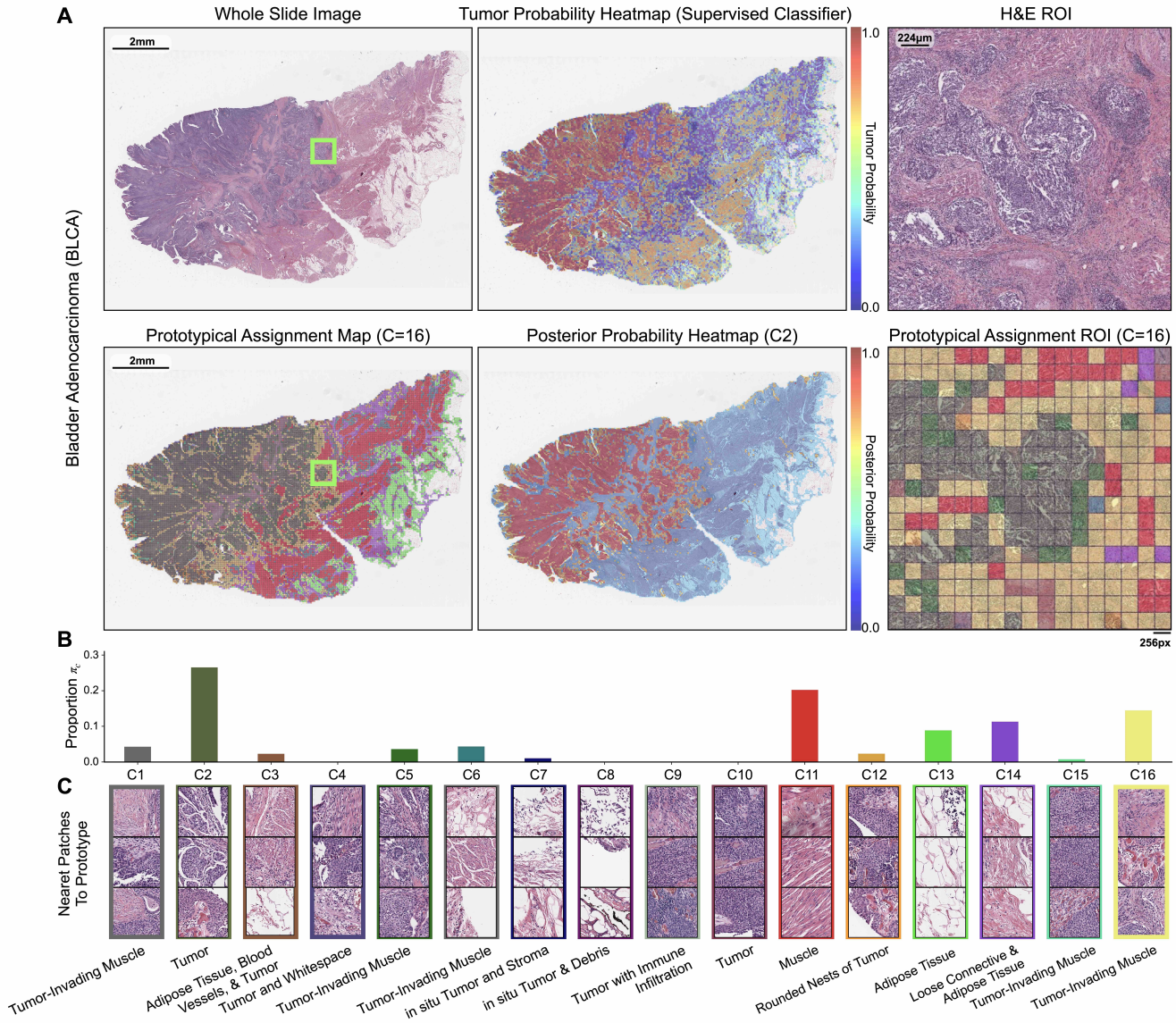

Figure S1. **Prototype-oriented heatmap interpretation of BLCA**. (A) Visualization of prototypical assignment map in an exemplar BLCA H&E WSI, with zoomed-in histopathology ROI of tumor-invading muscle (C2, C8, C11, C16). We show the posterior probability heatmap for the tumor-containing C2 prototype, which has strong concordance with a tumor probability heatmap obtained by a supervised patch-level classifier for BLCA tumor prediction. (B) Prototype distribution $\hat{\pi}_c$ of the exemplar slide. (C) Morphological annotations of all prototypes by a board-certified pathologist in the BLCA cohort.

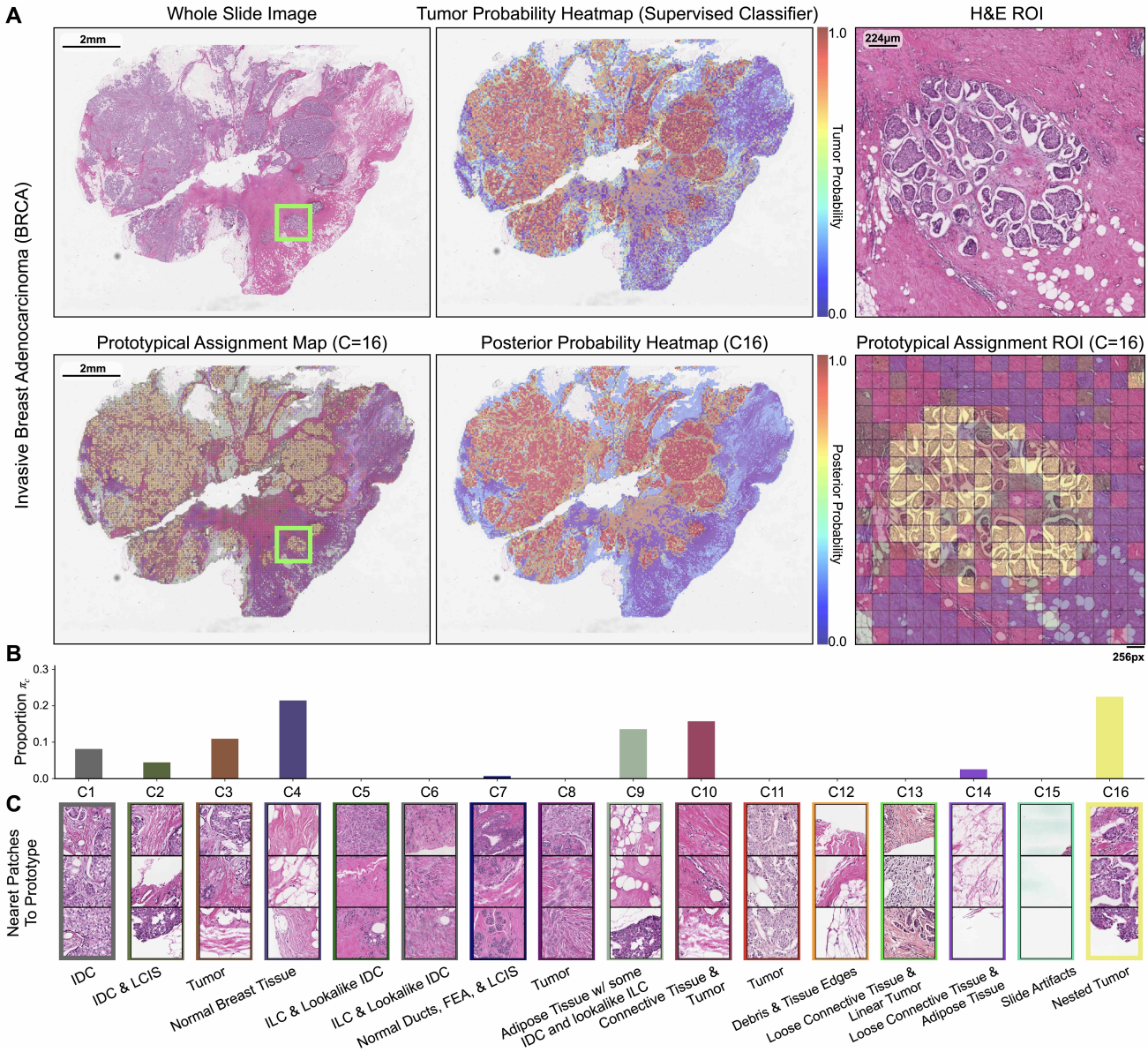

Figure S2. **Prototype-oriented heatmap interpretation of BRCA**. (A) Visualization of prototypical assignment map in an exemplar BRCA H&E WSI, with zoomed-in histopathology ROI of dense tumor nests (C16) with surrounding connective tissue (C10), adipose tissue (C9) with tumor presence (C3). We show the posterior probability heatmap for the tumor-containing C16 prototype, which has strong concordance with a tumor probability heatmap obtained by a supervised patch-level classifier for BRCA tumor prediction. (B) Prototype distribution $\hat{\pi}_c$ of the exemplar slide. (C) Morphological annotations of all prototypes by a board-certified pathologist in the BRCA cohort.

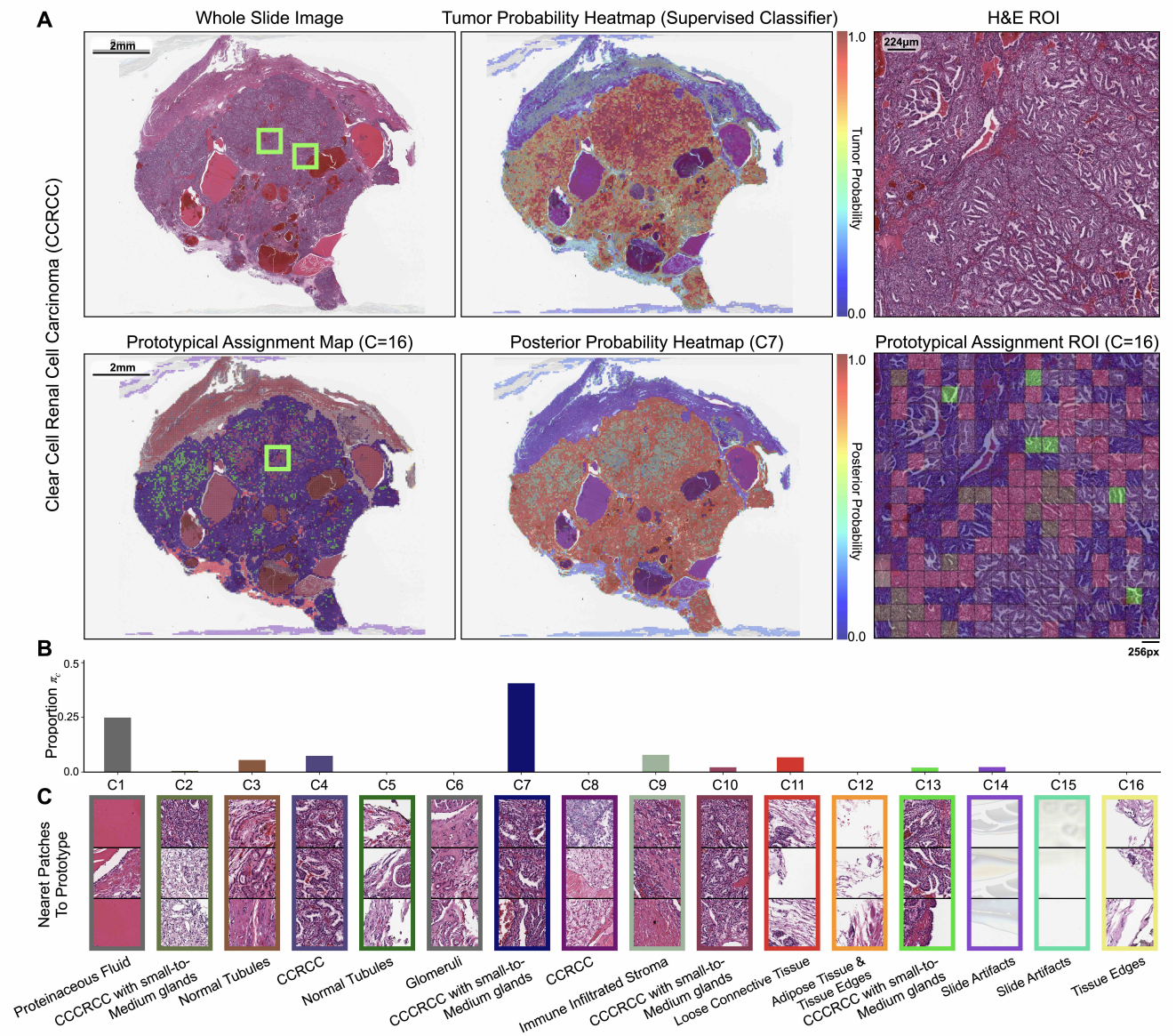

Figure S3. **Prototype-oriented heatmap interpretation of KIRC**. (A) Visualization of prototypical assignment map in an exemplar KIRC/CCRCC H&E WSI, with zoomed-in histopathology ROI of CCRCC in small-to-medium glands (C7, C10, C13). We show the posterior probability heatmap for the tumor-containing C7 prototype, which has strong concordance with a tumor probability heatmap obtained by a supervised patch-level classifier for CCRCC tumor prediction. (B) Prototype distribution $\hat{\pi}_c$ of the exemplar slide. (C) Morphological annotations of all prototypes by a board-certified pathologist in the KIRC cohort.

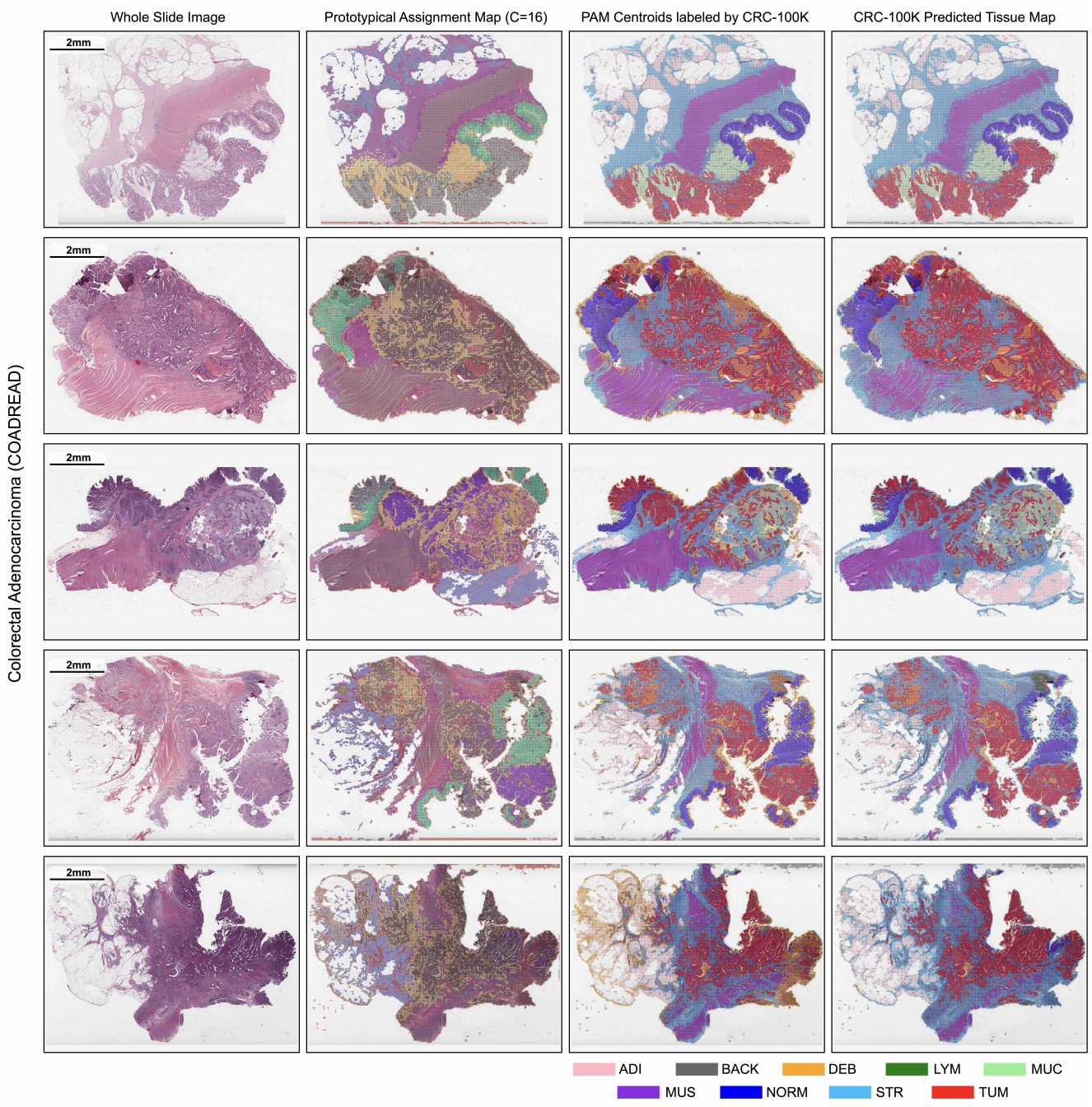

Figure S4. **Prototype-oriented heatmap interpretation of COADREAD and correspondence with CRC-100K**. For exemplar COADREAD slides, we visualize their prototypical assignment maps and their correspondence with tissue classes in CRC-100K. Using a supervised patch-level classifier for predicting the 9 tissue classes in CRC-100K, we predicted tissue classes in TCGA-COADREAD slides, shown in the **far-right column**. To match the prototypical assignment maps from PANTHER with the label distribution in CRC-100K, we applied the same classifier to predict CRC-100K tissue labels for the learned prototypes in PANTHER (**middle-right column**). Across all 9 classes, we find that PANTHER's prototypes correspond to morphologically-relevant and semantic histopathology tissue patterns annotated by supervised classifiers.